Contents

Notation *page* xii
Preface xvii

Part I Preliminaries 1

1 Molecular biology and high-throughput sequencing 3
1.1 DNA, RNA, proteins 3
1.2 Genetic variations 6
1.3 High-throughput sequencing 7
Exercises 9

2 Algorithm design 10
2.1 Complexity analysis 10
2.2 Data representations 12
2.3 Reductions 13
2.4 Literature 17
Exercises 17

3 Data structures 20
3.1 Dynamic range minimum queries 20
3.2 Bitvector rank and select operations 22
3.3 Wavelet tree 24
3.3.1 Balanced representation 24
3.3.2 Range queries 26
3.4 Literature 27
Exercises 27

4 Graphs 30
4.1 Directed acyclic graphs (DAGs) 30
4.1.1 Topological ordering 30
4.1.2 Shortest paths 31

4.2 Arbitrary directed graphs 33
4.2.1 Eulerian paths 33
4.2.2 Shortest paths and the Bellman–Ford method 34
4.3 Literature 38
Exercises 38

5 Network flows 41
5.1 Flows and their decompositions 41
5.2 Minimum-cost flows and circulations 45
5.2.1 The residual graph 47
5.2.2 A pseudo-polynomial algorithm 50
5.3 Bipartite matching problems 51
5.3.1 Perfect matching 52
5.3.2 Matching with capacity constraints 54
5.3.3 Matching with residual constraints 56
5.4 Covering problems 58
5.4.1 Disjoint cycle cover 58
5.4.2 Minimum path cover in a DAG 60
5.5 Literature 64
Exercises 65

Part II Fundamentals of Biological Sequence Analysis 69

6 Alignments 71
6.1 Edit distance 72
6.1.1 Edit distance computation 73
6.1.2 Shortest detour 76
*6.1.3 Myers' bitparallel algorithm 78
6.2 Longest common subsequence 83
6.2.1 Sparse dynamic programming 84
6.3 Approximate string matching 86
6.4 Biological sequence alignment 88
6.4.1 Global alignment 89
6.4.2 Local alignment 90
6.4.3 Overlap alignment 92
6.4.4 Affine gap scores 94
6.4.5 The invariant technique 97
6.5 Gene alignment 98
6.6 Multiple alignment 101
6.6.1 Scoring schemes 101
6.6.2 Dynamic programming 103
6.6.3 Hardness 103
6.6.4 Progressive multiple alignment 104

6.6.5 DAG alignment 105
6.6.6 Jumping alignment 107
6.7 Literature 108
Exercises 109

7 **Hidden Markov models (HMMs)** 113
7.1 Definition and basic problems 114
7.2 The Viterbi algorithm 118
7.3 The forward and backward algorithms 118
7.4 Estimating HMM parameters 120
7.5 Literature 122
Exercises 123

Part III Genome-Scale Index Structures 127

8 **Classical indexes** 129
8.1 *k*-mer index 129
8.2 Suffix array 132
8.2.1 Suffix and string sorting 133
8.3 Suffix tree 140
8.3.1 Properties of the suffix tree 142
8.3.2 Construction of the suffix tree 143
8.4 Applications of the suffix tree 145
8.4.1 Maximal repeats 145
8.4.2 Maximal unique matches 147
8.4.3 Document counting 149
8.4.4 Suffix–prefix overlaps 151
8.5 Literature 151
Exercises 153

9 **Burrows–Wheeler indexes** 157
9.1 Burrows–Wheeler transform (BWT) 158
9.2 BWT index 160
9.2.1 Succinct LF-mapping 160
9.2.2 Backward search 162
9.2.3 Succinct suffix array 163
9.2.4 Batched locate queries 165
*9.3 Space-efficient construction of the BWT 166
9.4 Bidirectional BWT index 171
*9.4.1 Visiting all nodes of the suffix tree with just one BWT 175
*9.5 BWT index for labeled trees 177
*9.5.1 Moving top-down 179
*9.5.2 Moving bottom-up 181

*9.5.3 Construction and space complexity 182
*9.6 BWT index for labeled DAGs 182
*9.6.1 Moving backward 185
*9.6.2 Moving forward 186
*9.6.3 Construction 187
9.7 BWT indexes for de Bruijn graphs 188
9.7.1 Frequency-oblivious representation 190
9.7.2 Frequency-aware representation 192
9.7.3 Space-efficient construction 193
9.8 Literature 194
Exercises 196

Part IV Genome-Scale Algorithms 199

10 Read alignment 201
10.1 Pattern partitioning 202
10.2 Dynamic programming along suffix tree paths 204
10.3 Backtracking on BWT indexes 204
10.3.1 Prefix pruning 206
10.3.2 Case analysis pruning with the bidirectional BWT index 208
10.4 Suffix filtering for approximate overlaps 209
10.5 Paired-end and mate pair reads 211
10.6 Split alignment of reads 212
10.7 Alignment of reads to a pan-genome 214
10.7.1 Indexing a set of individual genomes 214
*10.7.2 Indexing a reference genome and a set of variations 215
10.8 Literature 216
Exercises 217

11 Genome analysis and comparison 220
11.1 Space-efficient genome analysis 221
11.1.1 Maximal repeats 221
11.1.2 Maximal unique matches 223
11.1.3 Maximal exact matches 225
11.2 Comparing genomes without alignment 229
11.2.1 Substring and *k*-mer kernels 232
*11.2.2 Substring kernels with Markovian correction 238
11.2.3 Substring kernels and matching statistics 244
11.2.4 Mismatch kernels 251
11.2.5 Compression distance 253
11.3 Literature 255
Exercises 256

12 Genome compression 262

12.1 Lempel–Ziv parsing 263
12.1.1 Basic algorithm for Lempel–Ziv parsing 264
12.1.2 Space-efficient Lempel–Ziv parsing 265
*12.1.3 Space- and time-efficient Lempel–Ziv parsing 266
*12.2 Bit-optimal Lempel–Ziv compression 270
*12.2.1 Building distance-maximal arcs 275
*12.2.2 Building the compact trie 278
12.3 Literature 279
Exercises 280

13 Fragment assembly 282

13.1 Sequencing by hybridization 282
13.2 Contig assembly 284
13.2.1 Read error correction 285
13.2.2 Reverse complements 286
13.2.3 Irreducible overlap graphs 287
13.3 Scaffolding 291
13.4 Gap filling 297
13.5 Literature 299
Exercises 301

Part V Applications 305

14 Genomics 307

14.1 Variation calling 308
14.1.1 Calling small variants 308
14.1.2 Calling large variants 309
14.2 Variation calling over pan-genomes 313
14.2.1 Alignments on a set of individual genomes 313
14.2.2 Alignments on the labeled DAG of a population 314
14.2.3 Evaluation of variation calling results 315
14.3 Haplotype assembly and phasing 315
14.4 Literature 322
Exercises 323

15 Transcriptomics 325

15.1 Estimating the expression of annotated transcripts 325
15.2 Transcript assembly 329
15.2.1 Short reads 329
15.2.2 Long reads 330
15.2.3 Paired-end reads 335

15.3 Simultaneous assembly and expression estimation 337
15.4 Transcript alignment with co-linear chaining 342
15.5 Literature 345
Exercises 346

16 Metagenomics 350

16.1 Species estimation 351
16.1.1 Single-read methods 351
16.1.2 Multi-read and coverage-sensitive methods 353
16.2 Read clustering 357
16.2.1 Filtering reads from low-frequency species 357
16.2.2 Initializing clusters 359
16.2.3 Growing clusters 363
16.3 Comparing metagenomic samples 364
16.3.1 Sequence-based methods 365
16.3.2 Read-based methods 365
16.3.3 Multi-sample methods 366
16.4 Literature 366
Exercises 367

References 370
Index 386