

Contents

	<i>List of insights</i>	page xiii
	<i>Preface</i>	xv
	<i>Notation</i>	xxi
	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
2	Algorithm design	10
	2.1 Complexity analysis	10
	2.2 Data representations	12
	2.3 Reductions	13
	2.4 Literature	17
3	Data structures	21
	3.1 Dynamic range minimum queries	21
	3.2 Bitvector rank and select operations	23
	3.3 Wavelet tree	25
	3.3.1 Balanced representation	25
	3.3.2 Range queries	27
	3.4 Static range minimum queries	28
	3.4.1 From RMQs to LCAs through Cartesian tree	29
	3.4.2 From LCAs to special RMQs	30
	3.4.3 Solving short special RMQs	31
	3.4.4 From large special RMQs to shorter general RMQs	31
	3.4.5 Solving general RMQs	32
	3.5 Hashing	33
	3.5.1 Universal hashing	33
	3.5.2 Approximate membership query	36
	3.5.3 Rolling hash	37
	3.5.4 Minimizers	37
	3.6 Literature	38

4	Graphs	42
	4.1 Directed acyclic graphs (DAGs)	42
	4.1.1 Topological ordering	42
	4.1.2 Shortest paths	44
	4.2 Arbitrary directed graphs	45
	4.2.1 Eulerian paths	45
	4.2.2 Shortest paths and the Bellman–Ford method	47
	4.3 Literature	50
5	Network flows	53
	5.1 Flows and their decompositions	53
	5.2 Minimum-cost flows and circulations	57
	5.2.1 The residual graph	59
	5.2.2 A pseudo-polynomial algorithm	62
	5.3 Bipartite matching problems	63
	5.3.1 Perfect matching	64
	5.3.2 Matching with capacity constraints	67
	5.3.3 Matching with residual constraints	69
	5.4 Covering problems	71
	5.4.1 Disjoint cycle cover	71
	5.4.2 Minimum path cover in a DAG	72
	5.5 Literature	76
	Part II Fundamentals of Biological Sequence Analysis	81
6	Alignments	83
	6.1 Edit distance	84
	6.1.1 Edit distance computation	85
	6.1.2 Shortest detour	88
	*6.1.3 Myers' bitparallel algorithm	90
	6.2 Longest common subsequence	95
	6.2.1 Sparse dynamic programming	96
	6.3 Approximate string matching	99
	6.4 Biological sequence alignment	100
	6.4.1 Global alignment	101
	6.4.2 Local alignment	102
	6.4.3 Overlap alignment	104
	6.4.4 Affine gap scores	106
	6.4.5 The invariant technique	109
	6.5 Gene alignment	110
	6.6 Multiple alignment	113
	6.6.1 Scoring schemes	113
	6.6.2 Dynamic programming	115
	6.6.3 Hardness	115

	6.6.4 Progressive multiple alignment	116
	6.7 DAG alignment	117
	6.8 Alignment on cyclic graphs	120
	6.9 Jumping alignment	123
	6.10 Literature	124
7	Hidden Markov models	129
	7.1 Definition and basic problems	130
	7.2 The Viterbi algorithm	134
	7.3 The forward and backward algorithms	135
	7.4 Estimating HMM parameters	136
	7.5 Literature	138
	Part III Genome-Scale Index Structures	143
8	Classical indexes	145
	8.1 k -mer index	145
	8.2 Suffix array	148
	8.2.1 Suffix and string sorting	149
	8.3 Suffix tree	157
	8.3.1 Properties of the suffix tree	158
	8.3.2 Construction of the suffix tree	159
	8.4 Applications of the suffix tree	161
	8.4.1 Maximal repeats	161
	8.4.2 Maximal unique matches	164
	8.4.3 Document counting	164
	8.4.4 Suffix–prefix overlaps	166
	8.4.5 Approximate string matching in $O(kn)$ time	167
	8.5 Literature	168
9	Burrows–Wheeler indexes	174
	9.1 Burrows–Wheeler transform (BWT)	175
	9.2 BWT index	177
	9.2.1 Succinct LF-mapping	177
	9.2.2 Backward search	179
	9.2.3 Succinct suffix array	180
	9.2.4 Batched locate queries	182
	*9.3 Space-efficient construction of the BWT	183
	9.4 Bidirectional BWT index	188
	*9.4.1 Visiting all nodes of the suffix tree with just one BWT	192
	*9.5 BWT index for labeled trees	195
	*9.5.1 Moving top-down	198
	*9.5.2 Moving bottom-up	199
	*9.5.3 Construction and space complexity	199

x	Contents	
	*9.6 BWT index for labeled DAGs	200
	*9.6.1 Moving backward	203
	*9.6.2 Moving forward	204
	*9.6.3 Construction	204
	9.7 BWT indexes for de Bruijn graphs	206
	9.7.1 Frequency-oblivious representation	208
	9.7.2 Frequency-aware representation	210
	9.7.3 Space-efficient construction	211
	9.8 Literature	212
	Part IV Genome-Scale Algorithms	217
10	Alignment-based genome analysis	219
	10.1 Variation calling	221
	10.1.1 Calling small variants	222
	10.1.2 Calling large variants	222
	10.2 Pattern partitioning for read alignment	224
	10.3 Dynamic programming along suffix tree paths	226
	10.4 Backtracking on BWT indexes	226
	10.4.1 Prefix pruning	227
	10.4.2 Case analysis pruning with the bidirectional BWT index	230
	10.5 Suffix filtering for approximate overlaps	231
	10.6 Paired-end and mate pair reads	232
	10.7 Algorithmic approach to variant selection	233
	10.8 Literature	236
11	Alignment-free genome analysis and comparison	240
	11.1 <i>De novo</i> variation calling	241
	11.2 Space-efficient genome analysis	242
	11.2.1 Maximal repeats	242
	11.2.2 Maximal unique matches	244
	11.2.3 Maximal exact matches	247
	11.3 Comparing genomes without alignment	250
	11.3.1 Substring and k -mer kernels	253
	*11.3.2 Substring kernels with Markovian correction	259
	11.3.3 Substring kernels and matching statistics	264
	11.3.4 Mismatch kernels	272
	11.3.5 Compression distance	274
	11.3.6 Approximating Jaccard similarity using min-hash	276
	11.4 Literature	277
12	Compression of genome collections	284
	12.1 Lempel–Ziv parsing	285
	12.1.1 Basic algorithm for Lempel–Ziv parsing	286

	12.1.2 Space-efficient Lempel–Ziv parsing	287
	*12.1.3 Space- and time-efficient Lempel–Ziv parsing	289
	*12.2 Bit-optimal Lempel–Ziv compression	293
	*12.2.1 Building distance-maximal arcs	297
	*12.2.2 Building the compact trie	300
	12.3 Prefix-free parsing and run-length encoded BWT	301
	12.3.1 Prefix-free parsing	302
	12.3.2 Suffix sorting	303
	12.3.3 Construction of the run-length BWT	304
	12.4 Literature	306
13	Fragment assembly	308
	13.1 Sequencing by hybridization	308
	13.2 Contig assembly	310
	13.2.1 Read error correction	311
	13.2.2 Reverse complements	312
	13.2.3 Irreducible overlap graphs	313
	13.3 Scaffolding	317
	13.4 Gap filling	324
	13.5 Literature	326
	Part V Applications	331
14	Haplotype analysis	333
	14.1 Haplotype assembly and phasing	333
	14.1.1 Minimum error correction	333
	14.1.2 Hardness	335
	14.1.3 Dynamic programming	337
	14.2 Haplotype matching and positional BWT	340
	14.2.1 Haplotype matching in linear time	340
	14.2.2 Positional Burrows–Wheeler transform	341
	14.3 Literature	342
15	Pangenomics	344
	15.1 Overview of pangenome representations	345
	15.1.1 Colored de Bruijn graphs	346
	15.1.2 Founder graphs and founder sequences	348
	15.2 Aligning reads to a pangenome	351
	15.2.1 Indexing a set of individual genomes with a hybrid scheme	353
	15.2.2 Indexing a set of individual genomes with the r -index	354
	*15.2.3 Indexing a reference genome and a set of variants	357
	15.3 Variation calling over pangenomes	359
	15.3.1 Analysis of alignments on a set of individual genomes	360

xii	Contents	
	15.3.2 Analysis of alignments on the labeled DAG of a population	361
	15.3.3 Evaluation of variation calling results	361
	15.4 Literature	362
16	Transcriptomics	367
	16.1 Split alignment of reads	367
	16.2 Estimating the expression of annotated transcripts	369
	16.3 Transcript assembly	372
	16.3.1 Short reads	373
	16.3.2 Long reads	374
	16.3.3 Paired-end reads	379
	16.4 Simultaneous assembly and expression estimation	381
	16.5 Transcript alignment with minimizers and co-linear chaining	386
	16.6 Literature	389
17	Metagenomics	394
	17.1 Species estimation	395
	17.1.1 Single-read methods	395
	17.1.2 Multi-read and coverage-sensitive methods	397
	17.2 Read clustering	401
	17.2.1 Filtering reads from low-frequency species	401
	17.2.2 Initializing clusters	403
	17.2.3 Growing clusters	407
	17.3 Comparing metagenomic samples	408
	17.3.1 Sequence-based methods	409
	17.3.2 Read-based methods	409
	17.3.3 Multi-sample methods	410
	17.4 Literature	410
	<i>References</i>	414
	<i>Index</i>	439