Genome-Scale Algorithm Design

Biological Sequence Analysis in the Era of High-Throughput Sequencing

VELI MÄKINEN
DJAMAL BELAZZOUGUI
FABIO CUNIAL
ALEXANDRU I. TOMESCU

University of Helsinki, Finland
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
### Contents

#### Part I  Algorithms for Graphs

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

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
  8.2.1 Suffix and string sorting 133
8.3 Suffix tree
  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
  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
  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
      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
      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
      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
      11.2.1 Substring and k-mer kernels 229
      *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
### Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>Genome compression</td>
<td>262</td>
</tr>
<tr>
<td>12.1</td>
<td>Lempel–Ziv parsing</td>
<td>263</td>
</tr>
<tr>
<td>12.1.1</td>
<td>Basic algorithm for Lempel–Ziv parsing</td>
<td>264</td>
</tr>
<tr>
<td>12.1.2</td>
<td>Space-efficient Lempel–Ziv parsing</td>
<td>265</td>
</tr>
<tr>
<td>12.1.3</td>
<td>Space- and time-efficient Lempel–Ziv parsing</td>
<td>266</td>
</tr>
<tr>
<td>12.2</td>
<td>Bit-optimal Lempel–Ziv compression</td>
<td>270</td>
</tr>
<tr>
<td>12.2.1</td>
<td>Building distance-maximal arcs</td>
<td>275</td>
</tr>
<tr>
<td>12.2.2</td>
<td>Building the compact trie</td>
<td>278</td>
</tr>
<tr>
<td>12.3</td>
<td>Literature</td>
<td>279</td>
</tr>
<tr>
<td>13</td>
<td>Fragment assembly</td>
<td>282</td>
</tr>
<tr>
<td>13.1</td>
<td>Sequencing by hybridization</td>
<td>282</td>
</tr>
<tr>
<td>13.2</td>
<td>Contig assembly</td>
<td>284</td>
</tr>
<tr>
<td>13.2.1</td>
<td>Read error correction</td>
<td>285</td>
</tr>
<tr>
<td>13.2.2</td>
<td>Reverse complements</td>
<td>286</td>
</tr>
<tr>
<td>13.2.3</td>
<td>Irreducible overlap graphs</td>
<td>287</td>
</tr>
<tr>
<td>13.3</td>
<td>Scaffolding</td>
<td>291</td>
</tr>
<tr>
<td>13.4</td>
<td>Gap filling</td>
<td>297</td>
</tr>
<tr>
<td>13.5</td>
<td>Literature</td>
<td>299</td>
</tr>
<tr>
<td>14</td>
<td>Genomics</td>
<td>307</td>
</tr>
<tr>
<td>14.1</td>
<td>Variation calling</td>
<td>308</td>
</tr>
<tr>
<td>14.1.1</td>
<td>Calling small variants</td>
<td>308</td>
</tr>
<tr>
<td>14.1.2</td>
<td>Calling large variants</td>
<td>309</td>
</tr>
<tr>
<td>14.2</td>
<td>Variation calling over pan-genomes</td>
<td>313</td>
</tr>
<tr>
<td>14.2.1</td>
<td>Alignments on a set of individual genomes</td>
<td>313</td>
</tr>
<tr>
<td>14.2.2</td>
<td>Alignments on the labeled DAG of a population</td>
<td>314</td>
</tr>
<tr>
<td>14.2.3</td>
<td>Evaluation of variation calling results</td>
<td>315</td>
</tr>
<tr>
<td>14.3</td>
<td>Haplotype assembly and phasing</td>
<td>315</td>
</tr>
<tr>
<td>14.4</td>
<td>Literature</td>
<td>322</td>
</tr>
<tr>
<td>15</td>
<td>Transcriptomics</td>
<td>325</td>
</tr>
<tr>
<td>15.1</td>
<td>Estimating the expression of annotated transcripts</td>
<td>325</td>
</tr>
<tr>
<td>15.2</td>
<td>Transcript assembly</td>
<td>329</td>
</tr>
<tr>
<td>15.2.1</td>
<td>Short reads</td>
<td>329</td>
</tr>
<tr>
<td>15.2.2</td>
<td>Long reads</td>
<td>330</td>
</tr>
<tr>
<td>15.2.3</td>
<td>Paired-end reads</td>
<td>335</td>
</tr>
</tbody>
</table>
Contents

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