

Index

- 2-connected, 51
- 3-CNF Boolean formula, 19
- 3-SAT problem, 19
- 3-colorable, 380
- acyclic, *see* directed acyclic graph
- adenine, 4
- affine gap model, 106
- Aho–Corasick, 17
- algorithmic information, 274
- alignment, 8, 84
 - global, 101, 124, 141
 - local, 102, 124, 141, 240, 279, 386
 - semi-local, 102, 110, 219, 220, 226, 237
 - split-read, 367
- alignment on cyclic graphs, 120
- alignment-free comparison, 251, 277
- allele, 7
- alternative spliced transcript, 4
- alternative splicing, 4, 6, 367
- amino acid, 3
- amortized analysis, 11, 291, 405
- approximate membership query, 37, 39
- approximate overlaps, 231, 314
- approximate string matching, 99
- automaton
 - CDAWG, *see* directed acyclic word graph
 - DAWG, *see* directed acyclic word graph
 - deterministic, *see* finite automaton
 - KMP, *see* Knuth–Morris–Pratt
 - nondeterministic, *see* finite automaton
- average-case complexity, *see* complexity
- backward algorithm, 135
- backward search, 179
- backward step
 - bidirectional, 191
- base, 4
- batched locate, 243, 292
- Baum–Welch, 137
- Bellman–Ford algorithm, 47, 62, 63, 325
- biconnected, *see* 2-connected
- bidirectional BWT index, *see* BWT index
- big- O notation, 10
- binary search, 148
- binary search tree, 21
- bipartite graph, 64, 389
- bisulfite sequencing, 8, 221
- bit-manipulation, 18
- Bloom filter, 37, 39, 213, 216
- border, 12, 315
- bottleneck, 55, 56, 59, 78
- bubble, 310, 326, 352
- Burrows–Wheeler transform, 175, 212
 - space-efficient construction, 183, 212
- BWT, *see* Burrows–Wheeler transform
- BWT index, 314
 - bidirectional, 188
 - forward, 190
 - of a DAG, 200
 - of a de Bruijn graph, 206
 - of a graph, 213
 - of a tree, 197, 213
 - bidirectional, 212, 230, 243, 244, 248, 254, 260, 277, 316, 396, 401, 408
 - unidirectional, 192, 212
- case analysis pruning, 230
- Catalan number, 199
- CDAWG, *see* directed acyclic word graph
- central dogma, 5
- Chinese postman problem, 309, 326, 327
- ChIP-sequencing, 8, 221, 236
- chromosome, 4
- circulation, 58
 - minimum-cost circulation, 58
- clique problem, 16, 19
- codon, 5, 83, 111
 - start, 5, 127
 - stop, 5, 127
- colored de Bruijn graph, 346
- compact de Bruijn graph, 364
- complementary base pair, 4
- complexity
 - amortized, *see* amortized analysis
 - average case, xviii, 225, 229
 - worst case, xviii, 10
- composition vector, 251

- compressed de Bruijn graph, 364
- compressed representation, 13
- compressed suffix array, 212
- compression distance, 274
- consensus genome, 6
- contig assembly, 310, 326
- convex function, 64, 78, 382, 386, 391
- copy-number variation, 239, 411
- cosine kernel, 252, 260
- counting sort, 149
- cover
 - cycle, 71, 79
 - edge, 79
 - minimum path, 73, 373, 377, 389
 - path, 73, 374, 378, 380
 - vertex, 16, 125, 127, 234
- covering assignment, 234
- cross-entropy, *see* entropy
- cycle cover, *see* cover
- cytosine, 4

- D_2 statistics, 280
- DAG, *see* directed acyclic graph
- DAG alignment, 117, 118
- DAWG, *see* directed acyclic word graph
- de Bruijn
 - graph, 170, 206, 213, 279, 308
 - sequence, 279, 306
- de novo*
 - sequencing, 8
 - SNP detection, 241, 277
 - variation calling, 311
- decision problem, 15
- deleterious variant, 7
- δ -code, 146, 293
- deoxyribonucleic acid, 4
- depth-first search, 42, 379
- deque, 19
- descending suffix walk, 171
- DFS, *see* depth-first search
- Dilworth's theorem, 77
- dinucleotide, 5
- diploid organism, 7, 8, 219, 233, 309, 311, 333
- directed acyclic graph, 42, 50, 51, 293, 373, 374, 380, 382, 390, 392
- directed acyclic word graph, 169, 172
- distinguishing prefix, 200
- distributivity property, 275
- DNA, *see* deoxyribonucleic acid
- document counting, 165, 396, 410, 411
- donor genome, 219, 221, 359, 361
- double-stranded, 4
- doubling technique, 89
- dynamic programming, 10, 44, 51, 86, 87, 114, 125, 126, 129, 133, 134, 141, 226, 236, 237, 325, 327, 330, 342, 360, 385, 386, 392, 410
 - sparse, 96, 110, 115, 124, 126, 127
- dynamic range minimum queries, 21

- edge cover, *see* cover
- edit distance, *see* Levenshtein distance
- Elias codes, 146, 293
- enhancer module, 6
- entropy
 - k th-order, 280
 - cross-entropy, 265, 278
 - maximum-entropy estimator, 261, 277
 - relative, 104, 127
- enzyme, 3
- epigenomics, 9
- Eulerian
 - cycle, 45, 279
 - path, 45, 309, 326
- evolution, 6
- exact cover with 3-sets (X3C), 400
- exact path length problem, 327, 329
- exact string matching, 11
- exon, 4
- expression level, 6, 367, 370

- finite automaton
 - deterministic, 204
 - nondeterministic, 204
- finite history, *see* Markov chain
- fission, 7
- fixed-parameter tractable, 11, 337
- flow
 - conservation, 54, 57
 - decomposition, 54, 74, 78, 374, 382, 383, 385, 386
 - maximum, 53, 58, 78
 - minimum, 79
 - minimum-cost, 53, 374, 377, 382, 383, 390, 391, 399, 411
 - network, 53, 57
- FM-index, 212
- forward algorithm, 135
- forward BWT index, *see* BWT index
- founder graph, 348
- founder sequence, 350
- four Russians technique, 23, 38, 40, 147
- fractional cascading, 39
- fragment assembly, 8, 104, 166, 231
- full-text index, 145
- fusion, 7

- γ -code, 146, 293
- gap filling, 224, 324
- gapped kernel, 274, 282
- GC content, 130
- gene, 4
 - alignment, 110, 368
 - annotation, 369
 - duplication, 239

- generalized suffix tree, 260
 genetic code, 5
 genome, 4
 genome compression, 284
 genotyping, 352
 germline mutation, *see* mutation
 global alignment, *see* alignment
 Gotoh's algorithm, 108
 guanine, 4
 guide tree, 116
- Hamiltonian path, 15, 18, 51, 321
 Hamming
 - ball, 229, 272, 278
 - distance, 85, 124, 227, 272, 325
 haplotype, 7
 haplotype assembly, 333, 342
 - minimum error correction, 333
 haplotype matching, 340, 342
 hash function, 39
 hash functions, 33
 heterozygous variant, 7
 hidden Markov model, 129, 221, 352, 373
 high-throughput sequencing, 7
 Hirschberg algorithm, 104, 126
 HMM, *see* hidden Markov model
 homozygous variant, 7
 human genome, 5
 Hunt–Szymanski algorithm, 124
- implicit Weiner link, 158
 indel, 6, 100
 indel penalty, 100
 independent set, 19, 116, 125
 indexed approximate pattern matching, 220
 information retrieval, 145
 intron, 4
 invariant technique, 97, 109
 inverse suffix array, 148
 inversions, 7, 239
 inverted file, 145
 inverted index, 145
 irreducible overlap graph, 316
 irreducible string graph, 316, 327
- Jaccard similarity, 256, 276, 278, 280
 Jensen's inequality, 147
 jumping alignment, 123
- k*-errors problem, 99, 220, 226
k-means clustering, 408, 411
k-mer
 - kernel, 253
 - complexity, 253
 - composition, 401
 - index, 145, 229, 236
 - kernel, 251
 - spectrum, 253, 308
- k*-mismatches problem, 99, 227
 Kautz graph, 216
 kernel function, 253
 KMP algorithm, *see* Knuth–Morris–Pratt
O(*kn*) time approximate string matching, 167
 Knuth–Morris–Pratt, 11, 17, 291
 Kolmogorov complexity, 274, 278, 293
*k*th-order Markov chain, 139
 Kullback–Leibler divergence, 104, 127, 259, 266, 281
- labeled DAG, 361
 labeled tree, 195
 LCA
 - queries, *see* lowest common ancestor
 LCE, *see* longest common extension
 LCP array, *see* longest common prefix array
 LCS, *see* longest common subsequence
 learning theory, 278
 least absolute values, 382, 385, 389
 least-squares problem, 370, 381, 386
 Lempel–Ziv
 - bit-optimal, 292
 - compression, 363
 - parsing, 274, 284, 353
- Levenshtein distance, 85, 90, 124, 126, 220
 lexicographic rank, 148
 LF-mapping, 177
 linear gap model, 106
 LIS, *see* longest increasing subsequence
 local alignment, *see* alignment
 log-odds ratio, 103
 long read, *see* read
 longest common extension, 167
 longest common prefix, 288, 298
 longest common prefix array, 160
 - construction, 171
 longest common subsequence, 14, 17, 95
 longest common substring, 145
 longest increasing subsequence, 14, 17
 longest previous factor, 293
 lossless compressor, 274
 lossless filter, 224
 lowest common ancestor, 39, 161
 lowest common ancestor queries, 161
- mapping quality, 220
 Markov chain, 129
 - variable length, 270, 278
 Markovian process, 259
 matching, 65
 - bipartite, 53, 64, 389
 - many-to-one, 67, 69
 - maximum-cardinality, 79
 - perfect, 64, 66, 72

- matching statistics, 264, 265, 269, 272, 278, 396
- mate pair read, *see* read
- MAX-CUT problem, 335
- maximal exact match, 247, 277, 386, 407
- maximal exact overlaps, 314
- maximal repeat, 161, 242, 277, 279
 - k -submaximal, 406, 411
- maximal unary path, 310
- maximal unique match, 164, 241, 244, 277, 279
 - multiple, 245
- maximum-entropy estimator, *see* entropy
- measurement error, 220, 222, 231, 237, 238, 326
- MEM, *see* maximal exact match
- Mercer conditions, 254
- messenger RNA, 5
- metabolic network, 3
- metagenome
 - comparison, 408, 411
 - distributed retrieval, 411
- metagenomic sample, 394
- metagenomics, 9
- metric, 125
- microarray, 308
- min-hash, 278
 - fingerprint, 276
 - sketch, 276
- minimal absent word, 249, 260, 279
- minimizer, 37, 39, 237, 347, 386
- minimum mean cost cycle, 52, 63
- minimum mutation tree, 114, 125
- minimum path cover, *see* cover
- minimum-cost disjoint cycle cover, 322
- minimum-cost flow, *see* flow
- minimum-cost maximum matching, 322
- mismatch kernel, 272, 278, 282
- Morris–Pratt algorithm, *see* Knuth–Morris–Pratt
- multiple alignment, 113, 241
- multiple MUM, *see* maximal unique match
- multiple pattern matching, 17
- multiple sequence alignment, 349
- MUM, *see* maximal unique match
- mutation, 7, 221, 362
- Myers’ bitparallel algorithm, 90, 125, 226

- near-supermaximal repeat, 162
- Needleman–Wunch algorithm, 102
- negative cost cycle, 48
- neighborhood kernel, 274, 283
- next-generation sequencing, *see* high-throughput sequencing
- normalized compression distance, 275, 278
- normalized information distance, 278
- NP-hardness, 15, 16, 19, 55, 78, 116, 127, 234, 321, 322, 326, 335, 380, 385, 392, 393, 400, 410
- nucleotide, 4

- offset network, 383
- overlap alignment, 104
- overlap graph, 105, 310, 313, 326, 373

- pair HMMs, 141
- paired-end read, *see* read
- pangenome, 353, 359
- parsimony score, 114
- path cover, *see* cover
- pathway, 3
- pattern partitioning, 224
- pBWT, *see* positional Burrows–Wheeler transform
- peak detection, 221
- phasing, *see* haplotype assembly
- phylogenetic tree, 266
- pigeonhole principle, 224
- positional Burrows–Wheeler transform, 341, 342, 351
- positional de Bruijn graph, 348
- positional weight matrix, 237
- powerset construction, 204, 213
- pre-mRNA, 4
- prefix doubling, 153
- prefix-free parsing, 301, 306
- primary transcript, 4
- primer design, 268
- probabilistic suffix trie, *see* suffix trie
- profile HMMs, 139
- progressive multiple alignment, 116
- promoter area, 6
- proper coloring, 380
- proper locus, 158
- protein, 3
 - families, 123, 138, 395, 410
 - sequence, 9, 110, 123
- pseudo-polynomial algorithm, 11
- pseudocounts, 137
- PWM, *see* positional weight matrix

- q -gram index, 145
- queue, 12

- r -index, 354
- Rabin–Karp fingerprint, *see* rolling hash
- radix sort, 150, 177, 182
- RAM, *see* random access model
- random access model, 13
- random process, 259
- range counting, 38
- range maximum queries, 110, 124, 126, 388
- range minimum queries, 21, 39, 97, 287
- range queries
 - reporting, 27
 - two-dimensional counting, 27, 233
- rank, 23, 38
- re-alignment, 222, 236
- read, 7

- clustering, 401, 411
- coverage, 221, 222, 308, 311, 327, 337, 360, 369, 394, 397, 401, 404, 406, 407, 410, 411
- error correction, 311, 326
- filtering, 220
- long, 8, 367, 374, 386, 389
- mate pair, 223, 232, 238
- paired-end, 223, 232, 317–319, 321, 322, 329
- pileup, 8, 222, 223, 234, 236
- short, 8
- recombination, 7
- reduction, 14, 16, 53
- reference, 6
- reference taxonomy, 394
- regularization term, 385, 390
- regulation network, 3
- regulatory region, 6
- resequencing
 - targeted, 8, 221
 - whole genome, 8, 219
- residual graph, 59
- reverse complement, 312
- ribonucleic acid, 4
 - sequencing, 8, 367
 - transcript, 96, 367, 368
- RMaxQ, *see* range maximum queries
- RMQ, *see* range minimum queries
- RNA, *see* ribonucleic acid
- rolling hash, 37, 39
- run-length encoded Burrows–Wheeler transform, 304, 306, 354
- run-length encoding, 301, 306
- scaffolding, 317, 327
- segmentation problem, 130
- select, 23, 38
- self-delimiting codes, 146, 293
- self-indexing, 181
- semi-dynamic data structure, 13
- semi-local alignment, *see* alignment
- sequencing, 7
- sequencing by hybridization, 308
- sequencing error, 8
- set cover, 235
- sexual reproduction, 7
- shortest common supersequence, 326
- shortest detour, 88, 125
- shortest grammar problem, 306
- shortest path, 47, 51
- shortest substring array, 268
- shortest unique substring, 282
- shotgun sequencing, 7
- single-nucleotide polymorphism, 6, 222
- sliding window maxima, 110
- sliding window minima, 38, 39
- small parsimony problem, 114, 125
- Smith–Waterman algorithm, 103
- SNP, *see* single-nucleotide polymorphism
- somatic mutation, *see* mutation
- SP score, *see* sum-of-pairs score
- sparse dynamic programming, *see* dynamic programming
- speciation event, 7
- species estimation, 395, 410
 - coverage-sensitive, 398
- splice-site, 369, 373
- splicing, 4
- splicing graph, 369, 373, 381
- split read alignment, *see* alignment
- spurious arc, 310
- stack, 12, 192
- start codon, *see* codon
- stop codon, *see* codon
- strand, 4, 219, 232, 308, 309, 312, 313, 317, 322
- string graph, 327, 328
- string kernel, 250, 277
- string sorting, 151, 314, 328
- structural compression, 313
- structural variant, 7, 352
- subadditivity property, 275
- subdivision (arc), 54
- subdivision (vertex), 73, 370, 382
- subsequence, 84
- subset sum problem, 56
- substitution, 8
 - matrix, 100
 - score, 100
- substitution matrix, 103
- substring complexity, 256
- substring kernel, 254, 257, 264, 268
- succinct
 - data structure, 13, 174
 - de Bruijn graph, 313, 327
 - representation, 13
 - suffix array, 180, 212, 225, 230, 242, 287
 - text indexes, 12
- suffix array, 148
- suffix filtering, 231
- suffix link, 158
- suffix sorting, 154
 - linear time, 154
 - prefix doubling, 153
- suffix tree, 157, 226, 286, 315, 379
 - construction from suffix array, 159
 - Ukkonen’s online construction, 168, 172
- suffix trie, 171
 - probabilistic, 270
- suffix–prefix overlap, 166, 375, 379, 407
- suffix-link tree, 159, 162, 191, 195, 271, 278
- sum-of-pairs score, 113
- superbubble, 326, 328
- supermaximal repeat, 162

- targeted sequencing, 265
- taxonomic composition, 395, 411
- taxonomic marker
 - core, 410
 - crown, 410
- taxonomic rank, 394
 - family, 394
 - genus, 394
- TFBS, *see* transcription factor binding site
- thymine, 4
- tip, 310
- topological ordering, 42, 118
- transcript, *see* ribonucleic acid
- transcription, 4, 367
 - factor, 6
 - factor binding site, 6
- translation, 5
- translocation, 7, 239
- tree, 12
- trie, 157, 171, 270, 298
- tries, 278
- two-dimensional range counting, *see* range queries
- unambiguous contig, 310
- union-find, 405
- unitig, *see* unambiguous contig
- upstream, 6
- uracil, 4
- van Emde Boas tree, 38, 40, 365
- variable-length encoder, 293
- variation calling, 221
 - de novo*, 241, 311
 - evaluation, 362
 - over pangenomes, 359
- vertex cover, *see* cover
- Viterbi
 - algorithm, 134, 135
 - training, 137
- wavelet matrix, 342
- wavelet tree, 25, 38, 178, 181, 197, 202, 208, 210
- weakly connected graphs, 120
- Weiner link, 158, 260, 267
- worst-case complexity, *see* complexity