

# Index

- LF-mapping, 160
- $k$ th order Markov chain, 123
- 2-connected, 39
- 3-CNF Boolean formula, 19
- 3-SAT problem, 19
- 3-colorable, 336
- acyclic, *see* directed acyclic graph
- adenine, 4
- affine gap model, 94
- Aho–Corasick, 17
- algorithmic information, 253
- alignment, 8, 72
  - global, 89, 109, 125
  - local, 90, 109, 125, 220, 257, 343
  - semi-local, 90, 98, 201, 202, 204, 217
  - split-read, 212
- alignment-free comparison, 230, 255
- allele, 7
- alternative spliced transcript, 4
- alternative splicing, 4, 6, 325
- amino acid, 3
- amortized analysis, 11, 269, 361
- approximate overlaps, 209, 287
- approximate string matching, 87
- 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, 119
- backward search, 162
- backward step
  - bidirectional, 173
- base, 4
- batched locate, 222, 270
- batched lowest common ancestor queries, 144
- Baum–Welch, 121
- Bellman–Ford algorithm, 35, 50, 51, 298
- biconnected, *see* 2-connected
- bidirectional BWT index, *see* BWT index
- big- $O$  notation, 10
- binary search, 132
- binary search tree, 20
- bipartite graph, 52, 345
- bisulfite sequencing, 8, 307
- bit-manipulation, 18
- border, 12, 288
- bottleneck, 43, 44, 47, 66
- bubble, 284, 300
- Burrows–Wheeler transform, 158, 194
  - space-efficient construction, 166, 195
- BWT, *see* Burrows–Wheeler transform
- BWT index, 288
  - bidirectional, 171
  - forward, 171
  - of a DAG, 182
  - of a de Bruijn graph, 188
  - of a graph, 195
  - of a tree, 179, 195
  - bidirectional, 195, 208, 222, 223, 227, 233, 239, 255, 290, 352, 357, 364
  - unidirectional, 175, 195
- case analysis pruning, 208
- Catalan number, 182
- CDAWG, *see* directed acyclic word graph
- central dogma, 5
- Chinese postman problem, 283, 299, 301
- ChIP-sequencing, 8, 308, 322
- chromosome, 4
- circulation, 46
  - minimum-cost circulation, 46
- clique problem, 16, 19
- codon, 5, 71, 98
  - start, 5, 111
  - stop, 5, 111
- color-space, 111
- complementary base pair, 4
- complexity
  - amortized, *see* amortized analysis
  - average case, xix, 203, 207
  - worst case, xix, 10

- composition vector, 230
- compressed representation, 13
- compressed suffix array, 195
- compression distance, 253
- consensus genome, 6
- contig assembly, 284, 299
- convex function, 51, 66, 338, 342, 347
- copy-number variation, 218, 367
- cosine kernel, 231, 239
- counting sort, 133
- cover
  - cycle, 59, 67
  - edge, 67
  - minimum path, 61, 330, 333, 345
  - path, 61, 330, 331, 334, 336
  - vertex, 16, 109, 111, 311
- covering assignment, 311
- cross-entropy, *see* entropy
- cycle cover, *see* cover
- cytosine, 4
  
- $D_2$  statistics, 258
- DAG, *see* directed acyclic graph
- DAG alignment, 105, 106
- DAWG, *see* directed acyclic word graph
- de Bruijn
  - graph, 153, 188, 195, 257, 282
  - sequence, 257, 280
- de novo
  - sequencing, 8
  - variation calling, 285
- decision problem, 15
- deleterious variant, 7
- $\delta$ -code, 130, 271
- deoxyribonucleic acid, 3
- depth-first search, 30, 335
- deque, 19
- descending suffix walk, 155
- DFS, *see* depth-first search
- Dilworth's theorem, 65
- dinucleotide, 4
- diploid organism, 7, 201, 283, 285, 307, 310, 315
- directed acyclic graph, 30, 38, 271, 329–331, 336, 338, 346, 348
- directed acyclic word graph, 153, 156
- distinguishing prefix, 184
- distributivity property, 253
- DNA, *see* deoxyribonucleic acid
- document counting, 149, 352, 366, 367
- donor genome, 201, 308, 313, 314
- double-stranded, 4
- doubling technique, 77
- dynamic programming, 10, 33, 39, 74, 75, 102, 109, 110, 113, 117, 118, 125, 204, 216, 217, 298, 300, 303, 313, 323, 341, 342, 348, 366
  - sparse, 84, 98, 103, 108, 110, 111
- dynamic range minimum queries, 20
  
- edge cover, *see* cover
- edit distance, *see* Levenshtein distance
- Elias codes, 130, 271
- enhancer module, 6
- entropy
  - $k$ th order, 258
  - cross-entropy, 244, 256
  - maximum-entropy estimator, 240, 241, 255
  - relative, 92, 111
- enzyme, 3
- epigenomics, 8
- Eulerian
  - cycle, 33, 257
  - path, 33, 283, 299
- evolution, 6
- exact cover with 3-sets (X3C), 356
- exact path length problem, 300, 302
- exact string matching, 11
- exon, 4
- expression level, 6, 325, 326
  
- finite automaton
  - deterministic, 187
  - nondeterministic, 187
- finite history, *see* Markov chain
- fission, 7
- fixed-parameter tractable, 11, 319
- flow
  - conservation, 42, 45
  - decomposition, 42, 62, 65, 66, 330, 338, 339, 341, 342
  - maximum, 41, 46, 66
  - minimum, 66
  - minimum-cost, 41, 330, 333, 338, 339, 346, 347, 355, 367
  - network, 41, 45
- FM-index, 195
- forward algorithm, 118
- forward BWT index, *see* BWT index
- four Russians technique, 22, 27, 28, 131
- fractional cascading, 27
- fragment assembly, 7, 92, 151, 209
- full-text index, 129
- fusion, 6
  
- $\gamma$ -code, 130, 271
- gap filling, 297, 310
- gapped kernel, 252, 260
- GC content, 114
- gene, 4
  - alignment, 98, 212
  - annotation, 325
  - duplication, 218
- generalized suffix tree, 240
- genetic code, 5
- genome, 4

- genome compression, 262
- germline mutation, *see* mutation
- global alignment, *see* alignment
- Gotoh's algorithm, 96
- guanine, 4
- guide tree, 104
- Hamiltonian path, 15, 18, 38, 295
- Hamming
  - ball, 207, 251, 256
  - distance, 73, 108, 206, 251, 299
- haplotype, 7
- haplotype assembly, 315
  - minimum error correction, 316
- heterozygous variant, 7
- hidden Markov model, 113, 307, 308, 329
- high-throughput sequencing, 7
- Hirschberg's algorithm, 92, 110
- HMM, *see* hidden Markov model
- homozygous variant, 7
- human genome, 5
- Hunt–Szymanski algorithm, 108
- implicit Weiner link, 142
- indel, 6, 88
- indel penalty, 88
- independent set, 19, 103, 109
- indexed approximate pattern matching, 202
- information retrieval, 129
- intron, 4
- invariant technique, 85, 97
- inverse suffix array, 132
- inversions, 6, 218
- inverted file, 129
- inverted index, 129
- irreducible overlap graph, 290
- irreducible string graph, 290, 300
- Jaccard distance, 236, 258
- Jensen's inequality, 131
- jumping alignment, 107
- $k$ -errors problem, 87, 202, 204
- $k$ -means clustering, 364, 367
- $k$ -mer
  - kernel, 232
  - complexity, 232
  - composition, 357
  - index, 129, 207, 216
  - kernel, 230
  - spectrum, 232, 282
- $k$ -mismatches problem, 87, 206
- $k$ th-order Markov chain, 124
- kernel function, 232
- KMP algorithm, *see* Knuth–Morris–Pratt
- Knuth–Morris–Pratt, 11, 17, 269
- Kolmogorov complexity, 253, 256, 270
- Kullback–Leibler divergence, 92, 111, 238, 245, 259
- labeled DAG, 314
- labeled tree, 177
- LCA
  - queries, *see* lowest common ancestor
- LCP array, *see* longest common prefix array
- LCS, *see* longest common subsequence
- learning theory, 256
- least absolute values, 338, 342, 346
- least-squares problem, 327, 337
- Lempel–Ziv
  - bit-optimal, 270
  - compression, 216
  - parsing, 214, 253, 262
- Levenshtein distance, 73, 78, 108, 110, 202
- lexicographic rank, 132
- linear gap model, 94
- LIS, *see* longest increasing subsequence
- local alignment, *see* alignment
- log-odds ratio, 91
- long read, *see* read
- longest common prefix, 265, 276
- longest common prefix array, 143
  - construction, 154
- longest common subsequence, 14, 17, 83
- longest common substring, 129
- longest increasing subsequence, 14, 17
- longest previous factor, 271
- lossless compressor, 253
- lossless filter, 203
- lowest common ancestor, 144
  - batched queries, 153
- mapping quality, 201
- Markov chain, 113
  - variable length, 249, 256
- Markovian process, 239
- matching, 53
  - bipartite, 41, 52, 345
  - many-to-one, 54, 56
  - maximum-cardinality, 66, 67
  - perfect, 52, 53, 59
- matching statistics, 244, 247, 250, 256, 352
- mate pair read, *see* read
- MAX-CUT problem, 317
- maximal exact match, 225, 255, 342, 363
- maximal exact overlaps, 287
- maximal repeat, 145, 221, 255, 257
  - $k$ -submaximal, 362, 367
- maximal unary path, 284
- maximal unique match, 147, 221, 223, 255, 257
  - multiple, 223
- maximum-entropy estimator, *see* entropy
- measurement error, 202, 210, 218, 299, 308, 323

- MEM, *see* maximal exact match
- Mercer's conditions, 233
- messenger RNA, 5
- metabolic network, 3
- metagenome
  - comparison, 364, 367
  - distributed retrieval, 367
- metagenomic sample, 350
- metagenomics, 8
- metric, 109
- microarray, 282
- minimal absent word, 227, 239, 257
- minimum mean cost cycle, 40, 51
- minimum mutation tree, 102, 109
- minimum path cover, *see* cover
- minimum-cost disjoint cycle cover, 295
- minimum-cost flow, *see* flow
- minimum-cost maximum matching, 296
- mismatch kernel, 251, 256, 260
- Morris–Pratt algorithm, *see* Knuth–Morris–Pratt
- mr helswerftrtoruenwrbso, *see* Burrows–Wheeler transform
- multiple alignment, 101, 221
- multiple MUM, *see* maximal unique match
- multiple pattern matching, 17
- MUM, *see* maximal unique match
- mutation, 7, 308, 315
- Myers' bitparallel algorithm, 78, 109, 204
  
- near-supermaximal repeat, 147
- Needleman–Wunch algorithm, 89
- negative cost cycle, 37
- neighborhood kernel, 252, 260
- next-generation sequencing, *see* high-throughput sequencing
- normalized compression distance, 254, 256
- normalized information distance, 256
- NP-hardness, 15, 16, 19, 43, 66, 103, 111, 294, 295, 299, 311, 317, 336, 341, 347, 348, 356, 366
- nucleotide, 4
  
- offset network, 339
- overlap alignment, 92
- overlap graph, 93, 284, 287, 300, 329
  
- pair HMMs, 125
- paired-end read, *see* read
- pan-genome, 214, 313
- parsimony score, 102
- path cover, *see* cover
- pathway, 3
- pattern partitioning, 202
- peak detection, 307
- phasing, *see* haplotype assembly
- phylogenetic tree, 245
  
- pigeonhole principle, 202
- positional weight matrix, 217
- powerset construction, 187, 195
- pre-mRNA, 4
- prefix doubling, 137
- primary transcript, 4
- primer design, 247
- probabilistic suffix trie, *see* suffix trie
- profile HMMs, 123
- progressive multiple alignment, 104
- promoter area, 6
- proper coloring, 336
- proper locus, 142
- protein, 3
  - families, 107, 122, 351, 366
  - sequence, 9, 98, 107
- pseudo-polynomial algorithm, 11
- pseudocounts, 121
- PWM, *see* positional weight matrix
  
- q*-gram index, 129
- queue, 12
  
- radix sort, 134, 160, 165
- RAM, *see* random access model
- random access model, 13
- random process, 239
- range counting, 27
- range maximum queries, 98, 109, 110, 344
- range minimum queries, 20, 85, 153, 265
- range queries
  - reporting, 26
  - two-dimensional counting, 26, 211
- rank, 22, 27
- re-alignment, 309, 322
- read, 7
  - clustering, 357, 367
  - coverage, 213, 282, 285, 301, 307, 309, 313, 319, 350, 353, 357, 360, 361, 363, 367
  - error correction, 285, 299
  - filtering, 201
  - long, 8, 325, 330, 342, 345
  - mate pair, 211, 218, 309
  - paired-end, 211, 291–294, 296, 302, 309
  - pileup, 8, 308–310, 322
  - short, 8
- recombination, 7
- reduction, 14, 16, 41
- reference, 6
- reference taxonomy, 350
- regularization term, 341, 346
- regulation network, 3
- regulatory region, 6
- resequencing
  - targeted, 8, 307
  - whole genome, 8, 201

- residual graph, 47
- reverse complement, 286
- ribonucleic acid, 4
  - sequencing, 8, 212
  - transcript, 84, 212, 325
- RMaxQ, *see* range maximum queries
- RMQ, *see* range minimum queries
- RNA, *see* ribonucleic acid
- scaffolding, 291, 300
- segmentation problem, 114
- select, 22, 27
- self-delimiting codes, 130, 271
- self-indexing, 164
- semi-dynamic data structure, 13
- semi-local alignment, *see* alignment
- sequencing, 7
- sequencing by hybridization, 282
- sequencing error, 8
- set cover, 312
- sexual reproduction, 7
- shortest common supersequence, 299
- shortest detour, 76, 109
- shortest grammar problem, 279
- shortest path, 34, 39
- shortest substring array, 247
- shortest unique substring, 259
- shotgun sequencing, 7
- single-nucleotide polymorphism, 6, 308
- sliding window maxima, 98
- small parsimony problem, 102, 109
- Smith–Waterman algorithm, 91
- SNP, *see* single-nucleotide polymorphism
- SOLiD, 111
- somatic mutation, *see* mutation
- SP score, *see* sum-of-pairs score
- sparse dynamic programming, *see* dynamic programming
- speciation event, 7
- species estimation, 351, 366
  - coverage-sensitive, 354
- splice-site, 326, 329
- splicing, 4
- splicing graph, 326, 329, 337
- split-read alignment, *see* alignment
- spurious arc, 284
- stack, 12, 174
- start codon, *see* codon
- stop codon, *see* codon
- strand, 4, 201, 211, 282, 283, 286, 287, 291, 295
- string graph, 300, 302
- string kernel, 229, 255
- string sorting, 134, 288, 302
- structural compression, 287
- structural variant, 6
- subadditivity property, 253
- subdivision (arc), 42
- subdivision (vertex), 62, 326, 338
- subsequence, 72
- subset sum problem, 44
- substitution, 8
  - matrix, 88
  - score, 88
- substitution matrix, 91
- substring complexity, 236
- substring kernel, 233, 236, 244, 247
- succinct
  - data structure, 13, 157
  - de Bruijn graph, 287, 301
  - representation, 13
  - suffix array, 163, 194, 195, 203, 208, 221, 265
  - text indexes, 12
- suffix array, 132
- suffix filtering, 209, 210
- suffix link, 142
- suffix sorting, 138
  - linear time, 138
  - prefix doubling, 137
- suffix tree, 140, 204, 264, 288, 335
  - construction from suffix array, 143
  - Ukkonen’s online construction, 152, 155
- suffix trie, 155
  - probabilistic, 249
- suffix–prefix overlap, 151, 332, 335, 363
- suffix-link tree, 142, 147, 173, 177, 250, 256
- sum-of-pairs score, 101
- superbubble, 300, 301
- supermaximal repeat, 147
- taxonomic composition, 351, 367
- taxonomic marker
  - core, 366
  - crown, 366
- taxonomic rank, 350
  - family, 350
  - genus, 350
- TFBS, *see* transcription factor binding site
- thymine, 4
- tip, 284
- topological ordering, 30, 105
- transcript, *see* ribonucleic acid
- transcription, 4, 325
  - factor, 5
  - factor binding site, 6
- translation, 5
- translocation, 6, 218
- tree, 12
- trie, 140, 155, 249, 256, 276
- two-dimensional range counting, *see* range queries
- unambiguous contig, 284
- union-find, 360

- unitig, *see* unambiguous contig
- upstream, 6
- uracil, 4
  
- van Emde Boas tree, 27, 28, 219
- variable-length encoder, 271
- variation calling, 308
  - evaluation, 315
  - over pan-genomes, 313
  
- vertex cover, *see* cover
- Viterbi
  - algorithm, 118
  - training, 121
  
- wavelet tree, 24, 27, 160, 164, 179, 185, 191, 192
- Weiner link, 142, 240, 246
- worst-case complexity, *see* complexity