INDEX

Entries in bold text refer to a section of the book.

2-break operation *see* DCJ
2-colorability problem 271–272
3-colorability problem 276

abstraction 320
acceptor sites 68, 81
acyclic graphs 70
adenovirus 119, 119, 124
adjacencies 181, 184
adjacency 177
  list 294–295
  matrix 294–295
adjacency-based ancestral reconstruction 213–218
algorithms
  anchors 309
  choosing 333–334
  hashing 259–261
  polynomial-time 180
  rounding 25
  stopping rule 285
algorithms, specific
  DCJSORT 184
  Get-Predecessor-Successor (R) 217
  GRAAL (GRAph ALigner) 309–311
  GreedyReversalSort 175–176
  GRIMM-Synteny algorithm 209
  PathBLAST 309
alignment problem 66–67, 77
alignments **77–80**
  edit distance 173
  local 79
  matches and mismatches 67
  multiple 80
  phylogenetic trees 194
  whole-genome 209
alleles 4, 14
  bi-allelic marker 24
  complex 18–19, 18
  disease 101
  major and minor 24
Alu sequence 40
Amenta, N. *et al.* 261
amino acids 67
  amino terminus 127
  match weights matrix 80
  residues 291
  selection pressure 118
  signals in 129
  substitution matrices 91
analysis of variance (ANOVA) 13
ancestral karyotype reconstruction 211–212
ancestral reconstruction 214–216
  adjacency-based 213–218
  base-level 206–207
  rearrangement-based 212–217
anchors (algorithms) 309
animal influenza viruses 148–164, 155
antiviral drugs 150
approximation algorithms 176
arbitrary dependencies **138–139**
archaea 119, 190–191, 195
arcs (graphs) 44–48, 69–70
association(s)
  association test 16–17
  chromosome populations 14
  common disease 20
  epistasis, effect of 19
  vs. linkage 15–16
  Linkage Disequilibrium 10
Avian Flu 148, 150

Bacillus subtilis 118, 121, 124
backtracking (graphs) 72
bacteria replication 116

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-01146-5 - Bioinformatics for Biologists
Edited by Pavel Pevzner and Ron Shamir
Index
More information

bacterial genomes 113, 116–118
bait protein 293, 298
baker's yeast 303, 309, 311
base (nucleotide) 94, 94
base-level reconstruction 206–207
base-pair 96
Bayes' Law 100, 100–102, 101, 102, 105
Bayesian estimation of species trees (BEST) 254
Bayesian inference 102–103
  arbitrary dependencies **138**–**139**
  MrBayes 253, 254
  prior probability 103, 103, 103, 104
  uninformative priors 103
Bayesian information criterion (BIC) 337
Bayesian model 342
Bayesian posterior probabilities (BPP) 253
Bergeron, Ann 187, 180
BEST (Bayesian estimation of species trees) 254
bi-allelic marker 24
bias 293, 299
BIC (Bayesian information criterion) 337
big cats (*Panthera*) 248–263
binding affinity 140, 151
binding partners 126
binding sites (*see also* TF binding sites)
  clusters 142
  dependencies 143
  identification 143, 141
  positions 138
  prediction 140–141, 140
  search for 140–143
binding specificity 130
Bininda-Edmonds, O. R. P. 254
binomial probability distribution 102
bins (classes) 14
biochemical interaction networks 305
BioGRID database 303
Bioinformatics Algorithms 167
biological function, discovering **303**–**306**
biomolecules 126–127, 127, 151
bipartite graph 184
bipartitions **254**–**258**, 255–256, 258–263
bitstrings 256–263
Blanchette M. et al. 207
BLAST algorithm 309
blood pressure 13–14
Bombyx mori (silkworm) 169
Bonferroni correction 16
Boot Split Distance method (BSD) 195
bootstrapping 195, 253
Borellia burgdorferi 116–118
Boreoeutherian common ancestor 203, 207, 217
Boyd, S. 342
BPP (Bayesian posterior probability) 253
branch-and-bound algorithms 284
breakpoints 168, 177–178, 209, 212, 219
Brenner, Sydney 64
brewer's yeast (*Saccharomyces cerevisiae*) 126, 318
Bruijn, Nicolaas de 52, 55, 63
BSD (Boot Split Distance) 195

cancer 168, 220–221, 220
CARs (continuous ancestral regions) 217
cases and controls 4–5, 16
cats (felids) 229, 248–263, 250, 253
causal loci 8, 15
causal mutation 4–6, 8
cDNAs 319
CDRV (Common Disease Rare Variant) 20
ceiling function 177
cell division tree 192
cells 3–4
cellular interactions 126–129
centralities (networks) 296
Chargaff parity rules 124, 124
Charleston, Michael 234, 245
chimeric protein sets 191
chimpanzees 95, 205, 207
ChIP (Chromatin Immunoprecipitation) 130, 130, 143, 157
Chi-square (χ^2) statistic 135–136, 138
Chi-square (χ^2) test 11
chloroplasts 191
chromatid 157, 207
chromatin structure 143
chromosome painting 212
chromosomes 6–9, 94–95, 168–170, 207–208
  circular 118, 168, 181
  disease 100–101, **219**–**221**
  human genome 211
  intervals 94
  linear 180–181
  mammalian common ancestor 211
  paternity inference 103
  super-chromosome 180
cis-regulatory module (CRM) 142
Citoscape software 295
classes (bins) 14
Classical MultiDimensional Scaling (CMDS) 196–199
Clay Mathematics Institute 48
cluster analysis 196
clustering 296, 297, 297, 304
CMDS *see* Classical MultiDimensional Scaling
coalescent trees 7, 8
coding potential 68
codons 67, 118
coevolution 227, 228–229, 230–235, **233**–**235**, 244
collision (bitstrings) 261
common ancestor 6–7, 203, 207, 211, 217, 250
Common Disease Rare Variant (CDRV) 20
comparative genomics 202–206, 205–207

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-01146-5 - Bioinformatics for Biologists
Edited by Pavel Pevzner and Ron Shamir
Index
More information

comparisons, network 295–300
Complete Genomics 58
computation time 234 (*see also* run times of algorithms)
computational complexity
  large, noisy data sets 316
  objective function 322
  penalizing 337
computational problems 268–277, 320 (*see also* glycan motif finding problem; heuristic solutions; NP-hard problems)
  2-colorability 271–272
  3-colorability 276
  cophylogeny 229–233
  Fixed-tree Maximum Parsimony 278
  genome rearrangements **171**–**175**
  global alignment 77
  machine learning 333
  Median Problem 213
  motif finding 148
  network alignment **306**–**312**
  NP-completeness 76
  optimization 267
  regression 339
  tractable vs. intractable 48–49
"computational thinking" 250
conditional probability 97, 99–100, 138
confounding factors 16, 17–18
Congdon, P. 342
connected graphs 44, 70
consensus
  base 137
  methods 157–158
  model 132
  nucleotides 131, 135–136
  sequence 131–132, 140–141
consensus representation 131–132
consensus tree algorithm 256
consensus trees 248, 250–251, **254**–**263** (*see also* evolutionary histories)
consensus trees, majority 251–252, 254, 256, 258–259, 261–263, 262–263
conserved regions 40
conserved segment 209
constructive proof 61
contingency table 160
continuous ancestral regions (CARs) 217
continuous data (real-valued data) 338–339
controls and cases 4–5, 16
Cooties 228, 229–232
cophylogeny 245
cophylogeny data 241
cophylogeny reconstruction problem 227, 229–233, 232, 234, 239
  Jane software 235, 239
  jungles technique 234
cospeciation events 230–232, 242–243
cost (numerical)
  cophylogeny reconstruction **233**–**235**, 239, 239–241
  phylogeny estimation 278–283
  traveling salesman problem 235–237
  trees 278–283, 284–285
CRM *see cis*-regulatory module
cross-species genomic changes 121–122, 124, 190–193, 207–210 (*see also* horizontal gene transfer)
cumulative GC skew 114–115
cumulative skew diagrams **112**–**124**
cut based network 304
cycles (graphs) *see* acyclic graphs; Eulerian cycle; Hamiltonian Cycle Problem; supercycle
cyclic genomes 49
cytoplasm 4
cytosine nucleotide (C) 23, 118, 119

D-statistic 10
Dantzig, George 32
Darwin, Charles 189, 228, 245, 249
data
  de-noising 303
  noisy 293, 316
  normalized 319
  real-valued (continuous) data 338–339
data collection 293, 303
data sources, combining 339–341
databases
  BioGRID 303
  DIP (Database of Interacting Proteins) 292
  GenBank 113, 130, 253
  HPRD 303
  JASPAR 130
  largest molecular 253
  sequence data 268
  TRANSFAC 130
Davis, B.W. *et al*. 251, 253, 254
DCJ (double-cut-and-join) model 180, 180, 184, 218
DCJSORT algorithm 184
de Bruijn graphs 52–54, 61 (*see also* directed graphs)
de Bruijn, Nicolaas 52, 55, 63
deamination 118–119, 119
degree distributions 296, 300–302
degree of a node 296, 303, 304
degree of a vertex 43–45
deoxyribonucleic acid (DNA) *see* DNA entries
dependencies, arbitrary **138**–**139**
depth-first traversal 257
d_{HP} distance 180, 183
diameter of a network 297
dinucleotides 119, 134
DIP (Database of Interacting Proteins) 292
directed graphs 45–47, 59–60 (*see also* de Bruijn graph)

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-01146-5 - Bioinformatics for Biologists
Edited by Pavel Pevzner and Ron Shamir
Index
More information

diseases
  alleles 101
  cancer 168, 220–221, 220
  carriers 100
  chromosomal aberrations **219**–**221**
  complex 16
  development 167
  estimating risk 100–102
  genes 100
  parasites 245
  proteins 304
  recessive 100
  SNPs 94
  tests 98–99, 303
distance matrix 193, 196
distance metrics 171
  BSD method 195
  d_{DCJ} distance 183
  d_{HP} distance 180, 183
  edit distance 171–173, 173
  genome rearrangement **171**–**175**
  minimum-evolution method 252
  reversal distance 212–213
distribution degree 300–302
distribution law 84
diversification 250
DNA (deoxyribonucleic acid) 167–168
  cDNAs 319
  double-stranded DNAs 119, 119, 124, 167–168
  fragments 56–57
  horizontal transfer 121–122, 124
  motif 157
  replication 111, 191
  signals 129
  single-stranded 118
  structure 124
DNA and RNA, regulatory interactions 316–319
DNA sequencing 23, **36**–**40**, 63
  Complete Genomics 58
  the early days 49–50, 56
  largest molecular database 253
  modeling regulatory motifs 130
  motif finding problem 157
  next generation technologies 58
  and the overlap puzzle 36–40
  phylogeny estimation 277–285
  sequencing machines 40, 55
  WebLogo 133
Dobzhansky, Theodosius 6, 173–175, 207, 221
dogs 213–214
Dollo parsimony model 253
donor sites 68, 81
dot-plot 170–171, 180
double-stranded DNAs (dsDNAs) 119, 119, 124, 167–168
Double-Cut-and-Join see DCJ
drift, genetic 8, 95
Drmanac, Radoje 55–56, 58, 64
Drosophila pseudoobscura (fruit fly) 142, 173, 174, 207
drugs 150, 304, 305, 305–306
dsDNAs *see* double-stranded DNAs
Duffy locus 17–18
duplication events 242–243
dynamic programming 66–92, 91, 239, 282

"earthquakes" (genomic) 208
ECPs *see* Eulerian Cycle Problems
edge lists (adjacency lists) 294–295
edges (trees) 229, 239
edges (vertices) 271–272, 291
edit distance 171–173
efficiency of a method (*see* computational complexity; time complexity)
endosymbiotic events 191–192
epidemics 148
epistasis 18–19
epithelial cells 150, 155
equivalence of conditions 45, 59
Erdos–Renyi random graph model 300
Escherichia coli 116, 118, 118, 190–191
ethnicity 17–18
eukaryotes 68, 128, 142, 191–192, 207
Euler, Leonhard 40, 55, 63
Eulerian assembly 58
Eulerian cycle 45–48, 53–54, 60–61
Eulerian Cycle Problem (ECP) 43–44, 49, 50–52
Eulerian graphs 54
Eulerian path 45
Euler's Theorem for directed graphs 44–48, **58**–**61** (*see also* Königsberg Bridge Problem)
  Theorem I 45–47, 58, 59–60
  Theorem II 47, 59–61
evolution 111, 268
  and alignment 77
  mammalian 203–204
  and mutagenesis 119
  rates of 303
  simulation of 237
evolutionary conservation 142
evolutionary histories 250–251, 267 (*see also* consensus trees)
evolutionary trees 173, 248, **250**–**254**, 268, 277–286 (*see also* phylogenetic trees; phylogenies)
exhaustive searches 273, 274–276, 282–284
exons 68, 68, 81–83, 81, 81–82

false-negative error rate 16
false-positive error rate 16
family traits 15, 101, 168

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-01146-5 - Bioinformatics for Biologists
Edited by Pavel Pevzner and Ron Shamir
Index
More information

fast solutions 233, 234, 236, 259–261, 303 (*see also* heuristic solutions)
feasible region 31
felids (cats) 229, 248–263, 250, 253
figs (*Ficus*) 228, 228, 228–229, 241–243
finches (Estrildidae) 228–229, 241, 244
Fisher's exact test 152, 160
fissions 180, 218
Fitch's method 214–216, 216–217
fitness 111–112, 142, 236, 239–241
Fixed-tree Maximum Parsimony problem 278
flow-based network 304
fluorescence 57, 96, 103, 319
forensic DNA tests 94, 96
Forest of Life (FOL) analysis 193–199
FRAG_NEW 252
fragment assembly 37–40, 49–50
  directed graphs 45
  Eulerian Cycle Problem 50–52
  Hamiltonian Cycle Problem 49–51
  read multiplicities 54
Frank, A. C. 118
Frontiers at the Interface of Computing and Biology (NRC Committee) 250
fruit fly (*Drosophila pseudoobscura*) 142, 173, 174, 207
F-test 14
fungi 311
fusions 180, 181, 208, 218, 220

galactose (Gal) 126–127, 155
gap penalties 80
gaps in a network pathway 309
gaps in sequences 66–67, 206–207
Gaussian bell curve 338–339
GBPs *see* glycan binding proteins
GC-skew (guanine–cytosine) 112, 113–114, 114–115, 118, 119, 119
GDV (graphlet degree vector) 304–305, 309–310
GenBank 113, 130, 253
gene expression 58, 139, 318–319, 342
gene mapping 209
gene order data 212
gene pairs, corresponding 179
gene permutations 175–178
gene recognition 67, 68, 68, **81**–**83**, 91
generalized random graphs model 300
genes 4
genetic algorithms 234–237, 242, 244
genetic code *see* genotype
genetic fingerprint 95
genome assembly 49
genome rearrangement problem **171**–**175**, 173, 186 (*see also* rearrangements)
Applications of Genome Rearrangements 187
genome reconstruction 205–207
genome sequencing 37, 56, 190–193, 202–204, 209 (*see also* DNA sequencing)
genome sequencing projects 202, 220, 268
genome sorting problem 175–176
genomes 118, 119, 167–168, 173, 191
genomics (*see also* Tree of Life (TOL))
  changes 112, 112–113
  comparative 202–206
  "earthquakes" 208
  genomic anchors 179
  RNAi functional 305
genotype (genetic code) 3, 4–5, 5, 14, 94
genotyping cost, tag SNPs 24
Genscan software 135
geometric graphs 301–302, 302, 303
Get-Predecessor-Successor (R) algorithm 217
Gibbs sampling algorithm 334
Gilbert, Walter 38, 63, 113
global alignment problem 77
global network alignments 307–308
global optimum solution 285
global polarity switch 113–115
global properties (networks) 295–296
global-alignment (sequences) 77, 91
glycan arrays 148, 151, 156–157, 160, 161, 163
glycan binding proteins (GBPs) 155, 156, 158
glycan motifs 148, **157**–**161**
glycan structures 152–153, 160
glycans 153, 156–157
glycans and hemagglutinin interaction 148, 150–151, 151, 156–157, 156–157
glycans ligands 156–157
glycobiology 151, 163
glycoconjugates 153
glycoproteins 150, 153
glycosidic bond 152–153
glycosylations 153
GRAAL (GRAph ALigner) 309–311
graph isomorphism 295–296
graph theory 43, 63
GraphCrunch software 295, 300
graphlet count estimation 303
graphlet degree vector (GDV) 304–305
graphlets 298–299
graphs 43–48, **69**–**70**, 271–272, 291 (*see also* Eulerian graphs; networks)
  arbitrary dependencies 138
  binding site prediction 140–141
  connected 44, 70
  de Bruijn graphs 52–54, 61
  directed 45–47, 59–60
  exon–intron 81–82
  geometric 301–302, 303
  hypergraphs 86
  oriented 69
  RIGs (residue interaction graphs) 291

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-01146-5 - Bioinformatics for Biologists
Edited by Pavel Pevzner and Ron Shamir
Index
More information

segment 82
supercycle 60–61
greedy algorithms (greedy heuristics) 26, 28–30, 72, 236, 284, 308
GreedyReversalSort algorithm 175–176
GRIMM-Synteny algorithm 209
Groodies 229–232
guanine nucleotide (G) 23, 112–113, 119, 121
guilt-by-association (GBA) 334

H protein *see* hemagglutinin
H1N1 virus 150–151
Haeckel, Ernst (19C) 189
Haemophilis influenzae 63, 113, 115, 121
Hamilton, William 41, 55
Hamiltonian Cycle Problem (HCP) 43–45, 45, 48–50, 49, 50, 63
Hannenhalli, Sridhar 140, 180, 212
haplotype block 24, 25, 26, 27, 31–32
Hardy–Weinberg equilibrium 15
harmonic series 30
hashing 259–261
Hb (TF protein) *see* Hamiltonian Cycle Problem
Helicobacter pylori 121–124, 122, 122, 309
helix–coil transitions 91
hemagglutinin (HA) 148, 149, 155, 163–164
hemagglutinin–glycans binding specificity 155
hemagglutinin–glycans interaction 148, 150–151, 151, 156–157
Hemmer, H. 254
hemoglobin 94
hemophilia B 128, 128
heterozygous SNP 94
heuristic solutions 234, 276, 295, 303, 333–334 (*see also* fast solutions; greedy algorithm; NP-hard problems)
maximum parsimony 284–286
multiple genome rearrangements (MGR) 213
PAUP* software 253, 256
phylogeny estimation 267
stopping rule 285
HGP (Human Genome Project) 202, 220
HGT (horizontal gene transfer) 121–122, 124, 190–193, 195–198, 232
hidden event 100
Hidden Markov Models 91
hidden variables 98, 102, 103
higher-order PWM 134–135
high-LD regions *see* haplotype blocks
hill-climbing heuristics 334
Histone modifications 143
HIV 229, 232, 245
homologous gene sequences 171
homologous proteins 79, 309
homologous recombination 95
homologs (homologous traits) 191
homology 305
homozygous SNP 94
horizontal gene transfer (HGT) *see* HGT
host species 227
host specificity 155
host switches 148, 151, 155, 232, 239, 242–243
host trees (host phylogenies) 229, 238–239
HPRD database 303
HPV-IA 120
hubs (nodes) 297
human (*Homo sapiens*) 63, 169, 202–204, 207, 211, 214
chromosomes 207, 211
disease causes **219**–**221**
epithelial cells 155
influenza viruses 155, 161
population patterns 24
Human Genome Project (HGP) 202, 220
human viruses
adenovirus 119, 119, 124
cytomegalovirus 119
influenza virus 155, 161
hypergeometric distribution 160
Hyseq 58

Icosian Game 41–42, 43–44, 48
in vivo identification of binding sites 130
indegree of a vertex 47–48
indigo birds 228–229, 241, 244
induced subgraph 298–299
inference *see* network inference; paternity inference; regulatory network inference
inference (statistical) 342 (*see also* Bayesian inference)
infinite sites assumption 8, 10
influenza virus
animals to humans 148–164, 155
classification 150
host specificity 155
human 155, 161
strains 155
switches 148, 151, **151**–**157**, 155
transmission efficiency 155
types 149
vaccines 150
virion 149
Information Content 133
inheritance
chromosomes 6
DNA 4
natural selection 111, 237, 244
recessive 17
SNPs 94
insertion and deletion events 168, 207, 211
integer programming **30**–**32**, 31
integral constraint 32

integration, numerical 114
interaction maps 302
interaction specificity 127, 156
interactome detection 302–303
intergenic regions *see* adjacencies
International Union of Pure and Applied Chemistry (IUPC) 132
intractable problems 48–49, 213 (*see also* NP-completeness)
introns 67–68, 68, 81–83, 253
inversions *see* reversals
isomers 153
IUPC *see* International Union of Pure and Applied Chemistry (IUPC)

jaguar (*P. onca*) 248–263
JAK-STAT signal transduction pathway 127
Jane software 235, **237–245**
Janecka, J. E. *et al.* 254
JASPAR database 130
Johnson, W. E. *et al.* 252, 253
joint probability 97–98, 326
Jones, Neil 167
jungles technique 234

K12 genome 190–191
karyotypes 207, 211–212
Königsberg Bridge Problem 40, 43–45, 63

lagging DNA strand 116, 118, 120, 121
Laplace prior 133
large data sets 316
large populations 244
LD *see* Linkage Disequilibrium (LD)
leading DNA strand 116, 118, 120, 121
leaf nodes 152
least-cost solutions, dynamic programming 239
leopard (*P. pardus*) 248–263
Levy, S. 140
lice 228, 230, 241
ligands, glycans 156–157
likelihood models 102, 103, 340
likelihood of a model, model likelihood 106, 323–324, 333–334
linear chromosomes 116, 168, 180–181, 207
linear constraints 30–32, 31, 31
linear programming 30–32
linear regression 339
linkage 8, 15–16, 95, 152
Linkage Disequilibrium (LD) 10–12, 12, 15, 15, 18, 20, 24
Linkage Equilibrium 10, 15
links (graphs) 291
lion (*P. leo*) 248–263
l-mer 49–50, 54, 56, 58
Lobry, J.R. 118
local alignments (sequences) 79, 91
local network alignments 307
local network properties 298–300
loci (genetic) 4
  alleles 14
  causal 8, 15
  complex alleles 18–19
  Duffy locus 17–18
  orthologous gene 209
  polymorphic 14
logarithmic approximation ratio 32
Logo representation 133
long-range LD 15, 18
loops (graphs) 69
loss events 242–243
low-LD regions *see* recombination hotspots
l-treelet, glycan motif 158–161
l-tuple DNA motif 157

Ma, J. *et al.* 209–210, 213, 218
machine learning 316, 333, 342
MAF *see* minor allele frequencies (MAF)
major alleles 24, 25
majority consensus trees 251–252, 254, 256, 258–259, 261–263, 262–263
malaria 94, 245
mammalian genomes 202–204
Margoliash, Emanuel 190
markers 4, 15, 17–19, 175, 176, 209
Markov chain Monte Carlo method 334, 342
Markov model 135, 138
mass spectrometry 163
Massively Parallel Signature Sequencing (MPSS) 64
match scoring 140
MATCH software 140
match weights 80
matches 67
matching *see* alignment; sequences, similarity of
matrices (*see also* Position Weight Matrix)
  adjacency matrix 294–295
  amino acid match weights 80
  matrix vs. star models 293
  polymorphisms 8
  probability **132–135**
  tree distance matrix 193
matrix technique 91
Maxam, A. W. 113
maximum common subsequence problem 80
maximum common subword problem 80
maximum independent set (graphs) 274–276
maximum likelihood methods 195, 253, 323–328
maximum parsimony 277–286, 284–286
MDD (Maximum Dependence Decomposition) **135–138**
mean 338

CAMBRIDGE

Cambridge University Press
978-1-107-01146-5 - Bioinformatics for Biologists
Edited by Pavel Pevzner and Ron Shamir
Index
More information

Median Problem 213
meiosis 8
Mendel, Gregor J. 4
mental health 94
Merkle, Daniel 234
Methanosarcina (archaea) 191
methods (computational)
  Boot Split Distance 195
  clustering 304
  evolutionary histories 268
  exhaustive searches 273, 274–276
  Fitch's method 214–216
  guilt-by-association (GBA) 334
  jungles technique 234
  Markov chain Monte Carlo 334, 342
  maximum likelihood 195, 253
  minimum-evolution 252, 252
  MPSS 64
Metropolis–Hastings algorithm 334
MGR (multiple genome rearrangements 213
mice 202–203, 205, 207, 213, 214
microarrays 64, 96, 318–319 (*see also* nanoball arrays)
  analysis 157
  gene expression 342
  how they work 56–58
  paternity inference 103, 106
  probe sequence 96
microbial genomes 118, 119, 190, 193
microchips 5, 8, 15, 55–58, 56–58 (*see also* microarrays)
Middendorf, Martin 234
Millennium Problems 48
minimization 30–32
minimum cost reconstructions **233–235**
minimum test collection problem 25–26
minimum-evolution method 252
minor allele frequency (MAF) 16, 24
minor alleles 24, 25
Mirzabekov, Andrey 55–56, 64
mismatches
  in an alignment 67
  base pair strands 118
  in a network pathway 309
  pairs of amino acids 80
missing data 33
mitochondria 119, 124, 191, 253, 254
Mixtacki, Julia 167
model likelihood 106, 323–324, 333–334
modeling software *see* software packages
models (*see also* algorithms)
  classes of 317
  computational thinking 250
  likelihood models 102, 103, 340
  machine learning 316, 333, 342
  network **300–303**
  sensitivity 143
  sequence-based 143
modulo function 260
molecular dynamics 164
monosaccharide residues 153
monosaccharides 152–153, 157, 160
most recent common ancestor (MRCA) 6–7 (*see also* common ancestor)
motifs, regulatory 126–143, 133, 148
MPSS *see* Massively Parallel Signature Sequencing
MrBayes software 253, 254
MRCA (most recent common ancestor) 6–7, 15
mRNA 116, 128
mtDNA 252
multiple alignments 80, 193, 194, 206–207
multiple chromosomes **180–185**, 180
multiple genome rearrangements (MGR) 213
Murphy, W. J. 253, 253
mutation pressure 118, 119
mutations 4–6, 8
  drift 8, 95
  Factor IX 128
  genome rearrangements 168
  hemagglutinin 155
  point mutations 168
  single nucleotide 24
  SNPs 18–19, 95
  spontaneous deamination 118
  transcription-induced 120
mutualism 228

Nadeau, J. 209
nanoball arrays 58
National Research Council (NRC) 250
natural selection 111, 237, 244
Naughton, B. T. *et al*. 141
nDNA (nuclear genes) 253
nearest neighbor interchange (NNI) 284–285
Nearly Universal Trees (NUTs) 194, 195–199, 195–198
negative (purifying) selection 142, 202
negative skew 115
Neighbor Joining algorithm 252
neighbors (nodes) 297
neighbors (treespace) 252, 284–285, 285
Neofelis (clouded leopard) 248–263
"net of life" 190
network alignment **306–312**
network alignment algorithms 308–309
  analysis software 295, 300
  comparisons 295–300
  diameter 296
  flow 304
  growth 302
  inference 321, 334
  models **300–303**

network alignment algorithms (*cont.*)
  motifs 298–300
  projections 305
  properties 296
  structure 298–299
  topology 296, **303**–**306**, 311–312
networks 291 (*see also* graphs)
neuraminidase (N) gene 150, 150
“newspaper problem” 36–40
NNI (nearest neighbor interchange) 284–285
Nobel Prize 38
node degree (graphs) 296, 303, 304
nodes (graphs) 138, 229, 239, 291
noisy data 293, 303, 316, 323
non-coding regions (introns) 67–68, 168, 253
non-consensus nucleotides 135–136
non-oriented paths (graphs) 69
non-trivial bipartitions 255
normalization 98, 100, 135, 319
normally distributed data 338
NP-completeness 48, 76, 296
NP-hard problems 268–277, **275**–**277**, 283
  cophylogeny reconstruction 234
  genome sorting 176
  integer programming 32
  tag SNP selection 26
  traveling salesman 236
nucleic acids 112, 127, 152, 161, 319
nucleosomes 143, 143
nucleotide(s) 4, 167–168
  bases 94, 130
  combination letter codes 132
  consensus 131, 135–136
  counting 112–113, 124
  non-consensus 135–136
  relative frequencies 112
  string of (*l*-mer) 49–50
  substitutions of 168
null hypothesis 232–233
NUTs (Nearly Universal Trees) 194, 195–199

objective function 30, 322
observed event 100
observed variables 98, 102, 103
odds ratio (OR) 19, 105, 107
Okazaki fragments 116, 118
oligosaccharides 151, 153, 161
O(n2) time 271, 272
operations, counting 270–271
optimization problems 267, 277–286, 342
orderings 236–237, 237–241
organelles 4, 119
organismal trees 190
oriented graphs 69
origin (ori) of replication 115–116, 118, 122
Origin of the Species 189, 228, 249
orthologous genes (orthologs) 157, 193–194, 209
outdegree of a vertex 47–48, 53
overfitting 337
overlap puzzle 36–40
Oxford Gene Technology 58

p-value 11
paired-end reads 220–221
pandemics 148–149, 155
Panthera genus 248–263
papillomavirus 120
PAR *see* Population Attributable Risk
parasite tree 229, 239
parasites 227, 229, 245
parasitism 228, 229
parents 4, 6–8, 236–237, 237
parsimony 172, 213, 214, 218, 253, 277–286
partial subgraph 298–299
partition function 85
partitioning 18
paternity inference 96–107, **103**–**107**, 104–105, 106–107
paternity tests 93–94, 96
path score (graphs) 70
PathBLAST algorithm 309
pathogenic strain genomes 190
pathogenicity islands 122, 191
paths (graphs) 45, 69, 69, 70, 71
pattern matching *see* optimal alignment
Pauling, Linus B 190
PAUP* software 253, 256
penalties (negative weights) 67
permutations (gene) 175–178
Pevzner, Pavel 167, 180, 209, 212
pharmacology 305–306
phenotypes 3–5, 8, 12, 12–14, 190
phylogenetic analysis 212, 268
phylogenetic footprints 142
phylogenetic trees 193–199, 251–254, 267 (*see also* evolutionary trees; phylogenies)
  bipartitons 255–256
  coevolution 227
  early 189
  edit distance 173
  Fitch’s method 214–216
  Groodies and Cooties 229–234
  mammalian comparative genomics 203–204
  maximum likelihood methods 195
  pantherines 249, 251–254
  phenotypes 190
  phylogenetic relationships 306
  topology comparison 195
phylogenetics 248
phylogenies 248–250 (*see also* evolutionary trees; phylogenetic trees)
  estimating 267, **277**–**286**

GRAAL 311
host 238–239
MrBayes 253, 254
phylogenomics 192, 193–195
pigs 150, 155
pocket gophers 228, 230, 241
point mutations 168, 171, 171–173
points, related (graphs) 301
Poisson-distribution 15
polarity switch, global 113–115
pollination 228, 241
poly A sites 129
polymerase enzymes 128, 317
polymers **83**–**86**
polymorphic locus 14
polymorphic markers 4
polymorphisms 8, 12 (*see also* SNPs)
polynomial-time algorithm 180, 268–269, 271, 276
polyten chromosome reversals 173
Population Attributable Risk (PAR) 19
population size 241, 244
population substructure 17–18
Position Weight Matrix (PWM) **132**–**135**
binding site positions 143
binding site prediction 141
binding sites search 140–143
higher-order PWM 134–135
positive skew 115
posterior probability 103, 105
power 16–17, 16
power-law 296, 301–302
PPI networks 291–292, 294–295, 298, 302–306, 304, 305
predecessor synteny block 216–217
premiums (positive weights) 67
prey proteins 293, 298
primates, non-human 229, 245
prior probability 103, 103, 103, 104, 132, 335–337
probability
BPP 253
conditional 97, 99–100, 138
density function 338
distributions, binomial 102
joint 97–98
machine learning 316
matrix 132
models 323–324
Position Weight Matrix (PWM) 141
unconditional 97, 99
problems *see* computational problems
profile methods 157–158
prokaryotes 68, 190, 191–192
promoter region 317
protein function prediction 303
protein structure networks 291
protein-binding DNA microarrays 129, 157
protein-coding regions 67–68 (*see also* exons)
protein–DNA interaction 130
protein-/non-protein coding regions 67–68
protein–protein interactions *see* PPI networks
proteins 4, 152
chimeric 191
connectivity 303
disease-related 304
identifying features 127
regulatory 127
structure 161
trans-membrane 127
protists 311
pseudo count *see* prior probability
pull-down experiments 302
purifying selection 142, 202
PWM *see* Position Weight Matrix (PWM)

quadratic time 269

random walk (graphs) 60
randomized rounding 32
rare variants (RVs) 19–20
rats 213, 214
read generation 37–38, 49, 55
reads 37, 54–55, 220–221
real-valued data 338–339
rearrangements 186
ancestral reconstruction 212–213
fission and fusion 180, 208
inversions (reversals) 208
large-scale 207–210, 211
operation types 181, 208
recessive disease 100
recessive inheritance 17
reciprocal translocation 208
recombination events 8, 10, 15, 24, 95
reconstruction *see* ancestral genome reconstruction; cophylogeny reconstruction problem
reconstructions 230–235
recursive algorithm 158
regression 339
regulation 128, 317–319, 322
regulatory DNA and RNA interactions 316–319
regulatory motifs 126–143, 133
regulatory networked inference **337**–**338**
regulatory networks 139, 299, 315–342, 316
regulatory regions 142, 157
relative entropy 133
relative nucleotide frequencies 112
relative risk (RR) 12
replication 112, 119
DNA 111
fidelity 111
mechanism 115–116

replication (*cont.*)
  origin (ori) 118
  terminus (ter) 115–116
  and transcription 118, **120–124**
residue interaction graphs (RIGs) 291
residues 153, 157, 291
resolution of synteny blocks 209
respiratory system 150, 155
Restriction Fragment Length Polymorphisms (RFLP) 252
reversal distance 212–213
reversals (inversions) 170
  cumulative skew, HGT 122
  DCJ model 181, 218
  phylogeny reconstruction 173
  polytene chromosome 173
  signed reversals **178–180**
  sorting by reversals 212
  unsigned reversals **175–178**
reverse transcription 319
RFLP (Restriction Fragment Length Polymorphisms) 252
rhesus genome 214
r statistic 11
RIGs (residue interaction graphs) 291
RNA (ribonucleic acid) 4, 86, 111, 112
  DNA interactions 316–319
  folding 239
  regulatory interactions 316–319
  secondary structures 91
  viruses 119, 149
RNAi functional genomics 305
Robertsonian translocation 208
rooted trees 251, 256
rounding algorithm 25
rRNA 190, 191
run times of algorithms 234
  3-colorability 276–277
  estimating 270–271, 282–283
  heuristics 285–286
  polynomial-time 268–269, 271
  stopping rule 285
RVs *see* rare variants (RVs)

Saccharomyces cerevisiae (brewer's yeast) 126, 318
sampling issues 16–17
  bias 293
  correcting for unobserved data *see* prior probability
  with DNA microchips 8
  sample size 16, 138
  under-sampling 14
Sanger, Frederick 38, 56, 63, 113
Sankoff, D. 212
scale-free network 296, 300
Science (1988), DNA arrays 57
scoring 67
  alignment scores 77
  match scoring 140
  optimum solutions 285
  paths 70
  sequences 141
search algorithms 309
seed-and-extend approach 309–310
segment graph 82
segmental duplication 211
segmentation (sequence) 67
segregating sites 4
selection algorithms 258–259, 259–261
SELEX 129
sensitive models 143
sequence analysis
sequence insertions 121–122
sequence windows 113, 114
sequence-based models 143, 143
sequenced genomes 112
sequences 66, 66–67, 122, 131–132, 309
sequencing machines (DNA) 40, 55
serotonin 94
set-covering problem 26–30, 26, 28–30
Sex Life of Flowers 228
shortest path algorithms 234
sialic acids 155, 157
sickle-cell anemia 17–18, 94
signaling molecules 151
signatures (molecular) 126
signed permutations 180
signed reversals **178–180**, 180
silk worm (*Bombyx mori*) 169
simian virus 113
simplex algorithm (Dantzig) 32
simulation of evolution 237
single-copy genes 253
single-nucleotide polymorphisms *see* SNPs
single-stranded DNA 118
sink vertex 70
skew 118, 121 (*see also* GC-skew)
skew plot 113
skewed distributions 296
skin color 17
small-world networks model 300
snow leopard (*P. uncia*) 248–263
SNPs (single-nucleotide polymorphisms) 4, 93, 94–96 (*see also* haplotype blocks; tag SNPs)
  and disease 6, 8, 15
  paternity inference 103, 104–105, 106–107
software
  Citoscape 295
  FRAG_NEW 252
  Genscan 135
  GraphCrunch 295, 300
  heuristic 234
  Jane 235, **237–245**

  MATCH 140
  MrBayes 253, 254
  PAUP* 256
  Tarzan 234
  TreeMap 234
  WebLogo 133
software run times *see* run times of algorithms
sorting 175–176, 176, 184, 212, 258–259
source vertex 70
Southern, Sir Edwin 55–56, 58, 64
Spanish flu 148
speciation events 229, 230–231, 230–231, 237–241, 239
species trees 190
specificity of interaction 127, 156
splice sites 68, 81, 129
splicing 68, 128
spoke models 293
standard deviation 338
star trees 254, 261–263
statistical hypothesis testing 232
statistical inference 342
statistical tests
  of association 12
  case-control test 16
  Chi-square (χ^2) test 11
  correlation between two events **9**–**12**
  Fisher's exact test 152, 160
  F-test 14
Stats Mode (Jane software) 244
Stephens, P. J. *et al*. 221
stopping rule (algorithms) 285
strict consensus trees 251, 254, 256
string of (*l*-mer) 131
Student's *t* distribution 13, 14
Sturtevant, A. H. 173–175, 207
subgraphs 158, 296
subpath (graphs) 71
substitutions (nucleotides) 168
substrate specificity 127
successors 216–217
Sul, S.-J. *et al*. 261
super-chromosome 180
supercomputers 233
supercycle (graphs) 60–61
superstring nucleotides) 49–52
Swine Flu 148–149, 150–151, 150
synteny blocks 179, 209–210, 213, 216–217
systems biology 316
systems pharmacology 305–306

t-statistic 14
tag SNPs 23–33, 24, **25**–**29**, 31, 33, 33
tandem duplication 211
tanglegram 230
Tarzan software 234
Taylor, B. 209
telomeres 168, 176, 180–181, 207
terminal residues 157
terminus (ter) 115–116, 118
Tesler, Glenn 209
test accuracy 98, 99
test for associations 14
tests, statistical *see* statistical tests
test's power 12
tetraglucose 152
TFs (transcription factors) 126, 127–130, 142–143, 317–318
TF binding sites (TFBS) 127, 317 (*see also* binding sites)
  additional hallmarks 141–143
  destruction 19
  identification 141, 143
  models 129–134
  multiple rare variants 19
TF proteins 142, 143
TF-DNA 128–129, 133, 133, 141, 143
tiger (*P. tigris*) 248–263
time complexity (*see also* run times of algorithms)
  O(n2) time 271, 272
  polynomial-time 268–269, 271, 276
  quadratic time 269
topology 195, 296, **303**–**306**, 311–312
tractable vs. intractable problems 48–49
transcription 111, 112, 115–116, 116, 118, 119, 319
transcription factors *see* TF entries
transcriptional regulation 128, 317–319
transcription-induced mutations 120
TRANSFAC database 130
translocations 122, 180, 181, 218
trans-membrane protein 127
transmission efficiency, influenza virus correlation 155
transpositions 181, 211
Traveling Salesman Problem 76, 235–237
tree distance matrix 193
Tree of Life (TOL) **189**–**192**, 268
treelets 158–161
TreeMap software 234
treespace 284–285
trivial bipartitions 255, 257
true (network) alignments 308

unconditional probability 97
uninformative priors (Bayesian inference) 103
unique bipartitons 258–259
universal gene core 194
universal hashing functions 261–263
University of Leipzig, Germany 234
University of Sydney, Australia 234

CAMBRIDGE

Cambridge University Press
978-1-107-01146-5 - Bioinformatics for Biologists
Edited by Pavel Pevzner and Ron Shamir
Index
More information

unobserved data *see* prior probability
unrooted trees 251, 263, 278
unsigned reversals **175–178**

vaccines 150, 229
variables, observed and hidden 98, 102
variants 4–5
Venn diagram 97
vertices (graphs) 69–70, 271–272
  degree of a vertex 43–45
  indegree of a vertex 47–48, 53
  outdegree of a vertex 47–48, 53
  sink vertex 70
  source vertex 70
vessel theory of influence pandemics 155
viral genomes 113, 119
viral glycan-binding protein 155
viral RNAs 150
Virchow, Rudolf 192
virus replication 116

walks (graphs) 69
wasps 228, 241–243
WebLogo sequencing software 133
weighting, event cost 232
weights 67, 70, 79–80
Woese, C. R. and coworkers 190
word-based algorithms 157
World Health Organization 149
Wright Fisher model 7

X- and Y-linked DNA sequences 253

Yancopoulos, S. and colleagues 180
yeast 293, 295, 318

Zuckerkandl, Emile 190

© in this web service Cambridge University Press www.cambridge.org