ACV see average correlation value (ACV)

adaptive neural fuzzy inference system (ANFIS), 205–206

adjusted goodness-of-fit index (AGFI), 561

Affymetrix technology, 335–336, 336

alignment-free methods, 43

ANFIS see adaptive neural fuzzy inference system (ANFIS)

ANNs see artificial neural networks (ANNs)

archaea and bacteria domains, benchmark, 55

area under the ROC curve (AUROC)

coarse-complete setup, 535–536

interp-mRNA setup, 535–536

learning accuracy, 530, 535

network reconstruction performance, 529–530, 530, 537

population dynamics, 543, 543, 546–547

reference network P2010, 534–535, 535

synthetic data model, 544, 545

arrow plot

area under the curve (AUC), 400–401

collection, 403–404, 403

DE genes vs. special genes, 398–399

liv data set, 411

OVL, 402–403

ROC curve

classification rule, 401

definition and properties, 400, 401

eexpression levels, 401, 402

artificial neural networks (ANNs)

activation X_i, 201–202

back-propagation, 201

dihedral angles, 202

empirical risk-minimization principle, 202–203

feed-forward, 201

input, hidden, output layer, 195, 201

MLP, 202

nonlinear summing device, 201

PSIPRED protein structure prediction, 202

AUROC see area under the ROC curve (AUROC)

automated plausibility analysis

adjacent nodes, 466–467

asymptotic time, 469–470

LCA

Euler tour, 466–467

induced subgraph, 461–462

node from V_3, 462–463, 463

preorder traversal, 463–465

mega-phylogeny, 473, 474, 474–475

nontrivial bipartitions, 472

plausibility-check algorithm, 460–461

preprocessing phase, 465–466, 471
automated plausibility analysis (Continued)
real-world datasets, 472
rooted tree
  bipartition, 459
  Euler tour, 457, 458
  induced subgraph, 459
  induced subtree, 459
  inorder traversal, 457–459
  preorder traversal, 459
  Robinson–Foulds distance, 457–459
simulated data sets
  effective algorithm, 473–474, 475, 477, 478
  induced tree, 469, 472–473, 478
  runtime improvement, 474, 476
  speedup calculation, 474, 477
  total execution time, 477, 477
SmallTreeTaxa, 471
space complexity, 470
unrooted tree
  node height, 457
  preorder traversal identifier, 467–469, 468
  average common subword (ACS) approach, 46, 46
  average correlation value (ACV), 385–386, 386
background correction methods, 394–395
Barabasi–Albert model, 509
*BEAST program, 443
Benjamini and Hochberg (BH) procedure, 339
BEST program, 443
biclusters
  with coherent evolutions, 379, 381–382
  with coherent values, 379, 381
  with constant values, 379, 380
  with constant values on rows/columns, 379, 380–381
definition, 379
distribution parameter identification methods, 389
divide-and-conquer algorithms, 386–387, 387
example, 378
exhaustive bicluster enumeration methods, 387–389, 388, 389
greedy iterative search methods, 387
iterative row and column clustering combination methods, 386, 387
measurement of homogeneity
  ACV, 385–386, 386
  additive and multiplicative models, 382–383
  mean-squared residue score, 383–385, 386
  normalization, 389
bidirectional recurrent neural network (BRNN), 118
biological sequence segmentation
  ab initio, 121
  Bayes classifier, 116
  chaperonin-mediated protein folding, 116
  empirical inference, 116, 121
  HSMM, 122
  \( L \in \infty^* \) length, 115
  MSVMpred (see MSVMpred)
  ordered pairs of sequences, 115
  postprocessing, 119–120, 120
  protein data sets, 122
Boolean retrieval system, 319–320
CA maps see correspondence analysis (CA) maps
canonical adjacency matrix (CAM), 178
CCMS see Complete Cross Motif Search (CCMS)
CellMontage, 330–331
Chernoff discriminant analysis (CDA), 302
chi-square \( (\chi^2) \) method, 300
Chou–Fasman method, 101, 103–104, 255
chronic lymphocytic leukemia (CLL), 557–558
cis–trans isomerization (CTI)
  COPS algorithm, 237
cross-validation scheme, 242
evolutionary data set construction, 238–239
global sequence homology, 244–245
  intelligent voting approach, 239–241, 240, 241
  in vitro protein folding, 236
  MCC, 237, 243
  parameter tuning, 242
  protein secondary structure information, 239
  \( Q_2 \) accuracy, 237, 243, 244
randomized meta-learning, 241–242, 242
sensitivity $Sn$, 243
solvent accessibility, 245
specificity $Sp$, 243
STDEV, 242, 244
$SVM_{AB}$, 242, 243
$SVM_{LIB}$, 242, 243
SVM models, 237, 243
types I and II error rates, 244
wrapper feature selection algorithm, 237
Clark’s rule, 14–15
classification tree (C4.5), 360
CMS algorithm see Cross Motif Search (CMS) algorithm
comparative fit index (CFI), 562
comparative modeling (CM) method, 254
Complete Cross Motif Search (CCMS), 146–149, 147, 148
complete weighted minimum letter flip (CWMLF), 9
compositional index
average index values, 229
DomCut/SWISS-PROT protein sequence, 230–232, 231–232, 231
linker and nonlinker segments, 227
optimized threshold value, 229, 230
content-based retrieval, 318, 320–322, 322
correspondence analysis (CA) maps
basic principles, 405–406
cancer classification, 409, 413
classification error rates, 408
definition, 404
false discovery rate (FDR), 410
interpretation, 406–407
symmetric maps, 410, 412
cosine distance, 328
covariance model (CM), 160
Critical Assessment of Protein Structure Prediction (CASP) standard, 100, 258
Cross Motif Search (CMS) algorithm
brute-force approach, 143–144
CCMS, 146–149, 147, 148
greedy approach, 144–146, 145
optimization, 149–150
parallel approaches, 150–151
CTI see cis–trans isomerization (CTI)
data retrieval, 319
dChip, 357
DDI analysis see domain–domain interactions (DDI) analysis
de novo motif finding, 68
Deterministic Expectation Maximization (DEM) algorithm, 70, 71
Dictionary of Protein Secondary Structure (DSSP), 132
differentially expressed (DE) genes selection
absolute analysis, 337
BH procedure, 339
Bonferroni procedure, 339
data normalization, 338, 339
replicates, 337–338
SAM, 339
two-sample t-test, 338–339
WB-DEG selection algorithm (see within–between DE gene (WB-DEG) selection algorithm)
diploid organisms, 4
discrete cluster analysis (DCA), 363
Dnaml tool, 55
domain–domain interactions (DDI) analysis, 307, 307–309, 308
domain linker prediction
compositional index (see compositional index)
DomCut, 216–217
DROP, 219–220
EGRN (see enhanced general regression network (EGRN))
FIEFDom, 218–219
prediction assessment, 215–216
protein structure
primary structure, 213
secondary structure, 213–214
tertiary structure, 213–214
Scooby-Domain, 217–218
simulated annealing algorithm (see simulated annealing algorithm)
SVM, 219
technical challenges and issues, 214–215
domain linker prediction using optimal features (DROP), 219–220
domain-specific retrieval systems, 320
enhanced general regression network (EGRN)
enhanced general regression network (EGRN) (Continued)

Benchmark, 2, 220–221
compact domain profile, 221–222
prediction accuracy, 226, 226, 227
predictive performance, 227
semi-parametric models
architecture, 224, 224
auto-associative network, 224–225
centroid vector, 222
distribution weight, 225
GRNN, 222–224
large network training, 222
nonparametric model, 222
training, testing, and validation, 225–226

entropy-based information content

false-positive rate (FPR), 304
fast and accurate haplotype reconstruction
(FAHRR), 9
FastHare, 9
feature frequency profiles (FFP) method, 45
ferritin, 130
Fisher’s discriminant analysis (FDA), 302
fold-change (FC) method, 399
fragment assembly haplotype linking problems
see single individual haplotyping
fragments per kilobase per million mapped
(FPKM) value, 358
fuzzy decision tree (FDT), 206–207
Fuzzy Integration of Extracted Fragments
for Domains (FIEFDom), 218–219

GapsMis algorithm, 88–92
GapsPos algorithm, 92–94
Garnier, Osguthorpe, and Robson (GOR)
method, 101, 104, 255
GAs see genetic algorithms (GAs)
GEM-TREND web service, 331
genealogical network, 418
GeneChaser web server, 331
gene duplication and loss (GDL), 438
gene expression data
data set, 355, 355–356
experiment classification, 360–361

GELA
Alzheimer’s disease, 363, 365
breast cancer, 364, 366, 366
DCA, 363
PCA, 363
Pearson correlation analysis, 363

GenePattern, 362
genes clustering, 358–360
logic separating formulas, 364, 365

microarray technology
advantages, 353–354, 354, 355
applications, 348–349
definition, 348
disadvantages, 354, 355
in situ synthesis, 349
laboratory steps, 349–350
normalization methods, 356–357
processing steps, 348
robotic deposition, 349
multiple sclerosis and psoriasis data sets,
364, 366

RNA-Seq
advantages, 352, 353, 355
disadvantages, 352–353, 355
normalization methods, 357, 357–358
workflow phases, 350–352, 351

TM4 software suite, 361–362
transcriptome analysis, 347–348
WEKA, 362, 363

Gene Expression Logic Analyzer (GELA)
Alzheimer’s disease, 363, 365
breast cancer, 364, 366, 366
DCA, 363
PCA, 363
Pearson correlation analysis, 363

Gene Expression Omnibus (GEO), 324

GenePattern, 362
general regression neural network (GRNN),
222–224
gene set enrichment analysis (GSEA), 326
genetic algorithms (GAs)
ab initio protein structure prediction,
196–197
benefits, 192, 196
cubic lattice model, 194, 197
energy conformations, 197–198
fitness functions, 198–199
individual representation, 194, 195, 198
mutation and crossover point, 196, 199–200
tracking, 200–201
GFI see goodness-of-fit index (GFI)
Gibbs algorithm, 73–74, 74
Gini index, 240
global sequence alignment
aligned pair, 86
alphabet \( \Sigma \), 85
dynamic programming approach, 84–85
edit distance, 86
empty string, 85
gap
in \textit{Arabidopsis thaliana}, 84
in \textit{bacteriophage PhiX174}, 84
in \textit{Beta vulgaris}, 84
creating mechanisms, 84
definition, 83
gap-free sequence, 86
GapsMis algorithm, 88–92
GapsPos algorithm, 92–94
length in \textit{Homo sapiens}, 84, 85
sequence, 86
\textit{k-differences and }\ell \textit{-gaps} problem, 87–88
length of string, 85
Needleman–Wunsch algorithm, 84
optimal global alignment, 86
resequencing, 84
goodness-of-fit index (GFI)
AGFI, 561
CFI, 562
Hoelter’s critical N, 563
NFI, 562
parsimonious fit measures, 562–563
RMSEA, 562
GOR method see Garnier, Osguthorpe, and Robson (GOR) method
graph-coloring problem, 38–39
graph similarity matching
chemical relations, 272
Dice’s coefficient, 270
geometric relation, 272
graph edit distance, 269
graph isomorphism, 269
graph-theoretic approach, 270, 271
graph-transforming system, 270
Jaccard similarity coefficient, 270
MCS, 269
protein graph remodeling, 270, 271
statistical comparison methods, 269
greedy algorithm, 32, 32, 33
Hamming distance, 328
HapCompass, 11
HapCUT, 10
haplotype assembly for single human (HASH), 10
haplotyping
ad hoc algorithms, 4
population haplotyping (see population haplotyping)
single individual haplotyping (see single individual haplotyping)
happy graph, 5
Helix–Turn–Helix (HTH) motif, 78–80, 79
hemoglobin, 130
heteroscedastic discriminant analysis (HDA), 302
HGT see horizontal gene transfer (HGT)
HH model see hydrophobic–hydrophilic (HH) model
hidden semi-Markov model (HSMM), 122
Hilbert–Schmidt Independence Criterion (HSIC), 589
HMEC, 9
horizontal gene transfer (HGT), 438
aggregation function, 485
implementation, 489
intergroup variability, 483, 483–484
intragroup variability, 483, 483–484
maximum and minimum distance, 485–486
MSA
\( p \)-value estimation, 486–488
ranking strategy, 489
time complexity, 489
real prokaryotic data analysis, 498–499, 500
MCL algorithm, 496–497
NCBI Genomes ftp site, 496–497
sensitivity, 493, 497–499, 498–499
T-Rex website, 493, 496
synthetic data
analysis, 499–500
destination sequence, 490
logarithmic scale, 490
\( p \)-value ordering, 490, 493–496
\( Q \)-value ordering, 490, 491–492
hydrophobic–hydrophilic (HH) model, 191, 194, 194, 197
hydrophobic–polar (HP) model, 191, 194, 194, 197

immunoglobulin, 130
incomplete function (IF) ratio, 592
incomplete lineage sorting (ILS), 424, 438
indexing, 320
Infernal tool, 157, 160, 162, 162
Influenza A-H1N1, 54, 57, 58, 60
information gain (IG), 300
information retrieval, 318–320
interaction networks/regulation networks, 417

keratin, 130
k-nearest neighbor (k-NN), 303, 361
Kullback–Leibler information divergence, 47

lateral gene transfer (LGT) networks, 422, 424
LCA see lowest common ancestor (LCA)
least absolute shrinkage and selection operator (LASSO)
Bayesian models, 535, 536–537
Elastic Net method, 528
L1-norm penalty constraint, 528
nonzero regression parameter, 534
population dynamics, 544, 545, 546–547
synthetic data model, 544, 545
light–dark (LD) cycles, 526
linear additive models, 510
linear Bayesian (LB) classifier, 302
linear dimensionality reduction (LDR), 301–302
linear ensemble method (LEM), 118
logic data mining, 360–361
lowest common ancestor (LCA)
Euler tour, 466–467
induced subgraph, 461–462
node from V3, 462–463, 463
preorder traversal, 463–465

machine learning (ML) techniques, 516
major histocompatibility complex (MHC) proteins, 286
Manhattan distance, 328
Markov chain clustering (MCL) algorithm, 496–497
Markov chain Monte Carlo (MCMC) algorithm, 10
Mathews correlation coefficient (MCC), 237, 243
maximal common subgraph (MCS), 269
maximum common edge subgraph (MCES), 273
maximum fragments cut (MCF) problem, 11
maximum likelihood (ML) optimizations, 440
MaxSAT problem, 10
Metropolis algorithm, 72–73, 73
Michaelis–Menten kinetics, 526
Microarray Analysis Suite 5.0 (MAS 5.0), 357
microarray data see also gene expression data
ArrayExpress, 324
arrow plot (see arrow plot)
background correction methods, 395–396
bi-clustering, 323
CA maps (see correspondence analysis (CA) maps)
cell-state transit mechanism, 372
class comparison, 397
class discovery, 397
classification, 323
class prediction, 397
clustering, 323
data matrix, 371, 371
DE genes selection
absolute analysis, 337
BH procedure, 339
Bonferroni procedure, 339
data normalization, 338, 339
replicates, 337–338
SAM, 339
two-sample t-test, 338–339
WB-DEG selection algorithm (see within–between DE gene (WB-DEG) selection algorithm)
expression levels, 372
functional role of gene, 372
gene expression matrix, 322, 323, 326
gene ontology enrichment, 342, 343
gene-regulatory mechanism, 372
gene selection, 323–324
GEO, 324
mining informative patterns (see mining informative patterns)
normalization methods, 396–397
oligo DNA array image, 335–336, 336
preprocessing methods
cDNA microarray data sets, 407
class comparison context, 408–412,
410, 409, 411, 412
class prediction context, 408, 409
properties, 371
retrieving methods
CellMontage, 330–331
content-based information retrieval, 325
GEM-TREND, 331
GeneChaser, 331
GSEA, 326
keyword-based search, 325
logical view, 326
metadata-based search, 325
microarray experiment retrieval framework, 326, 327
openSESAME, 331
performance evaluation, 329–330
ProfileChaser, 331
querying, 324
SPIEDw, 331
SAM, 404–405
similarity metrics, 327–329
two-color cDNA array, 336, 337
two-color cDNA microarray experiment, 394
minimum conflict individual haplotyping (MCIH), 11
minimum error correction (MEC) model, 8–10
minimum fragment removal (MFR), 6, 8
minimum letter flip (MLF) see minimum error correction (MEC) model
minimum redundancy maximum relevance (mRMR) algorithm, 300
minimum set cover (MSC) problem, 32, 32–33
minimum SNP removal (MFR), 6, 8
mining informative patterns
clustering
biclusters (see biclusters)
consensus clustering, 375–377, 377
consensus distribution, 377–378
gene selection, 323–324
goal of, 375
irrelevant data set, 375, 376
iris data set, 375, 376
similarity measurement, 373–374
MLP see Multi-Layer Perceptron (MLP)
modification indices (MIs), 564
MSA see multiple sequence alignment (MSA)
MSVMpred
BRNN, 118
classification combination, 119
LEM, 118
MLP, 118–119
MSVMpred2
base classifiers, 121
configuration, 123
experimental results, 123–125, 124, 125
PSSM profiles, 121
M-SVMs, 117–118
NNs, 117–118
PLR, 118, 119
topology of, 117, 117
MSVMpred2, 121–122
multiclass support vector machines (M-SVMs), 117–118
multi-labeled trees (MUL-tree), 433–434, 434
Multi-Layer Perceptron (MLP), 118–119, 202
multiple sequence alignment (MSA)
$p$-value estimation, 486–488
ranking strategy, 489
time complexity, 489
multiple sequence local alignment
background model, 67
DNA cis-regulatory sequence motif finding, 66
maximum likelihood model, 68–70
missing data space, 66
motif finding algorithms
CRP binding sites, 76–78, 77
DEM algorithm, 70, 71
Gibbs algorithm, 73–74, 74
HTH motif, 78–80, 79
Metropolis algorithm, 72–73, 73
performance evaluation, 76
pseudo-Gibbs algorithm, 74, 74–75
WEM algorithm, 70–72, 72
multiple sequence local alignment
(Continued)
objective function, 67–68
oops and zoops models, 67
protein–DNA interaction binding sites, 66
time complexity, 75, 75
w-mer motif model, 67
multiple structural alignment by secondary
structures (MASS), 138
Munich Information Center for Protein
Sequence (MIPS), 512, 514
myeloid differentiation primary response
gene (88) (MyD88), 556
naive Bayes (NB) mechanism, 304
Needleman–Wunsch algorithm, 84
network inference
AUROC
coarse-complete setup, 536–537
interp-mRNA setup, 536–537
learning accuracy, 530, 535
network reconstruction performance,
529–530, 530, 537
population dynamics, 543, 543,
546–547
reference network P2010, 534–535,
535
synthetic data model, 544, 545
Bayesian regression model, 528–529
Bio-PEPA modeling, 532–533, 537
change-point process
Mondrian process, 540–542, 541
multiple global change-points,
539–540, 541
nonhomogeneous Bayesian regression,
531–532
population dynamics, 545, 546,
546–547
protein concentrations, 530–531
synthetic data model, 542, 544, 545
definition, 524
ecological systems, 526–527
gene regulatory network, 536–538, 537
gradient estimation, 532
LASSO
Bayesian models, 535, 535–536
Elastic Net method, 528
L1-norm penalty constraint, 528
nonzero regression parameter, 534
population dynamics, 544, 545,
546–547
synthetic data model, 544, 545
molecular systems, 526
mRNA profiles, 533–534, 537
notations, 527–528
ROC curve, 529, 530
species interactions
marginal posterior probability, 546,
547–548
niche model, 542, 543
partition-specific linear regression
model, 538–539
plant samples, 544
simulated data, 543–544, 545
stochastic population dynamics,
542–543, 543
neural networks (NNs), 100, 102, 102,
103, 117–118
normed fit index (NFI), 562
OPC-tree, 388, 388
openSESAME, 331
overlapping coefficient (OVL), 402–403
pathogen associated molecular patterns
(PAMPs), 555–556
PDB see Protein Data Bank (PDB)
Pearson’s correlation coefficient, 328
perfect phylogeny model, 19–21
persistent perfect phylogeny (P-PP) model,
21
personal information retrieval system, 320
PHD method, 101, 104–105
phylogenetic network, 419, 420
phylogenetic tree
characteristics, 418–420
computational intractability, 421
constructing networks, 431–433
definition, 420
fingerprints, 429–431, 431
internal nodes, 418
leaf nodes, 418, 419
LGT networks, 422, 424
MUL-tree, 433–434, 434
optimal graph, 421
patterns
character-block events, 423
character-state changes, 423
chromosome blocks evolution, 423
evolutionary convergence creation, 422
evolutionary divergence creation, 422
gene duplication—loss, 424
genetic information, 422
gene tree, 421, 425–429, 428–430
homoploid hybridization, 423, 424
horizontal components, 423
hypotheses, 424–425
ILS, 424
introgression, 423–424
parallelism creation, 422
polyploid hybridization, 423, 424
species phylogeny, 425–426, 427
vertical components, 423
reconstruction (see reconstruction of phylogenetic tree)
rooted phylogenetic networks, 418, 419
species network, 421
taxa, 421
tree nodes, 419, 420
tree space, 421
*Plasmodium*, 56, 59, 60, 60
Poisson–Boltzmann equation, 298
polytomous (multinomial) logistic regression (PLR) model, 118, 119
population haplotyping
Clark’s rule, 14–15
definition, 12
disease association, 21–22
genotypes, 12–14
perfect phylogeny model, 19–21
PHASE program, 12
PPH (see pure parsimony haplotyping (PHH))
XOR-genotypes, 22
Position-Specific Scoring Matrix (PSSM) profiles, 121
PPH see pure parsimony haplotyping (PHH)
PPIs see protein–protein interactions (PPIs)
primary structure of protein, 131
principal component analysis (PCA), 360
ProDM see Protein function prediction using Dependency Maximization (ProDM)
ProfileChaser, 331
Protein Data Bank (PDB), 134–135, 258
protein 3D-structure motifs mining candidates generation
Apriori-based approach, 177
BFS, 176, 176
closed subgraph, 178–179
DFS, 176, 176
maximal subgraph, 178–179
pattern subgraph, 178–179
pattern-growth approaches, 177–178
feature selection
Ant-motifs, 182–183
approximation, 181–182
characteristics, 183–184, 184
classification, 179
clustering-based subgraph selection, 181
discriminative frequent graph patterns, 182
MIPs, 182–183
problem statement, 180
sampling-based approach, 181
SkyGraph, 182–183
top-*k* selection approach, 180–181
frequent subgraph discovery
candidate subgraphs, 175, 175
graph isomorphism, 174
labeled graph, 174, 174
subgraph isomorphism, 174
unlabeled graph, 174, 174
graph mining, 172–173
protein graphs (see protein graphs)
subgraph mining, 173
Protein families (Pfam) database, 293
protein folding, 130
Protein function prediction using
  Dependency Maximization (ProDM) algorithm, 589–590
component analysis, 602, 603
experimental setup
data set, 590–591, 592
evaluation metrics, 592–594
IF ratio, 592
ProWL, 591–592
TMC, 592
WELL and MLR-GL, 592
multi-label learning approaches, 585–586
partially annotated proteins
  composite similarity, 588
cosine similarity, 587
HSIC, 589
incomplete annotations, 587–588
local and global consistency, 588–589
Protein function prediction using
Dependency Maximization (ProDM)
(Continued)
prediction tasks, 584–585
replenishing missing functions, 594,
595–597, 598
run time analysis, 602–603, 603
unlabeled proteins, 598, 599–601, 602
Protein function prediction using Weak-label
Learning (ProWL), 591–592
protein graph repository (PGR), 172
protein graphs
correctness, 170
Delanay triangulation technique, 170
Main Atom technique, 170–171,
171–172
PGR, 172
transformation techniques, 169–170
protein–protein interactions (PPIs), 506
application server, 514
assortativity, 508
average shortest path, 507
Barabasi–Albert model, 509
biological data, 510
boolean network, 510
centrality, 508
classification methods
k-NN, 303
LDR, 301–302
NB mechanism, 304
SMVs, 303
clustering coefficient, 507
curating protein interactions, 513–514
databases, 512
data sets, 304, 305
DDI analysis, 307, 307–309, 308
directed acyclic graph, 510
domain-based model, 293, 295
eigenvectors of matrices, 511
feature extraction
cutoff distance, 295
domain-based properties, 298–299
obligate and nonobligate protein
complexes, 296, 297
physicochemical properties, 296, 298
PseAAC, 295
3D structures, 295
feature selection
filter methods, 299–300
wrapper methods, 301
fitness of nodes, 509
FPR, 304
graph-based operations, 514–515
graph structure, 510
HITS algorithm, 511
interaction prediction, 515–517
linear additive models, 510
metabolic network, 506, 507
node degree, 506–507
PageRank, 511
Pfam, 293
PINs, 512, 513
prediction properties analysis, 304–307,
305, 306
protein complexes, 511–512
random networks, 509
RDF, 512–513
reciprocity, 508
remote web data sources, 514
ROC curve, 304
structural CATH domains, 293, 294
TPR, 304
WSD, 508–509
Protein Structure Motif Search (ProSMoS)
web server application, 135–136,
136
protein structure prediction
all-atom model, 191, 195, 195
ANFIS, 205–206
chemical structure, 191–192, 192
FDT, 206–207
fuzzy logic and sets, 204
fuzzy SVMs, 204–205
genetic algorithms (see genetic algorithms
(GAs))
primary sequence, 191–192, 192
secondary structure, 191, 193, 193
supervised machine learning algorithm
(see supervised machine learning
algorithm)
tertiary mesh, 191, 193, 193
pseudo Boolean optimization (PBO), 18–19
pseudo-Gibbs algorithm, 74, 74–75
pseudoknotted noncoding RNAs (ncRNAs)
arc-based representation, 157
BLAST, 160–161
bulge loop, 156
dot–bracket notation, 157
flowchart, 158
functions, 156
hairpin loop, 156
hypothetical Stockholm alignment, 158
Infernal tool, 157, 160, 162, 162
interior loop, 156
multibranched loops, 156
preparation of testing sets, 159, 159–160, 160
preparation of training sets, 158–159, 159
RNATOPS tool, 157, 160, 161, 161
secondary structure, 157
stems/helices, 156

PSIPRED method, 101, 104–105
PSSpred method, 105
pure parsimony haplotyping (PPH)
    APX-hard, 16
    CollHaps, 19
    combinatorial branch-and-bound approaches, 18
    integer programming formulations of exponential size, 16–17
    of polynomial size and hybrid formulations, 17–18
    NP-hard, 16
    objective, 15
    PBO, 18–19
    quadratic, semidefinite programming approaches, 18
quadratic Bayesian (QB) classifier, 302
quaternary structure of protein
    CM method, 254
databases, 258
decision tree model, 259
definition, 131, 252
folding steps, 255
hemoglobin, 252, 253
linear regression model, 260
molecular wedge allosteric drug-inhibited mechanism, 253
overfitting, 260
PDB, 258
performance evaluation, 260
predictor models and features, 259
pseudo amino acid composition, 259, 259
rotational symmetry, oligomeric proteins, 253, 254
schematic representation, 252, 253
sequences and structures, 255, 256
SG approach, 254
SVM, 260
Quat-2L, 260
Random Forest ensemble method, 239–241, 240, 241
ranked retrieval system, 320
reads per kilobase per million mapped (RPKM) value, 358
receiver operating characteristic (ROC) curve, 304, 529, 530
reconstruction of phylogenetic tree
    Bayesian methods, 440
distance-based methods, 440
diverse evolutionary processes, 445–447
GDL, 438
HGT, 438
homoplasy, 447
ILS, 438
ML optimizations, 440
Most Recent Common Ancestor (MRCA), 420
number of articles per year, 438, 439
objective, 420
recombination, 444–445, 445
species trees inferring, 440, 442, 442–443
substitution model, 443–444
updated programs, 440, 441
ReFHap, 11
replenishing accuracy (RAccuracy), 593
resource description framework (RDF), 512–513
RNATOPS tool, 157, 160, 161, 161
Robinson–Foulds (RF) distance, 56–57, 457–459
robust multiarray average (RMA), 357
root mean square deviation (RMSD), 267–268
root mean square error of approximation (RMSEA), 562
Saccharomyces cerevisiae PPIs (ScPPI), 590–591
secondary structure cooccurrences (SSC) algorithm
    clusters of vote points, 141–143
gEometric parameters, 139
secondary structure cooccurrences (SSC) algorithm (Continued)
local coordinate system, 139
optimization, 149–150
parallel approaches, 150–151
reference table, 140
search space, 138
voting procedure, 140–141
voting rule, 138
secondary structure elements (SSE)
α–helix, 131
β–α–β motif, 134
β–barrel, 134
β–hairpin, 133
β–meander, 133
β–sheet, 131, 132
DSSP, 132
Greek key, 133, 134
structural motifs (see structural motifs)
secondary-structure matching (SSM), 137, 137
secondary structure of protein
benchmark, 106–107
biological sequence segmentation (see biological sequence segmentation)
CASP standard, 100
Chou–Fasman method, 101, 103–104, 255
data sets, 109, 110
definition, 100, 251
development history of, 101
DSSP and STRIDE programs, 100
general flowchart, 102–103, 103
GOR method, 101, 104, 255
machine learning, 256
MaxHom alignment, 256
meta methods, 105, 106
NN trained models, 100, 102, 102, 103
pattern recognition-based methods, 100
performance comparison, 107–109,
107–109
PHD method, 101, 104–105, 256
PSIPRED method, 101, 104–105
PSSpred, 105
Q̄̂ score, 106
RELINFO, 106
SOV, 106
SPINE-X method, 105
statistical-based methods, 100

statistical learning, 256
SEM see structural equation modeling (SEM)
sequential floating forward search (SFFS) algorithm, 298–299
shotgun sequencing, 5
significance analysis of microarrays (SAM), 339, 404–405
simulated annealing algorithm
advantage, 229
conjunction, 229
DomCut/SWISS-PROT protein sequence, 230–232, 231–232, 231
recall and precision, 230
threshold values, 228, 228
single individual haplotyping
amplification phase, 5
BOP, 11
definition, 6
DNA sequencing, 5
EM algorithm, 11
experimental errors, 5
fragment conflict graph, 7
gapless fragment, 7
Gibbs sampling, 11
HapCompass, 11
happy graph, 5
HASH, 10
mate pairs, 5
MCF problem, 11
MCIH, 11
MCMC algorithm, 10
MFR problem, 6, 8
minimum error correction model, 8–10
MSR problem, 6, 8
single nucleotide polymorphism (SNP), 3–4
Spearman’s rank correlation coefficient, 328
SpeedHap, 9
SPIEDw, 331
SPINE-X method, 105
SSC algorithm see secondary structure cooccurrences (SSC) algorithm
SSE see secondary structure elements (SSE)
standard deviation (STDEV), 242, 244
string barcoding problem
applications of, 31
approximation algorithms
entropy-based information content technique, 34–35
subset, 33, 34
average length, optimal barcodes, 33
\( \varepsilon \)–approximate solution, 30
heuristic algorithms, 39–40, 40
inapproximability results
  graph-coloring problem, 38–39
  TSO\(^m\), 36–38
  MCP\(^2\)(r), 30
  MSC problem, 32, 32–33
relationships, 30, 30
SB\(^2\)(1), 28–29, 29
SB\(^2\)(\( \kappa \)), 29
Structural Classification of Proteins (SCOP) database, 135
structural comparison
  definition, 266
  graph spectra
    CATH codes, 284, 285
    clustering, 286, 286
    cospectrals/isospectrals, 282, 283, 283
    definition, 279
    Euclidean distance, 281
    Laplacian matrices, 281
    matrix selection, 281–282, 282
    MHC proteins, 286
    scoring, 283–284, 285
    tested antigen molecules, 284, 285
structural alignment
  DALI, 268
  RMSD, 267–268
structural similarity
  CATH codes, 277, 278
  display style, 277, 278
  graph theory, 268–272
  line graph construction, 276
  maximum clique detection, 277
  MCES, 273
  modular graph construction, 276, 277
  molecular descriptions, 277, 278
  P-graph, 274–276, 279, 280
  protein graph transformation, 273
structural equation modeling (SEM)
  aspects, 558
  assumptions, 559–560
  endogenous and exogenous variables, 558–559
  estimation methods, 560
  GFI
    AGFI, 561
  CFI, 562
  Hoelter’s critical N, 563
  NFI, 562
  parsimonious fit measures, 562–563
  RMSEA, 562
  imputation, 561
  latent variables, 559
  listwise deletion, 560–561
  modification indices, 564
  observed variables, 559
  pairwise deletion, 560–561
  residuals, 563–564
structural genomics (SG) approach, 254
structural motifs
  definition, 132
  geometrical description, 133
  line segment, 133, 133
  PROMOTIF algorithm, 136–137
  ProSMoS web server application, 135–136, 136
  retrieval algorithm (see cross motif search (CMS) algorithm; secondary structure cooccurrences (SSC) algorithm)
  topological description, 132–133
  supermatrix approach, 442
  supertree approach, 442–443
supervised machine learning algorithm
  ANNs
    activation Xi, 201–202
    back-propagation, 201
    dihedral angles, 202
    empirical risk-minimization principle, 202–203
    feed-forward, 201
    input, hidden, output layer, 195, 201
    MLP, 202
    nonlinear summing device, 201
    PSIPRED protein structure prediction, 202
  SVMs, 195, 203–204
support vector machines (SVMs)
  cis-trans isomerization, 237, 243
  domain linker prediction, 219
  gene expression data analysis, 361
  PPIs, 303
  quaternary structure of protein, 260
  supervised machine learning algorithm, 195, 203–204
Tanimoto distance, 328
tertiary structure of protein
  \textit{ab initio} methods, 257–258
definition, 131
homology modeling, 257
threading method, 257
test set with order with parameter \(m\) problem
  \((\text{TSO}^m)\), 36–38
TM4 software suite, 361–362
toll-like receptors (TLRs)
  CLL, 557–558
  flowchart, 564, 565
  KEGG pathway, 564–565
  PAMPs, 555–556
  signaling pathways, 556–557
  structure of, 556
TLR1/2 pathway
  M-CLL dataset, 565, 565–567, 566, 567
  U-CLL dataset, 567, 567–569, 568
TLR2/6 pathway, 569–571, 569–571, 570, 572
TLR7 pathway
  M-CLL dataset, 572–573, 573, 574–575
  MyD88 and IRAK1, 571–572, 573
  U-CLL dataset, 573, 573–575, 574, 575
TLR9 pathway
  M-CLL dataset, 573, 575–576, 576, 578
  U-CLL dataset, 573, 577, 577–578, 578, 578
transductive multi-label classifier (TMC),
  definition, 50
distance-like measure, 53–54
general structure, 50–51
length, 49, 60
selection of, 51–53
statistics, 60, 61
subset, 50
whole-genome phylogenies
  FFP\(_{ry}\) method, 56
genome data sets, 54–56
influenza A-H1N1, 57, 58, 60
neighbor-joining method, 56
  \textit{Plasmodium}, 59, 60, 60
prokaryotes, 57, 59, 60
  R–F difference, 56–57, 57
unicellular eukaryotes, benchmark, 56

Waikato Environment for Knowledge Analysis (WEKA), 362
weak label learning (WELL), 586, 592
web retrieval system, 320
weighted rank metric (WRM), 329
weighted spectral distribution (WSD), 508–509
weighted version of MLF (WMLF), 9
whole-genome sequence analysis
  ACS approach, 46
  alignment-free methods, 45
  counting procedures, 44
  Kullback–Leibler information divergence, 47
  popular alignment tools, 44
Winner-take-all Expectation Maximization (WEM) algorithm, 70–72, 72
within–between DE gene (WB-DEG)
  selection algorithm
  between condition comparison (BCC), 339–340
  within condition comparisons (WCC), 340
  FDR\(^{WB}\) control, 340, 341
Latin square data, 341
R/Bioconductor project, 340
sensitivity and accuracy, 341, 342

underlying approach (UA)
  inversions and complements, 53
  irredundant common subwords, 48–51
  underlying subwords