Ab initio protein structure prediction, 216–217
amino acids (AAs), 703
ASTRO-FOLD, 709–710
benchmark study, 217
CASP9, subset, 717–718
challenges/discussion, 712–717
with database information, 710
FB5 distributions, 216
force field (FF) energy models, 704
force fields models, 707
FRAGFOLD, 710–711
LINUS, 708–709
molecular dynamics (MD) simulation, 707–708
Monte Carlo (MC) simulation, 708
normal prion protein (PrPc) structures of, 704
protein-folding problem milestones, 705–706
pure, 706–707
ROSETTA’s principle, 711–712
TASSER, 711
Accelerating pairwise alignment algorithms, 476, 477
cell update per second (CUPS), 477
equation, 477
CUDAlign, 477
MUMmerGPU, 477
achievement, 477
Smith–Waterman (SW) algorithm, 476
accelerating using CUDA, 477
acceleration using OpenGL, 476, 477
implementation, 477
implementations, 476, 477
pixel shader, 477
SW-CUDA, 477
Swiss-Prot database, 477
Universal Protein Resource (UniProt) database, 477
Acquired immunodeficiency syndrome, 725
clinical decision, 725
human immunodeficiency virus, 725
genotype, 725
Acquired immunodeficiency syndrome (AIDS), 725
Acyclic graphs, 292, 635, 823
Adaptive Markov chain Monte Carlo (MCMC), 211, 304, 492, 1054
Additive epsilon indicator, 613
Adenine (A), 6–8, 95
Adjusted Rand index (ARI), 578
Advantages and limitations, inference of GRNs, 815
   inferred model, 815
   ordinary differential equations (ODEs), 815
Affine gap penalty, 972, 1011
Aho-Corasick algorithm, 858
Akaike information criterion (AIC), 487
Algorithms, mining ARs, 741–743
   APRIORI itemset, 741
   ARs APRIORI, 741
   k-frequent itemsets, 741
   pseudocode, APRIORI algorithm, 741
   equivalence class transformation
      (ECLAT), 742
   frequent-pattern (FP) growth, 742
   frequent-pattern tree, 742
   transaction database, 742
Aligners, 227
   algorithms, 227, 229
   available NGS, 228
   distributed, 229
      architectures, 233–234
      Crossbow, 234
      MapReduce, 233
      pbWA, 234
   hash-based, 229–230
   k-difference/mismatch problem, 227
   metrics, 227
   multistep procedure
      to accurately map sequences, 227
   prefix-based, 230–233
   short-read alignment programs, 229
   string matching, 227
   suffix-based, 229, 231
Alignment algorithms
   accuracy of, 207, 216, 264
   integration, 1085
   multiple-sequence, 1052
   pairwise, 971
      accelerating, 977
      global, 972
Alignment data visualization, 1083–1085
   alignment visualizations, 1084
      EBI ClustalW visualization, 1084
      consensus logo, Weblogo, 1085
      EBI ClustalW multiple alignment, 1084
      hidden Markov model (HMM)
         HMM-logos, 1085
      multiple-sequence alignment (MSA)
         algorithm, 1084
         alternative approach, 1085
         visualization, 1086
   pairwise sequence alignment, 1083
      BLAST, 1083
      NCBI BLAST, 1084
      searches, 1083, 1084
      residues, 1084
      conservation, 1084
      Shannon entropy, 1084, 1085
   Alleles, 266, 479, 480, 485
   Alignment
      full sensitivity, 230
      genome-wide, 984
      shapes, 62, 550
   Alphabet, defined, 254
   Alternative splicing, 5, 9–12, 23
      databases, 12. See also Splicing databases (SDBs)
   Alternative Splicing Queries (dbASQ), 6
   Amdahl’s law, 1055
   Amino acids (AAs), 703
      Cyrus Levinthal, 705
   Amplification, 1050, 1059
   Amplitude, 62, 64
   Analysis of formal concepts (AFC), 561–563
      assembly algorithms, 563
      batch algorithms, 562
      incremental algorithms, 562–563
   Analysis of variance (ANOVA), 371, 493, 510, 513, 514, 662
   Animal genome size database, 1038
   Annealing, 401, 492
      conformational space annealing (CSA), 709
      simulated, 492, 609, 829
   Apache-2.0 webserver, 17
   Archaea, 97
      classification, 97
      distribution of nucleotides, 98
      habitats, 97
   AR data preprocessing, 743–746
      discretization, 744–746
      global/local, 744
      static/dynamic, 744
      equal-frequency-interval discretization, 745
      equal-interval-width discretization, 745
      grouping, 746
      midrange-based cutoff, 745
      postprocessing four phases, 746
      postprocessing of ARs, 746
      pruning, 746
      summarizing, 746
threshold methods, 745
visualization, 746
supervised/unsupervised, 744
ARMA, 509, 510, 515
semiparametric QMLE estimator, 514
AR mining problem, 739–741
antecedent or left-hand-side (LHS), 739
AR items, database, 739
consequent or right-hand-side (RHS), 739
expression matrix, 740
frequent-itemset generation, 740
rule generation, 740
subtasks of AR mining problem, 740
ArrayExpress, 35
Artificial intelligence, 333
Artificial networks, 177
Artificial neural networks, 192, 664
AspAlt, 13
Assemblers, 234
algorithms, 235
available assemblers, 235
categories based on core algorithm, 234
de Bruijn graph (DBG), 234
formulations, 236
greedy assemblers, 234
overlap/layout/consensus (OLC), 234
programs, 235
Association rule, 737, 791
bioinformatics, 737
biological research, 737
data, computational biology, 791
data mining, association rule, 791
data-mining research, 737
processing, biological information, 791
relationships, biological components, 791
transaction database, 737
algorithms, 738
algorithms taxonomies, 738
APRIORI, 738
binary attributes, 739
data transformation, 739
mining algorithms, 739
Association rule groups, interestingness measures, 783, 784
computing group, 783
grouping, items and itemsets, 783
standard individual interestingness measures, 783
Association rule mining frameworks, 763–771
binary matrix representation, 764
antecedent, 765
closed-itemset framework, 765–767
frequent closed itemset, 766
frequent-itemset framework, 766
galois connection, 765
itemset lattice, 765
lattice theories, 765
levelwise traversal, 765
maximal itemset, 766
minsup, 765
closed itemsets, 763
free itemsets, 763
frequent itemsets, 763
regular itemsets, 763
Association rules, 76, 564, 762, 770, 775
Association studies, 176
Augmented web browsing, 872–876
advantage of Web technologies, 872
annotation, journal article, 873
Annotation of webpages, 872
biological entities, 872
life sciences researchers, 872
Autocorrelation, 113
AutoDB system, 14
Automated annotation of scientific documents
biological knowledge, 869–871
biomedical databases, 869
biomedical entities, 869
interrelated knowledge requirements, 869
nucleic acids research, 869
relevant knowledge sources, 869
Automated reasoning, 145
class consistency, 146
class subsumption, 146
conceptual data model consistency, 145
exploiting services, 145–147
finding new relationships, and classes by
using instances, 147–148
patterns for discovering implicit knowledge, 147–148
unconstrained type query, 148
graphical depiction of query patterns, 147
instance classification, 146
instance retrieval, 146
over ontology-driven conceptual data model, 146–147
querying, 147–148
refinement of multiplicities, and typing for
UML Class Diagrams, 146
Average correlation value, 595
Average correspondence similarity index ($E_{ACSI}$), 597
Average linkage clustering, 829, 1014–1017
Average similarity score (AVSS), 594
Average Spearman’s rho (ASR) function, 596
AVID, 973

Background distribution, 364
Backtracking, 253, 324
Bacterial genomes, 164, 262
Bandwidth, 503, 508, 1015
selection, cross-validation (CV), 508
Barcoding, 165
Bartlett’s tests, 478
Bases of association rules, 774–778
duquenne–guigues basis, 775
Luxenburger basis, 775
Min–Min basis, association rule, 777–778
Batch algorithms, 562
Bat sends sound signal, 679
Bayesian biclustering (BBC) model, 603
Bayesian classifier, 165, 664
Bayesian information criterion (BIC), 365, 579
Bayesian network (BN), 288–289
directed acyclic graph (DAG), 822
dynamical Bayesian networks (DBNs), 823
gene network, 823
cyclic Bayesian network, 824
hybrid Bayesian network (HBN), 824
Kullback–Leibler (KL), 823
modeling, gene network, 822
mutual information (MI), 823
parent–child dependencies, 822
probability distribution, 822
sparse candidate algorithm, 822
state-space models (SSMs), 823
Bayes rule, 210
BEAST, 1058
Benchmark 1 data sets
fastest measures on, 551
Bernoulli model, 559
BicAT, 564, 602
Bicluster enumeration (BE) approach, 600
Bicluster enumeration tree (BET), 600
Biclustering
algorithms
evaluation function, 594–598
stochastic search, 600–603
systematic search, 598–600
bioinspired algorithms, see Bioinspired
algorithms, for biclustering
gene expression data, 591, 609, 628
bioinspired algorithms, 609
dominated and nondominated solutions, 612
multiobjective model, see Multiobjective
model, for biclustering
group of, 593–594
microarray data, problem
mean row variance, 610–611
mean-squared residue, 610
types of, 592–593
Biclustering Analysis Toolbox (BicAT), 564
Biclustering multiobjective modeling, 609
Biclusters, 591
definition of, 591
groups of, 593
rows/columns, selection, 609
Bicluster validation, 603–605
BicOverlapper, 603, 1091
BiGG, 1037
Bimax algorithm, 602
Binary clustering with missing-values (BCMV), 649
problem, 645
Binary search, 254, 258
Binary tree, 1013, 1014
Binding, 957
events, 957
order of precedence, 957
rules, 957
Binding sites, 310, 312
BioCarta, 1037
Biochemical pathways, 1036–1039
BioCyc database, 157, 888, 1037
Bioentity tagging, 945–946
a biomedical named entity recognizer
(ABNER), 945, 946
examples, 945
open source chemistry analysis routines
(OSCAR3), 946
unified medical language system (UMLS), 946
ontology relationship, 946
BioEve system
BioNLP shared task, 958, 959
experiments/evaluations, 958, 959
BioGRID database, 83, 85, 87
Bioinspired algorithms, for biclustering
indicator-based multiobjective evolutionary
algorithm (IBEA), 617–618
MOBI algorithms, 615–616
nondominated sorting genetic algorithm II
(NSGA-II), 617
results, 618–622
Biological applications, in genomics, 785–788
  categories, genomics, 786
  analysis, genomes, 786
  functional genomic analysis, 787
  gene expression analysis, 787
  gene regulatory analysis, 787
  rowwise gene group analysis, 787, 788
  genes, 785
  expression data, 786
  investigation, genome, 785
Biological applications, in pregenomics, 785
  medical and environmental behaviors, 785
  medical data, 785
  patient, disease cure, 785
Biological applications, in proteomics, 788–790
  adenosine triphosphate (ATP), 788
  analysis, proteomic research, 788
  categories, proteomics, 789
    amino acid molecules, 789
    data analysis, protein expression, 790
    differential proteomics, 790
    homologies, proteomics prediction, 789
    protein expression analysis, 790
    protein structures, prediction, 789
    query processing, 790
    relationship, protein-sequencing, 789
    structural protein analysis, 789–790
  transcription, replication, 790
  gene expression, 788
  protein interactions, 788
Biological data, for GRN inference, 805, 806
  gene expression data, 805
    biological evaluation, GRN, 806
    data availability, 806
    data mining, GRNs, 806
    preprocessing, 806
    several issues of GRNs, 806
Biological data integration, 36
Biological data modeling, 129
  conceptual analysis, 129
  informal overview, 130
  languages, 130–131
  waterfall methodology, 130
Biological data types, 1074–1076
  expression data, 1075
    important repositories, 1075
    measurement, 1075
  functional data, 1075–1076
    functional annotations, 1076
    gene set enrichment analysis, 1075
  relational data, 1075
  sequence data, 1075
characteristics, 1075
DNA sequences, 1075
major goals, 1075
structure data, 1076
  Protein Data Bank (PDB), 1076
  protein structure, 1076
  four levels, 1076
Biological data visualization, 1076–1078
  asymmetry, 1077
  frequent visualization issues, 1077
  complexity, 1076–1077
    important characteristics, 1077
    simplification techniques, 1076
    tools for reducing complexity, 1076
  dimensionality, 1077
  3D graphics, 1077
  principal-component analysis (PCA), 1077
  diversity, 1077–1078
    complicated issue, 1078
    different types, 1077
    visualization linkage, 1078
  issues, 1076–1078
Biological domain knowledge, 905–909
  definition in ontology, 906
  discourse processing module, 906
  anaphora resolution, 906
  entities registered, 906
    in PASTA, 906
  domain-specific ontology, 905
    includes, 905, 906
  gene regulation domain, 907
    application, 908
    dependency structures, 907
    evaluation, 908
  Gene Regulation Ontology (GRO), 907
    semantic
      inference module, 907
      inference rules, 907, 908
      pattern, 907
      representation, 907
      state-of-the-art extraction system, 908
      structure, 907
      syntactic pattern, 907
  limited-domain ontology, IE system, 905
    Protein Active Site Template Acquisition (PASTA), 905
      templates example, 906
      phase-structure analysis, 905
      information, 905
  $P \rightarrow Q$ form rules, biological knowledge
    representation, 905
Biological homogeneity index (BHI), 580, 581
Biological information extraction (IE) systems, 901, 902, 903
extracting intrasentential information, 902
identifying intrasentential information, 902, 903
biological domain knowledge, use of, see Biological domain knowledge
linguistic knowledge, use of, see Linguistic knowledge
Biological knowledge, 1112–1116
discovery, 145, 354
novel approach for, 1112–1116
integration, see Integration
mapping graph, 1116
definition, 1116
illustration, 1116
mapping functions, 1116
measurement import nodes (MI-nodes), 1116
usages, 1116
multimodal biological data, based on, 1112–1116
novel approach for visualization of, 1112–1116
visualization pipeline, 1114
illustration, 1114
three parts, 1114
Biological macromolecules, 971
coding sequence, 971
Biological measures, 583
Biological network, 1086–1089
BioCarta repository, 1086
Wnt signaling pathway visualization, 1087
expression levels, 1088
force-directed layout, 1087
functional annotation analysis, 1088
KEGG repository, 1086
pathways characteristics, 1086
vs. BioCarta, 1086
Wnt signaling pathway visualization, 1087
large networks, 1087, 1088
cytoscape visualization, S. cerevisiae
kinase transcription network, 1088
useful characteristics, 1088
whole transcription network, 1087, 1088
3D force-directed layouts, 1089
visualization, 1088
network visualization tool, 1089
novel visualization technique, 1089
Linnet, 1089
pathways, 1086
reactome repository, 1086
Wnt signaling pathway visualization, 1087
visualize different properties, 1089
Biological network models, 163
challenges, 168
graph-theoretic and probabilistic-based protein interaction network models, 164–165
models in genetic interaction networks, 165–166
sequence-based approach, 163–164
Biological sequences, 101, 103
combined gene expression and, 751
comparison of, 971, 977
CUDA for, 1005
in databases increasing exponentially, 976
effectively predict polyadenylation sites in, 751
variance of, 192
whole genomes from next-generation sequencing, 309
Biological stability index (BSI), 580
Biological validation, of inferred GRNs, 813–815
biological experiments, 815
biological viability, inferred ARs, 813
cross-validation, 814
data-mining metrics, 813
gene ontology (GO), 814
Kyoto Encyclopedia of Genes and Genomes (KEGG), 814
Biomarkers, 333, 667, 668, 790
Biomedical data, 35, 41
cleaning, integrating, and warehousing, 41
liver genes and diseases, 41, 44–45
problems, 38
Biomedical entity recognizers, 920
Biomedical knowledge, 841–843, 864, 865
biomedical literature, 841
construction of a lexicon, 864
create entity names, 864
dictionary-based solutions, 864
implement an NER system, 864
information extraction (IE), 841
information retrieval (IR), 841
named entity recognition (NER), 841
normalization and disambiguation, 841
relation extraction, 841
ambiguous names, 842
approaches, automated NER systems, 842
biomedical entities and processes, terms for, 842
descriptive naming convention, 842
newly discovered entities, names of, 842
work flow of NER system, 843
work based solutions, 864
specific corpus, 864
string-matching techniques, 864
annotation heterogeneity, 865
hybrid approaches, 865
ML-based solutions, 865
output annotation, 865
text mining (TM), 841
Biomolecular event extractor, BIOEVE, 944–946
BioEve system architecture, 945
classification, 946
dependency parser, 945
dictionary-based labeler, 945
event trigger identification, 946
bioevent definition, 946
naive Bayes classifier, 945
Biomolecular interaction events, 943
issues, 943
Biomolecular interaction networks, 1030
Biomolecular networks, 1027, 1028
BioNLP shared task
at BioEve system, 958, 959
evaluation, 959
BioNLP’09 shared task, 944
Bio-ontologies, 53
BioRAT system, 944
BioSilico, 1038
Bipartitions, 629, 1057, 1058
BiVisu algorithm, 602
Black-box model, 663
BLAST, 48, 205, 228, 252, 1051, 1084
BLAT, 14, 17
Blocks substitution matrix (BLOSUM), 66, 972
BOND database, 157
Bonferroni correction, 662
Boolean networks, 826–828
building block, PBN, 827
dependence-sensitive Boolean network (cBN), 827
entity level, Boolean networks, 826
example cBN, 828
Kauffman, 826
network dynamics, 826
probabilistic Boolean networks (PBNs), 826
system state or global state entity, 826
Bovine pancreatic trypsin inhibitor (BPTI), 708
BOWTIE software, 229, 234, 259
Breast cancer–associated gene network (BCN), 167
Breast cancer, triple-negative, 659
BRENDA, 1037
Bulge loops, 688
Burrows–Wheeler alignment (BWA) algorithms, 983
Caenorhabditis elegans, 685
Calcineurin, see Protein, 3D structure
Canalization, 176
5’ Capping, 5
Cardinality-based greedy randomized adaptive search, 601
Case studies
feature selection instability, 336–339
related to RNA data analysis, 685
CASP7 target T0354
native structure of, 712
Cassette exon, 15
CD4 count data
estimated regression coefficients (COEF), 505, 515
estimated semiparametric covariance function, 511, 514
estimated time effects, 506, 516
functional analysis-of-covariance (fANCOVA) model, 514, 515
measurements of, 501
nonparametric covariance function estimator, 512, 513
randomly sample, treatment groups, 502
standard error (SE), 515
CDF curves, 549
cDNA, 284, 380
CE algorithm, 392, 393
CeleganSDB5, 16
Central nervous system (CNS), 566
Centre Européen de Recherche Nucléaire (CERN), 981
large hadron collider, 981
CHASS algorithm, 973
Chaperone-assisted folding model of, 715
Chein’s algorithm, 562
Chemical Entities of Biological Interest (CHEBI) Ontology, 53
Cheng and Church's (CC) algorithm, 616
ChiP-chip library, 365
ChiP enrichment, 365
ChiP-on-chip technologies, 1028
ChiP-seq analysis, 994, 995
CompleteMOTIFS, 994
CUDA-MEME, 994, 995
mUDA-MEME, 994, 995
PeakRanger, 994
ChiP-seq data, 365
ChiP-Sequencing (ChiP-Seq), 1028
Chi-square test, 493
Chromatin immunoprecipitation (ChiP) assays, 1028
genomewide location analyzing, 1028
Church's algorithm, 615
Circular string, 254
cis regulatory sequences, 10
Classification and regression tree-based approaches, 818–821
classification trees, 818
for gene CLN2, 819
decision tree, 818
temporal AR mining process, 819
treelike graph or model, 818
Classification EM (CEM) algorithms, 559, 561
Classification maximum likelihood (CML), 559
Classification trees, 663
Class representation
ML approaches, 849, 850
BIO encoding, 849, 850
BMEWO encoding, 850
class specification of sentence, 850
IO encoding, 849
token-by-token classification, 849
Clinical research directions, 729–730
HIV clinical decision making, 729
prognostic factors, 730
Cluster analysis
fourth short Haar wavelet transform, 119, 122, 123
wavelet coefficients of complex DNA representation, 121
Clustering affinity search technique (CAST), 647
Clustering algorithms, 522, 533
analysis of formal concepts (AFC), 561–563
for genomic applications, 534
internal validation measures, 544
benchmark 1 data sets, 550
consensus, 546–549
FC, 549–550
recommendations, 550–551
state of the art, 545–546
microarray data, see Microarray data
for microarray data, 570
model-based clustering algorithms, see Model-based clustering algorithms
NMF, 536–538
performance evaluation, criteria, 535–536
recommendations, 538–544
state of the art, 534–535
Clustering approaches, 1051–1052
analysis, 1052
quasi-biclique analysis, 1052
requirements, 1052
sequence comparison tool
Smith–Waterman algorithm, 1051
vs. seed approach, 1052
Clustering method, 1015–1018
average-linkage clustering, 1015
achievement, 1017
performance results, 1015
complete-linkage clustering, 1015
performance results, 1015
HCluster, 1017
OTUs observed, 1015
genus ground truth, 1018
species ground truth, 1017
trade-off, 1018
Clustering microarray expression data, 521
Clustering webtools, 563–566
biclustering analysis toolbox (BicAT), 564
formal concept analysis–based association rule miner (FAM), 564
lattice miner (LM), 564
Mixmod, 566
SPECLUST, 565–566
Cluster stability, 575
CNS rat
adjusted Rand index curves, 539
F-index curves for, 543
FM-index curves, 541
Cochran–Armitage trend test, 493
CoClust, 632
Coclustering algorithms, 626
gene ontology, 625
Coding exons, 15
Coding region, 39, 202
Color coding, 1079
Common ancestor, 80, 88
Comparative analysis of classification approaches, 962
classification approaches, summary, 963
CRF, 962
F-measure, 962
Comparative genomics
methods apply only to, 812
splicing databases, 13
Complementary DNA (cDNA), 645
Complementary RNA (cRNA), 284, 687
Complete-data likelihood function, 566
Computational biology, 784
Computational geometry, 976
Computational molecular biology, 421
Computational research directions, 730–731
additional features, 730
feature integration, 730
low-quality patient data, 730
data set noise and inconsistency, 730
Conceptual data modeling, 131
description logics languages, 131–133
assertional Box (ABox), 131
CMcom conceptual data modeling language, 133–138
EER, UML, and ORM 2, 138–140, 144–145
semantics of DLR and DLR_{us}, 131–133
terminological Box (TBox), 131
extensions of languages, 140
DLR_{us} syntax, 144
dynamically extends (RDEX), 143
MADS, 143
more expressive languages, 143–145
object migration, 143
ontology-driven modeling, 140–143
Condensed representations, association rules, 771–778
approximate association rules, 772
Armstrong’s axioms, 773
inference systems, 773
informative bases, 774
itemset covering, 774
object covering, 774
condensed representation, 771
exact association rules, 772
redundant association rules, 772
representation systems, 771
Conditional entropy $H(\Omega|C)$, 1015
Conditional random field (CRF), 192, 205
applications, 202
CRF-based ab initio Genefinder, 205
linear chain CRF model in gene prediction, 205
model for protein threading, 215–216
online resources, 204
pros/cons, 203
Confusion matrix, 530
ConnectedRule, 958
benefits, 958
definition, 958
event class rule, 958
Conrad gene prediction method, 206
features adopted from, 206–207
nonprobabilistic features, 207
condition maximum likelihood, 207
maximum expected accuracy, 207
segments define by SMCRF, 207
Conrad method, 206
Consensus, 546, 548
Conserved pairs proportion (CPP) index, 575
Constitutive exon, 15
CONTRAST de novo gene prediction method, 207
Copy number variation (CNV), 226, 478
Correlation metric construction (CMC) database, 1013
Correspondence similarity list (CSL), 596
Cost function, 195, 611, 681
cREAL, 262
circular mapping, 262
experimental results, 263–265
proposed algorithm, for solving problem, 262
circular text $C(t)$, 262
complexities, 263
linearizing circular text, 262
mapping, 262
CRiSPy-CUDA, 1010–1014
accuracy profiling
genus assignments, 1017
species assignments, 1016
limitation, 1014
parallel genetic distance computation, 1011–1013
banded alignment concept, 1012
CUDA global memory, 1013
CUDA implementation, 1012
on GPU, 1012, 1013
DP matrices, 1012
ATGAT/ATTAAAT, 1012
linear memory suboptimal formula, 1011
based on Needleman–Wunsch algorithm, 1011
multiple CUDA threads, 1012
CRiSPy-CUDA (Continued)

for parallelization, 1011
scoring matrix, 1011
parallel k-mer distance computation, 1010, 1011
sorting-based k-mer distance calculation
method, 1010
algorithm, 1010
performance evaluation, 1014
runtime profiling, 1018
bandwidth employed for alignment, 1018
vs. ESPRIT-Tree, 1020
vs. UCLUST, 1020
space-efficient hierarchical clustering, 1013–1014
Hcluster algorithm, 1013
hierarchical clustering, 1013
memory-efficient hierarchical full-linkage
clustering, 1013
sparse matrix, 1013
uses, 1013
Critical assessment of structure prediction
(CASP), 706
cRNA, 284
Cross-correlation, 677, 829
Cross-validation, 339, 343, 407, 467, 508, 814, 830
CUDA programming, GPUs, 1006–1007
CRiSPy-CUDA, 1007
based on ESPRIT algorithm, 1007
processing pipeline, 1007
CUDA-enabled processor, 1006
kernel, sequential function, 1006
parallel programming language, 1006
per-block shared memory (PBSM), 1006
single-instruction multiple thread (SIMT), 1007
CycADS, 1037
Cyclic Bayesian network, 824
CYTOSCAPE, 1087–1089, 1101
Cytosine (C), 95

Data banks, 35
Database-based searching, 59
Database profiling, 49
metadata, useful for, 49, 52
Databases, 6
dictionary-based approaches, 845, 846
biomedical vocabulary, 845
defined standards, 846
entity names, 846
specific entities, databases, 846
unified medical language system (UMLS), 845
work flow, dictionary-based NER system, 846
Data conflicts, 47
Data integration, 36, 163
biological entity resolution, 46
ceruloplasmin gene, 46, 47
from multiple sources, issues with, 163
record linkage, 46
Data limits, 1066–1067
biological dichotomy, 1067
large phylogenetic data sets, 1066
construction, 1066
molecular data, 1066
rbcL data sets, 1067
sequences representation per species, 1067
Data mining, 22
implementation of dbASQ
and utility of SDBs, 22–24
and inference of GRN, 804–805
data mining, GRN, 805
discrete components, GRN, 805
methods, 609
Data preparation, 36
Data quality, 36
aware solutions, 45–46
biomedical resources, problems occur in, 38
contaminated data, 41
cross-annotations with conflicting values, 40
inconsistencies, 38–40
irrelevancy, 40, 41
obsolescence, 41
naming ambiguities, 39–40
putative information, 40
redundancy, 38
semantic errors, 39
syntax errors, 39
undersized/oversized fields, 40
uninformative features or data, 40

Data aggregation, 47
Data avalanche, 1059–1064
data sets, 1059
ribulose-biphosphate carboxylase large (rbcL), 1059–1064
chloroplast gene region, 1059
examination, 1059
ribulose 1,5-bisphosphate
 carboxylase/oxygenase (RUBISCO), 1059

Data limits, 1066–1067
biological dichotomy, 1067
large phylogenetic data sets, 1066
construction, 1066
molecular data, 1066
rbcL data sets, 1067
sequences representation per species, 1067
Data mining, 22
implementation of dbASQ
and utility of SDBs, 22–24
and inference of GRN, 804–805
data mining, GRN, 805
discrete components, GRN, 805
methods, 609
Data preparation, 36
Data quality, 36
aware solutions, 45–46
biomedical resources, problems occur in, 38
contaminated data, 41
cross-annotations with conflicting values, 40
inconsistencies, 38–40
irrelevancy, 40, 41
obsolescence, 41
naming ambiguities, 39–40
putative information, 40
redundancy, 38
semantic errors, 39
syntax errors, 39
undersized/oversized fields, 40
uninformative features or data, 40
computing metadata for documenting biomedical sources before integration, 52
metrics, 49
metadata, useful for, 49, 52
problems
categorization of potential intrarecord data, 42
free-ruled data annotation, 36
heterogeneity of data sources, 36
inadequacy of data quality control mechanisms, 37
instrumentation/experimental errors, 36
major sources, 36
practical solutions to biological data, 43
scalability issues, 37
Data quality metadata, 36
Data set assembly, 1050–1053
aligned matrix, 1050
large-data-set assembly, 1050
large phylogenetic analyses, 1050
matrices construction, 1050, 1051
homology assessment, see Homology assessment
multiple-sequence alignment, see Multiple-sequence alignment
phylogenetic data sets, 1050
phylogenomics, 1050
plant phylogenies, 1050
studies, 1050, 1051
rapid molecular wet-laboratory techniques, 1050
Data sources, 36, 156
gene and protein interaction databases, 156
gene expression, 160–162
gene ontology (GO), 156–160
protein domain interaction databases, 162–163
Data tables, of SDBs, 15
Davies–Bouldin validity index (DBI), 573
DAWG matching algorithm, 231
dbASQ filters, 14
dbASQ system, 19, 25
alternative splicing, 19
mapping criteria, 19
dbASQ website, 17
DBG-based assemblers, 240–241
ABYSS, 244–245
ALLPATHS, 244
EULER-SR, 241–243
VELVET, 243–244
DDBJ, generalist data bank, 35
Decision function, 436
Decision trees, 663, 667, 726, 727, 819, 821, 944
Delayed associations, time-series data, 813
gene network, 813
reconstruction of GRNs, 813
time variation, protein, 813
De novo assemblers, 990–993
de Bruijn graph, 990
ABYSS, 991
vs. Velvet, 991
ALLPATHS-LG, 991
thread-level parallelism, 991
vs. SOAP de novo, 991, 992
Eulerian path, 991
overview, 991
PASHA, 991
function, 992
vs. ABYSS, 992
threading building blocks (TBB), 991
tools and algorithms using HPC, 987
uses, 991
Velvet, 991
YAGA, 991
function, 991
GPU-based solution, 993
MPI-based parallelism, 991
overlap graph, 990
Hamiltonian path, 990
overview, 990
PE-Assembler, 993
scalable assembler, 993
SSAKE, 993
De novo gene prediction method, see Conrad method; CONTRAST de novo gene prediction method
Diabetes, 166, 445, 469, 484, 785
type I, 166
type II, 484
Dictionary-based approaches
biomedical knowledge, 845–848
matching entries, chunks of text, 845
tasks in dictionary based approaches, 845
Dictionary matching, 853, 857, 860
Diffusion kernel, 165, 166
Dijkstra’s algorithm, 243
DIP database, 84, 156
Directed acyclic graph (DAG), 288, 599
Discrete Haar wavelet transform, 115
Discrete variables
redundancy, 406
relevance, 404–405
synergy, 406–407
Discretization, 388, 665, 744, 745
Discriminative model, 203
Disease Ontology, Systems Biology Ontology, 53
Disease risk, 478
Disordered regions (DR) prediction, 718
Distance functions
basic definitions, 527–529
microarray experiments, recommendations, 532–533
performance, criteria, 529–531
problem statement, 526–527
results of, 531–532
state of art, 525–526
Distance matrix, 1005
Distributional analysis, text mining, 902, 914–915
distributional characteristics, 914
hyperspace analogue to language (HAL), 915
information-theoretic tests, 914
intersentential relations recognition, 914
latent semantic indexing (LSI), 915
literature-based discovery, 914, 915
medical subject headings (MeSHs), 915
RE tasks, 914
co-occurrence-based methods, 914, 915
simple approach, 914
rule-based method, text-mining systems, 914
Divergences, 628, 629, 989, 1066
Divide-and-conquer (DAC) approach, 599, 742, 1053, 1059
DmelSDB5, 16
DNA-Chips, 36, 55
DNA clone classification, 652
DNA (deoxyribonucleic acid) sequence, 95, 251
bacteria/fungi/archaea, complete sequences analysis, 96–97
characterization, 95
complex root representation, 108–109
pseudorandom sequence on unit circle, 109
complex sequence, DNA walk on, 110
on pseudorandom and deterministic complex sequences, 110–112
variance, 113–114
expedient method to analyze influence of close bp, 96
indicator matrix, 100–101
measure of complexity and information, 102–103
complexity, 104–105
entropy, 105–108
fractal dimension, 105
patterns on indicator matrix, 98–100
test sequences, 101–102
DNA microarray technology, 35, 380, 427, 564, 569, 653
DNA–protein interaction, 365
Document crawling, 925
document-type definition (DTD), 925
implementation, 925
Java database connectivity (JDBC), 925
purpose, 925
world wide web consortium, 925
XML files, APIs types
event-based simple API to XML (SAX), 925
tree-based document object model (DOM), 925
Document preprocessing and parsing, 925, 926
filtering metalanguage tags, 925
partial list, sample sentences, 925
associated with PubMed abstracts, 925
POS analysis, 925
POS tags, 925
Stanford parser, 926
Downregulation, 294
Drosophila melanogaster, 19
Dunn index, 574
Dynamic Bayesian networks, 289
constructing from cyclic network, 290
muscle development network in Drosophila larval stage using, 302
Dynamic classifier fusion (DCF), 682
for better accuracy, 680
proposed model, 681
Dynamic programming (DP), 191, 973
applications, 191
Dynamic QPNs (DQPNs), see Qualitative probabilistic networks (QPNs), dynamics
DynMap, short-read alignment program, 265–267
algorithm, for solving problem, 267
creating lists, 267–268
dynamic update assume, 268–270
mapping, 268
remapping, 270–271
complexities, 271–272
experimental results, 272–275
INDEX 1139

e-CCC-biclustering, 603
Echo signal, 679
EcoCyc, 1037
Edit distance, 52, 227, 253, 847
Edit operation, 267, 268
Efficient semiparametric profile estimator with a nonparametric covariance function (ESPR-NPC), 515
EMBL, generalist data bank, 35
Empty string, 254
Ensembl, 35
Ensemble feature selection algorithms, 339–340
designing, key aspects, 345
added complexity, 345
appropriate methods, 345
ReliefF algorithm, 345
ensemble based on data perturbation, 340
different data partitioning, 340–341
Ensemble methods, 310–311
lexicographic properties, 310
motifs, 310, 311
with character classes, 313–315
experimental results, 326–329
minimal motifs, and motif priority, 316–320
motif and location list, 315
pattern occurrence, 314
on transitive properties of character classes, 315–316
Z-score, 327–328
Ensemble size, 343
and effects, 344
Ensembl genome browser, 1038
Entrez genome database search, 1038
Entropies, 1015
ENZYME, 1038
Epistasis, 175
biological, 176
statistical, 176
statistical and computational challenges, 176
techniques proposed, to reduce computational complexity, 176
alternative pruning method, 176
entropy and conditional entropy, 177
genetic variants, 177
information-theoretic measures, 176
yeast metabolism networks, 176
Error-aware hash function, 230
Escherichia coli, 901
ESPRIT-Tree, 1014
π₀ Estimation, in statistical significance assessment, 359–360
algorithm, 362
bias and influence on FDR calculation, 363
BUM estimator, 364
estimator proposed, 361–362
mixture model as basis for, 361
reduced bias using mixture modeling of, 364
Euclidean distance, 532, 559, 628
ROC curve for, 531
Eulerian path, 240, 991
Evaluation of application’s behavior of biomedical technique, 844, 845
false negative (FN), 844
false positive (FP), 844
F-measure, 845
precision, 844
predictions, application’s behavior, 844
recall relevant items, 845
true negative (TN), 844
true positive (TP), 844
Event extraction
breadth-first fashion search, 952
dependency parsing, 951, 952
bind, trigger word, 952
tree, 952
Stanford parser
tagged sentence, 951
typed-dependency representation output, 951
Event extraction module, 962–964
BioEve extraction module evaluation
one-pass extraction, 964
two-pass extraction, 964
evaluation measures, 962
semantic classification improvement, 962
Event extraction rules, 957
binding event, 957
BioEve system, 957
dependency parsing, 957
Event phrase labeling
classifier accuracy, 959
CRF sequence tagger, 960
relevant trigger phrases, 960
retrieved trigger phrases, 960
semantic classification, 959–962
weighted average precision, 960
equation, 960
weighted average recall, 960
equation, 960
Evolutionary computation (EC) approach, 601
Evolutionary distance, 164
Evolutionary framework, 769–771
  discretization method, 769
  genetic, 770
    differential evolution, 771
    genetic network programming, 771
    genetic programming, 770
    grammar guided genetic programming (G3P), 771
    Michigan approach, 770
    multiobjective optimization, 770
    pareto-optimal rule sets, 770
    pittsburgh approach, 770
    rule generation, 770
    rule selection, 770
Evolutionary reduct generation, 428
  basic steps of proposed algorithm, 430–431
  comparative performance on gene expression data
    using single-objective GA, 437
    \(d\)-distinction table, 429
  experimental results, 431–434
    colon cancer data set, 431
    comparison, 434–436
    leukemia dataset, 432
    lymphoma data set, 432
  principal-component analysis (PCA), 436
  reduct generation, with single-objective (classical) GA, 436
    selection of attributes/genes in reducts, 433
  redundancy reduction for microarray data, 428–429
    attributewise normalization, 428
    choose thresholds based on, 429
    converting attribute value table to binary (0/1) form, 429
  saliency analysis to support vector machines, for gene selection, 435
    \(t\)-test-based feature selection, with a fuzzy neural network, 435
      using MOGA, 430
  Exhaustive enumeration, 184, 186
  Exonic splicing enhancers (ESEs), 7
  Exonic splicing silencers (ESSs), 7
  Exons, 28, 29, 204, 206, 983
  Expectation maximization, 207, 559, 949
  Exponential distribution, 811
  Expressed sequence tags, 17, 1075
  Expression data visualization, 1092–1097
    one or two conditions, 1093
      evolution of expression, 1093
      expression behavior, 1093
      expression-measuring technology, 1093
      MA plots, 1094
      volcano plots, 1094
      scatter plots, 1094
    several conditions, 1093–1096
      commercial analysis suites, 1096
      dendrograms, 1095
      gene expression atlas, 1094
      GGobi, 1095
      heatmaps, 1093
        hierarchical clustering, dendrograms, 1095
        scientific standards, 1094
        parallel coordinates, 1094, 1095
        scientific standards, 1094
      tools, 1096
        HCE, 1095
      visualization of groups, 1096–1097
        biological knowledge, 1097
        blind data analysis, 1097
        expression matrix, 1096
        grouping techniques, 1095
        biclustering, 1095
        heatmap rows, 1095
        hierarchical clustering, 1096
        major interaction techniques, 1095
        principal-component analysis, 1097
        venn-like diagram, 1095
        visualization, simplifying, 1097
      vs. sequence visualization, 1092
  Express sequence tag (EST) alignment information, 205
  Extended entity-relationship (EER) for relational databases, 129
  External measures, 583
  Factor analysis, 489, 680
  False discovery rates, 60, 61, 65, 662
  Familywise error rate (FWER), 662
  FARMER software, 750
  FASTA format, 259, 272, 689
  Fast approximation algorithms, 545, 546
  Fast mutual information estimation, 413–414
    assuming normally distributed variables, 415
    discretizing variables, 414
    empirical estimation, 415
    Miller–Madow correction, 415
  Fatty acid biosynthesis, 638
Feasible biological relation identification, 929–932
ABNER, molecular biology text analysis tool, 929
conditional random fields (CRFs), 929
biological actors, 929
biological entities, 929
biomedical relation, 929–935
biomedicalRelationExtraction (L_{IC}) algorithm, 932, 937
definition, 929
feasibility analysis, 929
biological domain, 929
morphological variants and partial list of, 938
root verb, 932
Feature selection, 333, 422, 950
orthographic features, 950
based on regular expressions, 950
semantic features, 950
database-referenced lexicons, 950
feature selection, 950
hand-prepared lexicons, 950

Feature selection algorithms

categorization of algorithm, 334–335
embedded approach, 335
filter algorithms, 334
genetic algorithm (GA)–based selection, 335
and instability, 334
data partitioning, 338–339
potential causes, 335–339
sample order dependency, 337–338
small sample size, 336–337
wrapper approach, 335
advantages, 335

Feature selection stability, performance on, 341
data-partitioning problem, 343
sample order dependency problem, 342–343
small-sample-size problem, 341–342

Features, of ML method, 850–855
linguistic characteristics, 850
matrix of features, 850
biomedical entities, 853
conditional random field (CRFs), 855
dictionary matching, 853–855
external features, 851
feature induction, 854
feature selection, 854
features illustration window, 854
internal features, 851
local context, 851
model, feature selection search, 854
orthographic features, 852
part-of-speech (POS), 851
POS tagging, 851
shallow parsing, 851
structures of morphological types, 852
support vector machines (SVMs), 855
tokenizer works, 850
Federated databases, 37
Feedback loops, 813, 1077
Feedforward loops, 301
Fermi-based GPU, 1014
NVIDIA GTX 480, 1014
Field-programmable gate arrays (FPGAs), 982
Figure of merit (FOM), 576
Filtering, by means of underlying motifs, 320–323
algorithm for underlying representative set, 323–326
check for untied occurrences, 325
compare motifs, 326
MotifFiltering, 325
StoreCoverage, 325
experimental results, 326–327
Z-score, 327–328
Fingerprint, 646
Fingerprint classification methods, 645
algorithmic complexity results, 650–652
BCMV(p), 650–652
IECBMV(p), 651–652
definitions, 646
estimating missing values, combinatorial approach, 649–650
experimental results, 652–653
future research, 653
problem statements, 646
First-generation data integration systems, 37
Fisher–Bingham (FB5) distribution, 216
Fisher’s statistical interaction, 486
Fitch’s hierarchical clustering, 829
Fitting the model, 446–447
coeff column, 447
lcpue section, 446, 447
root MSE, 447
source section, 447
using regress command, 446
Fixed rank approximation algorithm (FRAA), 648
Flat file indexing, 37
FlexStem methods, 692, 697
Flux balance analysis (FBA), 829
Formal concept analysis–based association rule Miner (FAM), 564
FP-tree, 742, 816
Free energy, 213
Free-set framework, 767–768
  frequent generators, 768
F-test, 466
Full-genome sequences, 728
FunCat database, 83
Functional analysis-of-covariance (fANCOVA) model, 513
  CD4 count data, 514
  multiple treatment groups, modeling, 513–515
Functional data visualization, 1085–1092
  annotations, 1090–1092
    functional annotations, 1091
    functional knowledge, 1090
    gene ontology (GO), 1091
    novel visualization techniques, 1091
    OntologyMaps, 1092
    visualization approaches, 1091
    visualization technique, 1091
    Voronoi map, 1092
BicOverlapper
  tag cloud, 1092
biological network, see Biological network
controversy, 1086
phylogenetic trees, see Phylogenetic trees
Functional link artificial neural network (FLANN) classifier, 675, 678
Functional proteomic analysis, 790
  prediction, protein function, 790
  protein–protein interaction, 790
  protein subcellular localization, 790
FunSpec tool, 580
GAIN, see Gene association interaction network (GAIN)
Galois closure, association rule, 761–763
  condensed representations, 761
  interestingness measures, 761
  itemset, 761
  objective (data-driven), 761
  semantic (context-driven), 761
  subjective (user-driven), 761
Galois lattices, 562
Gap penalties, 972
GARLI, 1054
GAs, see Genetic algorithms (GAs)
Gaussian errors, 515
Gaussian mixture modeling, 363, 365, 561, 726
Gaussian-type responses, 506
G-compass, 1039
GD library, 17
GDS2508 data, 639
GDS2713 experimental setting, 638
GenBank, 35, 36, 37, 40, 44, 45, 264, 274, 848, 981, 1059
  HFE gene, 48–51
Gene annotation keywords, 17
Gene association interaction network (GAIN), 181, 182
GeneBank accession number, 17
Gene clustering, microarray data sets, 585
GENECODIS, 604
Gene–disease associations, 353
GeneEntry databank, 48
Gene expression, 379, 381, 746–750
  genes, 746
  genetic information, 746
  genomewide patterns of, 399
  protein, 746
  gene expression matrix, 747
  microarray, 746
  mining in horizontal-data format, 747
  mining in vertical-data format, 749
  serial analysis of gene expression (SAGE), 746
Gene expression data, 591
  multiobjective model, 611
Gene expression data, fusion technique classification, 675
  experimental evaluation, 681–682
  preliminaries, 678
  bat position, 680
  BAT, working procedure, 679
  frequency, calculation, 679
  proposed model, 680–681
  PSO/SA, metaheuristic algorithms, 676
  studies of, 676–678
Gene expression data matrix, 592, 610
Gene Expression Data Warehouse (GEDAW), 41, 47
  categories of mapping rules, 47
Gene expression discretization, 807–810
  binary discretization, 807
    absolute values, discretization, 808
    clustering, discretization, 809
    discretization problem, 807–808
    equal frequency principle, 809
    gene expression matrix, 807
    variations between time points, discretization, 808
    biological data analysis, 807
intrinsic nature, biological data, 807
machine learning algorithms, 807
Gene expression microarray experiments, 647
Gene expression process, 282
Gene–gene interactions, 176, 177, 182, 353
Gene-ID, 48
Gene match score (GMS), 597
Gene Ontology (GO), 46, 48, 53, 54, 78–79, 140, 581, 598, 814, 912, 1119
coclustering, 625
computational biology, 626
constrained, 629–632
gene expression data analysis, 627, 628, 629
Goodman-Kruskal’s coefficient, 627
metric-based method, 628
for pathway identification, 625
graph, 633
Gene Ontology Consortium (GOC), 604
Gene prediction, using conditional random
fields, 202–207
Generalized likelihood ratio (GLR) test, 515
Generative model, 203
Gene regulation, 957, 958
negative, 957, 958
normal, 957, 958
positive, 957, 958
regulation events, 958
Gene regulatory networks (GRN), 285, 286, 803, 804
Bayesian networks for learning, 287–288
construction, 286
cytoplasm, 803
free-scale network, 286
gene network modeling, 803
gene regulation, 803
learning, 290
messenger RNA (mRNA), 803
models for reverse engineering, 287
reverse engineering, 803
toward qualitative modeling of, 291
abundance of qualitative information, 291
computational efficiency, 292
motivating factors for, 291
possibility for extension, 292
transcription factors (TFs), 803
transcriptome data, 804
Gene selection methods, 381
Algorithm 16.2, 395
kernel-based feature selection, 382–384
redundancy-based gene selection, 384–391
unsupervised feature selection, 391–395
Gene set enrichment analysis (GSEA), 161, 366, 1090
Genetic algorithms (GAs), 422, 425
multiobjective, 422–423
operators, 425
Genetic distance computation module, 1015
different bandwidths using effects, 1016, 1017
Genetic markers
and fisher’s nonadditivity interaction, 484–486
in genomewide association analysis, 486–489
elastic net, 488–489
least absolute shrinkage and selection
operator, 488–489
linkage disequilibrium, 487
octagonal shrinkage and clustering
algorithm for regression, 488–489
penalized regression, 486
ridge regression, 488–489
stepwise regression, involving large
number of decisions/tests, 487
variable reduction, 486
Gene trees, 1050, 1064
GENIA event, major modules
experiments and evaluations, 958, 959
GENIES, 944
natural language processing (NLP)
engine, 944
Genome annotation databases, 1036–1039
Genome data visualization, 1079–1083
assembly inspection basis, 1079
circular genome visualization, 1082
circular genome visualizer, 1082
GenoMap, 1082
comparing two genomes, 1082
gene track, 1080
genome analysis, 1079
GBrowse, 1082
genome assembly, 1079
genome browsing, 1079
genome comparison, 1079
popular browsers, 1080
UCSC genome browser, 1081, 1082
histogram tracks, 1081
human genome chromosomes, 1083
Circos visualization, 1083
MizBee visualization, 1083
next-generation sequencing (NGS)
technologies, 1079
Genome data visualization (Continued)  
sequence comparison, 1082  
large-scale, 1082  
tools, 1080  
GenomeMapper, 266  
Genome on-line database (GOLD), 1039  
Genome reviews, 1039  
Genome-scale metabolic reconstruction (GSMR), 1035–1036  
aplications, 1035–1036  
bottom-up model, 1035  
complex systems theory, 1036  
holistic approach, 1036  
*E. coli*, 1035, 1036  
evolutionary relationships, 1036  
final goal, 1036  
genetic expression data, conjugation with, 1035  
growth, 1036  
metabolic engineering, 1035  
metabolic knowledge integration, 1037  
opportunity, 1036  
Genome-scale models, 1031–1033  
challenges, 1031  
future prospect, 1032  
generation, 1034–1035  
reconstruction process, 1034  
Thiele and Palsson protocol, 1034–1035  
genome-scale metabolic network, 1031  
arbacia, 1032, 1033  
bacteria, 1032, 1033  
eukaryotes, 1032, 1033  
*Haemophilus influenza*, model organism, 1031  
history, 1031–1033  
large-scale human metabolic networks, 1032  
multispecies stoichiometric metabolic model, 1032  
myocardial metabolic network, 1032  
protein expression data, 1032  
tissue-specific gene data, 1032  
tissue-specific models, 1032  
Genome sequencing, 12, 158, 252, 283  
Genomewide association (GWA) studies, 333, 486. See also Single genetic marker analysis of qualitative traits, 478  
copy number variants, 478  
dimension reduction, 490  
generalized linear model, 477  
genetic polymorphisms, 478  
latent variable construction, 489–491  
logic regression, 491–492  
MCMC algorithms, 492  
partial leastsquares (PLS) constructs, 490, 491  
principal-component regression, 490  
link function, 477  
logistic regression in, 477–478  
R packages, 493  
single-nucleotide polymorphism (SNP), 478, 479  
S-shaped sigmoid functions, 478  
variable reduction, 490  
Genomic analysis, 540, 546, 786, 788  
Genomic exon, 15  
Genomic information database (euGenes), 1039  
Genotype–phenotype relationship, 478  
GEO database, 162, 636, 637  
GHMM model, 206  
Giga cell updates per second (GCUPS), 1005  
Glucconeogenesis, 638  
GO, see Gene Ontology (GO)  
GO-driven coclustering  
constraints, definition of, 635–636  
determining suitable number, 634–635  
methodology instantiation, 636  
parameterless methodology for, 633–634  
validation of  
cluster association, 638–639  
column clusters, accuracy, 636–637  
row clusters, homogeneity, 637–638  
Golm metabolome database (GMD), 1038  
Goodman-Kruskal’s τ coefficient, 627  
Google Scholar, 570  
GP-Close algorithm, 940  
G-protein-coupled receptors family 3, 311  
GPU clusters, 1010  
CUDA programming, 1010  
MPI programming, 1010  
GPU programming, 1006–1010  
Gradient–based optimization algorithm, 207  
Grammar engineering approaches, 944  
context-free grammar (CFG), 944  
Graph, 178  
acyclic, 179  
adacency matrix, 179  
array of metrics  
to measure robustness of connectivity, 179–180  
centrality measurements, 180  
degree centrality, 180  
eigenvector centrality, 180
Katz centrality, 180–181
PageRank, 181
connectivity, 179
degree distributions, 181
directed acyclic graph (DAG), 178, 179
multiedges, 178
multigraph, 178
out-degree, 178
parallel arcs, 178
self-loops, 178
simple, 178
strongly connected, 179
theory, 178
undirected, 178
Graphical files, 17
Graphical models, 192
conditional random field, 198–200
general conditional random field, 202
inference, 201
parameter estimation, 200–201
resources, 202
vs. Markov random field, 201–202
directed, 192–193
discriminative vs. generative model, 194–195
label bias problem, 197–198
maximum-entropy Markov models (MEMMs), 197–198
sequential, 195–196
hidden Markov model, 196–197
undirected, 193–194
Graphics processor units, 974–976
applications, 974
data-parallel processing, 974
dynamic RAM (DRAM), 974
floating-point operations per second (FLOPS), 974
CPU, 975
GPU, 975
memory bandwidth, 975
CPU, 975
GPU, 975
multiprocessors, 974
consist of, 974
open graphics library (OpenGL), 974
CPU, 976
direct compute GPU using CUDA, 976
GPU, 976
scan performances, 976
thread, 974
uses, 974
vs. CPU organization, 974
Greedy assemblers, 236
QSRA, 237
SHARCGS, 236–237
SHORTY, 237
SSAKE, 236
VCAKE, 237
Greedy iterative search (GIS) approach, 599
Greedy randomized adaptive search procedure (GRASP), 609
GRID database, 211
GRN, see Gene regulatory networks (GRN)
Growth associated ATP maintenance (GAM) reactions, 1034
GSMR, see Genome-scale metabolic reconstruction (GSMR)
GSNAP program, 266
Haar scaling function, 114
Haar wavelet, 114, 115, 118, 124
Hadamard product, 536
Hairpin loop, 687, 688, 693
Hamiltonian path, 990
Hamming distance, defined, 254
Hash table, 227, 229, 240, 988, 1058
Helix–turn–helix supersecondary structures, 714
Hidden Markov model (HMM), 191, 192
pros/cons, 203
High-performance computing (HPC), 981
architectures, 985, 986
application-specific integrated circuit (ASIC), 985
cloud computing, 986
genereal-purpose processors, 985
Knights Ferry (KNF), 985
message-passing interface (MPI) specification, 985
threads, 985
next-generation sequencing data, 986–995
High-performance liquid chromatography (HPLC), 1029
High-throughput technologies, omic data, 1028, 1029
dye-based sequencing, 1028
genome sequencing, progress, 1028
metabolomic technologies, 1029
H-InvDB, 1039
HIV subtype, 727–728
circulating recombinant forms (CRFs), Kohonen map, 727
HIV-1, HIV-2, 727
HIV-1 virus, 727
coreceptor CCR5, CXCR4, 727
predicting coreceptor usage, 727
HMMR gene, 168
Homology assessment, 1051–1052
definition, 1051
homologous gene regions, 1051
clustering approach, see Clustering
approaches
common approaches, 1051
seed approach, 1051
homologous sequences, 1051
HotKnots, 692, 693, 697
HPC, see High-performance computing (HPC)
HPRD database, 156, 157
Hsa.1039 gene, 438
Hsa.8147 gene, 438
Hubert’s $\Gamma$-statistics, 578, 579
Human B-cell expression, 621
data set, 619
Human genome, 9, 19, 175, 232, 244, 919, 1073
Human Genome Project, 251, 281, 282
Human hepatocyte growth factor
structure of, 705
Human immunodeficiency virus (HIV), 725
drugs resistance, 725–726
antiviral drugs, 726
decision tree, 726
genotype, 725
Human metabolomic database (HMDB), 1038
Human pathway database (HPD), 1037
HumanSDB3, 16
Hybrid algorithms., 537, 609
Hybrid (H) approach, 116–118, 602
NER systems, 857
research works, 857
several approaches, 857
Hybrid cases, 863–864
annotations method, 863
characteristics systems, 864
content provider approaches, 870
intersection, 863
named entity recognition (NER), 871
performance system, 864
union, 863
Hybrid evolutionary algorithms, 614
Hyervolume indicator, 613
Hypothesizing binary relations, 910–912
data cleaning, 912
example, 912
false-positive triplets, 912
Gene Ontology (GO), 912
discovery pattern, 910
diagram, 911
fish oil and Raynaud’s disease, relation, 910
drug–drug interactions (DDIs), 911
AnsProlog rules, 911
drug development, 911
molecular mechanism, 911
literature-based discovery methodology, 910
MEDLINE, 911
parse tree database, 911, 912
Parse Tree Query Language (PTQL), 912
semantics, 910
binary relations, for literature-based
discovery, 910
triplets direct logic translation, 912
Hypothesizing complex biology concepts, 912–913
AnsProlog rules, 912
answer set, 913
binary relations, 912
domain-specific database, 913
drug metabolism, 912
building pathways, 912
temporal ordering, 912
time points, 912
Immunoprecipitation, see ChIP- Sequencing
(ChIP-Seq)
Implied information, identification, 902–903
biological domain knowledge, use of, 905–909
for intersentential information, 903
linguistic knowledge, use of, 903–905
rule-based systems, 909
In-degree, 178
Indeterminate string, 314
Indexing
Bradyrhizobium japonicum (9,105,828 bp), 264
Brucella melitensis 16M genome, 264
BWT compression, 253
*Drosophila melanogaster* chromosome 3L (24,543,557 bp), 260
mapping 24,163,065 76-bp-long real reads, 261
mapping 25,000,000 64-bp-long simulated reads, 260
NGS Aligners, 228
*Pseudomonas aeruginosa PAO1* (6,264,403 bp), 265
Individual association rules, interestingness measures, 779–783
inversion property, 780
measures, Piatetsky–Shapiro, 780
semantic measures, 781, 782
subjective measures, 781
null-addition property, 780
objective measures, 779
row/column scaling property, 780
symmetric measure, 779
symmetry property, 779
Inference, 201, 908
Armstrong axioms, 773
cross-validation (CV), 508
cross-validation methods, difficulties, 467
GRNs based on ARs, 815
GRNs from multiple data sources, 812
inference rules, 773
maximum-likelihood techniques, 729
phylogenetic, 1053, 1065
protein’s function, 752
statistical, 354
Viterbi algorithm, 206
Inference of gene regulatory networks (GRNs), 830–831
binding data, 812
diverse data-mining approaches, 830
extracting ARs among genes, 830
genomics data, 812
identify genes, 812
inference methodologies, 830
integrating, data sources, 812
methodological advantage, 830
multiple data sources, 812
reverse engineering, 830
Information components extraction, 926–929
bag-of-words representation, 926
text document, 926
concept, 926
conceptual graph
extracted relation instances, 928
model semantic structure of text, 928
definition, 929
dependencies output, parser, 926–929
generalized association-mining technique, 926
generated information component, 928
informationComponentExtraction(T) algorithm, 929, 932–935
noun phrases (NPs) identification, 926
partial list, extracted from sample sentences, 936
phrase structure tree generated by parser, 928
PubMed sentences sample, 927
Alzheimer disease, 927
phrase structure tree representations, Stanford parser, 927
sample biological sentence, 928
NP/VP nodes, types of relations, 926
semantic tree analysis, 929
phrase structure tree analyzing rules, 930–931
verb phrases (VPs) identification, 926
Information extraction (IE), 901, 943, 944
success, 944
Information retrieval (IR), 920
contextual queries, 920
Information-theoretic filters, 407
backward elimination and relevance criterion, 408–409
conditional mutual information maximization (CMIM) criterion, 410
fast correlation–based filter, 407
forward selection, and relevance criterion, 409–410
Markov blanket elimination, 409
minimum interaction–maximum relevance (MIMR) criterion, 412
minimum redundancy–maximum relevance (MRMR) criterion, 411–412
theoretical comparison of filters, 413
variable ranking, 407
Information theory, 399
backward-elimination search, 402–403
bidirectional search, 403–404
curse of dimensionality, 400
discrete variables
redundancy, 406
relevance, 404–405
synergy, 406–407
fast mutual information estimation, 413–414
assuming normally distributed variables, 415
Information theory (Continued)
  discretizing variables, 414
  empirical estimation, 415
  Miller–Madow correction, 415
information-theoretic filters, 407
  backward elimination and relevance
criterion, 408–409
  conditional mutual information
  maximization (CMIM) criterion, 410
  fast correlation–based filter, 407
  forward selection, and relevance criterion,
  409–410
  Markov blanket elimination, 409
  minimum interaction–maximum relevance
  (MIMR) criterion, 412
  minimum redundancy–maximum
  relevance (MRMR) criterion,
  411–412
  theoretical comparison of filters, 413
  variable ranking, 407
  variable selection criteria, 404
  variable selection exploration strategies,
  400–401
  optimal search strategies, 401
  sequential search strategies, 401
  stochastic search strategies, 401
Information visualization, 1074
  computational information design, 1074
  definition, 1074
  numerical analysis, 1074
  visual analytics, 1074
In-neighborhood, 178
Insertion/deletion events (indels), 226
Inside edge binary clustering with missing
values (IEBCMV), 649
Instability, 334
  feature selection algorithms and, 334
Integrated microbial genomes (IMG), 1039
Integration, 1114–1116
  measurements, 1114–1115
  images, 1115
  networks, 1115
  numerical measurements, 1114
  volumes, 1115
  metadata, 1115–1116
  definition, 1115
  importance, 1115
  types, 1115
  multimodal biological data, 1114–1116
  views, 1121–1123
  function, 1123
  overview, 1122
Interestingness measures, 778–784
  aim, association rules, 778
  hierarchical representation, 778
  interestingness measures categories, 778, 779
  types of criteria, 778, 779
  Intergenic distance, 626, 629
  Internal measures, 583
Intervals
  confidence, 447
  equal-frequency discretization scheme, 414
  equal-frequency-interval discretization, 745
  equal-interval-width discretization, 745
  ESPR-NPC method, 515
  fuzzy permutation, 677
  gene expression, 665, 666
  in MaxQuant, 61
Intrinsically unstructured proteins (IUPs), 716
Intrinsic separation ability, 532
Intron, 15, 41, 207
Intronic splicing enhancers (ISEs), 7
Intronic splicing silencers (ISSs), 7
Isoleucine, 99
Itemset mining, constrained, 741, 749, 751, 765, 766, 767
  Iterative signature algorithm (ISA), 602
  IUB/IUPAC nucleic acid codes, 311
Jaccard set-based index, 339, 345
Jackknife techniques, 545
Joined genomic exons (JGEs), 15
Kappa symmetric measures, 780
  KCNMA 1 gene, 9
  k-dependence Bayesian (kDB) classifier, 664
  Kendall distance, 528
  K-means algorithms, 535, 559, 745, 809
  k-nearest-neighbor classifier, 343
  KNImpute method, 648, 649
  KnotFold execution, 699
  Knuth–Morris–Pratt algorithm, 847
  Kronecker delta, 575
  Kronecker product, 505
  Kullback–Leibler (KL) divergence, 823
  Kyoto Encyclopedia of Genes and Genomes
      (KEGG), 35, 1036
Laboratory information management systems
      (LIMSSs), 37
LAGAN, 973
Large-scale
  genetic regulatory systems, 826
  genome sequence, 1082
phylogenetic analyses, 1053
protein analysis, 59
protein expression measurements, 285
RAM for, 1013
Latencies, 991, 1055
Lattice miner (LM), 564
Leukemia data set
  adjusted Rand index curves, 539
  F-index curves for, 543
  FM-index curves, 541
  ROC curve for, 531
Linear discriminant analysis (LDA), 65
Linear regression technique, 445
  leverage points, 446
  validation, 445
Linguistic knowledge, 903–905
  anaphora, 903
  algorithms, 903
  exploitation, 903
  recognition, 903
  coreference relations, 903
  event extraction system, 903
  F-score, event–argument relation
    identification, 903
  intrasentential event–argument relations
    identification, 903, 904
  postprocessing step, 904
  subtasks, 903
  support vector machine (SVM),
    pipeline-based classifier, 903
  fusion operation, 905
  information pieces combination,
    905
  performance, improvement of, 905
  heuristic rules, 904
  clause-level patterns, matching of,
    904
  performance, improvement of, 905
  working, 904
human readers vs. IE systems, 903
Markov logic network (MLN) model, see
Markov logic network (MLN) model
posttranslational modifications (PTMs),
  905
  recovering site information rules, 904, 905
  recovering theme information rules, 904
  discourse principles, 904
  relation extractor, biology, 903
  intrasentential relation, 903
Linkage disequilibrium (LD) blocks, 353
LLS impute method, 648
Local optimum, 401
Local-pooled-error (LPE), 667
Local search, 600
Logic, see Truth tables
  Boolean function, 168, 826
Logistic/linear regression model, 203
Log likelihood
  for CRF, 200
  EM algorithm, 560
  negative, 81
  regression coefficients, 487, 488
Log ratio, 380, 660, 1093
LOOCV classification accuracy, 668
Lossless, impact, 771
Lymphoblastic, 604
Lymphoma
  adjusted Rand index curves, 540
  F-index curves for, 544
  FM-index curves, 542
Machine learning, 333, 381, 860–864, 909
  biomedical domain, 860
  correct brackets, CRF model, 862
  external tools, 860
  Java frame work, 860
  MALLET’s feature definition, 860
  ML-based applications, 860
  orthographic features, 860
  postprocessing module, 860
  steps and resources train ML model, 861
MAD, 367, 680, 681
Madison metabolomic consortium database
  (MMCD), 1038
MAFFT, 1053
Mahalanobis distance, 528, 531
Making treatment related decisions, 729
  cross-drug resistance, 729
  treatment for patient, 729
Mammalian genomes, 225, 234, 240,
  244, 252
Manhattan/city block distance, 528
Mann-Whitney statistical test, 619, 620, 662
Mapping and assembly with quality (MAQ)
  algorithms, 983
Mapping rules, 47
  categories for GEDAW, 47
  cognitive mapping rules, 48
  semantic and cognitive mapping rules, 48
  structural mapping rules, 47
Marginal model plots, 446
Market basket analysis, 737, 757, 791
Markov chain Monte Carlo (MCMC), 211, 304,
  492, 1054
Markov logic network (MLN) model, 904

F-score, event–argument relation identification, 904

GENIA corpus, 904

performance, increase in, 904

saliency in discourse, concept, 904

Markov random field (MRF), 192

applications, 202

based on PPI network, 209

online resources, 204

pros/cons, 203

Mass spectrometry (MS), 1028

based proteomics studies, 333

hierarchical clustering of, 565

technology, 59

Matlab script, 524

Matrix factorization (MF), 536

Matrix representation, of partial least-squares regression, 494

Maximum a posteriori principle (MAP), 560

E-step/M-step, 560

Maximum clique, 199, 651

Maximum entropy, 197, 202, 203, 855, 947, 949, 961, 962, 963

Maximum-entropy Markov models, 203

Maximum independent set problem, 652

Maximum likelihood (ML), 559

analyses, 1064

estimate, 483, 1063

inferences, 1054

parameter estimation, 200

procedure, 483

techniques, 729

Maximum oligonucleotide mapping (MOM), 228, 986, 987

Maximum parsimony (MP), function, 1055

Mean-squared residue (MSR), 594, 610

Measuring expression levels, 283

Mediation systems, 37

BioKleisli, 37

Discovery Hub, 37

DiscoveryLink, 37

integration component, 37

internal query language, 38

limitation, 38

object protocol model, 37

Object-Web Wrapper, 37

ontology-based integration, 38

P/FDM, 37

TAMBIS, 37, 38

types, 37

Mediators, 37

MEDLINE, 919

medical subject headings (MeSHs), 919, 920

Memory

external memory algorithm, 1057

FM index, 232

human genome, 992

Mesenchymal stemlike (MSL), 659

Metabolic disease database, 1038

Metabolic network modeling, 1030–1031

genome-scale metabolic models, 1030, 1031

kinetic modeling, 1030, 1031

metabolic networks, 1030

structural modeling, 1030, 1031

flux balance analysis (FBA), 1030

topological analysis, 1030

Metabolomics, 1029

advantages, 1029

database classification, 1036

definition, 1029

MetaCyc, 1038

Metagenomic analysis, 1004

Metaplasia, 380

Metropolis criterion, 709

Microarray analysis

publications addressing, 569

steps, 584

techniques, 648

Microarray, and gene expression data, 427–428

gene expression levels, 428

logarithmic ratio, 428

Microarray clustering context

search procedure, 571

Microarray data (MAD), 381, 422, 557, 680

clustering evaluation measures, 582

Microarray data analysis

approaches of, 663–667

cancer classification and prediction, 657

clustering techniques, 625

distance function, 525

experimental design, 659–660

work flow of, 658

experimental study, 667–669

classification, 667–669

statistical tests, 667

normalization, 660–661

ranking, 661–662

Microarray data sets, 566–567

Microarray experimental design, 658
Microarray expression profiles, 165, 208, 281, 521
Microarray gene clustering, 570
Microarray gene expression data analysis, 281, 286, 423, 432, 787, 812
Microarray Interval Discriminant CLASSifier (MIDClass), 658
Microarray technologies, 283–285, 354, 380, 421, 625
data, 286–287
Microbial communities, 225, 984, 1003, 1004
types, 1004
Microbial rDNA clones, 653
Microbial richness estimation accuracy, 1018, 1019
assessment, 1018, 1019
MicroRNA noncoding RNAs, 685
posttranscriptional regulation of genes, 685
precursors, 685
Microsatellite, 1075
MIDClass operates, 665
Minimal spanning tree (MST) problem, 524, 526
Minimum clique partition (MCP), 651
Minimum free energy (MFE) model, 213, 686, 688, 692, 693
Minimum spanning tree (MST), 524
Minkowski distance, 527
MIPS database, 87, 208, 637
miRBase, 686
MirID
accuracy rates of, 692
feature-mining algorithm, 690
screenshot, 691
steps, 689
Mismatch-aware hash function, 229
Mixture modeling, 566
ML, see Maximum likelihood (ML)
ML-based approaches, 848–856
relations between ML approaches, 849
work flow of ML-based NER system, 849
MLColumnAssign, 632
MLRowAssign, 632
MOBI algorithm, 614
Model-based clustering, 559
biological measures, 579–581
data sets, 584–586
discussion, 581–584
EM algorithm, 560
classification of, 561
stochastic EM (SEM) algorithm, 561
evaluation measures
search procedure and classification, 570–571
external measures, 576
classification-oriented measures, 577–578
naive statistics, 579
similarity-oriented measures, 578–579
finite-mixture models, 560
geometric clustering algorithms, 557
hierachical, 557–558
partitioning clustering, 558–559
internal measures
cluster stability, evaluation of, 574–575
predictive power of, 576
preliminaries, 572
single cluster, evaluation of, 573
single gene, evaluation of, 572
whole clustering, evaluation of, 573–574
Model-based clustering algorithms, 557, 559–561
Model, ML techniques, 855, 856
concept ML technique model, 856
future data, 855
semisupervised learning, 855
graphical structure of CRFs, 856
supervised learning, 855
MOEAs, see Multiobjective evolutionary algorithms (MOEAs)
Molecular dynamics (MD) simulation, 707
Molten globule (MG), 715
free energy of, 716
Monte Carlo (MC) method, 708
Morphology
derivational, 921
inflection, 921
Motif finding, 310, 982, 984
Motif representation, 311–312
MouSDB5, 16
MS-based proteomics, 59
experimental procedure, 60
Multi-database query languages, 37
Multifactorial, 660
Multilabel sentence-level classification, 961
evaluation, 961
MALLET library, 961
maximum-entropy experiments, 961
multilabel, sentence-level results, 961
naive Bayes EM classifier, 961
Multilayer perceptron (MLP), 675
perceptrons, 678
Multimodal biological data, 1109–1112
approaches to discover knowledge
database-oriented approaches, 1111–1112
numerical data combination, 1112
tools to combine data, detailed
comparison, 1113
biological knowledge, 1109
accumulation cycle, 1110
distribution, 1109
data domains, 1110
illustration, 1112
tools selection, to combine data, 1112
graph-based modeling approaches, 1110
important data types, 1110
illustration of, 1111
modality, 1110
preprocessed data, 1110
analysis, 1110
visualization, 1110
spatial information, 1110
transformation into biological knowledge, 1110
Multiobjective evolutionary algorithms
(MOEAs), 602, 612
Multiobjective genetic algorithms, 425–426
crowding distance, 426–427
crowding selection operator, 427
nondomination, 426
NSGA-II, 427
Multiobjective model, for biclustering, 611–614
definition of, 614
optimization
Pareto dominance, 611
Pareto front, 612–613
Pareto optimality, 611
Pareto optimal set, 611–612
problem, 611
quality indicators, 613
Multiobjective optimization problem (MOP), 611
Multiple alignment, 971
Multiple-classifier fusion method, 677
Multiple-label, sentence-level classification, 948–949
biomedical abstracts, 948
event-type probability, determination, 948
GENIA corpus data, 948
MALLET library, 948
classification algorithms, 948
maximum-entropy
main aspects, 949
MaxENT classifier, 949
naive Bayes EM, 949
negative regulation, 948
plain-text sentence, 948
phrases italicized indicates, 948
PUBMED abstracts, 948
Multiple-sequence alignment, 1052–1053
advances, 1053
algorithms, 1052
challenge, 1053
concatenated matrices, 1052
consist of, 1052
divide-and-conquer approach, 1053
implementations, 1053
profile alignment, 1053
implementation, 1052
large data set assembly problems, 1053
major uses, 1052, 1053
Multiple treatment groups, modeling
functional analysis-of-covariance
(fANCOVA) model, 513–515
treatment effects, 515–516
MUSCLE, 1004, 1053
Mus musculus, 16, 261
Mutations, 380, 728–729
indentifying mutation selection, 728–729
genotype mutations, 728
mutation pressure, transcriptase enzyme, 728
Mutual information (MI), 526, 1015
Myoglobin, x-ray crystallography, 705
Naive Bayes model, 203
Naïve statistics, 577
Named entity recognition (NER), 901, 920
National Center for Biotechnology Information
(NCBI), 1059
database, 890
RefSeq, 157, 874, 879, 886, 890
National Library of Medicine (NLM), 919
NCI60 adjusted Rand index curves, 539
F-index curves for, 543
FM-index curves, 541
Nearest neighbor
approaches, 752
classification, 437
classifier, 343, 432, 665
clustering, 648
Nearest rule, 958
benefits, 958
definition, 958
event class rule, 958
Neighborhood, 165, 176, 208, 210, 244, 602
based algorithms, 609
function, 210
nodes, 199
search approach, 600
Neighborhood search (NS) approach, 600
Network, 177–178
for epistasis studies, 181–182
  GAIN, 182
  SEN, 181, 182
natural and artificial, 177
Network inference, 412, 414, 812
  algorithms, 812
  task, 412
Network models, in understanding disease, 166
  interactome network for disease prediction, 166–167
  network perturbation due to pathogens, 167
  network view of cancer, 167–168
Network topology, 823
Neural networks, 178, 287, 477, 727, 752
Neuro-fuzzy computing, 422
Neuro-genetic hybridization, 422
New hypotheses, prediction, 909–914
  AnsProlog rules, 912
  binary relations, used to instantiate discovery pattern, 914
c-o-currence-based statistical analyses, 910
data cleaning, 912
discovery pattern, 911
disjoint, sets of literature, 909, 910
future prospect, 914
Gene Ontology (GO) and Entrez Gene summary, use of, 912
hypothesized disease–drug relation, 914
hypothesizing binary relations, see Hypothesizing binary relations
  Hypothesizing binary relations
hypothesizing complex biology concepts, see Hypothesizing complex biology concepts
imperfect precision, cause of, 914
incorrectly inferred DDIs, 912
inference rules used in, 911
literature-based discovery, 910
logic representation of repaglinide metabolic pathway, 913
MEDLINE abstracts, with semantic annotations, 912
notion of discovery pattern, 910
parse tree query language (PTQL), 912
PharmGKB, 913
relation extraction systems, 914
text-mining systems, 914
Newton’s equation, 708
Next-generation sequencing (NGS) technology, 982–985
  applications, 982–985
  chromatin immunoprecipitation sequencing (ChIP-seq), 984
  epigenetics-methylation, bisulfite sequencing, 984
gene assembly, 982, 983
  algorithms, 982
  assembly definition, 982
  comparative assembly approach, 982
de novo approach, 982
  main challenges, 983
machines error correction, 995
  achieving parallelism, 995
  cloud computing, 995
metagenomics, 984, 985
Roche 454 Life Sciences sequencer, 982
Solexa genome analyzer, 982
transcriptome sequencing, RNA-seq, 983, 984
  computational challenges, 983, 984
definition, 983
NGAM (non-GAM) reactions, 1034
NGS data, 226
NLP techniques, 939
NMF scheme, 536, 537
NN classifiers, 667
Nodes, 80, 158, 178, 193, 199, 207, 211, 240, 286, 288, 633, 665, 990, 1055, 1088
Noncoding RNA (ncRNA), 687
Nondominated sorting genetic algorithm (NSGA-II), 423, 602
Nonnegative matrix factorization (NMF), 522, 524
Normalized mutual information (NMI), 636, 637, 1015
equation, 1015
Normalized unscaled standard error (NUSE), 660
Notable classifiers, 1003
Nova1 and Nova2 proteins, 11
Nova proteins, 11
NSGA-II, Pareto front projections, 622
Nuclear magnetic resonance (NMR) spectroscopy, 703, 1029
Null hypothesis, 364, 365
Object-role modeling (ORM), for object-oriented software, 129
OBO Foundry ontologies, 54
Obtaining control, and background estimation, 364–365
ConReg-R, 365–367
testing with explicit control data, 365
Omics era, 1027–1029
biomolecular networks, 1027
innovative technologies, 1027
omic data integration, 1028
omic data regeneration, 1028
OMMBID, 1037
One-pass extraction, 952–955
extraction algorithm, 952, 953
extraction rule, 952
multiple events, 953, 954
cases, per sentence, 953
disjoint events, 953
nested events, 953, 954
dependency parse, 953
extraction, 953
trigger word, 953
sample parse, extraction, 954, 955
Online Mendelian Inheritance in Man (OMIM), 35
Ontology-based approaches, 49
in life sciences, 53
Ontology-driven conceptual data modeling, 130
waterfall methodology, 130
Open Biological and Biomedical Ontologies (OBO) Foundry, 53
Open reading frames, 1034
Operational taxonomic units (OTUs), 653, 1003
Operon annotation, 626, 629
Optimization problems
heuristic techniques, 425
multiobjective, 770, 771
multiobjective genetic algorithms, 425
single-objective, 426
Orchini distance, 528
Order of magnitude, 390, 392, 550, 551
Order-preserving submatrix (OPSM) algorithm, 602
Organism database, 1038
Other techniques, inference of GRNs, 829–830
average linkage clustering, 829
clustering, 829
clustering methods, 829
dimension problem, 829
gene network, 829
graphical Gaussian modeling (GGM), 829
Out-degree, 178
Outside edge binary clustering with missing values (OEBCMV), 649
Overlap graph, 234, 237, 238, 240, 990, 991
Overlap–layout–consensus (OLC) assemblers, 237–238
CABOG, 239–240
EDENA, 238
Newbler, 239
SGA, 238–239
Oversampling, 982
Pairwise alignment algorithms, 971–974
pairwise global alignment algorithms, see Pairwise global alignment algorithms
pairwise local alignment algorithms, see Pairwise local alignment algorithms
Pairwise global alignment (PGA), 1005, 1006
ESPRIT, 1006
filtration, 1006
Needleman–Wunsch algorithm, 1005
based on dynamic programming (DP), 1005
genetic distances computation, 1005
sparse matrix representation, 1006
three steps, 1005
Pairwise global alignment algorithms, 972, 973
anchoring approach, 973
steps, 973
vs. dynamic programming approach, 973
dynamic programming approach, 972, 973
maximum-score path, 972, 973
Needleman–Wunsch algorithm, 972, 973
substitution matrix, matrix $M$ construction, 972
two step procedure, 972, 973
Pairwise local alignment algorithms, 973, 974
dynamic programming approach, 973
score calculation matrix equation, 973
time complexity, 973
vs. Needleman–Wunsch algorithm, 973
seeding approach, 973, 974
based on filtering notion, 974
characterization, 974
representation, 974
Pairwise methods of inference GRNs, 829–830
correlation metric construction (CMC), 829
distance matrix, 829
gene pairs by cross-correlation, 829
machine learning algorithm, GRNCOP, 829
prototype pairs, 829
regulation matrix, 829
relationships among genes, 829
Pairwise vs. many-to-one associations, 811
distribution and assumption, GRNs, 811
gene regulation, 811
many-to-one regulatory functions, 811
one-to-one regulatory functions, 811
Parallel hyperbolic sampling (PHS), 708
Particle swarm optimization (PSO), 675, 678
metaheuristic algorithms, 676
Particle swarm optimization functional link
artificial neural network (PSO-FLANN), 675
PathCase, 1038
Pathway ontology, 53
Pearson correlation, 572
Pearson distance, 528
Percent accepted mutations (PAMs), 972
Performance evaluation, 935–938
ABNER, 935
biological relation extraction process
evaluation results, 939
biological system, 935
F-score, equation, 938
performance judgement, 935
precision value (π), 937
extracted biological relations
correctness evaluation, 935
GENIA abstracts, 935
recall value (ρ), 937, 938
Perl script, 17
Permutation
in differential expression studies, 369–371
of DNA sequence, 123
fuzzy permutation, 677
QQ plots, 371
symmetry property, 779
p53 gene, 167
PHP database, 17
PHP scripts, 17
Phrase-level labeling, 949–950, 961, 962
CRF
annotations, 962
performance evaluation, 962
evaluation, 961, 962
event trigger phrase classification, 949
GENIA, 961
annotated development, 961
annotations, 962
corpus, 961
invalid event phrases considering context, 950
phrase-level classifier, 949
training data observations, 949, 950
result analysis, 962
selected event annotation, 949
valid event not labeled, 949
valid event phrases considering context, 949
Phylogenetic inference, 1053–1057, 1059–1060
computational inference, 1053
computational limits, 1064–1066
challenges, technical level, 1065
computer architectures, 1065
German Science Foundation (DFG), 1064
Γ model, 1065
likelihood function, 1065
major problem, 1064–1065
prolegomena, 1065
crunching many-taxon data sets, 1055
key problem, 1055
long run-time requirements, 1055
many-taxon/few-genes data sets, 1055
crunching phylogenomic data sets,
1053–1055
42 AMD Magny-Cours nodes, 1054
emerging issue, 1053
fine-grain approach, 1054
issue, 1055
low-latency interconnect network, 1054
Markov chain Monte Carlo (MCMC), 1054
maximum-likelihood inference, 1054
ML-based code, 1054
ML-based inference, 1054
ML-based phylogenetics, 1054
ML-based tree searches, 1054
randomized accelerated maximum
likelihood (RAxML), 1054
data avalanche, 1059–1060
data sets, 1060
experiments, 1064
PHLAWD, 1059
divide-and-conquer algorithm, 1059
reducing memory footprints, 1056–1057
algorithmic techniques, 1056
external memory algorithm, 1057
future phylogenomic data sets, 1057
large phylogenies computation, 1056
mesh-based approach, 1056
phylogenomic data sets, 1056
reducing used RAM, 1057
subtree equality vectors (SEVs), 1056–1057
Phylogenetic inference (Continued)
ribulose-biphosphate carboxylase large
(rbcL), 1059–1060
Amborella, sequence for, 1060
Viridiplantae, green plant, 1059
x86 vector intrinsics, 1055–1056
AVX intrinsics, 1055–1056
AVX vectorization, 1056
issue, 1055
Parsimonator, 1056
RAxML-Light, 1056
Phylogenetic likelihood function, 1056
Phylogenetic postanalysis, 1060–1064
alignment problem, 1063
bootstrap support, 1061
data set, 1061
distribution, 1061
hierarchical patterns, 1061
clades, 1062
data avalanche, 1060–1064
data sets, 1061, 1062
examination, 1062
flowering plants, 1064
“known” phylogeny, 1064
maximum-likelihood estimates, 1063
multigene angiosperm phylogeny, 1062, 1063
rbcL analyses
plot of new species, 1061
plot of support for, 1062, 1063
ribulose-biphosphate carboxylase large
(rbcL), 1060–1064
plot of species, rbcL data sets, 1060
trees, 1060–1063
unique sequence names, 1060
Phylogenetics, 1050
recent advances, 1050–1059
Phylogenetic trees, 1089–1090
large tree visualizations, 1090
iTOL, online tool, 1090
tree of life visualization, 1091
multiple-alignment output, 1089
tree maps, 1090
trees representation, 1090
linear dendrogram, 1091
traditional way, 1090
tree visualization, 1090
treevolution, 1090
visualization tools, 1090
Voronoi map, 1090
Phylogenomic alignments
under RAM restrictions, 1057
Phylogeny
divergence–time analysis, 1066
likelihood-based, 1055
molecular evolution and fossil information,
1058
multigene, 1051
angiosperm, 1061
multiple-sequence alignment, 1050, 1052
mutations, 653
PknotsRG, 697
PknotsRG methods, 692, 697
PMN, 1038
PNG files, 17
Poisson distribution, 64
Poisson responses, 506
3’ Polyadenylation, 5
Polynomial-time approximation algorithm, 652
Polynomial–time arc-traversal algorithm, 292
Polynomial–time sign propagation algorithm, 294
Polypeptides, 35, 162, 285
Position-specific scoring matrices (PSSMs), 311
Positive matrix factorization (PMF), 536
Positive predictive value (PPV), 698
Post-genomic era, clustering
experimental set-up
algorithms, 524
external indices, 524–525
generic data set, 522–524
hardware, 524
microarray expression data, 521
three-step process, 521
PostgreSQL database, 17
Postprocessing
entity names, 847, 848
ambiguous terms, 847
antilymphocyte globulin (ALG), 847
external information, 848
protein identifier, 847, 848
protein identifiers, 847
species ambiguity, 848
unique identifier, 848
ML solutions, 856
errors, 856
recognize more entity names, 856
remove recognition errors, 856
Postprocessing phylogenetic trees, 1057–1059
bipartition concept, 1057
binary tree, 1057
nontrivial bipartitions, tree, 1057
unrooted binary tree, 1057
computational limits, 1066
comparative methods, 1066
discrete postanalysis, 1066
divergence–time estimation, 1066
MRE trees, 1066
rogue taxon identification, 1066
computing consensus trees, 1058
bipartitions, 1058
building a consensus tree, 1058
main flavors, 1058
majority-rule consensus tree, 1058
reconstructing, 1058
rogue taxa, 1058
computing distances between trees, 1057–1058
quartet distance, 1058
Robinson–Foulds (RF)
topological distance, 1057, 1058
value, 1058
subtree prune-and-regraft (SPR) distance, 1058
divergence–time analyses, 1058–1059
Bayesian autocorrelated methods, 1058
Bayesian uncorrelated methods, 1058
BEAST, 1058, 1059
estimation problem, 1059
likelihood surfaces, 1058
maximum-likelihood phylogenies, 1058
nonparametric rate smoothing (NPRS), 1058
PATHD8, 1059
penalized likelihood (PL), 1058
postprocessed consensus trees, 1058
phylogenetic postprocessing, 1057
Posttranslational modifications (PTMs), 285, 905
Power law, 181, 184, 811
PPI, see Protein-protein interactions (PPI)
Preliminaries of biomedical knowledge, 843–845
corpora, 843
annotation procedure, 843
text documents, 843
Pre-miRNA classification, 687, 691
Principal-component analysis (PCA), 528, 675, 678
Prion protein (PrPc)
structures of, 704
Probabilistic Boolean networks (PBNs), 826
Probabilistic models, 855
Probabilistic neural networks, 422
Processing speed, 1019, 1020
assessment, 1019, 1020
runtime profiling, 1020
CRiSPy-CUDA, 1020
vs. ESPRIT-Tree, 1020
vs. UCLUST, 1020
Profile-kernel method, 505
Profiling CUDA codes, NVIDIA visual profiler, 1007–1009
CUDA optimization guidelines, 1008, 1009
genetic distance computation, 1008
occupancy profiling, 1009
k-mer distance
computation, 1008
occupancy profiling, 1009
processing, 1008
Proposed biological relation-mining system, 924–935
architecture, 924
document preprocessing and parsing, see
feasible biological relation identification, see
Document preprocessing and parsing
Information components extraction, see
Primary focus, 938
uniqueness, 938–939
contextual analysis, 939
feasible biological relations, 938, 939
GENIA corpus, 939
mining biological relations, 939
mining relational verbs, 939
primary focus, 938
text-mining-based approach, 938, 939
PROSITE database, 309, 312, 326
Protein Data Bank (PDB), 35, 704
extract peptide fragments, 710
PSI-BLAST, 711
Protein–DNA, 155, 156, 354, 790, 812, 813, 1029
Protein, 3D structure, 714
Protein-folding problem milestones, 705–706
Protein function prediction, using Markov
random fields, 208
function prediction, 208–211
integrating multiple data sources, 211–212
Markov random field for function prediction, 208–211
protein tertiary structure prediction, application to, 213
ab initio protein structure prediction, 217
Protein function (Continued)
  CRF model for protein threading, 215–216
  free-energy estimation, 213–215
  side-chain prediction, 213–215
Protein–gene interaction, 168
Protein identification, 59
  approaches to cleanse MS data for, 60
  approach for improvement, 61
  identification filtering approach for, 64–66
  new dynamic wavelet-based spectra preprocessing method, 63–64
  new target-decoy approach, 66
  preprocessing existing approaches, 61–63
  results, 60
evaluation
data set, 68
  new identification filtering method, 70–72
  of new preprocessing method, 68
  strategy, 68–70
false discovery rates (FDRs), 60
limitations, 61
protein Q05649 and its sequence, 65
target–decoy filtering approach, 66
unequal matching bias, 66
Protein interactions, 753–755
  antibodies, 753
  collagen, 753
  enzymes, 753
  oxygen, 753
  protein function, 754
  protein graph, 754
Protein–protein interactions (PPI), 11, 77
classification, accuracy, 85–86
data, 77
  classification method, 85
  false, identification of, 84–85
  filtering, as critical preprocessing step, 77
  reliability, 87
integration of GO data to assess validity, 77–78
semantic similarity, evaluation of, 78
  annotation-based methods, 81–82
  correlation with functional categorizations, 83
  edge-based methods, 80–81
  gene ontology, 78–79
  hybrid methods, 82–83
  node-based methods, 82
  selected PPIs with zero-valued, 88
survey of measures, 79–80
Protein structure prediction (PSP), 703, 704
Protein tertiary structure prediction, 213
ab initio protein structure prediction, 216–217
Fisher–Bingham (FB5) distribution, 216
  possible FB5 distributions, 216
  conditional neural field for, 214
CRF model for protein threading, 215–216
  factor graph, 213
  free-energy estimation, 213–215
  side-chain prediction, 213–215
Proteome, 35
Proteomic technologies, 1028, 1029
  mass spectrometry, 1028
  yeast two-hybrid (Y2H) systems, 1028
Pseudoknot prediction methods, 698
Pseudo pre-miRNA, 689
Publish or Perish (POP) tool, 570
PubMed, 919
  search result of, 920
  services, 920
P-value distribution, in statistical significance assessment, 359–360
  beta uniform-mixture (BUM) model, 363
Pyrosequencing technologies, 1003
Qualitative probabilistic networks (QPNs), 281, 291–294
dynamic, 295, 296
  aided learning with qualitative joint influences, 300
  computational experiments, and results, 300–303
constructing qualitative model, 297
dynamic instance, 295–296
  estimating time lag, 300
  extracting most likely regulators, 298–300
  gathering potential regulators, 298
generalized joint influences, 296–297
  muscle development network in Drosophila larval stage using, 302
  qualitative influences, 296
temporal snapshot, 295
  for gene regulation, 294–295
Quality assessment (QA) programs, 712
Quartile, 661
Quasi-maximum-likelihood estimator (QMLE), 510
Ramachandran plot, 705
Rand index, 578
Random access memory, 232, 994, 1018, 1055
Random forests, 664
Randomization, 345, 368, 821
Random walk, 110, 112, 117
analysis, 165
Ranking and clustering motifs, 310
Raò's score test, 483
Rate heterogeneity, 1054–1056, 1065
RatSDB2, 16
RDP database, 1014
Reactome, 1037
REad ALigner (REAL), 253, 254
algorithm, 254–258
complexities, 258–259
experimental results, 259–261
mapping short reads, 254
Recurrent neural networks, 829
RefSeq, 157, 874, 879, 886, 890
Regression analysis technique, 445,
477–479
leverage points, 446
validation, 445
Regression models for biological data,
445
fitting the model, 446–447
marginal model plots, 458–463
nonconstant variance, 451–458
patterns in residual plots, 463–466
validity of model, 447–451
variable selection, 466–473
bias corrected version of, 468
cross-validation, 466–467
leaps-and-bounds algorithm, 468, 469
nestreg, 466
partial \( F \)-test, 466
\( R^2 \) adjusted information criterion, 467
residual sum of squares, 466
\textit{vif} command, 469
\textit{vselect}, 466
variable transformation, 451–458
Regular framework, 768–769
extended representation, 769
frequent regular itemsets, 769
Regulatory regions, 169, 750
Regulatory RNA binding proteins (RBPs), 8
RegulonDB database, 908
Related work
described techniques, 848
databases, drugs, 848
detect, drug reactions, 848
drug–disease pair, 848
EU-ADR European Project, 848
gene dictionaries, 848
gene/protein normalization system, 848
proMiner, 848
Relational data visualization, 1085–1092. See
\textit{also} Functional data visualization
Relation extraction, 901
Relation extraction (RE), 901
Relative logarithmic expression (RLE), 661
ReliefF algorithm, 345
ReliefF-based algorithm, 339
Resequencing projects, 226
Reverse engineering gene regulatory networks,
286
Ribosomal DNA (rDNA) gene libraries, 645
Ribosomal processing, 637
RNA classification, 685
data sets, 686–687
experiments/results, 697–699
feature pool, 687–688
Pre-miRNAs, algorithm, 688–689
pseudoknots, in silico prediction of
algorithm, 695–697
ensemble approach, 692–693
secondary-structure prediction, stem-based
estimator for, 693–695
sequences, 685–686
Web server, 689
RNA data analysis, 685
RNA molecules, 5
RNA polymerases, 10, 282
RNA pseudoknots, 695–697
RNA-seq data analysis, 993, 994
depends on, 993
parallelism techniques use, 993
RseqFlow, MPI-based tool, 994
RUM, 994
tools, 993
whole transcriptome shotgun sequencing,
993
RNA splicing, 5
RNA transcripts, 5
Robust multichip average (RMA), 667
Rough-set theory, 422, 423
discernibility matrix, 425
granularity representation, 423
lower and upper approximations of rough set,
424
reducts, 424–425
RPROP algorithm, 207
\textit{Saccharomyces cerevisiae}, 627
Sample classification, performance, 343
Sample parse, extraction, 954, 955. See also
One-pass extraction
BioEve extracts, 954
dependency parser, 954
extraction module, 954, 955
extraction rule, 954, 955
sample sentence, 954
binding event extraction, 956
dependency parse tree, 954
negative-regulation event extraction, 955
trigger words, 954, 955
Sample size, 336–337, 339
Sandwich formula, 508, 509
Schwarz criterion, 579
Scoring
function, 215, 310, 686, 697, 710, 711, 2902
matrix, 1011
metrics, 60, 290
SDBs, see Splicing databases (SDBs)
Seeding approach, 973
Self-organizing maps (SOMs), 558, 629, 647
Semantic classification, 959–962
document-level classification, 959
event phrase labeling, 959–962
sentence-level single-label classification, 959
Semiparametric partially linear model
bandwidth selection/inference, 508–509
covariance estimation, 509–510
efficient semiparametric regression method,
505–508
model assumptions, 503
modeling within-subject covariance, 509
nonparametric covariance estimation,
510–513
profile estimator, 503–505
single treatment group, 502
Semisupervised learning
metric-based method, 628
SEN, see Statistical epistasis networks (SEN)
Sensitivity, 260, 261, 264, 274, 530, 698
FDR curves, 369
Sensor network, 629
Sentence-level classification
BioEve event extraction module, 946
bioevent extraction, 946
first step, 946
biomedical literature, 946
conditional random-field-based classifier, 950
conditional random fields (CRFs), 950
part-of-speech tagging, 950
dictionary-based semantic class labeling, 946
feature selection, see Feature selection
incremental approach, classification task,
946–947
classification approaches summarization,
947
naive Bayes algorithm, 946
labeled phrases, 946
classification, 946
labeled trigger, 946
multiple-label, sentence-level classification,
see Multiple-label, sentence-level
classification
phrase-level labeling, see Phrase-level
labeling
semantic labeling, 946–951
single-label, sentence-level classification, see
Single-label, sentence-level
classification
trigger phrase dictionary, 951
filtered trigger words, 951
gene expression, 951
GENIA training data, 951
stopword cleaning, 951
trigger words, 951
SEQMAP software, 253
Sequence alignment tools, 986–990
algorithmic approach, 986
AMD Opteron processor, 989
Burrows–Wheeler transform (BWT)
approach, 986
complication, 989
index, 989
CloudBurst, 988
cutting-edge graphics cards, 989
end-user parallelization, 990
FANGS, 988
GNU MAP, 988
limitation, 988
hash-based tools, 986
MapReduce programming model,
990
MPI-based parallelization, 988, 990
MPI-based tool pMap, 990
mrFAST, hash-based index, 990
MUMmerGPU, 989
Novoalign, 988
NVIDIA GeForce 8800 GTX, 989
configuration, 989
OpenMP, 988
parallelization, 986
RMAP’s algorithm, 988
sequencing data, 986
exponential increase, 986
SHRiMP2, 987
Smith–Waterman local sequence alignment algorithm, 990
SOAP, hash-based aligner, 989
suffix tree based approach, 986
thread-level parallelism, 989
tools and algorithms, NGS data, 987
using HPC, 987
Sequence data visualization, 1078–1085
alignment data visualization, see Alignment data visualization
basic sequence representation, 1078
genome data visualization, see Genome data visualization
major structure patterns, sequence, 1078
number of leucines on a 20-amino-acid, Drosophila melanogaster, 1079
SLIT protein sequence, Drosophila melanogaster, 1079
visualizing abundance, 1078
visualizing complex sequences techniques, 1078
Sequence(s), generally
alignment, 48, 191, 241, 727, 977, 986–990, 1016, 1083
analysis, 728, 1075
comparison, 1051, 1082
motifs, 309
Sequencing by ligation (SOLiD), 226
Sequencing by synthesis, 226
Sequencing technologies, 251, 354
automation, 251
Burrows–Wheeler transform, 253
denovo assembly of genome, 252
alignment programs, 252
mapping, 252
variants identification, 252
ELAND, SOAP, FM and MAQ indexes, 253
multiplex polony sequencing protocol, 252
partitioning into exact matches, 253
plus-and-minus method, 251
Sanger sequencing, 252
SBS technology, 251, 252
sequencing systems, and technology developments, 252
traditional methods, 251
Sequential data mining, 750–752
biological sequences, 750
chi emerging pattern, 752
exons, 750
introns, 750
mRNA, 750
Shortest path, 80, 179, 923
Significance analysis of microarrays (SAM), 662
Silence splicing, 8
Simian immunodeficiency virus (SIV), 167
Simulated annealing (SA), 675
Single-gene classifier (SGC), 667
Single genetic marker, 479–482
common quantities defined by, 481
and Fisher’s nonadditivity interaction, 484–486
in genomewide association analysis, 486–489
indicator variables, and scaled variable, 480
overview, 479–482
statistical tests, 482–484
Single-instruction multiple-data (SIMD), 977
Single-label, sentence-level classification, 947–948
bag-of-words model, 948
event trigger words boosted, 948
genome names and event trigger words boosted, 948
genome names boosted, 948
naive Bayes classifier, 947
equation, 947
training, 947
term frequency inverse document frequency (TF-IDF) representation, 947
WEKA, machine learning algorithms, 947
library token delimiters, 947
Single-label sentence-level classification, 960, 961
evaluation, 960, 961
GENIA development data, 960
naive Bayes multinomial classifier, 960
single-label, sentence-level results, 961
WEKA library, 960
machine learning algorithm, 960
Single-nucleotide polymorphisms (SNPs), 182, 226, 333, 353, 478, 982, 1097
Single-participant event, 958
Singular value decomposition (SVD), 370, 391–392, 537, 648
Smith–Waterman alignment, 229
SNP–SNP interactions, 354
Soft computing methodologies, 421
Software development, in bioinformatics, 129
Spearman distance, 528
Spearman’s rho function, 346, 596
Species richness estimation, 1004
  categories, 1004, 1005
  performance evaluations, 1005
SpliceCenter, 13
Spliceosome, 12
  assembly, 8
Splicing, 5
  alternative, 9
    correlation of input transcript numbers
    and, 27
    databases, 12–14
    genomic and transcriptomic sequence
    analyses, 12
    mechanism, 10
    overview, 9–10
    regulation, 10–12
    types, 9, 10
    conservation, 12
    databases, 6
    evolution, 12
    mechanism, 6–7
    regulation, 7–9
Splicing databases (SDBs), 13–22
  cluster analysis, 27
  construction for eukaryotic organisms,
  16–17
  database statistics for, 17–22
  database terminology, 15
  data schema, 16
  data tables, 15–16
  dbASQ computational pipeline, 14–15
  exon analysis, 28
    cassette exons, 28
  web access, 17
  web interface
    CeleganSDB5, 26–27
    DmelSDB5, 24
    for human SDB3, 18
    for MouSDB5, 20–21
    for RatSDB2, 22
Splicing graph, 13
Splicing regulatory proteins, 8, 11
Stability, 334
  of feature selection algorithms, 339
  metrics for stability assessment, 345–346
    evaluation, remarks on, 349
    rank-based metrics, 346–347
    set-based metrics, 347–348
    threshold in stability metrics, 348–349
State-of-the-art, biological relation mining,
  922–924
  biological entities association, 922
Biological Information Extraction and Query
  Answering (BIEQA) System, 923
  fuzzy biological relation, 923
  complete syntactic analysis, 923
  co-occurrence-based approach, 922
    disadvantages, 922
    lexicon-syntactic patterns, 922
    MEDLINE abstracts search, 922
    MEDLINE documents, 922
    tri-occurrence extraction method, 922
  dictionary look-up approach, 923
  protein identification, text documents, 923
GENIES, natural language processing
  system, 923
linguistic-based approach, 922
  vs. co-occurrence-based approach, 922
mixed-mode approach, 923
  protein–protein interaction (PPI), 923
  support vector machines (SVMs), 923
PASBio system, 923
  predicate argument structure (PAS), 923
PASTA system, 922
  parts-of-speech (POS) tagging, 922
  query processing module, 924
  relation set, 924
RelEx, 923
  extract relations, between proteins and
  genes, 923
Statistical analysis, 160, 355, 619, 814, 893, 929
Statistical epistasis networks (SEN), 182
  implications, 184–186
  network construction and analysis, 182–183
  network $G_{0.013}$, 185–185
  observations, 183–184
Statistical hypothesis testing, 354
  alternative hypothesis, 354, 355
  false discovery rate (FDR) control, 356–357
    conditional, 357–358
    2D-JFDR, 359
    local, 358–359
    positive, 357
  familywise error rate (FWER) control, 355
    procedures, 356
  null hypothesis, 354, 355
Statistical significance, in integrative analysis,
  371–372
  FDR-controlling meta-analysis, 373
  $p$-value-controlling meta-analysis, 372–373
  inv chi-square analysis, 373
  max analysis, 373
  min analysis, 372–373
Stochastic EM (SEM) algorithm, 561
Stoichiometrically balanced cycles (SBCs), 1035
STOP codon, 44, 99
Stopwords, biomedical knowledge, 844
corpus sizes, 844
deletion of words, biomedical techniques, 844
normalization processes, 844
String, 84
graph, 234, 239
searching, 845, 847, 858
transformed, 232
String matching
biomedical names, 846, 847
approximate matching, 847
exact matching, 847
String searching, matching process, 847
Structural data mining, 752–753
protein, 752
secondary-structure prediction, 752
structural bioinformatics, 752
Structure data visualization, 1097
binding analysis, 1099–1101
DES ligand, 1100
graphic tools, 1099
S2S visualization, 1100
STITCH 2, 1101
perspectives, 1101–1102
structure inspection, 1097–1099
Suffix tree, 229, 231, 232, 238, 973, 977, 986, 989
abilities, 232
arrays, 238
SUISEKI system, 944
Support vector machine (SVM), 192, 686, 688
data-mining system, 676
Support vector machine methods, 830
interest in bioinformatics community, 830
kernel functions, 830
optimal hyperplane, 830
RFE capable of determining, 830
Support vector machines (SVMs), 422, 663
Survey of tools, 871–880
life sciences researchers, 871
modern software technologies, 871
tools associated with user activities, 872
SVD impute method, 648
SVM-RFE algorithm, 339
SWISS-PROT, 35
Synteny detection algorithms, 1083
Systematic search algorithms, 599
TAMBIS project, 53
Tandem MS-based technologies (MS/MS), 59
TaO ontology, 53
Target-Decoy approach for improving protein identification, 66–67
Taxonomy-independent approach, 1003
Taxonomy-independent microbial community analysis, 1014
Techniques of inference, of GRNs, 815–830
APRIORI algorithm, 816
correlation relationships, frequent-itemset, 815
downward-closure property, 816
FP-GROWTH, 816
FPM methods, 817
frequent itemset, 815
frequent-itemset-based methods, 815–817
GENMAX, 817
MAXMINER, 817
transaction identification (TID), 815
Term Finder, GO website, 604
Tertiary structure (TS) prediction, 717
Test data distribution, 960
event-type, 960
Testing using permutation, 366
binding site functional analysis, bias in, 366, 368–369
bias for varying window sizes on, 369
bias in GO enrichment analysis, 368
bias in RNA-seq functional analysis, 369
permutation in differential expression studies, 369–372
distribution standardization, 366, 368
GSA, 366
GSEA, 366
Text mining, 755–756
with distributional analysis, 914–915
distributional analysis, 915
hyperspace analogue to language (HAL), 915
latent semantic indexing (LSI), 915
medical subject headings (MeSHs), 915
RE tasks, 914–915
Z-score, 915
genes, 755
proteins, 755
systems, 902
techniques, 921
knowledge distillation, 921
Text mining (Continued)

text-refining, 921

vs. existing text document processing techniques, 921
textome, 755

bioinformatics text mining (BTM), 756
hypothesis generation, 756
named entity recognition (NER), 756
PubMed, 755
relationship extraction, 756
synonym and abbreviation extraction, 756
text classification, 756

The Origin of Species, 1049

Thymine (T), 95

Time-delayed ARs with decision trees
classification tree for gene CDC20, 821
single time-delayed interactions, 821
time-delayed gene expression profile (TdE), 822

Time-delayed ARs with frequent-itemset
mining, 817, 818
discrete matrix, 817
temporal association rule mining (TARM), 818
time-delayed gene expression (TdE), 817
time-delayed matrix, 818

Time-delayed Bayesian network, 824–826
equation network transformation, 825
time-delayed relationships, 824
variables and edges, time-delayed network, 825

Time-delayed GRNs Boolean networks, 828–830
delays, transcription and regulation, 828
functions and presentation, temporal boolean
network, 828
temporal Boolean networks (TBoN), 828
Tokenization of biomedical knowledge, 843
divide natural language texts, 843
process of breaking text, 843
tokens, 843

Transcript-initial cassette (TIC) exons, 24, 25

Transcription, 5, 379

Transcription factors (TFs), 9, 23, 166, 354,
365, 368, 638, 786, 787, 803, 984,
1028
Transcription regulatory networks, 603
visualization, 603
Transcriptomes, 12, 16, 19, 21, 25, 1051
Transcriptomic, 787, 1032
data, 787
sequence analyses, 12

Transcript-terminal cassette (TTC) exons, 24,
25
Translation, 379
trans regulatory factors, 10

Trie
definition, 231
prefix, 237, 253
suffix, 231, 232
suffix/prefix, 227, 228, 230
Truth tables, 1017

T-test, 365, 661, 667

filter, 345

Tumor progression, 380

Two-dimensional thin-layer chromatography
(2D TLC), 1029

Two-pass extraction, 955, 956
algorithm, 956
vs. one-pass extraction, 955, 956

Types of expression data, 806, 807
equilibrium (steady-state) expression levels,
806
microarrays, 806
noisy microarrays, 807
time series expression levels, 806

UBLAST tool, 1014

UCSC genome, 17

UMBBBD, 1037

UMLS ontology, 48

Unary vs. binary hypervolume indicator,
613

Unified Medical Language System (UMLS),
53, 129

UniGene, 14

corpus, 17

database version numbers, 17

UniProtKB/TremBL, 713

Untranslated regions (UTRs), 6, 203

Use cases, biomedical knowledge,
857–864

biocreative II gene, 857
dictionary-based approach
cases, 858
implementation, 858
dictionary chunker, 860
exact string matching, 858
generate orthographic variants, 858
Java library, 858
species and biomedical terms, 859
steps and resources to create dictionary,
858

unannotated corpus, 857
Validatory entities, 921
van der Waals force, 214
Variable selection criteria, 404
Variable selection exploration strategies, 400–401
  optimal search strategies, 401
  sequential search strategies, 401
  stochastic search strategies, 401
Vectorizing, 1056
Vertebrate genome annotation, 1038
Vertebrates, 11, 261
Visual analytics
  information visualization, 1074
Visualization, 1117–1121
  brushing visualization, 1119–1120
    biological use, 1119
    parts, 1119
  screenshot showing segmented cross section and network, 1120
3D visualization, 1117–1118
  gray-value volume, 1117
  human brain MRT volume, 1117
  network representation, 1117, 1118
  rendering algorithm, 1117
  rendering typical volumetric data sets, 1117
  transparency effect, 1117
Euclidean space, 1117
  measurement, 1117
  function, 1117
image visualization, 1119
  interaction technique, 1119
  segmentation information, 1119
  segments covered, 1119
multimodal biological data, 1117–1121
  network visualization, 1118–1119
  gene ontology, visualize experiment data, 1119
KEGG pathways, 1119
  projection, 1118
  screenshot, 1118
scatterplot visualization, 1121
statistics visualization, 1121
  screenshot, 1121
Viterbi algorithm, 206

Wald test, 483
Wavelet analysis, 114
  cluster analysis of wavelet coefficients
    of complex DNA representation, 121–123
    discrete Haar wavelet transform, 115
    Haar wavelet basis, 114
    Haar wavelet coefficients, and statistical parameters, 115–116
    Hurst exponent, 116–118
    short Haar discrete wavelet transform, 118
    algorithm of, 118–120
    clusters of wavelet coefficients, 120–121
Web access, to SDBs, 17
Web interface, 17
Weighted average discrepant pairs (WADP), 575
Welch $t$-test, 661, 667
White noise, 109, 510
Whole-genome sequencing, 281
Whole Genome Shotgun (WGS), 37
Wilcoxon rank-sum test, 662
Wilks's likelihood-ratio test, 483
Wiring diagram, 827
Withfield data set, 586
Word normalization biomedical knowledge, 844
  lemma, 844
  lemmatization, 844
  morphological variants, 844
  semantic interpretations, 844
  stem, 844
Working independence estimator (WI), 514
XML, 37
xMotif algorithm, 566, 602
X-ray crystallography, 703, 705, 708, 1099
X!Tandem, as search algorithm, 71
YCAY clusters, 11, 12
Yeast cell cycle, 618
  four biclusters in, 621
Yeast data sets
  adjusted Rand index curves, 540
  $F$-index curves for, 544
  FM-index curves, 542
Yeast two-hybrid (Y2H) systems, 77, 1028
  experiments, 805
Zero-letter sequence, 254
ZOOM aligner, 230
Z-score, 311, 327, 328, 810