

## INDEX

---

- Acute lymphoblastic leukemia
  - (ALL) vs. AML, 124f
- Acute myeloid leukemia (AML)
  - vs. ALL, 124f
  - predicting treatment response, 36
- Adaptive quality-based clustering (AQBC), 55, 69–70, 74, 75t
- Additive white Gaussian noise (AWGN), 175
- Affymetrix U94A microarrays, 102f
- Aldesleukin, 269
- Algorithms, 73
- AlignACE, 83–84, 87t
- ALL. *See* Acute lymphoblastic leukemia (ALL)
- All-zero syndrome
  - average syndrome distance, 201f
- Amino acids
  - electron–ion interaction potential, 263t
- AML. *See* Acute myeloid leukemia (AML)
- Analyzing microarrays
  - VxInsight, 117–118
- ANN. *See* Artificial neural network (ANN); Artificial neural networks (ANN)
- AQBC. *See* Adaptive quality-based clustering (AQBC)
- Artificial neural network (ANN), 35, 248f
  - feed-forward, 247
- Artificial neural networks (ANN), 247
- AWGN. *See* Additive white Gaussian noise (AWGN)
  
- BACIIS. *See* Biological and Chemical Information Integration System (BACIIS)
- Basic sequence model, 82f
  
- Baum–Welch algorithm, 243
- Bayesian approach
  - model-based clustering, 68
- Bayesian classification scheme, 244, 245
- Bayesian information criterion (BIC), 68
- BIC. *See* Bayesian information criterion (BIC)
- Binary block code, 186f
- Binary random sequence, 187f
- Binary table-based convolutional code models
  - translation initiation, 202f
- Binding sites, 56, 58
  - common
    - coregulated genes, 76–87
- Binding vectors
  - to code words
    - inverse ECC systems, 199
- BioCyc, 6
- BioKleisli, 210, 222
- Biological and Chemical Information Integration System (BACIIS), 211–212
  - knowledge base, 222
  - architecture, 212–213, 212f
  - cost model, 223
- Biological coding theory applications, 203–204
- Biological data sources
  - functional genomics and proteomics, 5–6
- Biological queries, 16t–17t
- BioProspector, 85, 87t
- BLAST, 236
- Blocks substitution matrix (BLOSUM), 237, 239f
- BLOSUM. *See* Blocks substitution matrix (BLOSUM)

- Bootstrap method, 119f
- Boundedness, 20
- Bursty channel, 175
  
- Cancer simulation, 131–132
- CAST, 64
- CATH. *See* Class Architecture Topology Homology (CATH)
- cDNA. *See* Complementary DNA (cDNA)
- Cell cycle, 259f
  - constant radiosensitivity throughout, 144
- Cell nucleus, 258f
- Channel transition probability, 191t
- CHIME, 242
- Cis-regulatory elements
  - searching for modules, 85–86
- Class Architecture Topology Homology (CATH), 231
- Class predictors
  - constructing, 34–35
- CLUSTALX, 236
- Cluster analysis, 64
- Clustering, 56
  - algorithms, 65t, 70
  - expression profiles, 64–65
  - finding most representative, 111, 112f
  - gene expression profiles, 63–70
  - microarray data, 103–104
  - parameters, 108–109
  - quality-based, 64, 68–69
  - significance, 110–111
  - validation, 70–75
  - VxOrd, 104
- CLUSTLAW, 236
- Coding theory, 203–204
- Combinatorial approach
  - oligonucleotide frequency analysis, 79–80
- Common binding sites
  - coregulated genes, 76–87
- Communication channel, 175
- Communication system, 174f
  - characteristics, 186–187
- Comparative modeling, 228
- Complementary DNA (cDNA), 60
  - schematic overview, 61f
- Complex life science multidatabase
  - queries, 209–223
  - architecture, 212–213
    - multidatabase queries
      - future trends, 222
- Component ontologies
  - functional genomics and proteomics, 3–5
- Compound channel, 175
- Computational analysis of proteins, 227–252
  - classification and prediction, 242–247
  - databases, 229–232
  - future trends, 252
  - modeling, 241–242
  - natural language processing, 248–250
  - sequence alignment, 235–240
  - sequence motifs and domains, 232–235
- Configuration, 4
- CONSENSUS, 81, 87t
- Continuous Wavelet Transform (CWT), 280
  - Hsp70/Hsp90 protein, 283f, 284–285
  - predicting protein functional epitopes, 284
- Convolutional codes, 178
- Coregulated genes
  - common binding sites, 76–87
  - upstream region, 77f
- CPN tools, 5
- CWT. *See* Continuous Wavelet Transform (CWT)
  
- Databases
  - federated
    - cost models for query planning, 222
    - integrate life science Web databases, 214
    - weaknesses, 222
  - integrate life science Web federated database, 214
  - list generation, 216–217
  - MIPS, 74
  - sequence, 230
- Data mart, 219
- Data-warehousing, 209–210, 219
  - response times, 222
- Decoder, 175–176
- Decoding methodology, 178–179
- Definition of secondary structure of
  - proteins (DSSP), 244
  - secondary-structure formations, 229f

- Deoxyribonucleic acid (DNA), 25, 257  
 computing, 203–204  
 double-helix hybridization, 59
- Deterministic motifs, 234
- Dictionary model  
 genomewide motif identification,  
 157–172
- DiscoveryLink, 210, 222
- Discrete memoryless channel  
 (DMC), 191
- DMC. *See* Discrete memoryless channel  
 (DMC)
- DNA. *See* Deoxyribonucleic acid (DNA)
- Domains, 228  
 sequence  
 computational analysis of proteins,  
 232–235  
 structural, 233
- Double-helix hybridization  
 DNA, 59
- DSSP. *See* Definition of secondary  
 structure of proteins (DSSP)
- Dunn's validity index, 71
- Dynamic programming algorithms,  
 236–237
- EBI. *See* European Bioinformatics  
 Institute (EBI)
- Efficiency query execution plan  
 early intersection, 220f
- EM. *See* Expectation–maximization (EM)
- Empirical risk minimization (ERM), 248
- Encoder, 174
- Encoding methodology, 176, 179–180
- Enhancers, 58
- Entrez, 6, 232
- Enumeration approach  
 oligonucleotide frequency analysis,  
 78–79
- Enzyme Nomenclature, 217
- Equivalence classes, 149
- ERM. *See* Empirical risk minimization  
 (ERM)
- Error control  
 codes and genome, 173–206  
 coding methods for genomic sequence  
 and system analysis, 204  
 and communication, 173–202  
 in genomic sequence, 185–186
- Error-correcting codes, 174
- Escherichia coli* K-12  
 initiation regions, 189f  
 noninitiating intergenic regions, 188f
- EST. *See* Expressed sequence  
 tags (EST)
- Eukaryotic genes  
 mutation rate comparison to genome  
 size, 190f  
 structure, 58f
- Eukaryotic replication channels  
 capacity, 192f
- European Bioinformatics Institute  
 (EBI), 230
- E-value. *See* Expected value (E-value)
- Expectation–maximization (EM), 56,  
 82–83, 235
- Expected value (E-value), 237
- Expressed sequence tags (EST), 60
- Extended Gibbs sampling methods,  
 83–84
- FastA, 236
- Fast orthogonal algorithm (FOA), 28
- Fast orthogonal search (FSO), 25
- Feature selection, 118
- Federated databases  
 cost models for query planning, 222  
 integrate life science Web  
 databases, 214  
 weaknesses, 222
- Feed-forward artificial neural  
 networks, 247
- Figure of merit (FOM), 72
- Filtering, 63
- Fingerprint, 233, 235
- First-generation algorithms, 64
- FOA. *See* Fast orthogonal  
 algorithm (FOA)
- Fold recognition, 244
- FOM. *See* Figure of merit (FOM)
- Footprinting  
 phylogenetic, 86
- Fractionation scheme  
 standard, 145f
- Frame shifting, 2
- FSO. *See* Fast orthogonal search (FSO)
- Functional categories  
 enrichment, 73–75

- Functional genomics and
  - proteomics, 1–21
  - biological data sources, 5–6
  - component ontologies, 3–5
  - methods and tools, 3–5
  - modeling approach and results, 6–18
  - qualitative knowledge models, 1–21
  - querying the model, 16–17
  - representing abnormal functions and processes, 9–10
  - representing high-level clinical phenotypes, 15
  - representing levels of evidence for modeled facts, 16
  - representing molecular complexes, 9
  - representing mutations, 6–7
  - representing nucleic acid structure, 7
  - simulating the model, 19
- Functional inversion, 196–197
- Fuzzy C-means, 64
- GA. *See* Genetic algorithm (GA)
- Gapped motifs, 78
- Garlic, 222
- Gatlin's communication model, 181
- Gaussian infinite mixture model (GIMM), 68
- GBM. *See* Glioblastoma multiforme (GBM)
- GDMB. *See* Genome Database (GDMB)
- GeneCards, 222
- Gene Expression Datamart, 219
- Gene List Exploration Environment (GLEE), 127
- Genes
  - chips, 60
  - coregulated
    - common binding sites, 76–87
    - upstream region, 77f
  - eukaryotic
    - structure, 58f
  - expression profiles
    - clustering, 63–70
    - measuring, 59–61
    - prediction, 35–45
  - lists, 119f, 124f
    - comparing, 123
    - generating, 118
    - predicting medulloblastoma, 44t
    - regulation bioinformatics of
      - microarrays, 55–90
    - shaving, 64
    - structure, 58
- Genetic algorithm (GA), 200, 245–247, 246f
  - operator, 247f
- Genetic error control system, 184–202
- GENIES, 251
- Genome Database (GDMB), 217
- Genomewide motifs
  - identification, 157–158
    - dictionary model, 157–172
    - previous methods, 159–167
- Genomics. *See* Functional genomics and proteomics
- Genomics Unified Schema (GUS), 210, 219, 222
- Geometric cell, 149
- G-functionals, 27
- Gibbs sampling, 56, 87t, 235
  - algorithms, 82–83
  - extended, 83–84
- GIMM. *See* Gaussian infinite mixture model (GIMM)
- GLEE. *See* Gene List Exploration Environment (GLEE)
- Glioblastoma multiforme (GBM), 260
  - fractionation schemes, 142–144
- Gram–Schmidt process
  - Volterra series, 26
- Graph coarsening
  - adding to VxOrd, 112–117
- Guilt-by-association*, 55
- GUS. *See* Genomics Unified Schema (GUS)
- Halting, 2
- Heat shock protein (HSP), 60, 260, 283
  - tumour suppressor protein interactions, 280–281
- Heat shock proteins 70/90 (HSP70/90)
  - CWT, 283f, 284–285
- Hidden Markov models (HMM), 234, 242, 243f
  - classifier, 245f
- Hierarchical clustering, 64–65, 66f
- High cell loss factor, 144

- High-level clinical phenotypes  
 functional genomics and proteomics  
 representing, 15
- HMM. *See* Hidden Markov models (HMM)
- Homology identification  
 natural language processing (NLP), 250
- Homology modeling, 228
- HSP. *See* Heat shock protein (HSP)
- IBM SPLASH, 40
- IL-2, 269, 272
- INCLUSive Web tool, 56, 84, 87–88
- Incyte, 219
- Information integration layer  
 architecture, 213f
- Initiation complex, 2
- In silico radiation oncology, 131–150
- Integral object composition, 4
- Integrate life science Web databases  
 federated database, 214
- Inverse ECC models, 193–196
- Inverse ECC systems  
 from binding vectors to code  
 words, 199
- Inverse EC model II, 196–197
- Inverse error control coding models, 193
- Japanese International Protein  
 Information Database (JIPID), 230
- JC virus (JCV), 260, 265, 268  
 medulloblastomas, 268
- JIPID. *See* Japanese International Protein  
 Information Database (JIPID)
- K-means, 64, 75t  
 clustering, 66
- K-nearest neighbors (k-NN), 40, 47t
- Laguerre functions, 28
- Learning  
 supervised, 103  
 unsupervised, 103
- Lee–Schetzen method, 27
- Levels of evidence for modeled facts  
 functional genomics and proteomics  
 representing, 16
- Likelihood evaluation  
 algorithms, 164–167
- Linear block codes, 176
- Linear-quadratic (LQ) model, 140–141
- LinkDB, 222
- Link-driven systems, 222
- Liveness, 20
- LNL cascade, 30
- LQ. *See* Linear-quadratic (LQ) model
- May et al.'s communication model,  
 182, 183f
- MCLUST software, 68
- Mean-square error (MSE), 33
- Mechanical wounding, 88t  
 microarray experiment, 88t
- MEDLINE, 232
- Medulloblastomas, 260  
 JCV, 268  
 predicting clinical outcome, 40–41  
 predicting metastasis, 43–46
- Melatonin, 269
- Member–bunch composition, 4
- MEME. *See* Multiple Em for Motif  
 Elicitation (MEME)
- Memoryless channel, 175
- Messenger ribonucleic acid (mRNA), 2  
 leader  
 functional definition, 196–197  
 structure, 58f
- Microarrays  
 Affymetrix U94A, 102f  
 analyzing  
 VxInsight, 117–118  
 basic analysis, 103  
 clustering, 103–104  
 experiment  
 mechanical wounding, 88t  
 experiments, 100–101  
 gene regulation bioinformatics of,  
 55–90  
 nonlinear system identification, 25–48  
 constructing class predictors, 34–35  
 gene expression profiling prediction,  
 35–46  
 parallel cascade identification, 30–33  
 predictors comparison, 46–47  
 online integrated analysis, 87–88  
 preprocessing, 61–63  
 robust analysis, 99–128
- Microbial genome mutation rate  
 comparison to genome size, 190f

- Minorization–maximization algorithm  
   parameter estimation, 167–171
- MIPS. *See* Munich Information Center  
   for Protein Sequences (MIPS)
- Misreading, 2
- Missing-value replacement, 62–63
- MITOMAP, 6
- Model-based clustering, 64, 68  
   Bayesian approach, 68
- Modeling  
   functional genomics and proteomics,  
     6–18
- Molecular biology, 203
- Molecular complexes  
   functional genomics and proteomics  
     representing, 9
- Motifs, 56, 233  
   deterministic, 234  
   discovery, 235  
   finding, 56, 57f  
     algorithms, 82–83, 87t  
     recent advances, 85–86  
     software, 87  
   gapped, 78  
   genomewide  
     identification, 157–158, 157–172,  
       159–167  
   identification, 234–235  
   probabilistic, 234  
   probabilistic model of sequence,  
     80–81  
   sequence  
     computational analysis of proteins,  
       232–235  
     structural, 234
- MotifSampler, 56, 84, 87t
- mRNA. *See* Messenger ribonucleic acid  
   (mRNA)
- MSE. *See* Mean-square error (MSE)
- Multidatabase queries  
   complex life science, 209–223  
     architecture, 212–213  
     multidatabase queries, 222
- Multiple Em for Motif Elicitation  
   (MEME), 81, 84, 87t, 235, 243
- MULTIPROFILER, 79, 87t
- Munich Information Center for  
   Protein Sequences (MIPS),  
   74, 230
- Mutations  
   functional genomics and proteomics  
     representing, 6–7
- National Center for Biotechnology  
   Information (NCBI), 210, 222, 232
- Natural language processing (NLP), 127,  
   248–249  
   computational analysis of proteins,  
     248–250  
   gene function extraction, 252  
   homology identification, 250  
   protein identification, 249–250  
   protein–protein interaction  
     identification, 251f  
   relation/pathway identification, 250  
   synonym identification, 250
- NCBI. *See* National Center for  
   Biotechnology Information (NCBI)
- Neighbor GC tumor (NGCT), 142
- NGCT. *See* Six-neighbor GC tumor  
   (NGCT)
- NLP. *See* Natural language processing  
   (NLP)
- Nonlinear system identification  
   microarray data, 25–48  
     constructing class predictors, 34–35  
     gene expression profiling prediction,  
       35–46  
     parallel cascade identification, 30–33  
     predictors comparison, 46–47
- Nonlinear transformation, 62
- Nonparametric methods, 26
- Normalization, 62, 101–102
- Nucleic acid structure  
   functional genomics and proteomics  
     representing, 7
- Nucleotides  
   electron–ion interaction potential, 263t
- Oligonucleotides, 56  
   frequency analysis  
     combinatorial approach, 79–80  
     enumeration approach, 78–79
- OMIM. *See* Online Mendelian  
   Inheritance in Man (OMIM)
- Oncogene, 269
- Online integrated analysis  
   microarray data, 87–88

- Online Mendelian Inheritance in Man (OMIM), 6, 105, 211, 217
- Ontologies, 1  
 component  
 functional genomics and proteomics, 3–5
- Open reading frames (ORF), 228
- ORF. *See* Open reading frames (ORF)
- P53, 144–148
- Pairwise alignment, 232
- PAM. *See* Point accepted mutation (PAM)
- PAM250 substitution matrix, 240f
- Parallel cascade identification (PCI), 30–33, 31f, 47t
- Parallel cascade ranking  
 test expression profiles, 39t
- Participants/role model, 3
- Path generation, 217–218
- Pattern, 233
- PCI. *See* Parallel cascade identification (PCI)
- PDB. *See* Protein Data Bank (PDB)
- Peak frequency  
 protein groups, 268t
- Penalties, 237
- Petri Net, 3, 4, 18f  
 translation into, 5
- Philosophiae Tumorialis Principia Algorithmica*, 131–132
- Phylogenetic footprinting, 86
- PIR. *See* Protein Information Resource (PIR)
- Place–area composition, 4
- PlantCARE, 89
- Point accepted mutation (PAM), 237
- Position weight matrix (PWM), 234
- Prediction-based method, 245
- Predictors  
 data set, 46–48
- Preprocessing, 101–102
- Probabilistic methods, 80–81
- Probabilistic model of sequence motifs, 80–81
- Probabilistic motifs, 234
- Probabilistic sequence models, 56
- Probabilistic suffix trees, 235
- Profiles, 233
- Prokaryotic replication channels  
 capacity, 192f
- Promoter regions, 56, 58
- PROSITE, 230–231, 234
- Protégé-2000 knowledge-modeling tool, 4
- Protein Data Bank (PDB), 210, 231
- Protein Information Resource (PIR), 230
- Proteins. *See also* Computational analysis of proteins  
 classification, 235  
 configuration, 227  
 folding, 228  
 homology, 228  
 identification  
 NLP, 249–250  
 multiple cross-spectral functions of, 266f–267f, 269f–271f, 273f–274f, 276f–279f, 281f–282f  
 peak frequency and SNR, 268t, 272t, 274t, 283t
- Protein Sequence Database (PSD), 230
- Proteomics. *See* Functional genomics and proteomics
- PSD. *See* Protein Sequence Database (PSD)
- P53 tumor suppressor gene, 265
- PWM. *See* Position weight matrix (PWM)
- Qualitative knowledge models  
 functional genomics and proteomics, 1–21
- Quality-based clustering, 64, 68–69
- Query, 15f  
 decomposition, 215–216  
 decomposition tree, 215f  
 execution plans, 214–215  
 functional genomics and proteomics, 16–17
- Radiation therapy, 131–150  
 in vivo, 135–136
- RAMSOL, 241f, 242
- Rand index, 71–72
- Realistic sequence models, 77–78
- REDUCE, 87t
- Relation/pathway identification  
 NLP, 250
- Replication, 257
- Repressors, 58

- Rescaling, 63
- Resonant recognition model (RRM), 260–265, 284  
frequencies, 264t
- Response times  
data warehouses, 222
- Reverse engineering, 184–202
- Ribosomal interaction  
functional definition, 196–197
- Ribosome as block decoder, 193–196
- Rose scale, 34
- RRM. *See* Resonant recognition model (RRM)
- RSA tools, 87t
- SAM. *See* Sequence Alignment and Modeling (SAM)
- SARAH. *See* Simultaneously axially and radially aligned hydrophobicities (SARAH) scales
- SCOP. *See* Structural Classification of Proteins (SCOP)
- Second-generation algorithms, 67
- Self-organizing maps (SOM), 64, 67
- Self-organizing tree algorithm (SOTA), 64, 67
- Sensitivity analysis, 72–73
- SeqStore, 219
- Sequence Alignment and Modeling (SAM), 235, 243
- Sequence-based methods, 244
- Sequence databases, 230
- Sequence motifs and domains  
computational analysis of proteins, 232–235
- Sequence Retrieval System (SRS), 210, 222, 232
- Shannon's channel coding theorem, 187
- Signal-to-noise ratio (SNR)  
protein groups, 268t
- Signature, 233
- Silhouette coefficients, 71
- Similarity measure  
choosing, 104  
postprocessing, 105
- Simulated annealing, 64
- Simulating cancer on computer  
algorithmic principles, 131–135  
literature review, 133–134
- Simulation model  
functional genomics and proteomics, 19  
parametric testing, 142–144
- Simulation outline, 140–142
- Simultaneously axially and radially aligned hydrophobicities (SARAH) scales, 34
- Single-nucleotide background model based on base frequency (SNF), 84
- Six-neighbor GC tumor (NGCT), 142
- SNF. *See* Single-nucleotide background model based on base frequency (SNF)
- SNR. *See* Signal-to-noise ratio (SNR)
- Solid tumor in vivo  
biology, 137–138  
radiobiology, 140–150
- SOM. *See* Self-organizing maps (SOM)
- SOTA. *See* Self-organizing tree algorithm (SOTA)
- Space-filling models, 241, 241f
- SPARC/osteonectin, 43
- SP-STAR, 79
- SRM. *See* Structural risk minimization (SRM)
- SRS. *See* Sequence Retrieval System (SRS)
- Standard fractionation scheme, 145f
- Standardization, 63
- STRIPS-like planning, 219
- Structural Classification of Proteins (SCOP), 231–232
- Structural domains, 233
- Structural motifs, 234
- Structural profile, 234
- Structural risk minimization (SRM), 248
- Structure-based methods, 244
- Substitution matrix, 237
- Suffix trees, 80
- Sum-of-square criterion of K-means, 71
- Supervised learning, 103
- Supervised methods, 117–118
- Support vector machine (SVM), 40, 248
- Support vector machine recursive feature elimination (SVM RFE), 124
- SVM. *See* Support vector machine (SVM)
- SVM RFE. *See* Support vector machine recursive feature elimination (SVM RFE)
- Swiss-PdbViewer, 242

- Swiss-Prot, 230, 231
- Symmetric channel, 175
- Synonym identification  
   natural language processing (NLP), 250
- TAMBIS. *See* Transparent Access to Multiple Biological Information Sources (TAMBIS)
- T-antigen, 260
- Ternary complex, 2
- Tertiary-structure components, 7f
- Test expression profiles  
   parallel cascade ranking, 39t
- Testing cluster coherence, 70–71
- TFBS. *See* Transcription factor binding sites (TFBS)
- Threading, 228
- Threshold sampler, 85
- Tokens, 4
- Transcription, 58, 257
- Transcriptional regulation, 57–58
- Transcription factor binding sites (TFBS), 59, 78f
- Transcription factors, 56, 58
- Transcription processes  
   initiation, 59f
- Transfer ribonucleic acid (tRNA), 2, 7f
- Translation initiation  
   binary table-based convolutional code models, 202f
- Translation processes, 257  
   process diagram, 11f, 12f, 13f
- Transparent Access to Multiple Biological Information Sources (TAMBIS), 4, 7, 9, 210, 211, 222
- tRNA, transfer ribonucleic acid (tRNA)
- Tumor cells  
   cytokinetic model, 138f  
   three-dimensional visualization, 143f
- Tumor growth  
   data collection and preprocessing, 135  
   data visualization, 135–136  
   radiation therapy in vivo, 135–136
- Tumor suppressor gene  
   p53, 265
- Tumor suppressor proteins  
   interactions, 269  
   viral interactions, 265
- Tumor/tumor suppressor protein  
   interactions computation analysis, 257–285  
   methodology, 260–265  
   results, 265–270
- UEP. *See* Unequal error protection (UEP)
- UMLS. *See* Unified Medical Language System (UMLS)
- Unequal error protection (UEP), 200
- Unified Medical Language System (UMLS), 4
- Unified model, 160–164
- UniProt/Swiss-Prot Protein  
   knowledgebase, 6
- Unsupervised learning, 103
- Untranslated regions (UTR), 58
- Upstream region  
   coregulated genes, 77f
- UTR. *See* Untranslated regions (UTR)
- Variable selection, 118
- Viral suppressor proteins  
   interactions, 269
- Volterra series  
   Gram–Schmidt process, 26
- VxInsight  
   analyzing microarray data, 117–118
- VxOrd  
   adding graph coarsening to, 112–117  
   clustering, 104
- Weeder, 87t
- Well clustered, 71
- WINNOWER, 79
- Wire-frame models, 241
- Woflan Petri Net verification tool, 5
- Word-counting methods, 56
- Workflow model, 3
- YMF, 87t
- Yockey's communication model,  
   181, 182f

