

# Index

NOTE: Page numbers followed by f refer to figures, page numbers followed by t refer to tables.

- Ab initio* gene prediction, 117, 118–119, 127, 133  
*Ab initio* structure prediction, 241  
Absolute frequency, 118f  
Abstractions, 258  
Abstracts, scanning, 61–62  
Abstract view, 61  
Acceptance, defined, 301  
Accession number, 6, 14  
    in nucleotide sequence flatfiles, 8–9  
ACeDB database, 279  
Ace file, 354  
AceView, 42, 43f  
*ACHE* gene, 83, 88, 95. *See also* GeneView  
    genomic context of, 96f  
    Model Maker for, 97f  
    querying for, 85  
    symbol for, 99  
AC lines  
    in nucleotide sequence flatfiles, 8–9  
    in TrEMBL records, 19  
ActiveModules plug-in, 282, 286  
ADIT (AutoDep Input Tool), 226  
Adobe SVG Viewer, 155  
Affine gap penalty, 304  
Affymetrix GeneChip, 425. *See also* GeneChip arrays  
AH lines, in nucleotide sequence flatfiles, 15  
ALFRED (ALLele FREquency Database), 181–183  
Algorithm complexity, 147  
Algorithms. *See also* Computational entries; Computer programs; Programs  
    constrained by initial alignment, 159  
    not constrained by initial alignment, 151–161  
    pattern-driven, 129  
    sequence-driven, 129–131  
Algorithm scaling, 161  
Alifold program, 159  
    server output from, 160f  
AlignAce technique, 129  
Alignment  
    editing, 374, 375f  
    in homology modeling, 238  
    of phylogenetic sequence data, 373–374  
    quality of, 328  
Alignment algorithms, 394–396  
Alignment editing, 373  
Alignment programs, 327  
Align sequences, multiplying, 327–331  
Allele frequencies, submitted to dbSNP, 179  
Alleles, 172  
    as dbSNP submissions, 177–178  
    frequencies of, 173  
AllFuse database, 275  
Alphabet blocks, hyperlinked, 71  
ALSCRIPT software, 331, 332–333  
Alternative splicing, Model Maker and, 96–97  
Alt-Splicing track, 87  
Amino acids, secondary structures and, 199  
Amino acid sequences, 198  
AMPS software, 330  
Analysis of Multiple Aligned Sequences (AMAS), 333  
    output from, 334–335f  
    subfamily analysis with, 338  
ANAREA algorithm, 242  
“Annotation document,” 212  
Annotations, 49  
    building on previous, 133–134  
    correctness of, 109  
    Ensembl, 103  
    highlighting and retrieving, 87  
Annotation tracks, 83, 84–87  
Annotation transfer, 211  
ANOVA (analysis of variance)  
    techniques, 413, 426, 430, 433, 437f, 439f  
Apollo system, 135  
Applied Biosystems marker sets, 33  
Apweiler, Rolf, 3  
*Arabidopsis* genome, 42  
Arabidopsis Information Resource, The (TAIR), 266  
AraCyc database, 266  
Archival databases, 5  
Arithmetic operations, 478  
ARKdb (ArkDB) database, 31, 47  
Array CGH, 29–30  
ArrayExpress database, 414  
Array probes, annotating, 410  
Arrays, in Perl, 491–492  
Artificial neural networks, 131

*Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, Third Edition*, edited by Andreas D. Baxevanis and B.F. Francis Ouellette.  
ISBN 0-471-47878-4 Copyright © 2005 John Wiley & Sons, Inc.

- AS lines, in nucleotide sequence flatfiles, 15
- ASN.1 format, 236–237
- Assemblers, 345–346
  - function of, 342
- Assemblies
  - changes in, 108
  - with the Phred/Phrap/Consed program suite, 346
  - sorting, 357
- Assembly header (AH) lines, in nucleotide sequence flatfiles, 15
- Assembly (AS) lines, in nucleotide sequence flatfiles, 15
- Assembly View function, 357
- Assignment shortcut operators, 490t
- Atlas assembler, 353
- Atlas of Protein Sequence and Structure* (Dayhoff *et al.*), 4
- Author lists, hyperlinked, 59–61
- AUTHORS lines, in nucleotide sequence flatfiles, 10–11
- AutoFinish mode, 357
- Automated DNA sequencing, 343–344
- Average linkage clustering, 428
- AVID algorithm, 396, 399
  
- BAC assembly, 353
- BAC clones, 358–359
  - sequences for, 359f
- BAC DNA, 343, 353
- BAC fingerprinting data, 49
- BAC images, 73
- BAC libraries, 34
- Bacterial artificial chromosomes (BACs), 27–28. *See also* BAC entries; Human BAC Resource
- BAC tiling map, 82
- Bader, Gary D., 253
- Band assignments, 31
- Barton, Geoffrey J., 325
- Basepair View, 103, 105
- Basic Local Alignment Search Tool (BLAST), 27, 34, 35, 56, 240, 304–311. *See also* BLAST entries versus FASTA, 319–320
- Batch folding, 155
- Batch jobs, 148–149
- Baxevanis, Andreas D., 55, 295, 393
- Bayesian networks, 275
- Baylor Human Genome Sequencing Center, 82
- Bidirectional best hit, 273
- Binding sites, annotated, 129
- BIND molecular interaction record, 262f
- BioCarta Web site, 270, 279
- BioCyc databases, 264, 266
- BioCyc project, 276–279
- BioGraph program, 456
- Bioinformatics, 77, 224, 254
  - research in, 366
- BioLayout visualization tool, 281, 284
- Biological analysis, 476. *See also* Perl
- Biological context approaches, 274
- Biological databases, information retrieval from, 55–79
- Biological information, ix
  - hierarchical flow of, 198
- Biological network information, 258–259
- Biological pathways, prediction algorithms for, 271–275
- Biological replication, 412
- Biological significance, in microarray assays, 432
- Biology, advances in, ix–x
- Biomolecular interaction data, visualizing and analyzing, 282–284
- Biomolecular Interaction Network Database (BIND), 260–261, 268. *See also* BIND molecular interaction record
- Biomolecular interaction networks, 280
  - b-ions, 450
  - Blanco, Enrique, 115
  - BLAST2Sequences software, 311–312
  - BLAST2 software, 388
  - BLAST algorithms, 304–306
  - BLAST analysis, 389
  - BLAST cutoffs, 309–310
  - BLAST home page, 306f
  - BLAST-Like Alignment Tool (BLAT), 34, 83, 109, 314–315. *See also* BLAT query
  - BLASTN algorithm, 312, 395, 402
  - BLAST output, 308–309
  - BLAST page, 88
  - BLASTP “hit list,” 309f, 310
  - BLASTP hyperlink, 307
  - BLASTP query page, 307f
  - BLASTP search, 308–309, 312, 313–314
  - BLAST search (BlastSearch), 49, 50, 99, 305f, 306–308
    - artifacts in, 310–311
    - extension of, 305f
  - BLASTX searches, 119
  - BLASTZ method, 394–395, 396
  - BLAT query, 315f, 316f. *See also* BLAST-Like Alignment Tool (BLAT)
  - BLOCKS database, 213–214, 302
  - BLOSUM62 matrices, 302, 304
  - BLOSUM62 scoring matrix, 299–301
  - BLOSUM matrices, 302, 303, 377–378
  - “Blurrograms,” 226
  - Bonferroni correction, 425
  - Books link, 62
  - Boolean operators, 59, 61f
  - Boolean search statements, Entrez, 60f
- Bootstrap analysis, work flow for, 386f
- Bootstrapping, 383, 428
- Bottom-up proteomics, 450
- Bouffard, Gerard G., 341
- BRAF* oncogene, 174
- Branch-swapping algorithms, 382
- BRCA1* gene, 184, 187f, 188f
- Breadth-first search, 281
- Brinkman, Fiona S. L., 365
- British Columbia Genome Sequence Centre (BCGSC), 34
- Buried residues, prediction of, 333–336
  
- Caenorhabditis elegans* genome, 42
- Cancer Chromosome Aberration Project, 31
- CAP assembler, 349
- Capillary isoelectric focusing (capillary IEF), 467
- cappuccino* gene, 74f
- Captured gaps, 353
- CATH database, 245
  - example description for, 246f
- CC lines, in nucleotide sequence flatfiles, 11
- cDNAs, 135
- Celera Genomics, 352
- Cell Markup Language (CellML), 271
- Cell signaling network mapping, 254
- Cellular metabolism mapping, 254
- Cellular model, 254
- Center for Inherited Disease Research (CIDR) genotyping service, 33
- Center for Medical Genetics, 32
- Centre d’Etude du Polymorphisme Humain (CEPH), 32
- CE server, 246–247
- Character data, randomized, 382–383
- Characterization, 468
  - defined, 463
- Character states, 373
  - weight matrices for, 376
- Charge state deconvolution, 451
- Chime (CHemical mIME) plug-in, 236
- Chimeric read, 348
- Chomp function, 480
- Chromatograms, 343
- Chromatography, 255
- Chromosomal features map, 76f
- Chromosomal rearrangement, 172
- Chromosomal region, telescoping graphical view of, 99
- Chromosomal segments, shared, 42–43
- Chromosome abnormalities, 30
- CIBEX database, 414
- CIF format, 230–231
- Cis-control elements, ix
- Citations, in nucleotide sequence flatfiles, 10–11
- Cited in PMC link, 62

- Cladistics, 367. *See also* Phylogenetic analysis
- Classification algorithms, 431  
validation of, 431-432
- ClinicalTrials.gov page, 63-64, 67f
- Clipboard, 70
- Clone-based maps, 34
- Clone fingerprinting, 344f
- Clones, defining map positions from, 49-50
- Clone sequences, integrating in larger assemblies, 358-360
- Clustal guide tree, 372f
- Clustal program, 387  
sequence alignment with, 389
- ClustalW program, 327, 329-330, 336, 387
- ClustalX, 387, 389
- Cluster analysis, 328
- Clustering approaches, 426-429
- Clusters of Orthologous Groups (COGs) resource, 273
- Cn3D macromolecular visualization package, 70, 71f, 236-237
- Coalescent framework, for modeling mutation, 172
- Coaxial stacking, 157
- Coding exons, 133  
signals involved in defining, 117
- Coding sequences (CDS), in nucleotide sequence flatfiles, 12-14
- Coding statistics, 118
- Collision-induced dissociation (CID), 450
- Co-localization plots, 272f
- Color annotation, 165
- Comment keys, in UniProt database, 21t
- Comment lines, in nucleotide sequence flatfiles, 11
- Common integrator approach, 36
- Comparative analyses, applying gene predictions to, 400-402
- Comparative gene prediction, 117, 120
- Comparative gene resources, 42
- Comparative genomics  
computational approaches in, 393-408  
evolution and, 394
- Comparative Genomics tracks, 109
- Comparative mapping, 42-47  
resources for, 44, 47
- Comparative Mapping by Annotation and Sequence Similarity (COMPASS), 47
- Comparative promoter prediction, 129
- Comparative proteomics methods, 464-467
- Comparative sequence analysis  
basis of, 159  
RNA structure and, 145
- Complete linkage clustering, 428
- COMPOUND table, 268
- Computational analysis, in comparing 2D gels, 464
- Computational challenges, in biology, ix-x
- Computer programs, for building linkage maps, 31. *See also* Algorithms; Programs; Visualization tools
- Consed assembly view, 351f
- Consed software, 348, 350, 354-358. *See also* Phred/Phrap/Consed program suite  
customizing parameters in, 358
- Consed windows, 355-356f
- CONSENSE program, 385
- Consensus sequence, 351-352
- Conserved Domain Database (CDD), 307
- ConStruct approach, 159
- Constructed (CON) sequences, in nucleotide sequence databases, 9
- Contig consensus sequences, 352
- "Contigged" (CON) sequences, in nucleotide sequence databases, 9
- Contig ordering, 350f
- Contigs, 33. *See also* Sequence contigs  
breaking and joining, 355-357  
large sequence, 35
- ContigView, 99, 103f, 104f
- "Contingency" matrix, 430
- Cooperative Human Linkage Center (CHLC) markers, 32
- Co-purification, 255
- Co-regulated sequences, 129
- Core promoter, 128
- Coriell Cell Repositories, 31
- Correlation coefficient (CC), 126-127
- CpG islands, 83
- CRI-MAP program, 31, 33
- Critical Assessment of Structure Prediction (CASP), 241
- Cross-species sequence comparison, 274
- Cross-strain GL maps, 47
- Cubby storage service, 64  
storage area for, 67f
- Curated databases, 5, 14
- Customized maps, 44
- cVISTA alignment results, 402f
- Cytogenetic and Genome Research*, 47
- Cytogenetic banding, 28
- Cytogenetic band positions, 44
- Cytogenetic localizations, 31
- Cytogenetic mapping, resources for, 30-31
- Cytogenetic maps, 26, 29-30  
Web sites that display, 31
- Cytogenetic positions, 48, 50
- Cytogenetics, technical advances in, 29
- Cytoscape network visualization tool, 271, 281-282, 283f
- DIS243* polymorphic marker, 28
- DALI server, 246-247
- Dangling end, 146
- Database cross-references (db\_xref), in nucleotide sequence flatfiles, 11, 12-14
- Database of Interacting Proteins (DIP), 261, 268  
human protein-protein interactions in, 263f
- Database of Ligand-Receptor Partners (DLRP), 261
- Databases. *See also* Sequence databases  
availability of, 260  
formats for, 6-7  
hard link relationships between, 58  
relationships between entries in, 56-58  
specialized, 72-74  
standards for, 270-271
- Data exchange formats, scope of, 271f
- Data models, 258  
choosing, 375-378  
detail in, 258
- Data repositories, 31
- Datasets, integrating, 275
- Date, in nucleotide sequence flatfiles, 7-8
- Date (DT) lines, in TrEMBL records, 19
- DAVID annotation resource, 410
- dbSNP database, 35, 175-181, 178. *See also* Single nucleotide polymorphisms (SNPs)  
content of, 176  
postsubmission computed data released by, 179  
query modules for, 181  
resource integration by, 179-180  
SNP clustering by, 180  
submissions to, 176-179
- dbSNP summary, 176
- DDBJ/EMBL/GenBank records, 5-7  
formats of, 7-16
- DDJB flatfile format, nucleotide record in, 497-499
- Deep resequencing, 173-174
- DeepView software package, 237, 238, 244
- Definition line (def line), 6  
in nucleotide sequence flatfiles, 8

- Deleted read, 348  
Deletion and insertion polymorphisms (DIPs), 173  
DE lines  
  in nucleotide sequence flatfiles, 8  
  in TrEMBL records, 19-20  
Denaturing high-performance liquid chromatography (DHPLC), 174-175  
Dendrograms, 428  
*De novo* approach, 450  
Dense alignment surface (DAS) method, 210  
Description (DE) lines  
  in nucleotide sequence flatfiles, 8  
  in TrEMBL records, 19-20  
Descriptors, in nucleotide sequence flatfiles, 11  
Detailed View, 99-103, 104f, 105  
  around *ACHE*, 106f  
Developmental Genome Anatomy Project, 31  
Diallelic insert and deletion (indel) markers, 33  
Dideoxynucleotide phosphates (ddNTPs), 343  
Difference gel electrophoresis (DIGE), 467  
Difference gel expression (DIGE), 465f, 466  
Discontinuous MegaBLAST, 312  
Discriminant analysis, 124  
Disease association studies, 172  
Disease Browser, 99  
Disease genes, maps of transcribed sequences and, 34  
Distance-based tree-building methods, 378-380  
Distance metrics, 29, 424  
Distributed Annotation System (DAS) data, 103  
Division code, in nucleotide sequence flatfiles, 7  
DNA  
  as a blueprint, 198  
  BAC, 343, 353  
  coding in nucleotide sequence flatfiles, 7, 8  
  replication errors in, 172  
  sequencing of, 4, 5  
DNA clones, 27-28  
DNA Databank of Japan (DDBJ), 4, 5f.  
  *See also* DDBJ/EMBL/GenBank records  
DNADIST program, 385  
DNA fingerprinting, 49  
DNA markers, 27  
DNA *mfold* server, 148  
DNA microarray experiments  
  data collection and management in, 414  
  design of, 410-414  
DNA microarrays, 410  
  array probe annotation and, 410  
  estimating background for, 416  
DNAML program, 385  
DNA records, coding of, 6  
DNA sequences  
  finishing, 343  
  defining map positions from, 49  
  tools for searching, 35  
DNA sequence tract, 26  
DNA sequencing technology, 342  
DNA-to-protein pathway, molecular processes involved in, 116f  
Domain complexity, 274  
Dotlet software, 297, 298f, 299f, 300f  
Dot plot, 150-151  
Dotplots software, 296-297  
Dotter software, 297  
Dottup software, 297  
Double-hit analysis, 174-175  
DoubleScan method, 401  
Download (dl) link, 95-96  
DRAGON annotation resource, 410  
Dragon Promoter Finder, 131  
DRAWGRAM program, 384  
DRAWTREE program, 384  
DR lines, in nucleotide sequence flatfiles, 11  
*Drosophila* data, 72-73  
*Drosophila melanogaster* genome, 352  
*Drosophila sucinea* R2 element, structure of, 164f  
DSSP program, 242  
DT lines  
  in nucleotide sequence flatfiles, 7-8  
  in TrEMBL records, 19-20  
Dual-genome comparative gene finders, 127  
DUST program, 310  
Dye-reversal replicates, 413  
Dyalign dynamic programming algorithm, 161, 162  
  input form for, 166f  
  secondary structure prediction using, 165-166  
Dynamic programming, 119, 147, 161, 163  
Dynamic programming secondary structure methods  
  comparison of, 157-158  
  for RNA secondary structure prediction, 511-514  
EASE software, 430, 433  
  gene lists with, 285  
EcoCyc database, 264-266, 279  
  chemicals class hierarchy in, 267f  
  pathways and reactions in, 268f  
Edge attributes, 281-282  
eGenome, 27, 28, 31, 35, 36, 38, 39f, 49  
  search interfaces with, 38  
  Web site for, 35  
Electron Microscopy Database (EMDB), 233  
Electrospray ionization (ESI), 448  
EMBL flatfile format, nucleotide record in, 501-503. *See also* European Molecular Biology Laboratory (EMBL)  
EMBL genome reviews record, 504-507  
ENCODE project, 135  
  Web site for, 137  
Energy dot plots, 148, 151  
ENm012 region, 135  
Enright, Anton J., 253  
Ensembl database, output from, 186f  
Ensembl system, 31, 35, 82, 99-108.  
  *See also* GENEWISE program  
  accessing SNPs in, 184  
  gene predictions with, 99  
  home page for, 99, 100f, 101f  
  SNP report from, 185f  
  toolbar with, 103  
EnsMart, 99, 184  
  database query interface with, 185f  
Entrez Gene feature, 62  
Entrez Genomes, 82  
Entrez search space, navigating, 59-64  
Entrez SNP. *See also* Entrez system  
  query environment with, 181  
  query results from, 183f  
  search terms for, 182t  
Entrez system, 56-70  
  discovery pathway in, 59-70  
  relationships in, 58f  
  unified results page with, 59f  
Environment, genomes and, ix  
Environmental Genome Project, 174  
Enzymatic functions, 275-276  
Enzymatic reactions, 255  
Enzyme Commission (EC) numbers, 35  
ENZYME table, 267-268  
e-PCR, 34, 35, 38  
  Web interface for, 49  
Equilibrium distribution of conformations, 146  
Equilibrium stability, 146

- Equivalent residues, 244  
*Escherichia coli*  
  genome of, 461  
  metabolism encyclopedia for, 264  
EST\_GENOME program, 120  
EST searches, 133. *See also* Expressed  
  sequence tags (ESTs)  
EST sequences, 120  
EST tracks, 86–87, 89f  
Euclidean distance metrics, 424  
EUCLID method, 217  
euGenes, 41–42  
Eukaryotes  
  gene expression regulation in, 128  
  gene prediction in, 117, 122  
Eukaryotic genomes, 26  
  cloning and, 27–28  
  first complete sequence of, 82  
Eukaryotic Promoter Database, 129  
European Bioinformatics Institute (EBI),  
  4, 5t, 99, 263  
European Molecular Biology Laboratory  
  (EMBL), 4, 5f. *See also*  
  DDBJ/EMBL/GenBank records;  
  EMBL entries  
EVA server, 205, 207  
  Web site for, 238  
Evidence viewer (ev), 96  
Evolution, comparing genomes to  
  understand, 394  
Evolutionary alignment, 327  
Evolutionary analysis, 366  
Evolutionary distance, 380  
Evolutionary models, 368–369  
Evolutionary theory, in phylogenetic  
  analysis, 366  
ExoFish method, 401–402  
Exon-defining signals, prediction of,  
  117  
Exons, 86, 87  
  assembly of, 118–119  
  display of, 97  
  filtered, 124  
  GENSCAN, 123–124  
  novel, 109  
  predicting, 117, 118, 120  
  scoring, 117, 118, 121, 122  
Expert Protein Analysis System (ExpASY)  
  server, 455  
  Proteomics Tools with, 464  
EXPLORE hyperlink, 228  
ExportView, 99  
Expressed sequence tags (ESTs), 26.  
  *See also* EST entries ; Spliced  
  ESTs track  
  in nucleotide sequence databases,  
  9  
  pseudogenes and, 134  
Expression profile reliability (EPR),  
  261  
Expression vectors, 426  
“Extrinsic” gene prediction, 117  
FASTA algorithms, 308t  
  steps in, 316–317  
FASTA databases, 179  
FASTA files, 6, 190, 344–345  
FASTA format, 183  
  example of, 515  
FASTA program, 315–319  
  versus BLAST, 319–320  
FASTA search, 317–319  
  summary for, 318–319f  
FastDNAmI program, 387  
Features pull-down menu, 105f  
Feature table (FT)  
  in nucleotide sequence flatfiles, 7,  
  11–16  
  in protein sequence flatfiles, 19,  
  20t  
Feature table (FT) lines, in TrEMBL  
  records, 19, 20  
FGENESH program, 124  
  output from, 125f  
FGENES program, 124  
Filtered exons, 124  
FindMod program, 455, 463  
FindPept program, 455–456  
Fingerprinting, 34. *See also* Peptide  
  mass fingerprinting (PMF)  
Finished sequences, versus ordered  
  sequences, 358–359  
Finishing, 343  
  methods of, 353–354  
  standards for, 353  
FirstEF method, 131, 135  
Fisher's exact test, 284  
  using to assess significance, 430  
FITCH program, 384  
Flanking markers, 48  
Flanking nucleotide sequences, as  
  dbSNP submissions, 177, 179f  
Flatfiles, 6–7  
  format of, 7–16  
  in UniProt Knowledgebase, 19–21  
Flicker program, 466  
Flip-dye pairs, 413  
FlyBase, 72–73  
FlyGRID, 261  
FM method, 379, 380  
FOLDALIGN program, 161, 162  
Folding temperature, 155  
Fold recognition, 238–240  
*Fold RNA* button, 149  
FootPrinter algorithm, 402–403  
  analysis results with, 403f  
Footprinting, phylogenetic, 402–403  
Formats  
  database, 6–7  
  of nucleotide sequence flatfiles,  
  7–16  
FOXP2 gene, 135  
FPC software, 34  
Free energy, of conformations, 146  
Free form query structure, 181  
FT lines, in TrEMBL records, 19, 20  
FUGUE threading server, 240  
*Fugu* genome, 133, 137  
Full-shotgun sequences, versus oriented  
  sequences, 358–360  
Functional analysis, dbSNP, 180–181  
Functional classes, 216–217  
  prediction methods for, 217  
Functional divisions, in nucleotide  
  sequence databases, 9  
Function prediction methods, 218  
GA2E system, 134  
Gap penalties, 303–304  
Gaps, 303–304  
GBrowse, 189  
GDB Web site, 35, 42. *See also*  
  Genome Database (GDB)  
  searching, 36–38  
Gel electrophoresis, 446  
Gels, multiple, 464–466  
Gel separation, 255  
GenBank, 4, 5f, 35, 48. *See also*  
  DDBJ/EMBL/GenBank records  
  accession numbers for, 360  
  format for, 95  
GenBank flatfile format, nucleotide  
  record in, 499–501  
Gene attributes, 41  
Gene-based comparative maps, 43–44  
Gene boundaries, 133  
GeneCards, 28, 41  
GeneCensus site, 275  
Gene-centric information retrieval,  
  70–72  
Gene-centric Web resources, 39  
GeneChip arrays, 410. *See also*  
  Affymetrix GeneChip  
Gene expression. *See also* Expression  
  vectors; Significant genes  
  estimates of, 414–416  
  measures of, 416–417  
  microarrays for, 274  
  normalizing measurements of,  
  417–422  
  using DNA microarrays to assay,  
  409–444  
Gene expression data, integrating with  
  pathway information, 284–287

- Gene Expression Omnibus, 414  
Gene finders, published evaluation of, 127  
Gene function transfer, 274  
Gene fusion, 273-274  
Gene identification, biological meaning of, 429  
GeneID program, 119, 120-122, 133  
output from, 123f  
Gene integration resources, 38-41  
GeneLoc, 35, 40f, 41, 49  
GeneLynx, 35, 41  
GeneMap '99, 34  
Gene model, biological evidence supporting, 96  
Gene neighborhood, 272-273  
plots of, 272f  
Gene ontology, 280  
Gene Ontology (GO) project, 41, 217, 410. *See also* GO entries  
Gene prediction  
  *ab initio* methods of, 118-119  
  comparative, 120  
  discriminant analysis in, 124  
  example of, 135-137  
  hidden Markov models in, 119  
  methods of, 117-120  
  interpreting, 133  
  promoter prediction and, 128-129  
  searching by content and, 117  
  sequence similarity-based, 119-120  
  strategies and considerations for, 132-135  
Gene prediction accuracy, 126, 128f  
measures of, 127f  
Gene prediction programs, 117, 120-125, 133  
appropriate parameters for, 134-135  
Gene promoter regions, prediction of, 131-132  
General Feature Format (GFF), 135  
General Repository for Interaction Datasets (GRID) database, 259, 261  
Gene Recognition and Analysis Internet Link (GRAIL) program, 120. *See also* GRAIL entries  
output from, 121f  
Gene regulation networks, 257  
GeneRIF (Gene References into Function), 71  
Genes, ix, x  
  portalized view of, 41  
  predicting on top of previous annotations, 133-134  
Genes\_sequence map, 92, 95, 98f, 99  
Genes\_Sequence track, 109  
GENES table, 268  
Généthon GL maps, 29  
Généthon markers, 32  
Genetic algorithm, for RNA secondary structure prediction, 159  
Genetic linkage (GL) analyses, 28-29  
Genetic linkage mapping, resources for, 32-33  
Genetic linkage maps, 26, 31-32  
Genetic Location Database (LDB2000), 29, 31, 35, 38  
GeneView, 105-108, 109  
for *ACHE*, 107f  
GENEWISE program, 120, 124-125  
Genew project, 19  
GENIE, 119  
GenMAPP software, 284, 285f  
Genome assemblies, complete, 360  
Genome Browser, 35, 49. *See also* UCSC Genome Browser  
uses for, 87  
Genome Browser Gateway, 83, 84  
Genome browsers, 48, 82-83, 108-109  
data downloading from, 109  
SNP integration in, 184-187  
Genome Channel Web, 35  
Genome Database (GDB), 27, 35, 36-38, 49. *See also* GDB Web site  
Genome evolution, understanding, 403  
Genome regions, exploring between two features, 91-92  
Genome Reviews, 15-16  
Genomes, ix, x  
GenomeScan, 119, 123  
Genome Scanner utility, 46-47  
Genome scanning, 32  
Genome-sequencing projects, 82. *See also* Genomic sequence entries;  
Human genome entries  
large-scale, 198  
Genome survey sequences (GSS), in nucleotide sequence databases, 9  
GenomeVISTA feature, 399  
Genome-wide genotyping, 33  
Genome-wide physical mapping, 33  
Genome-wide RH maps, 34  
Genomic alignments  
  generating, 398-399  
  precomputed, 396-398  
Genomic analysis tools, 27  
Genomic annotations, 28, 83  
Genomic-based query pages, 44  
Genomic catalogs, 35  
Genomic coordinates, changes in, 108  
Genomic databases, 81-112  
Genomic data interfaces, 35  
Genomic elements, 48  
  cataloging of, 42  
  tabular layout for, 38  
Genomic information, ix  
  presenting, 29  
  use of, 47  
Genomic mapping, 50  
  bioinformatics challenges of, 26  
  components of, 26  
  data errors in, 28  
  resources for, 35-42  
  summaries for, 41  
Genomic maps, elements of, 27-28  
Genomic markers, naming, 28  
Genomic regions  
  defining, 48  
  determining, ordering, and characterizing contents of, 48-49  
Genomic repositories, 35  
Genomic resources, 38-39  
Genomic segment descriptions, 36  
Genomic sequences, downloading, 90  
Genomic sequence tracts, 26, 50  
Genomic sequencing, goal of, 360  
Genotype frequencies, submitted to dbSNP, 179  
Genotype-phenotype association analyses, 184  
Genotypes, submitted to dbSNP, 179  
Genotyping, 187-188  
  high-throughput technologies for, 32  
GenPept database, 17  
GENSCAN program, 119, 122-124, 133, 401  
output from, 123f  
GFF2PS, 135  
GFS database, 452  
GFS method, 459-463, 469-470  
output with, 462f  
Gibbs free energy of formation, 146  
Giddings, Morgan C., 445  
gi (geneinfo) numbers, in nucleotide sequence flatfiles, 14  
Global alignment method, 395  
Global normalization, 418  
Global sequence alignment methods, 296  
GlycoMod program, 455, 464  
GN lines, in TrEMBL records, 19  
GO annotation, 261. *See also* Gene Ontology (GO) project  
GoMiner software, 285, 430  
GO terms, 41, 42, 285, 286f  
GO vocabulary query selector, 42  
GRAIL2 program, 120  
GRAIL-EXP program, 120  
Graph algorithms, 281

- Graphical tools, 135  
Graph layout algorithms, 281  
Graph theory, 279–281  
GRID database, 282  
Guide tree, 374  
Guigó, Roderic, 115
- Haemophilus influenzae* genome, 352  
Hansen, Nancy F., 341  
Haplotype, 188  
Haplotype block, 188  
Haplotype map, 26  
Haplotype tag SNPs, 189  
HaploView, 191f  
HapMap Consortium, 173  
HapMap Project, 175, 187  
    sample analysis of, 190–192  
    Web site genome browser of, 189  
Hard link concept, 58  
Hashes, 491, 493  
Header, in nucleotide sequence flatfiles, 7–11  
Helical structure, 199  
Hexamer frequencies, 118  
*HHO1* gene, 75f  
Hidden Markov models (HMMs), 119, 202, 312. *See also* Markov entries; Pair HMMs; TMHMM program in protein analysis, 204  
Hierarchical alignment software, 329–330  
Hierarchical clustering, 427–428  
Hierarchical method, 224  
Hierarchical multiple alignment, 328–330  
High-resolution linkage maps, 33  
High-scoring segment pair (HSP), 305–306  
High-throughput genome (HTG) sequences, in nucleotide sequence databases, 9  
History hyperlink, 64–66, 69f  
Hole filling process, 276  
Holmes, Mark R., 445  
HomoloGene database, 47, 96  
    multiple map display and, 97–99  
Homologous chromosome segments, 43  
Homologous sequences, 369  
Homologues, structural, 228  
Homology, defined, 296  
Homology-based gene prediction, 117, 119  
Homology comparison, 124  
Homology modeling, 57–58, 163, 238  
    example of, 239f
- HsUniG map, 92  
HTC sequences, in nucleotide sequence databases, 9  
HUGO Gene Nomenclature Committee, 19  
Human BAC Resource, 30–31  
Human chromosome 2, 189f  
Human chromosome 22, 127, 128f  
Human chromosome maps, 42  
Human cytogenetic mapping information, repositories for, 30–31  
Human genetic variations, 172–173  
Human genome, 82  
    methods used to annotate, 117  
    release v20.34c.1 of, 99  
Human genome assemblies  
    changes in, 108  
    naming conventions for, 91–92  
Human Genome Browser, 109  
Human Genome Organization (HUGO) Genome Nomenclature Committee, 28  
Human Genome Project, ix  
    goal of, 56  
Human Genome Variation database (HGVBBase), 181–183, 184  
Human Genome Variation Society, 184  
Human immunodeficiency virus (HIV), 159  
Human linkage maps  
    global coordination of, 32  
    polynomial transformation approach to, 32  
Human/Mouse/Rat/Chicken track, 87  
Human mucin, 297  
Human obese gene, promoter region of, 131f, 137  
Human Protein Reference Database (HPRD), 261–263, 270  
    protein information records in, 264f  
Human SNPs, in dbSNP, 176f  
Hybridization, 426, 427f  
Hydrophobicity, 208
- Icelandic families linkage map, 32  
Identification (ID) lines, in TrEMBL records, 19  
Identifiers, 6  
ID line, in nucleotide sequence flatfiles, 7  
IHMF structure, 70f, 71f  
ImageMaster program, 466  
Image processing, 414–416  
IMAGE software, 34  
Incorrect results, generating, 366  
Infobiogen, 31
- Information retrieval  
    from biological databases, 55–79  
    gene-centric, 70–72  
    integrated, 56–70  
Informative sites, 373  
Initial alignment  
    algorithms constrained by, 159  
    algorithms not constrained by, 159–161  
Initial exons, 118  
*In silico* mapping, 47  
*In silico* two-hybrid method, 274  
Institute for Genomic Research, The (TIGR), 49. *See also* TIGR assembler  
IntAct database, 263  
    yeast protein-protein interaction records in, 265f  
Integrated mapping resources, 35–36  
Integration tools, 135–138  
Interaction databases, strategies for navigating, 268–270  
Interaction prediction, resources for, 275–279  
Intermolecular interactions, prediction algorithms for, 271–275  
Internal exons, 118, 133  
International Haplotype Map Project, 188–190  
International Nucleotide Sequence Database Collaboration, 4, 5  
Internet-accessible software, 388. *See also* Web entries  
InterPro resource, 212–213  
    summary page for, 215f  
“Intrinsic” gene prediction, 117  
Intronless gene predictions, 134  
Ion suppression, 451  
IPI database, 21–22  
IPU data sets, 21–22  
iSCANPS program, multiple alignment by, 330–331  
Isoelectric focusing (IEF), 467  
Isotope coded affinity tags (ICAT), 450  
IUPAC/UBMB codes, for nucleic acid bases, 516
- Jack-knifing, 428  
JalView alignment, sample, 337f  
JalView software, 330, 336  
    subfamily analysis with, 338  
Japanese Millennium Genome Project, 173  
Japanese SNP (J SNP) database, 183. *See also* Kyoto Encyclopedia of Genes and Genomes (KEGG)  
JNET, 333, 336

- JOURNAL lines, in nucleotide sequence flatfiles, 10-11  
Jpg file format, 150  
JPred program, 205, 206f, 207  
Junk DNA, 404
- Karlin-Altschul equation, 306
- Keys  
comment, 21t  
in nucleotide sequence flatfiles, 8-9
- Keyword (KW) lines, in nucleotide sequence flatfiles, 10
- Kishino-Hasegawa test, 386
- k*-means clustering, 428
- k*-nearest neighbors (kNN) approach, 431
- Known Genes track, 87, 88f, 109
- KW lines, in nucleotide sequence flatfiles, 10
- Kyoto Encyclopedia of Genes and Genomes (KEGG), 46, 266-268.  
*See also* Japanese entries  
ENZYME database record with, 269f  
metabolic pathway reconstruction from, 277f  
Web site for, 279
- Large-insert clone ends, sequencing of, 34
- Large-scale data alignment, algorithms for, 394-396
- Large-scale genotyping, 33
- Large sequence contigs, 35
- Lawrence Livermore National Laboratory Web site, 225
- LDB2000. *See* Genetic Location Database (LDB2000)
- Leave *k* out cross-validation (LKOCV), 431-432
- Leptin structure, 240
- LIGAND database, 267
- Likelihood ratio tests, 383
- Limited Area Global Alignment of Nucleotides (LAGAN) algorithm, 395-396
- Limits hyperlink, 64-66, 68f
- Linear discriminant analysis, 124
- Linear gap penalty, 304
- Linkage disequilibrium (LD), 188  
rates of, 38
- Linkage maps, computer programs for building, 31
- LINKAGE program, 31
- LinkOut link, 63
- Linkout URLs, submitted to dbSNP, 179
- Liquid chromatography, 449
- LiveDIP, 261
- LOC3D database, 216
- Local alignment regions, finding, 297
- Local normalization, 418
- Local sequence alignment methods, 296
- LOChom method, 216
- LOCKey method, 216
- Locus ID, 71
- LocusLink, 70-72, 187  
Variations page with, 188f
- LOCUS name identifier, 6, 7
- Lod (logarithm of the odds) score, 31
- Log<sub>2</sub>(ratio) measurements  
dynamic range of, 422f  
limits of, 420-422
- Log-det transformation, 376-377
- Log-likelihood ratio (LLR) calculations, 349
- Log odds ratio, 301
- "Look-up" gene prediction, 117
- Loop design, 413
- Loops, 485  
combining with input, 486  
initiation energies of, 146
- Low-complexity regions, finding, 297-298
- Lowess (locally weighted linear regression), 418
- Low-intensity data filtering, 420
- Low-resolution molecular mechanics calculations, 163
- Lutefisk program, 450
- MacClade program, 387
- Machine-learning algorithms, 200, 201, 217
- MacPerl, 487. *See also* Perl
- Macromolecular Structure Database (MSD), 233
- Macromolecular visualization programs, 235-237
- Macromolecules, in phylogenetic analysis, 368
- MADAM data entry tool, 415f
- MAFFT package, 330
- MALIGN program, 374
- Mammalian Genotyping Service, 33
- Manhattan distance, 424
- MANIP program, 162
- MA plot, 418
- MAPMAKER program, 31
- MAP-O-MAT Web server, 31, 33
- MAPP files, 284-285
- MAPPFinder software, 285, 430
- Mapping, complexities and pitfalls of, 28-29. *See also* Maps
- Mapping resources. *See also*  
Comparative mapping resources integrated, 35-36  
practical uses of, 47-50
- Map position, defining, 49-50
- MAP program, 31
- Maps  
descriptions of, 36  
eGenome, 38  
multiple display of, 97-99  
non-sequence-based, 109  
types of, 29-35
- Maps & Options box, 92, 95f
- Maps column, 92
- Maps Displayed box, 92
- MapView, 38, 99  
for human chromosome 7, 102f
- Map Viewer, 29, 35, 97-99. *See also*  
NCBI Map Viewer  
human, 82
- Marker-based linkage map server, 33
- Markers  
cytogenetic positions of, 48, 50  
integration of, 48  
nonredundant, 48  
renaming of, 28  
types of, 27
- Marker sequences, BLASTing, 48
- Markov cluster algorithm (MCL), 280
- Markov models, 118. *See also* Hidden Markov models (HMMs)
- "Marshfield identifier" (MFD), 32
- Marshfield screening sets, 33
- Mart-Search, 99
- Mascot Distiller, 455
- Mascot Parser, 455
- Mascot software, 468-469  
match score from, 453-455  
query search results from, 454f  
search engine with, 452-455  
submission screen with, 453f
- Masked sequence data, 133
- Mass spectrometers, 451
- Mass spectrometry (MS), 446. *See also*  
MS entries  
for protein analysis, 448-450
- Master map, 92, 95
- MatchMiner software, 284
- Mathematical challenges, in biology, ix-x
- Mathews, David, 143
- Matise, Tara C., 25
- Matrices, choice of, 302-303
- Matrix-assisted laser desorption ionization (MALDI), 448
- Matrix operations, 280
- MaxHom algorithm, 201, 207

- Maximal dependence decomposition, 122
- Maximum likelihood (ML) method, 31  
analyses using, 380, 381  
tree building with, 376
- Maximum parsimony analyses, 380–381
- MAXTREES setting, 385–386
- MC-SYM program, 162
- McKusick-Kaufman syndrome (MKKS), allelic variants for, 75, 76f
- Medical databases, 74–76
- MEDLINE, 57
- MEDLINE lines, in nucleotide sequence flatfiles, 10
- MEDLINE/MEDLARS layout, 61
- MEDLINEplus page, 66f
- MegaBLAST, 312, 394  
example using, 320–322
- MEGA program, 378, 380
- Meiotic maps, 31–32
- Melanie program, 466
- MEME technique, 129, 135
- Mendelian Inheritance in Man (MIM) numbers, 35
- MeSH terms (medical subject headings), 61
- Messenger RNA (mRNA), 144
- Metabolic pathways, 254, 257  
databases of, 264–268  
maps of, 267  
reconstruction of, 275–279
- Metabolism modeling, 265
- Metacharacters, 488t
- MetaCyc database, 27, 264
- Metaservers, 205, 240
- Methionine residues, 452–453
- Methods, submitted to dbSNP, 178
- METREE package, 380
- Mfold* RNA server, 148–151, 149–150f, 157  
color annotation and, 165  
output from, 149, 150, 151f, 152f
- MIAME standard, 414
- Microarray assays  
example of, 433–440  
“noise” in, 432
- Microarray-based comparative genomic hybridization (array CGH), 29–30
- Microarray data analysis, 415f
- Microarray experiments  
categories of, 413–414  
classification problem associated with, 429–432  
replication and, 412  
two-color, 411–414
- Microarray Gene Expression Data (MGED) Society, 414
- Microarray gene expression object model (MAGE-OM), 414
- Microarray image processing software, 416
- Microarray measurements, 420–422
- Microsatellite-based linkage screening sets, 33
- MIDAS software, 433
- Minimal tiling path, 352
- Minimum evolution (ME) method, 379, 380
- Mirror tree method, 274
- MMDBBIND, 260
- Modeling, of RNA tertiary structures, 162–163
- Modeller software package, 238
- Model Maker, 96–97
- Model organisms, sequencing of, 56
- Molecular cross-linking, 255
- Molecular INteraction (MINT) database, 263–264  
protein record in, 266f
- Molecular interaction databases, 254, 255–256, 260–264
- Molecular interaction networks, 254
- Molecular mechanics calculations, 163
- Molecular Modeling Database (MMDB), 233
- Molecular signatures, identification of, 26
- Molecular weight search (MOWSE) algorithm, 452, 459
- MolMol program, 244
- MOLPHY shareware package, 387
- Mol type, in nucleotide sequence flatfiles, 7
- Monte Carlo test, 328
- Moonlighting proteins, 211
- Mouse genome assembly, 360, 394
- Mouse Genome Database (MGD), 19, 44, 46f, 72
- Mouse Genome Informatics (MGI) database, 36, 44–45, 72
- Mouse genomic sequence, 82
- Mouse obese gene promoters, 131f
- mRNA, sources of, 81
- MS-Bridge function, 456, 459
- MS data analysis tools, 463. *See also* Mass spectrometry (MS)
- MS-Fit function, 456
- MS/MS ion query, 455
- MS/MS (tandem MS) method, 449–450, 461
- MS/MS spectrum, 450
- MS-Product function, 456
- Mullikin, James C., 171
- Multiexon predictions, 134
- Multi-LAGAN program, 395–396
- MultiMap program, 31
- MultiPipMaker program, 399, 404–407.  
*See also* PipMaker program; Pips (“percent identity plots”)
- Multiple alignment. *See also* Multiple protein sequence alignments; Multiple sequence alignment hierarchical, 328–330  
STAMP, 331–332
- Multiple FASTA sequences, 345
- Multiple protein sequence alignments, 325–340. *See also* Multiple alignment; Multiple sequence alignment  
generating multiple structure alignments, 331–332  
multiplying align sequences, 327–331  
sample analysis of, 336–338  
structural and evolutionary, 326–327  
tools to assist in, 332–336
- Multiple RNA sequences, predicting secondary structure common to, 159–161
- Multiple sequence alignment, 296. *See also* Multiple alignment; Multiple protein sequence alignments  
best method for, 332  
defined, 326  
hierarchical methods for, 328–330  
prediction of buried residues from, 333–336  
by PSI-BLAST and *i*SCANPS, 330–331
- Multiple sequences, finding alignment and common secondary structure for, 159–161
- Multiple structure alignments, generating, 331–332
- Multiple testing problem, 425
- Multipoint linkage analysis, 31
- Multispecies alignment generation, example of, 404–407
- Multispecies comparative mapping resources, 44
- Multispecies conserved sequences (MCSs), 404
- MULTIZ method, 396
- Munich Information Center for Protein Sequences (MIPS), 73
- MUSCLE software, 330
- Mutation data, 184

- Mutations, 172  
  major types of, 400  
  in protein-protein interactions, 274
- mVISTA program, 399  
  alignment results from, 401f
- National Cancer Institute (NCI) Cancer Chromosome Aberration Project, 31
- National Center for Biotechnology Information (NCBI), 4, 5f, 35, 56.  
*See also* Entrez system; LocusLink entries; NCBI entries
- National Human Genome Research Institute (NHGRD), 82
- NCBI Gene resource, 34-35
- NCBI Genomic Biology page, 35
- NCBI Map Viewer, 87-99. *See also* Map Viewer  
  home page for, 88, 91f
- NCBI TraceArchive, 173
- Nearest-neighbor free energy parameters, 146-147
- Needleman-Wunsch algorithm, 395
- Neighborhood, 304
- Neighboring concept, 56-58
- Neighbor joining (NJ) algorithm, 379, 380
- NEIGHBOR program, 384
- Neighbors, 63f
- NetPhos program, 464
- Network attributes, 281-282
- Network visualization tools, 279-284
- Neural networks, 201, 207
- Neutral mutations, 400
- NMR spectroscopy, 225-226, 241
- NMT utility, 464
- Node attributes, 281-282
- Nodes, 367
- Node visual attributes, 282
- Nomenclature committees, 28
- Nonaffine gap penalty, 304
- Noncoding regions, 118
- Non-Euclidean metrics, 424
- Nonhierarchical multiple alignment methods, 330
- Nonneutral mutations, 400
- Non-sequence-based maps, 109
- Nonsynonymous substitution, 180
- Normalization factor, 417
- "Novel" motifs, 135
- NUCheck, 226
- Nuclear magnetic resonance (NMR) methods, 162, 255
- Nucleic acid bases, IUPAC/UBMB codes for, 516
- Nucleic acids, one- and three-letter codes for, 516
- Nucleotide link, 62
- Nucleotide records  
  coding of, 5-7  
  in DDBJ flatfile format, 497-499  
  in EMBL flatfile format, 501-503  
  in GenBank flatfile format, 499-501
- Nucleotide scoring matrices, 303
- Nucleotide sequence databases, 5-7
- Nucleotide sequences, 119  
  in nucleotide sequence flatfiles, 7
- Numbers, Perl, 478
- Numeric comparison operators, 482t
- Obese gene promoters, 131f
- OC lines  
  in nucleotide sequence flatfiles, 10  
  in TrEMBL records, 19, 20
- Ofran, Yanay, 197
- OG lines, in TrEMBL records, 20
- OLD-MEDLINE, 57
- Oligomer-target binding, 155
- Oligonucleotides, 155, 156
- OligoWalk, 155-156, 157f
- OMIM link, 62
- OMIM numbering system, 74-75
- Online Mendelian Inheritance in Man (OMIM), 35, 42, 74-75. *See also* OMIM entries
- OpenGL (Open Graphics Library), 236
- Open reading frames (ORFs), 275, 279, 459
- Operons, 272
- Orchid assembly viewer, 350f
- Ordered sequences, versus finished sequences, 358-360
- Organelle (OG) lines, in TrEMBL records, 20
- Organismal divisions, in nucleotide sequence flatfiles, 7, 8t
- Organism classification (OC) lines  
  in nucleotide sequence flatfiles, 10  
  in TrEMBL records, 19, 20
- ORGANISM lines, in nucleotide sequence flatfiles, 10
- Organisms, home pages for, 99
- Organism source (OS) lines  
  in nucleotide sequence flatfiles, 10  
  in TrEMBL records, 19, 20
- Organism-specific search, 99
- Oriented sequences, versus full-shotgun sequences, 358-360
- Orthologous genes, searches for, 44
- Orthologous interactions, 274
- Orthologs, 296  
  rules for identifying, 370  
  versus paralogs, 369
- Orthology mappings, 273
- OS lines  
  in nucleotide sequence flatfiles, 10  
  in TrEMBL records, 19, 20
- Osprey network visualization tool, 282
- Overgo probes, 359
- Overlap SNPs, 186-187
- Overrepresentation analysis, 284-285
- Oxford Grid, 44-45
- OX lines, in TrEMBL records, 19, 20
- P1-artificial chromosomes (PACs), 27-28
- PAC clones, 34
- Pair HMMs, 401
- Pairs of close bidirectional best hits (PCBBHs), 273
- Pairs of close homologs (PCHs), 273
- Pairwise sequence similarity assessment, 173, 295-324. *See also* BLAST entries  
  BLAT and, 314-315  
  Dotplots and, 296-298  
  scoring matrices and, 298-304
- PAM matrices, 303, 377, 378
- Paralogous verification method (PVM), 261
- Paralogs, 296  
  rules for identifying, 370  
  versus orthologs, 369
- Parameters, in gene prediction programs, 134-135
- Parametric bootstrap, 383
- Partition function, 151-153  
  calculation with, 156-157
- Patent (PAT) sequences, in nucleotide sequence databases, 9
- PathBLAST plug-in, 282
- Pathologic system, 276
- Pathway databases, 254, 255-256  
  data quality of, 259  
  using, 259
- Pathway informatics, 254
- Pathway information, integrating gene expression data with, 284-287
- Pathway language, universal, 259
- Pathway maps, 267
- Pathway organization, 256-257
- Pathway Processor software, 284
- Pathway representation, 259
- Pathways, prediction algorithms for, 271-275
- Pathway Tools software, 264, 265, 278-279f, 284

- Pathway visualization tools, 278–279f, 279–284
- Pattern-driven algorithms, 12
- Pattern extraction, 490–491
- Pattern matching, 488–490
- Patterns, defined, 212
- PCR-based markers, 27
- PCR products, 353
- PDB files, 228–230, 232. *See also* Protein Data Bank (PDB) format for, 228, 230 Web site for, 226
- PDB ID codes, 226
- PDB Query Result Browser, 228
- PDB Structure Explorer, 228
- PDBSum database, 233
- PDQuest program, 466
- Pearson correlation coefficient ( $r$ ), 273, 424, 431
- Peptide analysis, 449
- Peptide bonds, 198
- Peptide mass fingerprinting (PMF), 449
- PeptideMass program, 456
- PeptIdent database, 452, 455–456, 469 search results from, 457f
- Perl  
arrays and hashes in, 491–492, 493  
biological analysis using, 475–496  
conditional blocks in, 483–484  
decision making in, 481–483  
error handling in, 481  
filehandles in, 480–481  
input and output in, 487, 479–480  
interpreters for, 476, 477  
lists in, 492  
loops in, 485–486  
pattern extraction in, 490–491  
pattern matching in, 488–490  
problem solving example using, 493–496  
sequence file length in, 487–488  
split and join functions in, 492–493  
truth in, 484–485
- Permutation tail probability (PTP) test, 382
- Permutation testing, 382–383, 423
- Pfam collection, 212, 214  
summary page for, 214f
- PFold program, 157, 159  
output from, 161f
- PHDacc, 207
- PHD files, 355  
format of, 515
- PHDhtm, 208
- PHD program, 336
- PHDsec, 200–201
- “Photo galleries,” 237
- Phrap assembly, 346–352. *See also* Phred/Phrap/ Consed program suite
- Phrap parameters, 347t, 351
- Phrap processing steps, 346
- Phred/Phrap/Consed program suite, assembly with, 346
- Phred program, 344  
processing steps in, 346
- PhyloBLAST software, 388
- PHYLODENDRON software, 388
- Phylogenetic alignment, 373–374
- Phylogenetic analysis, 365–392  
alignment procedure for, 374  
assumptions underlying, 369  
example of, 389–390  
interpretation in, 368–373  
multiple interpretations in, 369–373  
software for, 384–388  
terminology related to, 366–368
- Phylogenetic Analysis Using Parsimony (PAUP) package, 378, 385–386
- Phylogenetic footprinting, 129, 402–403
- Phylogenetic inference, ML approach to, 380, 381
- Phylogenetic profiles, 273
- Phylogenetic relationships, detecting, 369–373
- Phylogenetic trees, 367, 368f, 390f  
constructing, 373–383  
evaluating, 382–383  
rooting, 382  
searching for, 381–382
- Phylogeny Inference Package (PHYLIP), 330, 374, 384–385  
bootstrap analysis with, 386f
- PhyloVISTA feature, 399
- Physical maps, 33
- PipMaker program, 359, 398–399
- Pips (“percent identity plots”), 399, 400f
- Plug-ins, 236, 282
- PMF query, 454f, 455
- Png file format, 150
- Point accepted mutation (PAM) matrices, 301–302
- Polymerase chain reaction (PCR) resequencing, 173. *See also* e-PCR entries; PCR entries
- Polymorphic markers, 27
- Polymorphism data, 184
- Polymorphism databases, 190
- Polymorphisms, 172  
evolution and origins of, 172  
types of, 172–173
- PolyPhred program, 173–174
- Pooling samples, 412
- Populations, mutations in, 172
- Population samples, dbSNP database, 178
- Portals, 35, 42
- Position-specific scoring matrices (PSSMs), 312, 313f
- Position weight matrices (PWMs), 117
- PostScript file format, 150
- POV-Ray (Persistence of Vision-Ray Tracing) software package, 237
- PRECHECK, 226
- PredicNLS method, 216
- Prediction methods, 200–206. *See also* Predictive methods  
disagreement among, 205–206  
performance of, 202–206  
solvent accessibility, 207–208  
for transmembrane segments, 208–211
- Predictive methods. *See also* Prediction methods  
efficacy of, 126–128  
predicting function, 211–217  
strategies and considerations for, 132–135  
using protein sequences, 197–221  
using RNA sequences, 143–170  
visualization and integration tools, 135–138
- Predictome Web-based tool, 275, 282–283
- PredictProtein service, 200, 207, 208
- Prefinished genomic sequences, 26–27
- Preview browser, 99
- Primary databases, 5, 14–15
- Primary keys, in nucleotide sequence flatfiles, 8–9
- PRIMARY lines, in nucleotide sequence flatfiles, 15
- Primary structure, 199, 224
- Primers, custom, 357–358
- Primer walking, 342
- Principal component analysis (PCA), 429
- PRISM linkage mapping sets, 33
- Probability dot plot, 153, 155
- Processing signals, 463
- PROCHECK NMR, 242
- PROCHECK software, 226, 242  
output data from, 243f
- PROCLAME software, 464
- PROCRUSTES program, 120
- /product qualifier, in nucleotide sequence flatfiles, 14
- PROFacc, 207
- Profiles, defined, 212

- PROF output, 202–203f  
PROFsec, 200  
ProfTMB, 208  
Programs, for RNA secondary structure prediction, 148–157. *See also* Algorithms; Computer programs  
Prokaryotes, gene prediction in, 117  
Proline residue, 201  
PROMO collection, 129  
Promoter analysis, 128–132  
Promoter elements, 128  
Promoter Inspector, 131  
Promoter prediction, 128–129, 135, 137–138  
Promoter regions, 119  
prediction and characterization of, 129, 131–132  
structure of, 128f  
ProSight PTM software, 464  
PROSITE database, 212, 214  
search output from, 213f  
ProSup Web server, 244  
PROTDIST program, 377, 384  
Protein analysis  
chemical separation and, 449  
hidden Markov model in, 204  
via mass spectrometry, 448–450  
Protein arrays, 467  
Protein-based comparative maps, 44  
Protein breakdown, 449f  
Protein characterization, 446, 463–464  
Protein-coding genes, locating, 116  
Protein-coding regions, 120  
compositional bias in, 118  
Protein crystallography, 225  
Protein Data Bank (PDB), 226. *See also* PDB entries  
purpose of, 228  
Protein domain families, collection of, 212  
Protein expression, complexity of, 448f  
Protein folding, 198–199  
databases for, 245  
Protein function, 216, 370f  
Protein identification  
data preprocessing and, 451–452  
example of, 468–470  
major programs for, 452–463  
Protein Information Resource (PIR), 4  
Protein interaction prediction  
example of, 287–288  
phylogenetic-profile-based, 273  
using dataset integration, 275  
Protein (pr) link, 62, 95  
Protein modifications, classifying, 463  
Protein multiple sequence alignments.  
*See* Multiple protein sequence alignments  
ProteinProspector software, 456–459, 469  
output from, 460f  
submission screen with, 458f  
Protein-protein interactions, 255  
databases of, 263  
methods for predicting, 271–275  
Protein Quaternary Structure (PQS)  
query form, 233  
Protein records, coding of, 6–7  
Proteins  
analysis of, 446  
informational content of, ix–x  
predicting function of, 211–217  
relationships between, 57  
sequencing of, 4, 5  
solvent accessibility of, 206–208  
Protein sequence alignments, 244  
Protein sequences  
databases of, 16–22  
predictive methods using, 197–221  
Protein structure evaluation software, 242–243  
Protein structure prediction/analysis, 223–251  
*ab initio*, 241  
example of, 247–249  
homology modeling, 238  
protein structure comparison, 243–247  
protein structure evaluation, 241–243  
with structure databases, 231–233  
threading, 238–240  
three-dimensional visualization packages for, 235–237  
visualizing proteins in, 233–234  
Protein structures  
complexity of, 241  
databases of, 226–231  
determination of, 225–226  
Protein visualization programs, 236, 244  
Proteome Analysis Database, 21–22  
Proteome sets, UniProt, 21–22  
Proteomic analysis, 450  
algorithms used for, 467  
Proteomic research, database choice and, 451–452  
Proteomics, 446. *See also* Comparative proteomics methods  
Proteomics Standards Initiative (PSI), 270  
ProtFun method, 217  
PROTPARS parsimony program, 384–385  
Provenance, 259  
Pseudoatoms, 163  
Pseudogenes, 134  
Pseudoknots, 147  
PSI-BLAST (position-specific-iterated BLAST), 312–314  
multiple alignment by, 330–331  
searches with, 201, 240  
PSI-MI data model, 270f  
PSIPRED, 201–202, 333  
PSORT method, 216, 464  
Public databases, 175–187  
PubMed, 57, 59  
records for, 56  
PUBMED lines, in nucleotide sequence flatfiles, 10  
“Putative transmembrane helices,” 208  
PUZZLE program, 387  
*p* value cutoffs, 423. *See also* Westphal and Young stepdown *p* values  
Quackenbush, John, 409  
Quadratic discriminant analysis, 124  
Qualifiers, in nucleotide sequence flatfiles, 12, 14  
Quality assurance, in sequence databases, 22  
Quantifiers, 489t  
Quartet puzzling, 382  
Quaternary structure, 225  
Queries, LocusLink, 71  
Query interfaces  
GeneLynx, 41  
LDB2000, 38  
Rat Genome Database, 46  
Query modules, dbSNP, 181  
Query results, 109  
Query word, 304  
QuickPDB, 228, 230f  
Radiation hybrid (RH) analyses, 29  
maps for, 34  
RA lines, in nucleotide sequence flatfiles, 11  
Ramachandran plots, 233–234  
Ramkissoon, Kevin R., 445  
Random coil regions, 224  
Randomized character data, 382–383  
Randomized trees, 382  
RasMol (RAStEr MOLEcule) software package, 236  
Rat cytochrome C5, 205  
Rat Genome Database (RGD), 36, 45–47  
Rat genomic sequence, 82  
Ratio-intensity (R-I) plots, 418, 419f, 421f

- Rat mitochondrial membrane cytochrome B5, 200f
- Ray-tracing package, 237
- RD0260 structure, 166f
- REACTION table, 268
- Read pair viewer, 350
- Read-read matrix, 348f
- Recombination events, 31
- Reference blocks, in nucleotide sequence flatfiles, 10-11
- Reference comparisons, 413
- REFERENCE lines, in nucleotide sequence flatfiles, 10-11
- RefSeq database, 15, 16, 17
- RefSeq Genes, 83, 86f
- RefSNPs, 176, 179, 181
- Regulatory elements, 128
- Regulatory network reconstruction, 286-287
- Related Articles hyperlink, 61
- Relevance pairs model of retrieval, 58
- RepeatMasker program, 132, 310
- Repeats  
  identifying, 297  
  searching for, 132-133
- Replication, in microarray experiments, 412
- Report File* button, 155-156
- Reports, Ensembl SNP, 185f
- Report view, LocusLink, 71
- Representation system, 258
- Research Collaboratory for Structural Bioinformatics (RCSB), 226
- Residues. *See also* Buried residues  
  functionally important, 211  
  predicting features of, 198-211  
  structural characteristics of, 199
- RESOURCERER annotation resource, 410
- Restriction fragment length polymorphisms (RFLPs), 27
- Rhodopsin protein sequence prediction, example for, 218-219
- RL lines, in nucleotide sequence flatfiles, 11
- RNA  
  coding in nucleotide sequence flatfiles, 7  
  reference samples of, 413  
  sequencing of, 5
- RNA folding. *See also* RNA tertiary structures  
  partition function for, 151-153  
  thermodynamics of, 145
- RNAfold program, 158
- RNA mfold version 2.3 server, 148
- RNAML-compliant programs, 150
- RNA secondary structure prediction  
  accuracy of, 147-148  
  dynamic programming method for, 511-514  
  example of, 165  
  genetic algorithm for, 159  
  programs for, 148-157
- RNA secondary structures, 145f, 163. *See also* RNA secondary structure prediction  
  interactively drawing, 162  
  thermodynamics of, 146-147
- RNA sequences, predictive methods using, 143-170
- RNA structure, 144-145. *See also* RNA secondary structure entries; RNA tertiary structures
- RNAstructure algorithm, 155-156, 157
- RNA tertiary structures. *See also* RNA folding  
  modeling, 162-163  
  predicting, 162-163, 163-165
- RnaViz software, 162
- RN lines, in nucleotide sequence flatfiles, 11
- Root mean square deviation (RMSD), 244-245, 247
- ROSETTA program, 241
- Rost, Burkhard, 197
- RP lines, in nucleotide sequence flatfiles, 11
- RSATools, 137
- RT lines, in nucleotide sequence flatfiles, 11
- Rutgers University linkage map, 32-33
- rVISTA program, 399
- RX lines, in nucleotide sequence flatfiles, 11
- Saccharomyces cerevisiae*, 82  
  genome of, 461
- Saccharomyces* Genome Database (SGD), 73-74
- SAM graph, 436f. *See also* Significance analysis of microarrays (SAM)
- Samples, multiple proteins in, 451
- Sample size fields, in dbSNP, 178
- SAM-T99 (SAMt99) threading server, 202, 240  
  output from, 205f
- Sanger Institute, 42, 99, 103
- SCOP database, 245-246
- Scoring matrices, 298-304  
  biological factors in, 299  
  log odds ratio and, 301
- Scripts, 476  
  functioning of, 477
- Search engine, InterPro, 212-213
- Searches  
  MGI, 44  
  OMIM, 75
- SearchFields, 226-228
- SearchLite, 226, 227f
- Secondary databases, 5
- Secondary structure, 172, 224. *See also* Dynamic programming secondary structure methods; RNA secondary structure entries; RNA structure common to multiple RNA sequences, 159-161  
  suboptimal, 148  
  types of, 199
- Secondary structure elements (SSEs), 247
- Secondary structure prediction, 199-206, 217, 333-336  
  accuracy of, 205
- SEG program, 310
- Self-organizing map (SOM), 428-429
- Semimetric distances, 424
- Sensitivity (Sn), of predictive methods, 126
- SEQBOOT program, 385
- Sequence alignments, global versus local, 296
- Sequence analysis methods, 212-214
- Sequence assembly, 345-353
- Sequence assembly/finishing methods, 341-363  
  example of, 361  
  integrating clone sequences, 358-360  
  shotgun sequencing, 343-345
- Sequence-based maps, 34-35, 44
- Sequence comparisons, 296
- Sequence contigs, 48-49. *See also* Contig entries
- Sequence databases, 3-22  
  data flow between, 5f  
  growth of, 198  
  history of, 4-5  
  non-NCBI, 72-74  
  primary and secondary, 5, 14-15  
  protein, 16-22  
  quality assurance in, 22  
  submitting sequences to, 16
- Sequence-driven algorithms, 129-131
- Sequence elements, identifying, 211
- Sequence finishing, 353-358
- "Sequence first" approach, 35
- Sequence information, expansion of, 77
- Sequence masking, 132-133
- Sequence motifs, 211-212

- Sequence polymorphisms, 171-193  
  genotyping and, 187-188  
  International Haplotype Map Project and, 188-190  
  as resources, 190
- Sequence position, 26
- Sequence records  
  coding of, 5-7  
  formats of, 7-16
- Sequences. *See also* Genomic sequences; Nucleotide sequences; RNA sequences; Sequencing  
  downloading, 87  
  subcellular localization of, 216
- Sequence signals  
  analysis of, 117  
  identifying, 116-117
- Sequence similarity-based gene prediction, 119-120
- Sequence tag, 455
- Sequence-tagged clones (STCs), 28
- Sequence-tagged sites (STSs), 26. *See also* STS markers  
  in nucleotide sequence databases, 9
- Sequence version (SV) lines, in nucleotide sequence flatfiles, 9-10
- Sequence viewer (sv), 95
- Sequencing. *See also* Sequences  
  of DNA, 4, 5  
  of proteins, 4, 5  
  relationship to mapping, 26-27  
  of RNA, 5
- Sequencing gaps, 353
- Sequencing projects, large-scale, 173
- Sequest package, 450, 463
- Sequin program, 360
- SFCHECK, 226
- SFold, 156-157  
  server output from, 158f
- SGP-2 program, 120
- "Shatter" libraries, 353
- Sherry, Stephen T., 171
- Short sequence length polymorphisms (SSLPs), 27
- Short tandem repeats (STRs), 27, 172.  
  *See also* STR entries  
  genotyping, 187
- Shotgun cloning, 342f
- Shotgun proteomics, 450
- Shotgun sequencing, 82, 342, 343-345, 352
- Signaling pathways, 255. *See also* Signal transduction pathways
- SignalP, 476
- Signal peptides, 463
- Signal transduction pathways, 257
- Signature, defined, 212
- Significance, using Fisher's exact test to assess, 430
- Significance analysis of microarrays (SAM), 425, 430, 433. *See also* SAM entries
- Significance assessment, 449
- Significant genes, finding, 423-426
- SIM4 program, 120
- Simian immunodeficiency virus (SIV), 159
- Similarity, defined, 296, 369
- Simulation studies, 381
- Single linkage clustering, 427
- Single nucleotide polymorphisms (SNPs), 27. *See also* dbSNP entries; Human SNPs; Overlap SNPs; SNP entries; ssSNPs  
  genotyping, 187, 188  
  haplotype tag, 189
- Single-resource maps, 29
- Single sequence RNA, secondary structure prediction of, 148-157
- Single-strand conformational polymorphism (SSCP), 174-175
- Sir-graph software, 162
- Skewness test, 382
- SLAM program, 120, 401
- Smith-Waterman algorithm, 396
- Smith-Waterman alignment, 347
- SNP-based genotyping, 33
- SNP-based screening set, 32
- SNP Consortium, 173, 184
- SNP databases, 181-184  
  annotated list of, 184
- SNP discovery methods, 173-175
- SNP integration, in genome browsers, 184-187
- SNP Properties link, 105
- SNPView, 105, 106f
- Solvent accessibility, 206-208  
  prediction methods for, 207-208
- Sorting signals, 463
- SOSUI program, 208-209
- SOURCE annotation resource, 410
- Source feature, in nucleotide sequence flatfiles, 12-14
- SOURCE lines, in nucleotide sequence flatfiles, 10
- Southeastern Regional Genetics Group, 31
- Specificity (Sp), of predictive methods, 126
- Spliced alignments, 120
- Spliced ESTs track, 89f
- Spliced transcripts, 87
- Splice signals, predicting, 132
- Split and join functions, 492-493
- Spring-embedded algorithm, 281
- ssSNPs, 176, 177
- STAMP multiple alignments, 331-332
- Standard Finishing Practices and Annotation of Problem Regions for the Human Genome Project*, 353
- STAR (Self-defining Text Archival and Retrieval) method, 231
- States, 119
- Stein, Lincoln D., 475
- STR alleles, 172
- String comparison operators, 482t, 483
- String interpolation, 479
- STRING resource, 275, 276f
- Strings, Perl, 477-478
- STR polymorphisms (STRPs), 27, 187
- Structural alignment, 326-327
- Structural bioinformatics, 224
- Structural biology, 247
- Structural classification services, 246
- Structural Neighbors link, 228
- Structure  
  concept of, 216  
  three-dimensional, 66-70
- Structure comparison, 243-247
- Structure databases, 226, 231-233
- Structure Explorer, information obtained from, 229f
- Structure prediction, 58, 217-218
- Structure superposition method, 243-244
- STS content mapping, 29, 33-34
- STS creation, sequences suitable for, 34
- STS markers, 27, 93-94f, 181  
  region between, 92
- Student's *t* test, 423
- Subcellular localization, 216  
  prediction methods for, 216
- Subfamily analysis, with JalView and AMAS, 338
- SUBLOC method, 216
- Suboptimal secondary structures, 148
- Substitution models, 375-378  
  choosing, 378
- Substitution rate matrix, 376-377
- Substitution rates  
  between amino acids, 377-378  
  between bases, 376-377
- SuperPose Web server, 244
- Superposition process, mathematical approaches related to, 244
- Support vector machine (SVM)-based system, 202
- SV lines, in nucleotide sequence flatfiles, 9-10
- SWISS-2DPAGE program, 466
- Swiss Institute of Bioinformatics (SIB), 455

- SWISS-MODEL server, 238  
SwissPDB-Viewer, 237  
Swiss-Prot entry, 508-511  
Swiss-Prot protein sequence database, 4, 18-19, 73, 87, 452, 455  
Synonymous substitution, 180  
Synthetic lethal interaction, 257  
Systems Biology Markup Language (SBML), 271
- Table Browser, 83  
TAIR. *See* Arabidopsis Information Resource, The (TAIR)  
Tandem MS (MS/MS) method, 449-450, 461  
Tandem repeats, 297  
TargetDB database, 233  
"Targeted" region assemblies, 359-360  
TargetP method, 216  
TATA box, 137  
Taxonomic divisions, in nucleotide sequence flatfiles, 7, 8t  
Taxonomy information, in nucleotide sequence flatfiles, 10  
Taxons, 367  
TBLASTX algorithm, 402  
T-Coffee program, 330  
Technical replicates, 412  
"Template" gene prediction, 117  
Terminal exons, 118  
Tertiary structure, 224. *See also* RNA tertiary structures  
Tertiary structure prediction, future of, 163  
*Tetrahymena* sequence, 162  
TextEdit, 476  
TextSearch, 99  
*Thermus thermophilus* sequence, 163  
Third-party annotation (TPA), in nucleotide sequence flatfiles, 14-15  
Thomas, Pamela Jacques, 341  
Threading, 238-240  
Threading services, 240  
Three-dimensional structures, 66-70  
  neighbors in, 181  
Three-dimensional threading, 240  
Three-dimensional visualization packages, 235-237  
TIGR. *See* Institute for Genomic Research, The (TIGR)  
TIGR assembler, 349  
Tiling Path Format (TPF) file, 360  
Time of flight (TOF) method, 448. *See also* TOF mass spectrometer  
TITLE lines, in nucleotide sequence flatfiles, 10-11
- TMHMM program, 209-210, 211  
  output from, 210f  
TOF mass spectrometer, 447f. *See also* Time of flight (TOF) method  
Top-down proteomics, 450  
Topology-dependent test, 383  
TopPred program, 208, 211  
  output from, 209f  
Total intensity normalization, 418  
TPF Processor, 360  
TraceArchive, 173  
Training set, 159  
Transcribed sequence maps, 34  
Transcription, regulation of, 128  
Transcriptional enhancers, 128  
Transcription start site (TSS), 128  
  determining the beginning of, 135  
  locating, 131  
Transcript (RNA) map, 95, 96f  
Transcripts, Ensembl display of, 100-103  
Transcript/translation summary, 108  
TRANSFAC database, 129, 135, 137  
TRANSFAC matrix entries, 130f  
TransferRNA (tRNA), 144, 148, 149.  
  *See also* tRNA sequence RD1140  
Transition probability, 119  
Translated nucleotide searches, 119  
Transmembrane proteins, 208  
Transmembrane segment prediction methods, performance of, 210-211  
Transmembrane segments, 208-211  
TransView, 109  
TreeAlign program, 374  
Tree-building methods, 378-382  
  character-based, 380-381  
  distance-based, 378-380  
Tree-calculating program, 328  
Tree-drawing software, 388  
Tree graphs, 280  
TREE-PUZZLE program, 387  
TREEVIEW software, 388  
TrEMBL (translation of EMBL nucleotide sequences) database, 4, 18, 19  
tRNA sequence RD1140, 153, 155, 157  
  minimum free energy structure for, 155, 156f  
TSC database, 184  
TSC SNP discovery project, 173  
TSSG method, 131  
TSSW method, 131  
*t* statistic, 423  
*t* test, 425  
TWINSCAN (TwinScan) program, 120, 401  
Two-color microarray experiments, 411-414
- 2D gel databases, 466-467  
2D gel methodology, 464  
Two-dimensional electrophoretic gel separation, 449  
Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE), 446, 464, 465-466, 467  
Two-dimensional threading, 240  
2DWG Image Meta-Database, 466
- Ubiquitin protein structures, 235f  
UCSC Genome Bioinformatics Group, 35, 83  
  home page for, 83  
UCSC Genome Browser, 34, 83-87, 135, 396, 397f. *See also* Genome Browser  
UCSC Human Genome Browser, 184-187  
Uncaptured gaps, 353  
Unfinished sequences, in nucleotide sequence databases, 9  
UniGene clusters, 35  
UniGene database, 28, 35  
Uninterpreted MS/MS spectral matching, 450  
UniParc (UniProt Archive) database, 17-18  
UniProt database, 17, 18f  
UniProt Knowledgebase, 15-16, 18-22  
UniProt proteome sets, 21-22  
UniRef50 database, 21  
UniRef90 database, 21  
UniRef100 database, 21  
UniRef database, 17, 18f, 21  
UniSTS database, 27, 48  
  Web site for, 35  
Universal Protein Resource. *See* UniProt entries  
University of California Santa Cruz. *See* UCSC entries  
Unix-based assemblers, 346  
Unmasked sequence data, 133  
Unstructured regions, 224  
Unweighted pair group method with arithmetic mean (UPGMA), 379, 380  
Uroporphyrinogen decarboxylate (URO-D) gene, 120, 121f, 122f, 123f, 125f  
  masking of sequence of, 132-133  
US BIONET, 4  
Use-case, 258  
U.S. Meat Animal Research Center, genome maps from, 47  
Validation information, submitted to dbSNP, 179

- Validation Suite, 226  
Validation technique, 432  
Valine biosynthesis pathway, 278f  
Variable interpolation, 479  
Variable number of tandem repeat units (VNTRs), 27  
Variables, Perl, 478  
Variance regularization, 418-420  
Variation databases, 172-173  
VAST server, 246-247  
VAST table, 57  
VCMMap, 47  
Vector Alignment Search Tool (VAST), 56-58. *See also* VAST entries  
Vega transcripts, 101  
Verification technique, 432  
Verify3D, 243  
VERSION lines, in nucleotide sequence flatfiles, 9-10  
Version number, 6, 14  
Vertebrate Genome Annotation (VEGA) database, 101  
Vienna RNA Package, 151-155, 157, 158, 159  
    Web server output from, 154f, 155  
Vienna Server input form, 153f  
Views, in Ensembl, 99  
Virtual reference sequences, 82  
VisANT applet, 282-284  
VISTA Browser, 396, 397f, 398f  
Visual attributes, 282  
Visual styles, 282  
Visualization tools, 135-138  
    network and pathway, 254, 279-284  
Volume, Area, Dihedral Angle Reporter (VADAR), 242-243  
Vtrace program, 345f  
Washington University Genome Sequencing Center (WUGSC), 34  
Web-accessible services, 238  
Web-based query forms, dbSNP, 181  
WebMol software package, 236  
WEBPHYLIP software, 388  
Web sites, for mapping, 28  
Weighted key terms, 58  
Weighted parsimony, 376  
Westphal and Young stepdown *p* values, 425. *See also* *p* value cutoffs  
WGS assembly, 342  
WGS reads, 360  
WHATIF software package, 238  
White, Peter S., 25  
Whole-genome alignments, precomputed, 396f  
Whole-genome assemblers, 352  
Whole-genome mapping resources, 42  
Whole genome shotgun (WGS) sequences, 343, 352-353. *See also* WGS entries  
    in nucleotide sequence databases, 9  
Windows Notepad, 476  
Wishart, David, 223  
WIT (What Is There?) database, 276-279  
Wolfsberg, Tyra G., 81  
Word matching, 347-348  
WormGRID, 261  
WormPep database, 476  
Xenologs, 369  
    rules for identifying, 370-373  
X-ray crystallography, 225  
Xrna software, 162  
Yammp software, 163  
YeastGRID, 261  
y-ions, 450  
Z3 program, 466  
Z-scores, 328  
Zuker, Michael, 143