Subject Index

Ab initio prediction, 621  
ABI SOLiD, 385  
Accepted point mutations (PAM), 79, 80, 81  
alternatives to, 91–94  
mutation probability matrix, 82–84  
PAM250 and other PAM matrices, 84–88  
practical usefulness of PAM matrices, 91  
Accession numbers, 34–36  
Accuracy, definition, 490  
Acquired Immune Deficiency Syndrome (AIDS), 766–770  
Acrocentric chromosomes, 348  
Adenine, 435  
Adenosine 5′ phosphosulfate (APS), 385  
Adenosine triphosphate (ATP), 564  
Advanced database searching, 167–168  
advice for students, 198  
BLAST-like alignment tools for rapid DNA searches, 186–194  
domain enhanced lookup time accelerated BLAST (DELTA-BLAST), 177–178, 197–198, 226  
finding distantly related proteins, 171–181  
hidden Markov models (HMMs), 181–186  
next-generation sequencing (NGS) alignment to reference genome, 194–197  
organism-specific BLAST sites, 168–170  
pattern-hit initiated BLAST (PHI-BLAST), 179–181  
perspective, 197  
pitfalls, 197–198  
position-specific iterated BLAST (PSI-BLAST), 171–177, 197–198  
profile searches, 181–186  
specialized BLAST sites, 168–171  
Arabidopsis thaliana, 28, 643, 910–913  
ARB project, 444  
Archaea, 7, 24, 700  
classification, 798–811  
genome size and geometry, 801–805  
genomes, 797–798  
lifestyle, 805–808  
molecular sequences, 810–811  
phylogenetic diversity, 810  
protein-encoding genes, 804  
ribosomal RNA sequences, 809–810  
Archaeal genome analysis, 814–830  
challenges, 825  
finding genes, 819–825  
gene annotation, 825–827  
lateral gene transfer (LGT), 827–830  
nucleotide composition, 817–819  
Archaeal genome comparisons species with closely related strains determined, 831  
Arginine (Arg), 76  
Array comparative genomic hybridization (aCGH), 356, 357  
Arthropods, 920  
Asexual reproduction, 309  
Ascomycetes, 869  
Asparagine (Asn), 76  
Aspartic acid (Asp), 76  
Aspergillus, 871  
Association of Biomolecular Resource Facilities (ABRF), 542–543, 551, 565  
Autism, 1024–1025  
Babesia bovis, 898  
Bacteria, 7, 24, 700  
classification, 798–811  
genome size and geometry, 801–805  
genomes, 797–798  
human disease relevance, 808–809  
lifestyle, 805–808  
molecular sequences, 810–811
Bacteria (continued)
morphological criteria, 800–801
phylogenetic diversity, 810
protein-encoding genes, 804
ribosomal RNA sequences, 809–810
vaccine-preventable disease, 808
Bacterial genome analysis, 814–830
Bacterial genome analysis (continued), 815
search concepts, 145
search overview, 151
search principles, 146–151
search statistics, 141–142
search steps, 124–138
search strategies, 145–155
search summary, 133
selecting appropriate program, 124–126
selecting database, 126–127
specialized sites, 168–171
specifying sequence of interest, 124
stand-alone version, 135–138
taxonomy report, 154
BLASTn tool, 70
threshold value, effect of, 140
web resources, 160
Basidiomycetes, 870
Bayesian inference tree-building method, 290–293, 294
BED files, 56–57, 327, 361, 414–416, 424
BEDTools, 413–417
Beta globin gene (HBB) see Globins
Bifurcating rooted trees, 263
Binomial coefficient, 289
BioConductor, 11
BioGrid network map, 677
Bioinformatics definitions, 3–4
other informatics disciplines, 15
overview, 5–8
software see Software for bioinformatics
Biological databases see Databases
Biomer project, 20, 54, 314–320
Birth-and-death evolution model, 353, 354
Bit score, 75, 91–92, 132, 134, 137, 143, 175, 438
BLAST-like alignment tools for rapid DNA searches, 186–194
benchmarking to assess genomic alignment performance, 187–188
BLAST-like tool (BLAT), 192, 193
BLASTZ, 188–191
discontinuous MegaBLAST, 191
Enredo, 191
Limited Area Global Alignment of Nucleotides (LAGAN), 192–194
MegaBLAST, 191–192
PatternHunter, 188, 189
Pecan, 191
Sequence Search and Alignment by Hashing Algorithm (SSAH2), 194
BLAST-like tool (BLAT), 192, 193
BLAST-related algorithms, 170
see also Domain enhanced lookup
WU-BLAST 2.0, 170
BLASTZ, 188–191
Block substitution matrix (BLOSUM), 91–94
Boa constrictor, 311
BodyMap-2, 459
Bos taurus, 25, 28, 718, 736, 744, 934
Bowtie-2 program, 196–197
Branches of phylogenetic trees, 259–262
Burrows–Wheeler Transform (BWT) alignment, 196–197
Butterflies, 922–923
BWA program, 196–197
C-value, 310–312, 322–323
protein-coding gene paradox, 342
Caenorhabditis elegans, 311, 643–644, 918–919
Cancer, 718, 721, 760, 762, 1033
Candida albicans, 175, 871–872
Canis familiaris, 311, 934
Cap analysis gene expression (CAGE), 467
Carcharias obscurus, 311
Caseins, 85, 86
Catalogue of somatic mutations in cancer (COSMIC), 1033
CATH database, 613–615
search results, 615
CCAAT box, 336
CEL files, 496
CEL definition file (CDF), 505
input, 506
Central bioinformatics resources, 31–34
Centroid axes, 500
Chagas’ disease, 892
Charge coupled device (CCD) camera, 385
Chemical mutagenesis, 665
Chi-squared analysis, 256
Chlorophyta, 908–910
Chloroplast genomes, 714–715
Chloroplasts, 703, 714–715, 723, 743, 902
Chromalveolates: *Plasmodium falciparum*, 895–898
Chromatin diminution, 349
Chromatin immunoprecipitation sequencing (ChIP-seq), 345
Chromosomal abnormalities, 349, 351–353, 355–359, 1050–1051
Chromosomal aneuploidy frequency, 1025
Chromosome territories, 314
Chromosomes
analysis by ENCODE project, 320–321, 323–324
analysis in genome browsers, 314
analysis using BioMart and biomaRt, 314–317
deletions, 349, 352
duplications, 351, 353
fission, 348
fusion, 348
gene density, 336
gene family models, 353–354
human, 313, 959–963, 979–986
inactivation, 349
inversions, 351
measuring change, 355–359
single nucleotide polymorphisms (SNPs), 354–355
translocation, 348
yeast, 854–860
Ciliophora, 899
*Paramecium tetraurelia*, 899–901
*Tetrahymena thermophila*, 902
*Ciona intestinalis*, 925–926
Circular binary segmentation (CBS) method, 356
Cis-regulatory modules (CRMs), 342
Clades, 262
Cladograms, 262
Classical structure biology, 601
ClinVar database, 421
ClustalW program, 208, 209–214
Clusters
agglomerative hierarchical clustering, 512–513
divisive hierarchical clustering, 512–513
partition methods, 516–517
relatedness, 515
self-organizing maps, 517
Clusters of Orthologous Groups (COGs) database, 573
*Escherichia coli*, 639
Coding sequence (CDS), 40
Codons of DNA, 83
COILS program, 559, 563
Co-immunoprecipitation, 672
Column score (CS), 223
Command line software tools, 10, 11–12, 317–320, 824
NCBI access, 42–49
Comparative genome hybridization (CGH), 355
Comparative genome hybridization, array (aCGH), 356, 357
Comparative modeling, 618–619
Complete Genomics self-assembly DNA nanoarrays, 387
Complex disorders, 1024–1025
Concerted evolution model, 353–354
Confidentiality, 390, 1044
Consensus Coding Sequence (CCDS) project, 37
Consent, informed, 390
Conservative substitutions, 74
Conserved Domain Database (CDD), 226
Conserved Domain Database of profile HMMs (Pfam), 223–224, 225, 226
Critical Assessment of Genome Interpretation (CAGI), 668
Critical Assessment of Protein Function Annotation (CAFA), 672
Critical Assessment of Techniques for Protein Structure Prediction (CASP), 621–622, 623
Cross-linking of proteins, 673
*Cryptococcus neoformans*, 872–873
*Cryptosporidium hominis*, 898
CuffLinks sample protocol, 523–524
assembling transcripts, 525
differential expression, 525–526
CummeRbund sample protocol, 526–527, 528
Curation of databases, 227
Cyclic reversible termination, 382–384
Cysteine (Cys), 76
Cystic fibrosis, 622–624
Cystic fibrosis transmembrane regulator (CFTR), 624
Cytosine, 435
Dayhoff Model of protein scoring, 79
step 1 – accepted point mutations (PAM), 79, 80, 81
step 2 – frequency of amino acids, 79, 81
step 3 – relative mutability of amino acids, 80–82
step 4 – mutation probability matrix for 1 PAM evolutionary distance, 82–84
step 5 – PAM250 and other PAM matrices, 84–88
step 6 – relatedness odds matrix, 88
step 7 – log-odds scoring matrix, 89–91
Dayhoff’s protein superfamilies, 77
DbEST database, 455
De Bruijn graphs, 395–396, 398
DeepView software, 593, 594
Degree of divergence, 274
Deletions, 349, 409–410
chromosomes, 352
mutations, 78
Descriptive statistics for microarray data analysis, 511
classification of genes or samples, 517–519, 520
clustering strategies, 517
confusion matrix, 521
data visualization methods, 518
hierarchical cluster analysis, 511–516
k-means clustering, 516–517
multidimensional scaling (MDS) compared with principal component analysis (PCA), 517
self-organizing maps (SOMs), 517, 518

Dictyostelium discoideum, 916–917
Dideoxynucleotide sequencing, 19, 380
Difference gel electrophoresis (DIGE), 547
Digital Differential Display (DDD) tool, 456, 457
Diploid cells, 309, 348
Direct protein sequencing, 543
Edman degradation, 544
Disability-adjusted life years (DALYs), 1016–1017
Discontinuous MegaBLAST, 191
Distance measures compared with similarity measures, 213
Divergent evolution model, 353
Divisive hierarchical clustering, 512–513
DNA
chromosomal variation, 347–355
codons, 83
functional genomics, 668
interspersed repeats, 325–326
noncoding and repetitive sequences, 323–325, 327
processed pseudogenes, 326–331
rDNA, 444
segmental duplications, 331–333
simple sequence repeats, 331
structure, 434
tandemly repeated sequences, 333–334
transcription, 434
transposon-derived repeats, 325–326
DNA Database of Japan (DDBJ), 21, 22
data types, 26–27
organisms, 24–25
DNA next-generation sequencing (NGS) analysis, 387–421
experimental design, 389–390
FASTQ files, 391–394
genome assembly, 394–398
interpretation biological significance of variants, 417–421
overview, 387–389
SAM/BAM format files and SAMtools, 402–408
sample preparation, 390
sequence generation, 390–391
storing data, 421
variant calling, 408–410
VCF format and VCFtools, 410–413
visualization and tabulating data, 413–417
workflow chart, 388, 389
DNA next-generation sequencing (NGS) technologies, 379, 382
ABI SOLID, 385
compared with Sanger sequencing, 382
Complete Genomics self-assembly DNA nanorays, 387
cyclic reversible termination, 382–384
decline in costs, 383
Illumina, 382–384
Ion Torrent, 387
ligation sequencing, 385
Pacific Biosciences DNA sequencing, 387
pyrosequencing, 384–385, 386
DNA sequence databases, 20–24
content, 24–31
data types, 26–27
enomic DNA, 27
growth, 22
range of file sizes, 23–24
scales of basepairs, 23
dNA trees, 268–270
DNase I, 345
Documentation, 14
Domain enhanced lookup time accelerated BLAST (DELTABLAST), 177–178, 226
assessing performance, 179
pitfalls, 197–198
Domains, 552–559
definition, 553, 554
Homo sapiens, 554
methyl-binding domains, 557
multidomain proteins, 556–557
yeast, 853
Dotplots, 104–105, 783–785, 832–834
Down syndrome, 313, 349, 480, 1020, 1025
Drosophila genome database (FlyBase), 566
Drosophila melanogaster, 311, 645, 919–921
Dysidea crawshagi, 311
E value, 142–143
reltion to p values, 143–144
Ebola virus, 775–776
EcoCyc database, 825–826
Edge branch of phylogenetic trees, 259–262
EDirect, 21, 42, 44–49, 137, 160, 236, 541, 578, 744, 878, 1002, 1018–1019
Edman degradation, 544
ELANDv2 program, 196
Electron crystallography, 600
Encephalitozoon cuniculi, 873
ENCODE Genome Annotation Assessment Project (EGASP), 339–340
ENCODE project, 447, 467
chromosome analysis, 320–321
critiques of, 322–323
functional elements catalog, 323
gene definition, 335–336
Enredo program, 191, 231, 232
Ensembl, 11, 32–33, 34, 50–52
BioMart, 316
chromosomes, 314, 315
genomic multiple sequence alignments, 231, 232
Human Genome Project, 959–961
Ensembl BLAST, 168–170
Ensembl Genomes, 710
Entrez, 20, 31
command-line access, 45, 46
usage tips, 32
Environmentally caused disease, 1029
Enzyme Commission (EC) numbers, 38, 567, 573–574
Epicellular bacteria, 806
Epstein–Barr virus (EBV), 762
Equilibrium dialysis, 673
Errors
DNA alignment, 400
DNA sequencing error rates, 379, 382, 385, 391, 393
genome annotation, 826–827
genome assembly, 397–398
Microsoft Excel, 42, 485
protein structure modeling, 619
spelling, 827
variant calling, 404, 420
web-based versus command-line analysis, 320
Erysiphe cichoracearum, 311
Escherichia coli, 639–640
genome, 715–716
phylogenetic relationships of strains, 815
Ethical considerations, 390
N-Ethyl-N-nitrosurea (ENU), 665
Euclidean distance, 513
Euglenozoa, 893
Eukaryotes, 7, 24, 700
first chromosome, 715
first genome, 715
first organellar genome, 712–714
gene annotation, 738–742
genomes, 228
origins, 704
phylogeny, 888
ribosomal DNA, 444
Eukaryotic chromosome, 308
advice to students, 360
algorithms for finding genes, 338
chromosomal variation in individual genomes, 349–355
comparison of eukaryotic DNA, 346–347
deletions, 352
differences compared with Bacteria and Archaea, 308–309
duplications, 351, 353
dynamic nature, 347–349
EGASP competition, 339–340
features, 309, 310–323
gene content, 334–342
gene family models, 353–354
genes, finding, 336–339
genomes, 310
inversions, 351
measuring change, 355–359
organization, 312–314
perspective, 359
pitfalls, 359–360
protein-coding gene paradox, 342
protein-coding gene study resources, 340–342
regulatory regions, 342–346
repetitive DNA content, 323–324
single nucleotide polymorphisms (SNPs), 354–355
structural variations, 351–354
variation in chromosomal DNA, 347–355
web resources, 360
whole-genome duplication, 347–349
Eukaryotic genomes
advice for students, 941
bioinformatics, 889
biological principles, 889
cataloguing information, 888–889
Chromalveolates, 895–906
comparative genomics, 889
databases, 913
introduction, 887–889
metazoa, 916–940
perspective, 940–941
pitfalls, 941
plant genomes, 906–916
protozoans, 890–892
sequence analysis, 889
unicellular pathogens, 892–895
web resources, 942
European Bioinformatics Institute (EBI), 11, 21, 32–34
BLAST-related algorithms, 170
sequence similarity search tools, 171
European Molecular Biology Laboratory (EMBL)-Bank Nucleotide Sequence Database (EMBL-BANK), 20–21, 22, 32
data types, 26–27
organisms, 24–25
European Molecular Biology Open Software Suite (EMBOSS), 100
European Nucleotide Archive (ENA), 14, 393, 465
Evolution of life, 78–79
Evolutionary distance, 82–84
Evolutionary divergence between sequences, 275
Exons, 336, 337
Expert Protein Analysis System (ExPASy), 546, 562
Expressed Sequence Tags (ESTs), 28, 122, 337, 454, 455
Expression quantitative trait loci (eQTLs), 459, 468, 469
Expresso program, 220–221
Extracellular bacteria, 806
Extreme value distribution, 120, 141–142
Extremophilic microbes, 806
Facultatively intracellular bacteria, 806
False discovery rate (FDR), 495, 502
FASTA format, 40–41, 73, 169, 381
FASTA local alignment algorithm, 103–104
FASTG files, 394
FASTQ, 733–734
FASTQ files, 391–392
finding files, 392–393
format, 387
format conversion, 392
Illumina 1.3+ format, 391
quality assessment, 393–394
quality scores, 392
Sanger format, 391
Solexa format, 391
viewing files, 393
Feng–Doolittle method for progressive sequence alignment, 208–209, 210, 214, 218
File formats, 57
Find-a-gene project, 156–159
Fisher’s exact test, 457, 458
Fission of chromosomes, 348
Fluorescence in situ hybridization (FISH), 355
Fluorescence resonance energy transfer (FRET), 673
Fold recognition, 619–620
Forbidden mutations, 863–864
Forward genetics, 649
chemical mutagenesis, 665
compared with reverse genetics, 665–666
Fragile sites, 349
Frequency of amino acids, 79, 81
Frogs, 929
Functional annotation of microarray data, 528–529
Functional Annotation of the Mouse (FANTOM) project, 459
Functional genomics, 5
advice for students, 686
approaches to function, 666, 667
bioinformatics resources, 682–685
central dogma, 666–670
definitions of function, 666–667
DNA, 668
forward and reverse proteomics, 671
integrating information, 668
introduction, 635–638
model organisms, 638–647
pairwise interactions and protein networks, 678–682
perspective, 685
pitfalls, 686
protein–protein interactions, 672–678
proteins, 672
reverse and forward genetics, 672–678
RNA, 668–670
web resources, 686
Fungal genome analysis, 869–870
Aspergillus, 871
Candida albicans, 871–872
Cryptococcus neoformans, 872–873
Encephalitozoon cuniculi, 873
Neurospora crassa, 873–874
other genomes, 876
Phanerochaete chrysosporium, 875
projects using Ascomycetes, 869
projects using Basidiomycetes, 870
Saccharomyces pombe, 875–876
Fungal genomes, 847–848
advice for students, 877
perspective, 876
pitfalls, 877
web resources, 877
Fungi, 24
description and classification, 848–849
pathogens, 876
taxonomy, 849
Fusion of chromosomes, 348
G protein-coupled receptor (GPCR), 599, 624
Galaxy program: genomic multiple sequence alignments, 229–231
Galaxy program, 20, 57–58
Galaxius domesticus, 311
Gamma distribution, 278
Gaussian distribution, 107
GCRMA normalization, 488–489
GenBank, 11, 20, 22, 52–53
data types, 26–27
organisms, 24–25
GENCODE project, 328–329, 339, 340–342, 447
Gene discovery, 155–159
Gene expression, 5–6, 450–466
Gene Expression Omnibus (GEO), 14, 465
Gene Ontology (GO) Consortium, 551, 552, 566–570
controlled vocabulary, 569
evidence codes, 568
participating databases, 567
Gene Set Enrichment Analysis (GSEA), 529
Gene silencing, 662–664
Gene symbols, official, 27, 317–320
Gene trapping, 657–660
strategies, 659
Gene trees, 266–268
GeneCards, 1041
General linear models, 504
Generic Model Organism Project (GMOD), 913
Genes
classification, 517–519
definition, 334–336
eukaryotic chromosomes, 334–342
identification, 335
models for creating families, 353–354
single nucleotide polymorphisms (SNPs), 354–355
Genetic code, 83
Genetic footprinting, 661
Genome, 5
Genome analysis projects
advice for students, 743
ancient DNA, 725–727
annotation, 737–738
annotation in eukaryotes, 738–742
applications of sequencing, 720
assembly, 730–733
challenges, 735–737
FASTQ, 733–743
HTGS archive, 730
introduction, 720–727
large-scale projects, 721–722
mapping contigs, 734–735
metagenomics projects, 727
perspective, 742
pitfalls, 742–743
resequencing projects, 725
role of comparative genomics, 724–725
selection criteria, 722–724
sequence completed, 735
sequencing centers, 728
trace archive, 728–730
Genome Analysis Toolkit (GATK), 388, 401, 403–404
Genome Assembly Gold-standard Evaluation (GAGE), 397
Genome assembly using NGS data, 394–396
completion standards, 398
performance evaluation, 396–398
software, 395
Genome browsers, 49–52
Genome Portal of DOE, 710
Genome Reference Consortium (GRC), 49, 395
Genome sequence alignment, 399–400
Genome Analysis Toolkit (GATK), 401
repetitive DNA, 400–401
software, 399, 1039–1040
Genome Survey Sequences (GSSs), 27
Genome Workbench software, 417, 418, 855–856, 1039–1040
Genomes On Line Database (GOLD), 710
Genomes
Archea, 797–798
Bacteria, 797–798
bioinformatics aspects, 701
biological principles, 701
cataloguing comparative genomic information, 701
cataloguing genomic information, 701
chomosomal variation in individual genomes, 349–355
compared, 310
currently sequenced, 709
features of eukaryotic genomes, 310–323
human disease relevance, 701
introduction, 700–710
life on Earth, 705
molecular sequences, 705–709
multicellular organism, 716
organization of eukaryotic chromosomes, 312–314
size variation, 310–312
sizes, 309–310, 801
systematics, 701–704
taxonomy, 709–710
web resources, 710
whole-genome duplication, 347–349
Genome-sequencing projects, 711
1981, 712–714
1986, 714–715
1992, 715
1996, 715
1997, 715–716
1998, 716
1999, 716
2000, 716
2001, 716–717
2002, 717
2003, 717
2004, 717–718
2005, 718
2006, 718
2007, 718
2008, 718
2009, 718
2011, 719
2012, 719
2013, 719
2014, 719
2015, 720
chronology, 711
Genome-wide association studies (GWAS), 468, 643, 1047–1050, 1052
Genomic disorders, 1025–1028
chromosomal aneuploidy frequency, 1025
molecular mechanisms, 1028
Genomic DNA databases, 27, 227–234
assessment of whole-genome alignment methods, 231–234
Ensembl program, 231, 232
Galaxy program, 229–231
UCSC Genome Bowser, 229, 230
Genomic Evolutionary Rate Profiling (GERP), 229
Genomic multiple sequence alignments, 227–234
assessment of whole-genome alignment methods, 231–234
Ensembl program, 231, 232
Galaxy program, 229–231
UCSC Genome Browser, 229, 230
Genomic promoter regions software, 343
Genomic regulatory factors
databases, 342–345
nonconserved elements, 346
ultraconserved elements, 345
Genotype and phenotype, 637–638
Genotype-Tissue Expression (GTEx)
project, 459
GENSCAN program, 338
GEO dataset analysis using R analyses, 504–505
CEL file input, 506
identifying differentially expressed genes, 508–510
microarray analysis and reproducibility, 510–511
microarray analysis plots, 507
RMA normalization, 506–508
GEO2R resource at NCBI, 482
corrections for multiple comparisons, 494–495
data normalization, 486–488
R scripts, 482–485
robust multiarray analysis (RMA) normalization, 488–490
statistical tests, 490–494
Giant viruses, 782–783
Giardia lamblia, 891–892, 942
Gibbs free energy, 590
GLEAN gene model combiner, 339
GLIMMER gene finding algorithm, 820–825
Global alignment algorithms, 96
Needleman and Wunsch algorithm, 96–100
statistical significance, 106–108
websites and URLs, 110
Globins, 6–8
see also Myoglobin
beta globin gene, 20, 650–653
beta globin gene mutants, 654
beta globin BLAST example, 136–138
Dayhoff subfamilies, 248
hemoglobin, 6, 7
phylogeny tree, 247
structural biology, 591
substitution frequencies, 111
Glutamic acid (Glu), 76
Glutamine (Gln), 76
Glyceraldehyde-3-phosphate dehydrogenase (GAPDH), 85
Glycine max, 28, 910, 915
Green algae, 908–910
Green fluorescent protein (GFP), 645
Guanine, 435
GXW motif, 590
Haemophilus influenzae, 19
Hamming distance, 274, 277
Hammondia hammondi, 899
Haploid cells, 309, 348
C value, 310–312
HaplotypeCaller, 408
HapMap browser, 988
HapMap project, 354, 717
Hash tables, alignment based upon, 194–196
Hemiascomycetes, 865
functional element identification, 868–869
whole-genome duplication, 866–868
Hemoglobin see Globins
Hepatitis A, 761
Hepatitis B virus (HBV), 761, 762
Hepatitis C virus (HCV), 258, 762
Herpesvirus, 776–780
Heterokonta, 905
Hive algorithms, 77
HGNC see HUGO Gene Nomenclature Committee
Hidden Markov models (HMMs), 181–186, 566
HMMER program, 184–186, 187
Protein Family database of profile HMMs (Pfam), 223–224, 225, 226
Hierarchical cluster analysis, 511–516
Highly homologous sequences, 75
High-throughput gene expression data acquisition, 462
High-Throughput Genomic Sequences (HTGS), 27
H-Invitational Database, 459
Histidine (His), 76
Histones, 52–53
HIV-1 pol, 53–54
BLAST multidomain protein search, 151–155
HMMER program, 184–186, 187, 223, 224
Homogentisate 1,2-dioxygenase (HGD), 1014
Homo sapiens, 25, 28, 311
chromosome sequenced, 716
chromosomes, 313
draft sequence of genome, 716–717
protein domains, 554
variation, 647
HomoloGene, 270
compared with NCBI Gene, 42
Homology
definition, 70–74
evolution of life, 78–79
history, 72
modeling, 618–619
Honeybee, 921, 923
Hordeum vulgare, 25
Hubbard plots, 623
HUGO Gene Nomenclature Committee,
20, 27, 54, 60, 61, 62, 314, 316–317, 319
Human chromosomes, 313, 979–981
Group A, 981–982
Group B, 982–983
Group C, 983
Group D, 983
Group E, 984
Group F, 984
Group G, 984
groups, 980
Human disease-associated genes and loci, 1046
chromosomal abnormalities, 1050–1051
genome-wide association studies (GWAS), 1047–1050
linkage analysis, 1047
Human disease categories, 1020–1036
allele frequencies and effects sizes, 1020–1021
cancer, 1033
complex disorders, 1024
environmentally caused disease, 1029
genetic background, 1030
genomic disorders, 1025
mitochondrial disease, 1030–1032
monogenic disorders, 1021–1024
somatic mosaic disease, 1032–1033, 1035
Human disease databases, 1036
amino acid substitutions, 1045–1046
ClinVar, 1040–1041
GeneCards, 1041
Human Disease Mutation Database (HGMD), 1039
Integration of Disease Database, 1041
limitations, 1045
Locus-Specific Mutation Databases, 1041–1044
Online Mendelian Inheritance in Man (OMIM), 1036–1039
PhenCode project, 1044–1045
Human disease gene functional classification, 1060–1063
disease characteristics, 1062 protein products, 1061
Human disease genes in model organisms, 1055
nonvertebrate species, 1056–1057
primates, 1059–1060
rodents, 1058–1059
Human Disease Mutation Database (HGMD), 1039
Human diseases
advice for students, 1063
perspective, 1063 pitfalls, 1063
Human Gene Mutation Database (HGMD), 417, 420
Human genetic disease
bioinformatics perspective, 1012–1014, 1014
causes of death in USA, 1016
classification, 1015–1017
consequence of DNA variation, 1011–1019
Garrod’s view of disease, 1014–1015
ICD classification system, 1018
MeSH terms, 1017–1019
mutation mechanisms, 1013
projected global deaths, 1016
Human genome
advice for students, 1001
gateway to access, 959–964
introduction, 957–958
perspective, 999–1000
pitfalls, 1000–1001
statistics, 961
Human Genome Organization (HUGO)
Gene Nomenclature Committee (HGNC), 27, 317
Human Genome Project, 964 assemblies, 966–968
background, 964
broad genomic landscape, 968–970
centromeres, 974
comparative proteome analysis, 975–977
complexity of proteome, 978–979
CpG islands, 969–970
exons, 977
gene characteristics, 976
gene content, 974–979
genetic and physical distance compared, 970–971, 973
global statistics, 968
goals, 965
introns, 977
issues addressed, 967
long-range variation in GC content, 969
main conclusions, 958–959
noncoding genes, 976
noncoding RNAs, 975
paralogous genes, 979
protein-coding genes, 975
repeat content, 971–974
segmental duplications, 973–974
simple sequence repeats, 973
simple sequence repeats, 974
strategic issues, 966
transposon-derived repeats, 972–973
Human genome sequencing, 1051–1055
complex disorders, 1051–1052
conditions, genes, and variants, 1053–1054
disease-causing variants in otherwise healthy people, 1054–1055
incidental findings, 1052–1054
monogenic disorders, 1051
research versus clinical sequencing, 1052–1054
Human genome variation, 986
1000 Genomes Project, 995–998
haplotype phasing, 996
major conclusion of HepMap project, 994
sequencing individual genomes, 998–999
SNPs, haplotypes, and HapMap, 986–988, 989
viewing and analyzing, 988–993
viewing and analyzing, 990
Human Immunodeficiency Virus (HIV-1), 765–770
Human leukocyte antigen (HLA), 258
Human microbiome, 811–814
bacterial taxa, 813–814
fungi, 870
Human Microbiome Project (HMP), 708, 710, 721, 727, 811–814
Human mitochondrial genome, 985–986 haplogroups, 986
Human papillomavirus (HPV), 762
Human Protein Reference Database (HPRD), 542, 565
Human Proteome Organization (HUPO), 542
Human T-lymphotropic virus-I (HTLV-I), 762
Huntingdon disease, 624
Hydropathy index, 565
Identity, definition, 75
Illumina, 382–384
Indels, 408
Influenza, 761
Influenza Genome Sequencing Project (IGSP), 773
Influenza virus, 771–774
genes, 772
Informed consent, 390, 1052, 1054
Insertion mutations, 78
Insulin, 248–250
Integrated Microbial Genomes (IMG) website, 816
Integration of Disease Database, 1041
Integrative Genomics Viewer (IGV), 407, 988, 990
International Cancer Genome Consortium (ICGC), 1033
International Committee of Taxonomy of Viruses (ICTV), 756–758
International Gene Trap Consortium (IGTC), 658
International Human Genome Sequencing Consortium (IHGSC), 957
International Mouse Phenotyping Consortium (IMPC), 647
International Nucleotide Sequence Database Collaboration (INSDC), 21
Interpolated context model (ICM), 822–823
Interpolated Markov models (IMMs), 820–821
InterPro database, 226–227
Interspersed duplication, 409, 410
Introns, 336
Inversion of chromosomes, 409, 410
Ion Torrent, 387
IPrOClass database, 226–227
Isoelectric focusing, 544
Isoleucine (Ile), 76
Iterative alignment, 214–218
JIGSAW program, 339
Jukes–Cantor correction, 277
Kaposi’s sarcoma herpesvirus (KSHV), 762
Kappa caseins, 85, 86
Karlin–Altschul statistics, 142
Kazusa mammalian cDNA set (KIAA), 459
KDEL sequence, 570
Kimura two-parameter model, 277
Klenow fragment, 379
K-means clustering algorithm, 516–517
Knockout Mouse Project (KOMP), 647, 653
Kyoto Encyclopedia of Genes and Genomes (KEGG), 682–685
Lactase-phlorizin hydrolase (LPH), 468
Last universal cellular ancestor (LUCA), 766
Lateral gene transfer (LGT), 827–830
Laurasia, 705, 706, 777, 778
Leishmania, 894–895
Leucine (Leu), 76
Levinthal’s paradox, 598
Life on Earth, 705
geological history, 706, 707–708
evolution of, 78–79
Ligation sequencing, 385
Likelihood mapping, 290
Likelihood ratio test, 279
Lilium formosanum, 311
Limited Area Global Alignment of Nucleotides (LAGAN), 192–194
algorithm, 193
Linkage analysis, 1047
Linux, 11
commands, 43–44
Literature access, 58–59
Local alignment, 74
Local alignment algorithms, 96
dotplots, 104–105
FASTA and BLAST, 103–104
Smith and Waterman algorithm, 101–103
statistical significance, 108
websites and URLs, 110
Locus Reference Genomic (LRG), 37
Locus-Specific Mutation Databases, 1041–1044
Locusta migratoria, 311
Logarithms in base-2, 499
Log-expectation (LE) score, 219
Log-odds ratio, 92
Log–odds scoring matrix, 89–91
Long interspersed elements (LINEs), 326, 327
Long-branch attraction, 287–289
Los Alamos National Laboratory (LANL) HIV databases, 769, 770
Lymphoblastoid cell lines (LCLs), 468
Lysine (Lys), 76
MA plots, 498
Major histocompatibility complex (MHC), 50
Malaria, 895–898
Mammalian Gene Collection (MGC), 459
Mann–Whitney test, 493
MAQ program, 195
Markov chains, 181, 183, 292–293
Markov models, 181
see also hidden Markov models (HMMs)
MASCOT® software, 551
Mass spectrometry (MS), 547–551
mass-fragmentation (MALDI) mass spectrometry, 548
matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry, 548, 551, 675
protein complexes, 675–676
tandem affinity purification mass spectrometry (TAP), 676
triple quadrupole (QQQ), 548
Mass Spectrometry protein sequence Data Base (MSDB), 548
Massive Open Online Courses (MOOCs), 13
Matrix multiplication method, 87
Matrix-assisted laser desorption ionization (MALDI) mass spectrometry, 548
Matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry, 548, 551, 675
Maximal unique matches (MUM), 784
Maximum likelihood methods, 289–90, 291
Maximum parsimony methods, 287–289
Mean (average), 107, 491
Measles, 761
Measles virus, 774–775
proteins, 775
Medical Subject Heading (MeSH) terms, 1017–1019
Medical Subject Heading (MeSH) terms, 1017–1019
MEDLINE, 31, 58
MEGA software, 259
MegaBLAST, 191–192
Meiotic errors, 348
Mendelian genomic disorders, 1026
Merkel cell polyomavirus (MCV), 762
Messenger RNA (mRNA), 434, 450
full-length cDNA projects, 459
gene expression analysis in cDNA libraries, 455–458
gene expression studies, 450–452
low- and high-throughput studies, 452–455
measuring gene expression across the body, 459
RNA export, 451, 452
RNA processing, 451, 452
RNA surveillance, 451, 452
transcription, 450–451, 452
Metadata, 14
Metaphase, 312
Metazoans, 24, 916, 917–918
Anopheles gambiae, 921–922
butterflies, 922–923
Caenorhabditis elegans, 918–919
Ciona intestinalis, 925–926
Dictostelium discoideum, 916–917
Drosophila melanogaster, 919–921
fish genomes, 926–929
frogs, 929
honeybee, 923
insect genomes, 923–924
mammalian genomes, 933–934
mouse, 934–937
opposum, 931–933
platypus, 931–933
primate genomes, 937–940
rat, 934–937
reptiles, 929–931
sea urchin, 924–925
silkworms, 922–923
Methionine (Met), 76
MicroArray Quality Control (MAQC) project, 495, 511
Microarray RNA expression data analysis, 479–482
advice for students, 531
descriptive statistics, 511–519
functional annotation, 528–529
GEO dataset analysis using R, 504–511
GEO2R at NCBI, 482–495
Partek Genomics Suite, 495–504
perspective, 529–530
pitfalls, 530–531
Microarray RNA expression measurement, 460–466
biological confirmation, 465
data acquisition, 464–465
data analysis, 465
databases, 465
experimental design, 461, 494
further analysis, 466
probe preparation, 464
radioactive probes, 463
sample preparation, 461–464
Microbe, 700
Microorganism, 700
MicroRNA (miRNA), 445–447
Minimum Information About a Microarray Experiment (MIAME), 465
MiRBase, 446
Mitochondrial disease, 1030–1032
Mitochondrial Eve, 985
MitoSeek software, 421–422
Mitotic errors, 348
Mobile-element insertions, 409, 410
ModelTest program, 279
Molecular barcodes, 653–657
Molecular clock hypothesis, 250–254
amino acid substitution rates, 253
corrected number of amino acid changes per 100 residues, 252
nucleotide substitution rates, 254
Molecular evolution, 245–246
Molecular Genetics Evolutionary Analysis (MEGA) program, 282
bootstrap procedure, 295
maximum parsimony method, 287
Molecular Modeling Database (MMDB), 32, 608
Molecular phylogeny and evolution advice for students, 296
five stages of phylogenetic analysis, 270–295
goals, 246–247
historical background, 247–250
molecular clock hypothesis, 250–254
neutral theory of molecular evolution, 258–259
perspective, 295
phylogenetic trees, 259–266
pitfalls, 254–258
Tajima’s relative rate test, 255, 256
types of phylogenetic trees, 266–270
web resources, 297
Monogenic disorders, 1021–1024
examples, 1022
Monte Carlo Markov Chain (MCMC) command, 292–293
Mosquito-borne human diseases, 922
Moss, 916
Motifs, 552–559
characteristic of proteins, 557–559
definition, 553
Mouse Genome Informatics Database (MGD/GXD), 566, 646
Mouse knockouts, 650–653
MrBayes program, 282, 292, 294
MSD, 32
Multidimensional scaling (MDS), 518
compared with principal component analysis (PCA), 517
Multiparent Advanced Generation Inter-Cross (MAGIC), 913
Multiple Alignment using Fast Fourier Transform (MAFFT) program, 214–218
Multiple sequence alignment advice for students, 235
algorithm assessment, 207–208
benchmarking, 207–208
benchmarking studies, 221–222
consistency-based alignment, 218–220
databases, 222–227
definition, 206–207
evaluation, 223
exact approaches, 208
genomic regions, 227–234
introduction, 205–208
iterative alignment, 214–218
main approaches, 208–221
perspective, 234
pitfalls, 234
practical strategies, 207
progressive sequence alignment, 209–214
structure-based alignment, 220–221
typical uses, 207
Multiple Sequence Comparison by Log Expectation (MUSCLE) program, 215–219
profile–profile alignment, 219
MUMmer program, 346–347, 833–834
virus genomes, 783–785
Mumps, 761
Mus musculus, 25, 28, 646–647, 935–937
knockout mice, 650–653
MUSCLE program, 208, 215–219
Mutageneic Insertion and Chromosome Engineering Resource (MICER), 658
Mutation probability matrix for 1 PAM evolutionary distance, 82–84
Mutations, 78, 269
forbidden, 863–864
mechanisms, 1013
transition substitutions, 270
transition substitutions, 270
Mycobacterium tuberculosis, 101
National Center for Biotechnology Information (NCBI), 11, 21
access via gene resource, 38–42
BLAST-related algorithms, 170
command-line access, 42–49
command-line access to Entrez databases, 45
compared with HomoloGene, 42
compared with UniGene, 41–42
Conserved Domain Database (CDD), 177, 226
EDirect access, 45–49
Ensembl, 50–52
genome browsers, 49–52
Genomes, 710
GEO2 for RNA gene expression analysis, 482–495
Human Genome Project, 959
introduction, 31–32
Map Viewer, 52, 314
PDB access, 606–609
National Human Genome Research Institute (NHGRI), 398
Human Genome Project, 961–963
National Library of Medicine (NLM), 31, 58
Needleman and Wunsch global alignment algorithm, 96–100
Negative selection, 254–258
Neighbor-joining tree
p-distance correction, 276
Poisson correction, 276, 280
Neurospora crassa, 311, 873–874
Next-generation sequencing (NGS), 19, 359, 378–379, 1051
advice for students, 423
alignment to reference genome, 194–197
BLAST-related algorithms, 170–171
Burrows–Wheeler Transform (BWT) alignment, 196–197
DNA sequencing analysis, 387–421
DNA sequencing technologies, 379–387
hash table alignment, 194–196
perspective, 422–423
pitfalls, 423
short read alignment strategies, 195
specialized applications, 421–422
web resources, 424
Nicotiana tabacum, 311, 908
Nodes of phylogenetic trees, 259
Noncoding RNA, 436
long noncoding RNA (lncNA), 447–448
microRNA (miRNA), 445–447
other noncoding RNAs, 448
Rfam database, 436–438
ribosomal RNA (rRNA), 441–444
short interfering RNA (RNAi), 447
small nuclear RNA (snRNA), 445
small nucleolar RNA (snoRNA), 445
splicosomal RNAs, 445
transfer RNA (tRNA), 438–441
UCSC Genome and Table Browser, 448–449
Nonconserved elements, 346
Nonparametric bootstrapping, 293–295
Nonparametric tests, 488, 493
Normal distribution, 107
Novel sequence insertions, 409, 410
Novel species, 348
Nuclear magnetic resonance (NMR) spectroscopy, 600
Nucleomorphs, 902–904
Nucleotides
Limited Area Global Alignment of Nucleotides (LAGAN), 192–194
sequence databases searchable via BLAST, 127
step matrix, 270, 271
substitution rates, 254
transition substitutions, 270
transversion substitutions, 270
Null hypothesis, 106
Obligately intracellular and parasitic bacteria, 807–808
Obligately intracellular and symbiotic bacteria, 806–807
Odds ratio, 89
Odorant-binding protein (OBP), 572–573
One-based counting, 57, 58
Online Mendelian Inheritance in Man (OMIM), 32, 417, 419, 420, 1021, 1036–1039
numbering system, 1038
Oomycetes, 905
Open reading frames (ORFs), 336, 447
Saccharomyces cerevisiae, 640
Open Regulatory annotation (ORegAnno) database, 342–343
Operational taxonomic unit (OTU), 259–264
microarray data, 514
number of rooted and unrooted trees, 263, 265
tree-building by neighbor-joining (NJ), 285–287
Opposum, 931–933
Organism-specific BLAST sites, 168–170
Ortheus program, 231, 232, 233
Orthologous proteins or genes, 70–71, 319, 552, 568, 571, 573, 575, 642, 645, 650, 673, 680, 865, 866, 874
Oryza sativa, 311, 913–914
Oxytocin, 250
P value, 491–492
Pacific Biosciences DNA sequencing, 387
Pairwise sequence alignment, 69–79
advice for students, 112
gaps, 78
global and local alignment algorithms, 96–106
homology and evolution of life, 78–79
limits of detection, 94–96
perspective, 110–112
pitfalls, 112
scoring matrices, 79–96
statistical significance, 106–110
websites, 112
Pairwise Sequence Comparison (PASC) tool, 780–782
Pan troglodytes, 91
Pangaea, 705, 706, 778
Parabasala, 891
Paralogous proteins or genes, 70–71, 73, 320, 552, 642, 643, 656, 679, 864, 866, 876
Paramecium aurelia, 311
Paramecium caudatum, 311
Paramecium tetraurelia, 899–901
Parametric tests, 488
Parascaris equorum, 311
Parkinson disease, 624
Parsimony analysis, 287–289
maximum parsimony principle, 288
Partek Genomics Suite, 495–496
ANOVA, 501–504
data analysis, exploratory, 498–501
data import, 496
log2 transformed microarray data, 498
MA plots, 498
principal components analysis (PCA), 498–501
quality control, 496–497
sample histogram, 498
sample information, adding, 497
scatter plots, 498
t-test, 503–504
PartTree program, 215
Pattern-hit initiated BLAST (PHI-BLAST), 179–181
choosing a pattern, 180
PatternHunter, 188
nonconsecutive seeds, 189
P-distance correction, 276
Pearson correlation coefficient, 510
Pecan program, 191, 231, 232
Peptide bonds, 593
Percent identity, 108–109
Percent similarity, 75
Perfect match (PM) probesets, 488
Perl, 11, 627, 728, 1032
Phaeophyceae, 906
Phanerochaete chrysosporium, 875
PHAST package, 229
PHD program, 596
PhenCode project, 1044–1045
Phenotype and genotype, 637–638
Phenylalanine (Phe), 76
Phosphorylation, 564–565
PHRED scores, 391, 392
Phylogenetic analysis, 270–295
amino acid substitution, 272–281
DNA models, 272–281
multiple sequence alignment, 271–272
pitfalls, 295–296
sequence acquisition, 270–271
tree-building methods, 281–293
Phylogenetic Analysis Using Parsimony (PAUP) program, 282
Phylogenetic trees, 257 259
DNA trees, 268–270
efficient, 263–266
globins, 247, 248, 260, 276, 280, 291, 294
number of rooted and unrooted trees, 263, 266
protein-based trees, 268–270
RNA trees, 268–270
roots, 262–263
search strategies, 263–266
species trees versus gene/protein trees, 266–268
topologies and branch length, 259–262, 286
types, 266–270
Phylogenetic trees of species
Phylogenetic trees of species (continued)
lentiviruses, 767
mammals, 932
metazoans (animals), 917
origin of life, 704
plants, 908
primates, 938
reptiles, 930
Phylogenies, inconsistent, 890
Phylogeny, 246
Phylogeny Inference Package (PHYLIP) program, 282
Phylograms, 262
Pinus resinosa, 311
PipMaker software, 346–347
Plant genomes, 906–908
Arabidopsis thaliana, 910–913
Chlorophyta, 908–910
detecting ancient whole-genome duplications, 912
evolution of plants, 907
giant genomes, 915
land plants, 915
moss, 916
Oryza sativa, 913–914
phylogeny, 908
Populus trichocarpa, 914
tiny genomes, 915
Vitus vinifera, 915
Plasmodium falciparum, 895–898
Plastids, 902
Płatypus, 931–933
PLINK, 992
Ploidy see Polyploidy
Point-and-click web-based software, 10–11
Poisson correction, 276
Poisson distribution, 274
Poliomylititis, 761
Polyacrylamide gel electrophoresis (PAGE), 543–547
Polymerase chain reaction (PCR), 384, 660
Polymorphism Phenotyping-2 (PolyPhen) software, 417–420
Polypeptides, 540
amino acids, 591–594
phi and psi angles, 593
polyploidy, 347, 348, 863
plants, 911
Populus trichocarpa, 914
Position-specific iterated BLAST (PSI-BLAST), 171–177
assessing performance, 179
corruption problem, 177
hits from human beta globin, 174, 175
homologous matches, 176
matrix view, 173
pitfalls, 197–198
reverse position-specific BLAST (RPS-BLAST), 177, 178, 226
target frequencies, 173
Position-specific scoring matrix (PSSM), 172–173, 564
Positive selection, 254–258
Posterior probability, 219, 282, 289, 291–293, 295
Post-translational modifications to proteins, 560, 561
Precision, definition, 490
Primary structure of proteins, 591–594
Principal components analysis (PCA), 498–501
axes, 501
compared with multidimensional scaling (MDS), 517
plots, 500
Privacy, 390
Probability matrix, 82–84
Probability that base not sequenced, 405
ProbCons program, 215, 217
consistency-based alignment, 219–220
Profile Alignment (PRALINE) program, 215–218
structure-based alignment, 220–221
Profile searches, 181–186
Profile sum-of-pairs (PSP) scoring method, 219
Profile–profile alignment, 219
Progenote, 700
Programming, learning, 13–14
Programs, 77
Progressive sequence alignment, 209–214
Prokaryote, 700, 702
Proline (Pro), 76
PROSITE database, 226, 558–559
Protein alignment, 70
Protein analysis, 539–540, 543
direct protein sequencing, 543
mass spectrometry, 547–551
polyacrylamide gel electrophoresis (PAGE), 543–547
Protein complexes, 675–676
Protein Data Bank (PDB), 30, 32, 590, 602–617
accessing entries, 606–609
CATH database, 613–615
classification, 603
comparison of resources, 617
Dali Domain Dictionary, 617
number of searchable structures, 604
protein folding, 609–610
search results, 605
Structural Classification of Proteins (SCOP) database, 610–613
structure retrieval, 609
visualization tools, 606, 607
Protein databases, 29–31
content, 24–31
Protein homologs, 832, 833
Protein Information Resource (PIR), 30
Protein networks, 678–680
accuracy, 680
data, 680
experimental organism, 680–681
map categories, 681–682
pathway variation, 681
Protein Research Foundation (PRF), 30
Protein structure
advice for students, 625
disease, 622–625
high-resolution structures, 599
overview, 589–591
perspective, 625
pitfalls, 625
primary structure, 591–594
principles, 591–602
protein-folding, 598–600
quaternary structure, 592
secondary structure, 592, 594–598
sequence and structure, 590
structural genomics, 600–601
target selection, 602
tertiary structure, 592, 598–600
Protein Structure Initiative (PSI), 601–602
Protein structure prediction, 617–618
ab initio prediction, 621
accuracy, 620
approaches, 618
fold recognition, 619–620
homology modeling, 618–619
progress assessment, 621–622
websites, 620
Protein trees, 266–268
Protein-based trees, 268–270
Protein–protein interactions, 672–680
accuracy, 680
data, 680
databases, 676–678
experimental organism, 680–681
map categories, 681–682
pathway variation, 681
Proteins
amino acid relative abundance, 76
coding gene study resources, 340–342
databases, 540–542
Dayhoff’s superfamilies, 77
finding distantly related proteins, 171–181
functional genomics, 672
functions, 570–573, 574, 575
homology, 70–74
interactions, 672–678
intrinsically disordered, 622
localization, 570
methyl-binding domains, 557
modular nature, 552–559
motif characteristic, 557–559
multidomain proteins, 556–557
multiple copies of distinct domains, 555
orthologous, 70–71
overview, 552
pairwise alignment, 72–74
pairwise interactions, 678–682
paralogous, 70–71, 73
perspectives, 551–573
physical properties, 559–566
post-translational modifications, 560, 561
Protein Family database of profile
HMMs (Pfam), 223–224, 225, 226
sequence databases searchable via
BLAST, 126
sharing common domains, 556
Proteomics, 539–540
accuracy of prediction programs, 564–565
advice for students, 574–575
Association of Biomolecular Resource
Facilities (ABRF), 542–543
community standards, 542
definitions, 553
domains and motifs, 552–559
forward and reverse, 671
function of proteins, 570–573, 574, 575
gene ontology websites, 568
localization of proteins, 570
perspective, 573–574
phosphorylation, 564–565
physical properties, 559–566
pitfalls, 574
transmembrane regions, 565–566
Gene Ontology (GO) Consortium, 566–570
web resources, 576–578
Proteomics Identifications (PRIDE)
database, 549–550
Protostomes, 890
Genome database, 549–550
Proteomics, 539–540
accuracy of prediction programs, 564–565
advice for students, 574–575
Association of Biomolecular Resource
Facilities (ABRF), 542–543
community standards, 542
definitions, 553
domains and motifs, 552–559
forward and reverse, 671
function of proteins, 570–573, 574, 575
gene ontology websites, 568
localization of proteins, 570
perspective, 573–574
phosphorylation, 564–565
physical properties, 559–566
pitfalls, 574
transmembrane regions, 565–566
Gene Ontology (GO) Consortium, 566–570
web resources, 576–578
Proteomics Identifications (PRIDE)
database, 549–550
Protostomes, 890
*Giardia lamblia*, 891–892
*Trichomonas*, 890–891
ProtTest program, 280–281
Pseudogene Decoration Resource
(psiDR), 329
Pseudogenes, 326–331
PubMed, 31, 59
Pyrosequencing, 384–385, 386
Python, 11, 112, 114
Quaternary structure of proteins, 592
R programming language, 482–485
biomaRt, 317
seqinr, 818
RStudio, 504
Ramachandran plots, 594, 596
Random insertional metagenesis, 657–660
Rapid Annotations using Subsystems
(RAST) server, 825, 827
Rattus norvegicus, 25, 311, 935, 937
Receiver operating characteristic (ROC)
curves, 356, 421
Redundancy of coverage, 405
Reference sequence (RefSeq) project, 36–37, 340–342
RefSeqGene, 37
Regulatory factor databases, 342–345
Relatedness odds matrix, 88
Relative entropy, 109–110
Relative mutability of amino acids, 80–82
RepeatMasker software, 325, 326, 329
Reproducible research in bioinformatics, 14
Reptiles, 929–931
Rett syndrome (RTT), 555, 1023
Reverse genetics, 649
compared with forward genetics, 665–666
gene silencing, 662–664
β-globin gene, 650–653
insertional mutagenesis, 660–662
molecular barcodes, 653–657
mouse knockouts, 650–653
random insertional mutagenesis, 657–660
techniques, 658
yeast knockouts, 653–657
Reverse position-specific BLAST (RPSBLAST), 177, 178, 226
Reverse transcription polymerase chain
reaction (RT-PCR), 452–453
Rfam database, 436–438
Ribosomal Database Project (RDB), 444
Ribosomal RNA (rRNA), 435, 441–444
genome numbers in selected organisms, 443
major forms in bacteria and eukaryotes, 443
Risk allele frequencies, 1021
RNA
advice to students, 470
 disrupting, 662–664
functional genomics, 668–670
gene expression analysis method
overview, 481
gene-wide gene expression, 460–466
introduction, 433–436
messenger RNA (mRNA), 450–459
noncoding RNA, 436–449
perspective, 469–470
pitfalls, 470
structure, 434
transcription of DNA, 435
web resources, 470
RNA analysis interpretation, 466
relationship between DNA, mRNA,
and protein levels, 466–467
transcription, 467–468
RNA databases, 27–29
content, 24–31
RNA-inducing silencing complex (RISC), 447, 662
RNA interference (RNAi), 662
RNA-seq expression data analysis, 479–482, 519–523
advice for students, 531
CuffLinks sample protocol, 523–524
CuffLinks to assemble transcripts, 525
CuffLinks to determine differential
expression, 525–526
CummeRbund sample protocol, 526–527, 528
perspective, 529–530
pitfalls, 530–531
RNA-seq Genome Annotation
Assessment Project (RGASP), 527–528
TopHat reference genome, 524–525
TopHat sample protocol, 523–524
workflow chart, 522
RNA-seq expression measurement, 460–466
biological confirmation, 465
data acquisition, 464–465
data analysis, 465
databases, 465
experimental design, 461
RNA-seq expression measurement (continued)
  further analysis, 466
  probe preparation, 464
  sample preparation, 461–464
RNA-seq Genome Annotation
  Assessment Project (RGASP), 527–528
RNA trees, 268–270
Robertsonian translocation, 348
Robust multiaarray analysis (RMA), 488
normalization for accuracy and precision, 488–490
normalization for CEL files, 506–508
Root mean square deviation (RMSD), 488–490
Roots of phylogenetic trees, 262–263
Roots of phylogenetic trees, 262–263
Rosetta method, 621
Rubella, 761
Rotovirus, 761
Saccharomyces cerevisiae, 23, 311, 640–643, 849, 850
  chromosomes, 854–860
  features, 851–854
  gene duplication, 860–865
  gene nomenclature, 856
  genome duplication, 860–865
  genome sequencing, 851
  insertional mutagenesis, 660–662
  introns, 853
  knockout genes, 653–657
  multiple yeast genome browsers, 857
  NCBI Genome Workbench, 855
  NCBI Map Viewer, 854
  polyploidy, 863
  protein domains, 853
  proteins, 852
  secretion proteins, 642
Saccharomyces Genome Database (SGD), 566, 640–643, 856
Saccharomyces pombe, 875–876
SAM/BAM format files and SAMtools, 402–405
  CRAM file format, 406–408
  file anatomy, 403, 404
  finding files, 405–406
  mandatory fields, 402
  read depth calculation, 405
  viewing files, 406
Sanger sequencing, 19, 35, 379–382
  compared with next-generation sequencing (NGS) technologies, 382
dideoxy nucleotide sequencing, 380
denovo genome DNA, 381
quality scores, 381
Scatter plots, 498
Scoring matrices for proteins, 79
  accepted point mutations (PAM), 79, 80, 81
  block substitution matrix (BLOSUM), 91–94
  frequency of amino acids, 79, 81
  limits of detection, 94–96
  log-odds scoring matrix, 89–91
  mutation probability matrix for 1 PAM evolutionary distance, 82–84
  PAM250 and other PAM matrices, 84–88
  practical usefulness of PAM matrices, 91
  relatedness odds matrix, 88
  relative mutability of amino acids, 80–82
  Sea urchin, 924–925
Secondary structure of proteins, 225, 592, 766–767
Self-organizing maps (SOMs), 517, 518
Selected reaction monitoring (SRM), 548
Selection, 254–258
Selected reaction monitoring (SRM), 548
Selection, 254–258
Sexual reproduction, 309
Serine (Ser), 76
Sexual reproduction, 309
Short interfering RNA (RNAi), 447
Short read alignment strategies, 195
Silkworms, 922–923
Simplified Modular Architecture Research Tool (SMART) database, 224
Simple Modular Architecture Research Tool (SMART) database, 224
Simple sequence repeats in DNA, 331
Small subunits (SSUs), 707
Small nucleolar RNA (snoRNA), 445
Small nuclear RNA (snRNA), 445
Single nucleotide polymorphisms (SNPs), 354–355, 408
  microarrays, 356–358
  Single nucleotide variants (SNVs), 408
Slime molds, 916–917
Small nuclear RNA (snRNA), 445
Small nucleolar RNA (snoRNA), 445
Small subunits (SSUs), 707
Small-angle X-ray scattering, 600
Smallpox, 761
Smith and Waterman local alignment algorithm, 101–103, 122
  rapid heuristic versions, 103–104
SNARE protein, 641, 864
Sodium dodecyl sulfate (SDS), 544
Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE), 544–545, 547
Software for bioinformatics, 10–14
  bridging the two cultures, 12–13
  command-line, 10, 11–12
  programming, learning, 13–14
  reproducible research, 14
  web-based, 10–11
Solanum lycopersicum, 25
Solexa quality score, 391
Somatic mosaic disease, 318, 349, 378, 422, 1032–1033, 1035
Sorting Tolerant from Intolerant (SIFT) software, 417–420
Species trees, 266–268
Splitting of chromosomes, 348
SRA Toolkit, 392–393
Standard deviation, 107
Standard genetic code, 83
Statistical significance of pairwise alignments, 106
  global alignments, 106–108
  local alignments, 108
  percent identity, 108–109
  relative entropy, 109–110
Statistics
  average, 491
  binomial coefficient, 289
  bit score, 75, 91–92, 132, 134, 137, 143, 175, 438
  Chi-squared analysis, 256
  covariance models, 439, 441
  E value, 142–143
  false discovery rate (FDR), 495, 502
  Fisher’s exact test, 457, 458
  gamma distribution, 278
  Gaussian distribution, 107
  logarithms in base-2, 499
  log-expectation (LE) score, 219
  log-odds ratio, 92
  log–odds scoring matrix, 89–91
  Mann–Whitney test, 493
  mean (average), 107, 491
  multidimensional scaling (MDS), 518
  MDS compared with principal components analysis (PCA), 517
  nonparametric bootstrapping, 293–295
  nonparametric tests, 488, 493
  normal distribution, 107
  null hypothesis, 106
odds ratio, 89
p value, 491–492
parametric tests, 488
p-distance correction, 276
Pearson correlation coefficient, 510
Poisson correction, 276
Poisson distribution, 274
precision, definition, 490
probability matrix, 82–84
probability that base not sequenced, 405
receiver operating characteristic (ROC) curves, 356, 421
standard deviation, 107
variance, 107, 491
Wilcoxon test, 493
Z scores, 107–108
Step matrices, 270, 271
Stochastic context-free grammar (SCFG), 439, 441
Stramenopila, 904–906
Strongylocentrotus purpuratus, 311, 924
Structure, 591
Structural Biology Knowledgebase (SBKB), 602
Structural Classification of Proteins (SCOP) database, 610–613
release notes, 612
Structural Classification of Proteins-extended (SCOPe) database, 613
Structural genomics, 600–601
Structural polymorphisms, 1027
Structural variants, identifying, 409–410
Structure-based alignment, 220–221
Substitution mutations, 78
Sum-of-pairs scores (SPS), 223
Surface plasmon resonance, 673
Sus scrofa, 25, 28, 934
SWISS-PROT, 30, 31, 32
Synonymous Nonsynonymous Analysis Program (SNAP), 258
Synteny, 346
Systematics, 701–704
Tajima’s relative rate test, 255, 256
Tandem affinity purification mass spectrometry (TAP-MS), 676
Tandem duplication, 409, 410
Tandemly repeated sequences in DNA, 333–334
Target frequencies, 90
TATA box, 336
Taxonomy, 709–710
Taxonomy Browser, 52
Taxons (taxa), 259–262
TaxPlot, 830–833
TBLASTN tool, 70
T-COFFEE program, 215, 217
consistency-based alignment, 220
Expresso program, 220–221
structure-based alignment, 220–221
Telomeres, 312, 333, 348, 855, 860, 868, 894
Template-free modeling, 621
Tertiary structure of proteins, 592, 598–600
Thalassemias, 1023
The Arabidopsis Information Resource (TAIR), 643, 644, 911, 913
The Cancer Genome Atlas (TCGA), 1033–1034
Theileria annulata, 896
Threaded Blockset Aligner (TBA), 229
Threading, 619–620
Threonine (Thr), 76
Thuja occidentalis, 311
Thymine, 435
TMHMM program, 566
Tool makers, 15
Tool users, 15
TopHat sample protocol, 523–524
reference genome, 524–525
Toxoplasma gondii, 896–899
Transcription activator-like effector nucleases (TALENs), 664
Transcriptome, 5
Transcripts of uncertain coding potential (TUCP), 448
Transcript-specific variance, 491
Transfer RNA (tRNA), 434–435, 438–441
identification, 440
Transition substitutions, 270
Translocation, 348
Transposon tagging, 662
Transposons, 325–326
Transversion substitutions, 270
Tree bisection reconnection (TBR) approach, 266
Tree of life, 7, 246, 812, 828
global, 703
inferred tree, 246
molecular sequences, 705–709
nomenclature, 700
true tree, 246
Tree-building methods, 281–282
Bayesian inference method, 290–293, 294
branch lengths, 286
distance-based, 282–287
evaluating trees, 293–295
maximum likelihood, 289–90
maximum parsimony, 287–289
neighbor-joining (NJ), 285–287
unweighted pair group method of arithmetic averages (UPGMA), 283–285
TREE-PUZZLE program, 282
maximum likelihood method, 289–290, 291
TrEMBL, 32, 540
Trichomonas, 890–891
Triple quadrupole mass spectrometry (QQQ), 548
Triploid cells, 348
Trisomy 21 (Down syndrome), 313, 349, 480
Triticum aestivum, 25, 28, 908, 915
tRNAscan-SE, 439–441, 825
Truturus cristatus, 311
Trypanosoma, 892–894
Tryptophan (Trp), 76
T-test statistic, 491, 492–493
Tyrosine (Tyr), 76
UCNEbase database, 345
Ultraconserved elements, 345
UniBuild
human cluster sizes, 455
human cluster sizes, ten largest, 456
nonhuman cluster sizes, ten largest, 456
Unicellular pathogens, 892–895
Leishmania, 894–895
Trypanosoma, 892–894
UniGene, 28
compared with NCBI Gene, 41–42
phyla and organisms, 29
UniMes database, 540
UniParc database, 540
Uniparental disomy, 349
UniProt, 29, 31, 540
UniProtKB database, 540, 548, 550, 553
UniRef database, 540
University of California at Santa Cruz (UCSC), 11
Genes, 340–342
Genome Bioinformatics site, 321
Genome Browser, 50, 190, 229, 230, 314, 710
Human Genome Project, 961
Table Browser, 54–56, 229, 327, 328
Unix, 11, 12, 15, 42–47, 379
3′-Untranslated region (3′UTR), 451
5′-Untranslated region (5′UTR), 451
Unweighted pair group method of arithmetic averages (UPGMA), 211, 218, 282, 513
distance-based method, 283–285
Uracil, 435
VAAST software, 420–421
Vaccine-preventable bacterial disease, 808
Valine (Val), 76
Variance, 107, 491
Variant call format (VCF) files and
VCFtools, 388, 407, 408, 410–413
file columns, 411
file description, 412
finding and viewing files, 413
globin VCF, 996–998
Variant Effect Predictor (VEP) program, 419
Variant surface glycoprotein (VSG), 895
Variants, biological significance of, 417–421
Varicella, 761
VAST program, 611
Vector alignment search tool (VAST), 32
Velvet software, 734
Vertebrate Genome Annotation (VEGA) project, 37, 314
Vertebrate Multiz Alignment and Conservation, 229
*Vitis vinifera*, 915
Viral diseases, 761
Viridiplantae, 24
Virus genomes, 763
advice for students, 786
bioinformatics approaches, 765–766
International Committee of Taxonomy of Viruses (ICTV), 756–758
introduction, 755–758
metagenomics and diversity, 764–765
MUMmer comparisons, 783–785
Pairwise Sequence Comparison (PASC) tool, 780–782
perspectives, 785–786
pitfalls, 786
web resources, 786–787
Viruses, 24, 704, 1029
cancer-causing, 762
classification, 758–765
diversity and evolution, 762–764
ebola virus, 775–776
genome size, 758–760, 761
genomes, 711–712
giant viruses, 782–783
herpesvirus, 776–780
Human Immunodeficiency Virus (HIV-1), 765–770
influenza virus, 771–774
measles virus, 774–775
morphology, 758
nucleic acid composition, 758, 760
VISTA program, 346–347
Web-based software tools, 10–11
Wellcome Trust Sanger Institute (WTSI), 168, 170
Human Genome Project, 964
Whole-genome duplication, 347–349
graphical representation, 902
Whole-genome shotgun (WGS) strategy, 22, 395
Wilcoxon test, 493
WoLF PSORT program, 570, 571
WU-BLAST 2.0, 170
X chromosome dosage compensation, 349, 983
*Xenopus laevis*, 311, 927, 929
*Xenopus tropicalis*, 28
X-ray crystallography, 598, 599
X-ray free-electron lasers (XFELs), 599
Yeast knockouts, 653–657
Yeast two-hybrid system, 673–675
Z scores, 107–108
*Zea mays*, 25, 28, 910
Zebrafish Information Network (ZFIN), 645
Zero-based counting, 57, 58
Zinc fingers, 664