

# Index

## • *Symbols and Numerics* • • *A* •

- : (colon), 292
- > (the definition line), FASTA program, 48
- 1000 bp (kb), 23
- . (period), 292
- <PRE> parasite character, 52
- // (slash marks, two), 77
- \* (star), 292
- 3-D protein structure
  - additional structural features, predicting, 334–336
  - computer, folding in, 351
  - described, 329–330
  - guessing, 340–342
  - homology modeling, 351
  - interactions, predicting, 352
  - interactive exploration, 344–349
  - interplay between multiple alignments and structural analysis, 343–344
  - local segments, 330
  - in movement, looking at, 352
  - patterns, identifiable, 178
  - PDB structures, 350–352
  - from primary to, 336–337
  - retrieving and displaying from PDB site, 337–340
  - sample, illustrated, 16
  - secondary structure, predicting, 330–334
  - sequence and structure, interactive analysis, 349–350
  - sequence/PDB structure relationship, interactive exploration, 344–349
  - sequences, analyzing, 14–16
  - similar shapes, finding proteins with, 350
- 3'-terminus, 18
- 5'-terminus, 18
- A or G, IUPAC code, 19
- accession number
  - GenBank entry, 74, 81
  - Swiss-Prot, 111–112
- AceDB suite, 154
- Acembly engine, 154
- adenine (A)
  - IUPAC code, 19
  - RNA nucleotide sequence letters, 21
- Aeropyrum pernix*, 96–97
- alanine, 11
- Align pairwise alignment program, 263
- alignments, 238, 263
- alignments, BLAST. *See also* PSI-BLAST
  - alignments, 213, 215–216
  - biological questions, asking, 218–219
  - complementary property, 20
  - described, 57–58, 199, 413
  - DNA sequences, overview, 216–218
  - EMBNet blastp, 207–209
  - graphic display, 211–212
  - hit list, 212–213
  - homologues, 214
  - hybridizing primers, 138
  - NCBI blastp, 202–207
  - output, 209–210, 224–225
  - parameters, 216, 219–220, 223–224
  - protein domains, discovering and using, 230–231
  - protein sequences, handling, 201–202
  - results, 60–61
  - sequence masking, 220–223
  - servers, alternative, 231–233
  - starting, 58–60
- alignments, local
  - benefits of using, 255
  - described, 238, 254

- alignments, local (*continued*)
    - Align output, interpreting, 258–261
    - Align to find ten best, 256–258
    - methods, choosing, 255–256
  - alignments, multiple sequence
    - ClustalW, 282–287, 300
    - common ancestor, 266
    - common ancestor, sequences without, 297–299
    - described, 265–266
    - DNA or protein sequences, 272
    - evolutionary constraints, revealing, 294–297
    - guidelines for selecting, 271
    - Internet resources, 299–302
    - interpreting, difficulties of, 291–292
    - method, choosing, 281
    - motif-finding methods, addresses listed, 301–302
    - MUSCLE, crunching large datasets with, 291
    - naming correctly, 275
    - number, choosing right, 272–273
    - online BLAST servers, 275–281
    - phylogenetic analysis, 380–382
    - protein alignment, recognizing good parts, 292–293
    - research, helping, 267–270
    - selecting correct sequence, 270
    - similarity versus new information, 273–274
    - Tcoffee, 287–291
    - when not to use, 267
  - alignments, multiple sequence, editing and publishing. *See also* formatting
  - beautifying tools, 325
  - Boxshade utility, 319–321
  - described, 303–304
  - editing packages, 323–324
  - Logos, generating high-impact pictures with, 322–323
  - tools for extracting information, 324
  - Alion pairwise alignment program, 263
  - ALN multiple sequence alignment format, 307, 309–310
  - Amas tool, 324
  - ambiguity, 13
  - amino acid
    - ambiguity, 13
    - lost in reformatting, 312
    - patterns of conservation, 293
    - protein sequences, analyzing, 10–12
  - ancestor, multiple sequence alignment with common, 266
  - ancestor, sequences without common conserved patterns, searching, 299
  - described, 297–298
  - Gibbs sampler, 298
  - annotation, lost in reformatting, 312
  - applications. *See also individual programs listed by name*
    - described, 412
    - listed, 413
  - archaea, 70
  - arginine, 11
  - ArrayExpress resource locator, 414
  - asparagine, 11, 13
  - aspartic acid, 11, 13
  - Assembler from The Institute for Genome Research (TIGR), 154
  - assembling fragments for single DNA sequence
    - CAP3 documentation, 155–157
    - machines, limitations of, 153
    - public software, managing large projects with, 154–155
  - author's names, searching PubMed by, 32–35
- **B** ●
- bacterial genomes, 92–94
  - Bairoch, Amos (ExpASy server creator), 42, 175
  - base pairs (bp), 23
  - Basic Local Alignment Search Tool. *See* BLAST
  - BCM ClustalW server, 300
  - Belvu package, 323
  - beta-strands, 330
  - Bielefeld resource locator, 414
  - binding sites, 21
  - biochemistry, computer
    - ExpASy server, 160
    - protease digestions, 166
    - Swiss EMBnet, 160
  - biochemistry sites, 125
  - Bioedit package, 324

- bioinformatics, defined, 9–10  
 Bioinformatics Web site, 415  
 Bio-informer resource locator, 414  
 biological ancestor, multiple sequence alignment with common, 266  
 biological ancestor, sequences without common  
     conserved patterns, searching, 299  
     described, 297–298  
     Gibbs sampler, 298  
 biological dot plot analysis, 249–254  
 BioNJ phylogenetic tree interface, 401  
 Bioprospector motif-finding method, 302  
 BLAST (Basic Local Alignment Search Tool). *See also* PSI-BLAST  
     alignments, 213, 215–216  
     biological questions, asking, 218–219  
     complementary property, 20  
     described, 57–58, 199, 413  
     DNA sequences, overview, 216–218  
     EMBnet blastp, 207–209  
     graphic display, 211–212  
     hit list, 212–213  
     homologues, 214  
     hybridizing primers, 138  
     NCBI blastp, 202–207  
     output, 209–210, 224–225  
     parameters, 216, 219–220, 223–224  
     protein domains, discovering and using, 230–231  
     protein sequences, handling, 201–202  
     results, 60–61  
     sequence masking, 220–223  
     servers, alternative, 231–233  
     starting, 58–60  
 Blast2seqs pairwise alignment program, 263  
 blastn, 217  
 blastx, 217  
 BLAT database search engine, 232, 233  
 Blockgap tool, 324  
 BLOCK-Maker motif-finding method, 302  
 BLOCKs domain collection, 183  
 Blocks tool, 324  
 Boehringer site, exploring biochemical pathways through, 125–126  
 Bork, Dr. Peer (European Molecular Biology Laboratory senior scientist), 262  
 bottom cursor, Dotlet, 247  
 Boxshade tool, 319–321, 325  
 bp (base pairs), 23  
 branch, phylogenetic tree, 399  
 breast-cancer-susceptibility gene of type 1 (Brca1), 237
- C ●
- C (cytosine)  
     composition, analyzing single DNA sequence, 138–139  
     IUPAC code, 19  
     RNA nucleotide sequence letters, 21  
 CAP3 documentation, 155–157  
 cap3 sequence assembly tool, 267  
 CATH (Class, Architecture, Topology, Homologous superfamily), 127  
 CAZy database, 128  
 CBS (Center for Biological Sequence Analysis), 171–172  
 CBS protein sequence analysis, 195  
 CD (Conserved Domain) server of NCBI  
     described, 187–190  
     protein sequence analysis, 195  
 cDNA, 147  
 CENSOR software tool, 145  
 Center for Biological Sequence Analysis (CBS), 171–172  
 Center for Information Technology, NIH, 158  
 chain peptide, 120  
 ChemIDplus, 125  
 chemistry sites, 125  
 chips, 142  
 chromosomes, 72  
 CINEMA package, 323  
 Class, Architecture, Topology, Homologous superfamily (CATH), 127  
 ClustalW multiple sequence alignment  
     computing tree, 384–387  
     described, 282, 284–286, 413  
     history, 282–283  
     output, 64–65  
     parameters, tuning, 286–287  
     phylogenetic tree, building, 371  
     servers listed, 300

- ClustalW multiple sequence (*continued*)
  - starting, 62–63
  - Tcoffee versus, 291
- ClustalX color scheme, 315
- Clusters of Orthologous Groups (COG)
  - database, 128, 183
- coding regions, DNA
  - described, 23–24
  - position, beginning with different, 25–26
  - protein sequence, translating into, 24–25
  - standard genetic code, table of, 25–26
  - topics covered by chapters, 26
- codon, 141
- Coffee Corner resource locator, 414
- COG (Clusters of Orthologous Groups)
  - database, 128, 183
- coiled-coil regions
  - computer, identifying by, 166
  - primary structure analysis, 174
- collection, protein domains, 182–183
- colon (:), 292
- comments section
  - EGFR, 114–116
  - GenBank entry, 75
- common ancestor, multiple sequence
  - alignment, 266
- common ancestor, sequences without
  - conserved patterns, searching, 299
  - described, 297–298
  - Gibbs sampler, 298
- comparative genomics, 88
- comparisons, pairwise. *See also* dot plot
  - described, 235
  - local alignments over Internet, 254–261
  - method, choosing, 237–239
  - proteins and DNA, aligning, 262
  - sequences, choosing, 236–237
  - servers, listed, 262–263
- complementary property, BLAST, 20
- composition, analyzing single DNA
  - sequence
    - EMBOSS modules, 142
    - G+C content, 138–139
    - genome-specific repeats, identifying, 145
    - internal repeats, finding, 142–144
    - long words, counting, 140–141
    - words, counting, 139–140
- Comprehensive Enzyme Information
  - System BRENDA, 126
- computer
  - biochemistry using, 160–166
  - protein 3-D structures, folding in, 351
  - sequence analysis, roots of, 12
- computer, finding known protein domain
  - CD server of NCBI, 187–190
  - collection, choosing right, 182–183
  - described, 180–181
  - Internet tools, 194–195
  - InterProScan results, interpreting, 185–187
  - InterProScan server, 183–185
  - Motif Scan, 190–193
  - new domains, finding, 194
- computer, primary structure analysis
  - coiled-coil regions, 174
  - properties revealed by, 166
  - “sliding windows” technique, 167–168
  - transmembrane segments, 168–174
- computer, ProtParam program
  - described, 161–163
  - extinction coefficient, 165
  - half-life, 165
  - instability, 165
  - molecular weight, 164–165
- conferences Web site, 415
- confidence line (Conf), 332
- conservation, patterns of, 293
- Conserved Domain (CD) server of NCBI
  - described, 187–190
  - protein sequence analysis, 195
- conserved patterns, searching, 299
- contig, 155
- CORE tool, 287, 290
- covariance phenomenon, 361
- CpG rich region finder, 142
- cross-references, PIR (Protein Information Resource), 116
- C-terminus, 14
- cysteine, 11
- cytosine (C)
  - composition, analyzing single DNA
    - sequence, 138–139
  - IUPAC code, 19
  - RNA nucleotide sequence letters, 21



DALI software, 413  
 Database of Interacting Proteins (DIP), 117

- databases, addresses listed, 412. *See also individual databases listed by name*
- date range, limiting dUTPase search to, 40
- DDBJ ClustalW server, 300
- DEFINITION, GenBank entry, 74, 81
- definition line (>), FASTA program, 48
- deoxyribonucleic acid (DNA)
- described, 17
  - and proteins, aligning, 262
  - regulatory elements, 269
- deoxyribonucleic acid (DNA) coding regions
- described, 23–24
  - position, beginning with different, 25–26
  - protein sequence, translating into, 24–25
  - standard genetic code, table of, 25–26
  - topics covered by chapters, 26
- deoxyribonucleic acid (DNA) sequence analysis
- described, 17, 216–218
  - double helix, 18–20
  - IUPAC code, 19
  - nucleotide, 23
  - palindromes, 20–21
  - reading, 17–18
- deoxyribonucleic acid (DNA) sequences, retrieving
- introns and exons, 51
  - from protein sequences, 52–53
  - relevant to my protein, 53–56
- Dialign server
- described, 301
  - multiple sequence alignment, 301
- digestions, protease, 166
- DIP (Database of Interacting Proteins), 117
- disease genes, finding with coding SNPs using BioMart data-mining system, 102–104
- distance matrix, 380
- DNA (deoxyribonucleic acid)
- described, 17
  - and proteins, aligning, 262
  - regulatory elements, 269
- DNA (deoxyribonucleic acid) coding regions
- described, 23–24
  - position, beginning with different, 25–26
  - protein sequence, translating into, 24–25
  - standard genetic code, table of, 25–26
  - topics covered by chapters, 26
- DNA (deoxyribonucleic acid) sequence analysis
- described, 17, 216–218
  - double helix, 18–20
  - IUPAC code, 19
  - nucleotide, 23
  - palindromes, 20–21
  - reading, 17–18
- DNA (deoxyribonucleic acid) sequences, retrieving
- introns and exons, 51
  - from protein sequences, 52–53
  - relevant to my protein, 53–56
- Dnadot program, 240
- DNASTAR Lasergene, 154
- documentation, CAP3, 155–157
- DoE (U.S. Department of Energy) whole-genome database, 96–97
- domain
- identification, 269
  - Swiss-Prot, 120–121
- domain, protein
- CD server of NCBI, 187–190
  - collection, choosing right, 182–183
  - described, 180–181
  - Internet tools, 194–195
  - InterProScan results, interpreting, 185–187
  - InterProScan server, 183–185
  - Motif Scan, 190–193
  - new domains, finding, 194
- dot plot. *See also* Dotlet
- biological analysis, 249–254
  - described, 143–144, 238, 239–240
  - inverted repeats, identifying, 144
  - low-complexity regions in proteins, finding, 253
  - programs, different types of, 240
  - tandem repeats, identifying, 250–252
- Dotlet
- downloading, 241–242
  - entering sequence in, 242–244
  - fine-tuning, 245–248
  - nucleic acids, analyzing with, 253–254
  - results, interpreting, 248–249
- Dotter program, 240
- Dottup program, 240
- double helix, 18–20

## • E •

- E. coli* (*Escherichia coli*)  
 DNA sequence, retrieving, 53–57  
 GenBank entry, 73–77  
 researching, 42–45
- EBI ClustalW server, 300
- EBI (European Bioinformatics Institute), 105
- editing multiple sequence alignments. *See also* formatting  
 beautifying tools, 325  
 Boxshade utility, 319–321  
 described, 303–304  
 editing packages, 323–324  
 Logos, generating high-impact pictures with, 322–323  
 tools for extracting information, 324
- editing packages, multiple sequence alignment, 323–324
- EGF receptor entry, deciphering, 110–111
- EGFR (epidermal growth factor receptor)  
 Comments section, 114–116  
 Cross-References section, 116–118  
 deciphering entry, 110–111  
 Features section, 119–123  
 general information about entry, 111–112  
 Keywords field, 118–119  
 name and origin of protein, 112–114  
 References section, 114  
 sequence section, 123
- Eisenberg Scale, 171
- e-mail address, 332, 389–390
- EMBnet  
 blastp, 207–209  
 ClustalW server, 300
- EMBOSS server (Pasteur Institute)  
 G+C content, establishing, 138–139  
 modules, 138–139  
 word frequency, computing, 140–141
- eMotif motif-finding method, 301
- Encyclopedia of *E. coli* Genes and Metabolism, 126
- energy dot plot, mfold, 359–360
- Ensembl project  
 described, 98, 412  
 disease genes, finding with coding SNPs using BioMart data-mining system, 102–104
- Human DUT ID card, getting complete, 101–102  
 Swiss-Prot cross-reference, 118  
 Web site, starting at, 98–101
- Entrez/Gene resource, NCBI server  
 bacterial genomes, 92–94  
 described, 413  
 LOCUS, 86–88  
 viral genomes, 89–92
- Enzymes database, 412
- epidermal growth factor receptor (EGFR)  
 Comments section, 114–116  
 Cross-References section, 116–118  
 deciphering entry, 110–111  
 Features section, 119–123  
 general information about entry, 111–112  
 Keywords field, 118–119  
 name and origin of protein, 112–114  
 References section, 114  
 sequence section, 123
- Escherichia coli* (*E. coli*)  
 DNA sequence, retrieving, 53–57  
 GenBank entry, 73–77  
 researching, 42–45
- ESPrnt tool, 325
- ESTs (expression sequence tags), 154
- eukaryotes, 70, 72–73
- eukaryotic genomes, gene parsing for, 151
- eukaryotic mRNA entry, GenBank  
 calling, 78–79  
 FEATURES section, 81–84  
 fetching, 80  
 gene sequence, 79  
 KEYWORD line, 79  
 keywords, 81  
 related, working with, 84–85  
 retrieving without accession numbers, 85–86
- European Bioinformatics Institute (EBI), 105
- E-value (expectation value)  
 cutoff point, 225–226  
 described, 200  
 hit list, 212  
 Lalign output, 259  
 Web-based servers, 408
- evolutionary constraints, multiple  
 sequence alignment, 294–297
- evolutionary similarity, 268
- exceptional amino acids, code for, 13

- exons
    - described, 72
    - DNA sequences, retrieving, 51
    - GenBank entry, 83
    - internal, 149–151
    - vertebrate, 150
  - ExPASy (Expert Protein Analysis System)
    - server
      - described, 42–43
      - entry parts, 43–45
      - FASTA format, 48, 51
      - parasite characters, warning about, 52
      - protein sequence analysis, 195
      - related protein sequences, 48–50
      - resource locator, 414
      - restricted searches, 45–47
      - selecting sequences on, 276–279
      - similarity searches, 160
  - expectation value (E-value)
    - cutoff point, 225–226
    - described, 200
    - hit list, 212
    - Lalign output, 259
    - Web-based servers, 408
  - experiments, 10
  - Expert Protein Analysis System. *See* ExPASy server
  - experts, finding through PubMed, 36, 38
  - expression, 70
  - expression sequence tags (ESTs), 154
  - EXPRESSO tool, 287, 290
  - extended strands, 330
  - extrapolation, 269
- F •**
- FASTA
    - database search engine, 232
    - format, 48, 51
    - multiple sequence alignment format, 306, 308
  - features section
    - EGFR, 119–123
    - GenBank entry, 55
    - GenBank table, 75, 76–77, 81–84
  - fields, searching PubMed by, 35–38
  - 5'-terminus, 18
  - flat-file GenBank entry, 73
  - fntseq sequence text converter, 310, 311
  - folds, UniProtKB/Swiss-Prot database, 109–110
  - formatting
    - converting, 309–311
    - correct, working with, 307–309
    - losing data, 312
    - publications, 307
    - variety of, 305–307
  - formatting, Jalview
    - described, 313, 413
    - features, 318
    - obtaining, 323
    - phylogenetic tree, 401
    - saving alignment, 318–319
    - starting, 314–315
  - fragments, assembling for single DNA
    - sequence
      - CAP3 documentation, 155–157
      - machines, limitations of, 153
      - public software, managing large projects with, 154–155
  - From field, Swiss-Prot, 113
  - functional signatures, 64
  - functional similarity, 268
  - functions, UniProtKB/Swiss-Prot database, 109–110
- G •**
- G (guanine)
    - IUPAC code, 19
    - RNA nucleotide sequence letters, 21
  - G (guanosine)
    - composition, analyzing single DNA
      - sequence, 138–139
    - IUPAC code, 19
  - gap
    - described, 13
    - penalties/cost, 223
    - type, lost in reformatting, 312
  - gap-extension penalty
    - ClustalW parameter tuning, 286
    - described, 258
  - gap-opening penalty
    - ClustalW parameter tuning, 286
    - described, 257
  - Garavelli, John (RESID database maintainer), 124
  - Gascuel, Olivier (mathematician), 397

- GenBank eukaryotic mRNA entry
    - calling, 78–79
    - FEATURES section, 81–84
    - fetching, 80
    - gene sequence, 79
    - KEYWORD line, 79
    - keywords, 81
    - related, working with, 84–85
    - retrieving without accession numbers, 85–86
  - GenBank prokaryotic entry
    - FEATURES table, 76–77
    - header, reading, 74–75
    - sample gene, fetching, 73–74
    - Sequence section, 77
  - GenBank/DDBJ/EMBL database, 412
  - gene density, 71
  - gene name, Swiss-Prot, 113
  - gene order formula, 82
  - gene tree, 377–379
  - Genebee server, 400
  - gene-centric database, 69–70
  - GeneMark, 148–149
  - genes
    - eukaryotes, 72–73
    - parsing for eukaryotic genomes, 151
    - prokaryotes, 70–72
    - sequence, GenBank eukaryotic mRNA entry, 79
  - Genetic Information Research Institute, 145
  - Genetics For Dummies* (Robinson), 70
  - Genomatix, 139–140
  - GenomeNet ClustalW server, 300
  - genomes
    - eukaryotes, 72–73
    - first sequence determined, 26–27
    - genomics, 27–28
    - prokaryotes, 70–72
    - repeats, identifying specific, 145
    - topics covered by chapter, 28
  - GenomeScan, 151–153
  - genomics, 27–28
  - GenScan software, 413
  - Gibbs Sampler
    - common ancestor, sequences without, 298
    - motif-finding method, 301
  - Gibson, Toby (ClustalX color scheme developer), 315
  - global alignments, 238, 254, 261–262
  - glutamic acid, 11, 13
  - glutamine, 11, 13
  - Glycan Structure Database, 125
  - glycine, 11
  - GlycoSuiteDb, 117
  - graph-align pairwise alignment analysis, 263
  - Graphic display, CD server, 189
  - greater than sign (>), FASTA program, 48
  - guanine (G)
    - IUPAC code, 19
    - RNA nucleotide sequence letters, 21
  - guanosine (G)
    - composition, analyzing single DNA sequence, 138–139
    - IUPAC code, 19
  - Guindon, Stéphane (mathematician), 397
- H ●
- header, GenBank prokaryotic entry, 74–75
  - Heiman, Max (Webcutter tool developer), 134
  - Helix-Turn-Helix (HTH) domain, 298
  - hellices, 330
  - Hemophilus influenzae* genome, 26
  - Hidden Markov Models, 330–331
  - Higgins, Des (ClustalW software developer), 282
  - histidine, 11
  - hit list
    - BLAST, 212–213
    - CD server, 189
  - Hits protein sequence analysis, 195
  - HIV-1 (type-1 human immunodeficiency virus), 89–92
  - Hogeweg, Paula (ClustalW software developer), 282
  - homologues
    - BLAST, 214
    - described, 200
    - protein 3-D structures modeling, 351
    - search engines, 233
  - HTH (Helix-Turn-Helix) domain, 298
  - HUGO (Human Genome Organization Gene Nomenclature Committee), 117
  - Human Brain Database, 128
  - Human DUT ID card, getting complete, 101–102
  - human genome, 97–98. *See also* Ensembl project

hybridizing primers, 138  
 hydrophilic, 15  
 hydrophilic stretches, 166  
 hydrophobic, 15  
 hydrophobic regions, 166

## • I •

identity, percentage of, 213  
 IMGT (International Immunogenetics database), 128  
 Improbizer motif-finding method, 302  
*in vitro* experiments, 10  
*in vivo* experiments, 10  
 The Institute for Genome Research (TIGR) Assembler, 154  
   bacterial genomes, 94–95  
 internal exons, finding in vertebrate genomic sequences, 149–151  
 internal repeats  
   composition, analyzing single DNA sequence, 142–144  
   pairwise comparisons, 237  
 International Immunogenetics database (IMGT), 128  
 International Union of Biochemistry and Molecular Biology (IUBMB), 126  
 International Union of Pure and Applied Chemistry (IUPAC) code  
   RNA sequences, analyzing, 21–22  
   tables listing, 11, 19  
 InterPro protein sequence analysis, 195, 412  
 InterProScan server, 183–185  
 introns  
   DNA sequences, retrieving, 51  
   gene density, 71  
 inverted repeats  
   described, 142  
   dot plot, 144  
 isoleucine, 11, 13  
 IUBMB (International Union of Biochemistry and Molecular Biology), 126  
 IUPAC (International Union of Pure and Applied Chemistry) code  
   RNA sequences, analyzing, 21–22  
   tables listing, 11, 19

## • J •

Jalview  
   described, 313, 413  
   features, 318  
   obtaining, 323  
   phylogenetic tree, 401  
   saving alignment, 318–319  
   starting, 314–315  
 Java applet, Dotlet  
   downloading, 241–242  
   entering sequence in, 242–244  
   fine-tuning, 245–248  
   nucleic acids, analyzing with, 253–254  
   results, interpreting, 248–249  
 Java applet, Jalview  
   described, 313, 413  
   features, 318  
   obtaining, 323  
   phylogenetic tree, 401  
   saving alignment, 318–319  
   starting, 314–315  
 Journal of Virology, 34

## • K •

Kalign  
   multiple sequence alignment, 301  
   server listed, 301  
 Kalignview package, 323  
 kb (1000 bp), 23  
 KEGG (Kyoto Encyclopedia of Genes and Genomes), 126, 412  
 keywords  
   EGFR, 118–119  
   GenBank entry, 74, 79, 81  
 Kimura, Motoo (neutralism, elaboration of), 375  
 Koonin, Eugene, 379  
 Kyte & Doolittle Scale, 171

## • L •

Lalign  
   interpretation difficulties, 291  
   local alignments, 256–258  
   output, interpreting, 258–261  
   pairwise alignment, 263

- lalnview pairwise alignment analysis, 263
  - Lama tool, 324
  - Lasergene (DNASTAR), 154
  - lateral transfer, 377
  - leucine, 11, 13
  - licensing issues, 410
  - Lipid Bank, 125
  - Lipman, D.J. (FASTA program creator), 48
  - local alignments
    - benefits of using, 255
    - described, 238, 254
    - Lalign output, interpreting, 258–261
    - Lalign to find ten best, 256–258
    - methods, choosing, 255–256
  - locus
    - Entrez/Gene resource, NCBI server, 86–88
    - GenBank entry, 74, 81
    - name, 55
  - Logos tool
    - described, 413
    - editing package, 324
    - high-impact pictures, generating, 322–323
  - long words, counting in single DNA
    - sequence, 140–141
  - loops, 23
  - low-complexity
    - regions in proteins, finding, 253
    - segments, 215
  - lysine, 11
- M •**
- macromolecules, 11
  - MAFFT
    - multiple sequence alignment, 301
    - server listed, 301
  - match details, Motif Scan, 192–193
  - match map, Motif Scan, 191–192
  - mature transcript (mRNA)
    - described, 53n
    - entry fields, 83
    - eukaryotes, 72–73
    - gene order formula, 82
  - mature transcript (mRNA), GenBank
    - eukaryotic
      - calling, 78–79
    - FEATURES section, 81–84
    - fetching, 80
    - gene sequence, 79
  - KEYWORD line, 79
  - keywords, 81
    - related, working with, 84–85
    - retrieving without accession numbers, 85–86
  - Mb (mega-bp), 23
  - McKusick, Victor (Online Mendelian Inheritance in Man database owner), 106
  - MCOFFEE tool, 287
  - Medline record, internal structure of, 37
  - MEME motif-finding method, 302
  - MEROPS database, 128
  - methionine, 11
  - Mfold software
    - described, 355–356
    - forcing interaction, 361–362
    - interpreting results, 359–361
    - obtaining, 413
    - sample, 356–359
  - miRNAs
    - described, 367–368
    - resource locator, 414
  - mismatches, 365
  - ModBase database, 116
  - modification, post-translational
    - described, 174–175
    - ORFs, 108
    - other tools, 180
    - output, understanding, 177–179
    - patterns, looking for, 175–177
    - short patterns, 179
    - species information, 179–180
    - weak patterns, eliminating, 180
    - weak signals, 180
  - molecular docking, 352
  - Motif Scan, 190–193
  - mRNA (mature transcript)
    - described, 53
    - entry fields, 83
    - eukaryotes, 72–73
    - gene order formula, 82
  - mRNA (mature transcript) entry, GenBank
    - eukaryotic
      - calling, 78–79
    - FEATURES section, 80–84
    - fetching, 80
    - gene sequence, 79
    - KEYWORD line, 79
    - keywords, 81

- related, working with, 84–85
  - retrieving without accession numbers, 85–86
  - multiple sequence alignments
    - ClustalW, 282–287, 300
    - common ancestor, 266
    - common ancestor, sequences without, 297–299
    - described, 265–266
    - DNA or protein sequences, 272
    - evolutionary constraints, revealing, 294–297
    - guidelines for selecting, 271
    - Internet resources, 299–302
    - interpreting, difficulties of, 291–292
    - method, choosing, 281
    - motif-finding methods, addresses listed, 301–302
    - MSF format, 306, 308
    - MUSCLE, crunching large datasets with, 291
    - naming correctly, 275
    - number, choosing right, 272–273
    - online BLAST servers, 275–281
    - phylogenetic analysis, 380–382
    - protein alignment, recognizing good parts, 292–293
    - research, helping, 267–270
    - selecting correct sequence, 270
    - similarity versus new information, 273–274
    - Tcoffee, 287–291
    - when not to use, 267
  - multiple sequence alignments, editing and publishing. *See also* formatting
    - beautifying tools, 325
    - Boxshade utility, 319–321
    - described, 303–304
    - editing packages, 323–324
    - Logos, generating high-impact pictures with, 322–323
    - tools for extracting information, 324
  - Munich Bioinformatics Center, 158
  - MUSCLE
    - crunching large datasets with, 291
    - multiple sequence alignment, 301, 413
    - server listed, 301
  - mutual ancestor, multiple sequence alignment, 266
  - mutual ancestor, sequences without conserved patterns, searching, 299
    - described, 297–298
    - Gibbs sampler, 298
  - Mview tool, 325
- *N* ●
- N (nucleotide)
    - DNA sequences, analyzing, 23
    - IUPAC code, 19
    - lost in reformatting, 312
    - RNA nucleotide sequence letters, 21
  - name
    - alignments, 213
    - author's, searching PubMed by, 32–35
    - EGFR, 112–114
    - entry, 111
    - gene, 113
    - multiple sequence alignment, 275
    - protein, Swiss-Prot, 113
  - National Institute of Health (NIH)
    - Center for Information Technology, 158
    - database, 127
  - NCBI (National Center for Biotechnology Information)
    - described, 131–132
    - phylogenetic tree primer, 402
    - primers resource locator, 414
    - structure-structure similarity search service, 350
  - NCBI (National Center for Biotechnology Information) BLAST (Basic Local Alignment Search Tool). *See also* PSI-BLAST
    - alignments, 213, 215–216
    - biological questions, asking, 218–219
    - complementary property, 20
    - described, 57–58, 199, 413
    - DNA sequences, overview, 216–218
    - EMBnet blastp, 207–209
    - graphic display, 211–212
    - hit list, 212–213
    - homologues, 214
    - hybridizing primers, 138
    - NCBI blastp, 202–207
    - output, 209–210, 224–225
    - parameters, 216, 219–220, 223–224

- NCBI (*continued*)
- protein domains, discovering and using, 230–231
  - protein sequences, handling, 201–202
  - results, 60–61
  - sequence masking, 220–223
  - servers, alternative, 231–233
  - starting, 58–60
- NCBI (National Center for Biotechnology Information) CD (Conserved Domain) server
- described, 187–190
  - protein sequence analysis, 195
- NCBI (National Center for Biotechnology Information) PubMed database
- author's names, searching by, 32–35
  - described, 412
  - fields, searching by, 35–38
  - items not available in, 41
  - limits, searching using, 38–40
  - protein, finding by name, 30–31
  - queries, making the most of, 41
  - saving multiple summaries, 31–32
- NCBI (National Center for Biotechnology Information) server Entrez/Gene resource
- bacterial genomes, 92–94
  - described, 413
  - LOCUS, 86–88
  - viral genomes, 89–92
- neo-Darwinism, 375
- neutralism, 375
- NIH (National Institute of Health) Center for Information Technology, 158
- database, 127
- nonoverlapping alignments, 258–259
- NR database, 412
- nsSNP analysis, 269
- N-terminus, 14
- Nuclear Receptor Signaling Atlas (Nursa), 128
- Nucleic Acid Research Web site, 415
- nucleotide (N)
- DNA sequences, analyzing, 23
  - IUPAC code, 19
  - lost in reformatting, 312
  - RNA nucleotide sequence letters, 21
- nucleotide sequence databases
- genes and genomes, reading into, 70–73
  - historical perspective, 69–70
  - human genome, 97–98
  - NCBI gene-centric database, 86–88
- nucleotide sequence, GenBank eukaryotic mRNA entry
- calling, 78–79
  - FEATURES section, 81–84
  - fetching, 80
  - gene sequence, 79
  - KEYWORD line, 79
  - keywords, 81
  - related, working with, 84–85
  - retrieving without accession numbers, 85–86
- nucleotide sequence, GenBank prokaryotic entry
- FEATURES table, 76–77
  - header, reading, 74–75
  - sample gene, fetching, 73–74
  - Sequence section, 77
- nucleotide sequence whole-genome database
- complete bacterial genomes, 92–94
  - complete viral genomes, 89–92
  - described, 88–89
  - DoE, 96–97
  - TIGR bacterial genomes, 94–95
- number, accession
- GenBank entry, 74, 81
  - Swiss-Prot, 111–112
- number, multiple sequence alignment, 272–273
- Nursa (Nuclear Receptor Signaling Atlas), 128
- ○ ●
- OMIM database, 412
- 1000 bp (kb), 23
- online BLAST servers
- characterized and uncharacterized, integrating, 275–276
  - ExpASy server, selecting sequences on, 276–279
  - multiple-alignment methods, Web addresses for, 276
  - Swiss-Prot server, gathering known collection of sequences from, 280–281
- ontology, 117
- Operational Taxonomic Units (OTUs), 399

- ORF (open reading frame)  
 described, 145  
 protein sequence, 53  
 UniProtKB/Swiss-Prot database, 107–108
- ORGANISM, GenBank entry, 75, 81
- origin of protein, epidermal growth factor  
 receptor (EGFR), 112–114
- orthologous genes, 373
- orthologs, 377
- OTUs (Operational Taxonomic Units), 399
- **p** ●
- pairing rules, 365
- pairwise comparisons. *See also* dot plot  
 described, 235  
 local alignments over Internet, 254–261  
 method, choosing, 237–239  
 proteins and DNA, aligning, 262  
 sequences, choosing, 236–237  
 servers, listed, 262–263
- pairwise projection, 381
- Pal2nal pairwise alignment program, 263
- palindromes, 20–21
- paralogs, 377
- parameters, BLAST, 216
- parasite characters, warning about, 52
- Parsimony package, 401
- Pasteur Institute  
 protein alignment tool, 262  
 resource locator, 414
- Pasteur Institute EMBOSS server  
 G+C content, establishing, 138–139  
 modules, 138–139  
 word frequency, computing, 140–141
- PatScan, finding RNA patterns with, 363–367
- pattern identification, 269
- pattern matching, 12
- patterns, looking for post-translational  
 modifications, 175–177
- PAUP phylogenetic tree package, 401
- Pbil protein sequence analysis, 195
- PCR (polymerase chain reaction)  
 analysis, 269  
 primer, 135–138
- PDB (Protein Data Bank) site  
 described, 412  
 protein 3-D structures, 337–340, 351  
 protein families, 127, 412
- Pearson, W.R. (FASTA program creator), 48
- penalty, gap-extension  
 ClustalW parameter tuning, 286  
 described, 258
- penalty, gap-opening  
 ClustalW parameter tuning, 286  
 described, 257
- period (.), 292
- PfamA domain collection, 182
- phenylalanine, 11
- Phred and Phrap sequence assembly tool,  
 267
- Phred/Phrap/Consed, 154
- Phylip software  
 computing tree, 387–395  
 multiple sequence alignment format, 307  
 obtaining, 413  
 phylogenetic tree, 371, 401  
 resource locator, 414
- phylogenetic analysis. *See also* ClustalW  
 multiple sequence alignments; Phylip  
 software; PhyML software  
 analysis application, 269  
 computing tree, 383–384  
 described, 373–374  
 displaying tree, 399–400  
 DNA or protein sequences, 374–375  
 gene tree or species tree, 377–379  
 generic resources, 401–402  
 jargon, 398–399  
 multiple sequence alignment, 380–382  
 online resources, 400–401  
 orthologous genes, 402  
 perfect set, creating, 379–380  
 purpose of, 372–373
- PhyML software  
 computing tree, 396–398  
 obtaining, 413  
 phylogenetic tree, building, 371
- PIR (Protein Information Resource)  
 ClustalW server, 300  
 cross-references, 116  
 described, 62–63  
 multiple sequence alignment format, 306  
 protein sequence analysis, 195
- PKR (Protein Kinase Resource) database, 128
- plot, dot. *See also* Dotlet  
 biological analysis, 249–254  
 described, 143–144, 238, 239–240

- plot, dot (*continued*)
  - inverted repeats, identifying, 144
  - low-complexity regions in proteins, finding, 253
  - programs, different types of, 240
  - tandem repeats, identifying, 250–252
- polymerase chain reaction (PCR)
  - analysis, 269
  - primer, 135–138
- positions, number of, 23
- post-translational modification
  - described, 174–175
  - ORFs, 108
  - other tools, 180
  - output, understanding, 177–179
  - patterns, looking for, 175–177
  - short patterns, 179
  - species information, 179–180
  - weak patterns, eliminating, 180
  - weak signals, 180
- Pratt motif-finding method, 301
- <PRE> parasite character, 52
- prediction line (Pred), 332
- predictions, importance of, 168
- primary structure analysis
  - coiled-coil regions, 174
  - properties revealed by, 166
  - “sliding windows” technique, 167–168
  - transmembrane segments, 168–174
- primary transcript, 53
- Primer3, 136–137
- PRINTs domain collection, 183
- Probcons
  - multiple sequence alignment, 301
  - server listed, 301
- PRODOM domain collection, 183
- profiles, Swiss-Prot, 118
- programs. *See also individual programs listed by name*
  - described, 412
  - listed, 413
- prokaryotes, genes and genomes, 70–72
- prokaryotic entry, GenBank
  - FEATURES table, 76–77
  - header, reading, 74–75
  - sample gene, fetching, 73–74
  - Sequence section, 77
- proline, 11
- promoter, 72
- PROSITE database
  - described, 174–175
  - other tools, 180
  - output, understanding, 177–179
  - patterns, looking for, 175–177
  - short patterns, 179
  - species information, 179–180
  - weak patterns, eliminating, 180
  - weak signals, 180
- PROSITE-Profile domain collection, 182
- Protal2dna pairwise alignment program, 263
- protease, 165
- protease digestions, 166
- protein
  - discovery, 145
  - and DNA, aligning, 262
  - family databases, 127–128
  - finding by name, PubMed, 30–31
  - name, Swiss-Prot, 113
- Protein Data Bank (PDB) site
  - described, 412
  - protein 3-D structures, 337–340, 351
  - protein families, 127, 412
- protein domain, finding known
  - CD server of NCBI, 187–190
  - collection, choosing right, 182–183
  - described, 180–181
  - Internet tools, 194–195
  - InterProScan results, interpreting, 185–187
  - InterProScan server, 183–185
  - Motif Scan, 190–193
  - new domains, finding, 194
- Protein Information Resource (PIR)
  - ClustalW server, 300
  - cross-references, 116
  - described, 62–63
  - multiple sequence alignment format, 306
  - protein sequence analysis, 195
- Protein Kinase Resource (PKR) database, 128
- protein maturation, 108
- protein sequence
  - amino acids, 10–12
  - chapters, topics covered by individual, 16–17
  - codes for ambiguity or exceptional amino acids, 13

- DNA coding regions, translating into, 24–25  
 history of sequence analysis, 12  
 reading, 13–14  
 3-D structures, 14–16
- protein structure databases, 126–127
- protein 3-D structures  
 additional structural features, predicting, 334–336  
 computer, folding in, 351  
 described, 329–330  
 guessing, 340–342  
 homology modeling, 351  
 interactions, predicting, 352  
 interactive exploration, 344–349  
 interplay between multiple alignments and structural analysis, 343–344  
 local segments, 330  
 in movement, looking at, 352  
 PDB structures, 350–352  
 from primary to, 336–337  
 retrieving and displaying from PDB site, 337–340  
 secondary structure, predicting, 330–334  
 sequence and structure, interactive analysis, 349–350  
 sequence/PDB structure relationship, interactive exploration, 344–349  
 similar shapes, finding proteins with, 350
- protein-coding regions, finding for single DNA sequence  
 described, 145  
 gene parsing for eukaryotic genomes, 151  
 GeneMark, 148–149  
 GenomeScan, 151–153  
 internal exons, finding in vertebrate genomic sequences, 149–151  
 ORFing, 145–147
- Protogene Web server, 262
- ProtParam program  
 described, 161–163  
 extinction coefficient, 165  
 half-life, 165  
 instability, 165  
 molecular weight, 164–165
- Protscale results, interpreting, 170–171
- Protscale, running, 168–170
- prss pairwise alignment analysis, 263
- PSI-BLAST  
 errors, avoiding, 228–230  
 protein domains, discovering and using, 230–231  
 protein sequences, 226–228  
 servers, alternative, 231–233
- PsiPred software, 413
- PSSMs, building, 194
- publishing multiple sequence alignments.  
*See also* formatting  
 beautifying tools, 325  
 Boxshade utility, 319–321  
 described, 303–304  
 editing packages, 323–324  
 Logos, generating high-impact pictures with, 322–323  
 tools for extracting information, 324
- PubMed database  
 author's names, searching by, 32–35  
 described, 412  
 fields, searching by, 35–38  
 items not available in, 41  
 limits, searching using, 38–40  
 protein, finding by name, 30–31  
 queries, making the most of, 41  
 saving multiple summaries, 31–32
- purine (R)  
 IUPAC code, 19  
 RNA nucleotide sequence letters, 21
- pyrimidine (Y)  
 IUPAC code, 19  
 RNA nucleotide sequence letters, 21
- pyrrolysine, 13
- *Q* •
- query  
 PubMed, making the most of, 41  
 sequence, 203
- *R* •
- R (purine)  
 IUPAC code, 19  
 RNA nucleotide sequence letters, 21
- RALEE package, 324
- random coils, 330

- Rasmol software, 413
  - reading frames, 26
  - READSEQ sequence text converter, 310
  - REBASE database, 128, 134
  - references section
    - EGFR, 114
    - GenBank entry, 55, 75
  - repeats, internal
    - composition, analyzing single DNA sequence, 142–144
    - pairwise comparisons, 237
  - repeats, inverted
    - described, 142
    - dot plot, 144
  - repeats, tandem
    - described, 142
    - dot plot, 250–252
  - replicates, number of, 391
  - research, multiple sequence alignment, 267–270
  - RESID<sup>(r)</sup> database, 124–125
  - residue
    - described, 12, 13
    - Swiss-Prot, 121–122
  - resource locators, 414
  - restriction enzymes, 21
  - restriction map, computing and verifying, 134–135
  - reverse-complement, 144
  - Review package, 324
  - ribosomal RNA (rRNA), 369
  - Rickettsia conorii* genome, 27–28
  - RNA. *See also* Mfold software
    - databases and genomes, searching, 362–367
    - described, 353–354
    - generic resources, 370
    - miRNAs and siRNAs, 367–368
    - predicting, modeling, and drawing, 354
    - rRNA, 369
    - secondary structures, 355
    - small, non-coding, 369–370
  - RNA sequences, analyzing
    - DNA versus, 21
    - IUPAC codes, 21–22
    - nucleotide, 23
    - sticky strands, 22–23
  - RNA World resource locator, 414
  - Roberts, Richard J. (Restriction Enzyme Database owner), 106
  - Robinson, Tara Rodden (*Genetics For Dummies*), 70
  - robustness, 409
  - rooted phylogenetic tree, 399
  - Rosen, Steve (Primer3 developer), 136
  - rRNA (ribosomal RNA), 369
- S ●
- same ancestor, multiple sequence alignment, 266
  - same ancestor, sequences without conserved patterns, searching, 299
    - described, 297–298
    - Gibbs sampler, 298
  - San Diego Supercomputer Center, 158
  - Sanger, Alfred, 17
  - scaffold sequence signatures, UniProtKB/Swiss-Prot database, 109–110
  - SCOP (Structural Classification Of Proteins), 127
  - screen capture, 248, 408
  - Seaview package, 323
  - security, Web-based servers, 406
  - SEGMENT, GenBank entry, 81
  - selenocysteine, 13
  - Selex multiple sequence alignment format, 307
  - SeqCheck sequence text converter, 310
  - sequence analysis, DNA
    - computers, 12
    - described, 17, 216–218
    - double helix, 18–20
    - IUPAC code, 19
    - nucleotide, 23
    - palindromes, 20–21
    - reading, 17–18
  - sequence fragments, assembling for single DNA sequence
    - CAP3 documentation, 155–157
    - machines, limitations of, 153
    - public software, managing large projects with, 154–155
  - sequence name, lost in reformatting, 312
  - sequence of protein, 14
  - Sequence Retrieval System (SRS), 185, 413

- sequence section
  - EGFR, 123
  - GenBank entry, 55
  - GenBank prokaryotic entry, 77
- sequence similarity, 268
- sequence-identification numbers, 407
- sequences, DNA, retrieving
  - introns and exons, 51
  - from protein sequences, 52–53
  - relevant to my protein, 53–56
- Sequencher (Gene Codes), 154
- sequencing human genome, 97–98
- serine, 11
- servers, online. *See also individual servers listed by name*
  - advantages, 405
  - alignments, 408
  - borderline results, checking, 409
  - E-values, 408
  - fresh data, importance of, 409
  - licensing issues, 410
  - parameters, 407
  - recording sequence-identification numbers, 407
  - results, saving, 407–408
  - security of data, 406
  - software, installing your own, 410
  - unpublished methods, 409
  - version, server, and database version, 406
- short patterns, 179
- SIB (Swiss Institute of Bioinformatics), 105
- signal peptide, 119
- silencing RNAs (siRNAs), 367–368
- similarity. *See also* BLAST
  - described, 160
  - importance of, 200–201
  - multiple sequence alignment, 273–274
- single DNA sequence
  - entry points, additional, 157–158
  - importance, 129
  - PCR primer, 135–138
  - restriction map, computing and verifying, 134–135
  - skills, necessary, 130
  - UniVec matches, 133–134
  - vector sequences, removing, 130–133
- single DNA sequence, analyzing
  - composition
  - EMBOSS modules, 142
  - G+C content, 138–139
  - genome-specific repeats, identifying, 145
  - internal repeats, finding, 142–144
  - long words, counting, 140–141
  - words, counting, 139–140
- single DNA sequence, assembling
  - sequence fragments
    - CAP3 documentation, 155–157
    - machines, limitations of, 153
    - public software, managing large projects with, 154–155
- single DNA sequence, finding protein-coding regions
  - described, 145
  - gene parsing for eukaryotic genomes, 151
  - GeneMark, 148–149
  - GenomeScan, 151–153
  - internal exons, finding in vertebrate genomic sequences, 149–151
  - ORFing, 145–147
- single protein sequence
  - biochemistry using computer, 160–166
  - described, 159–160
- siRNAs (silencing RNAs), 367–368
- size, protein molecules, 15
- Skaletsky, Helen (Primer3 developer), 136
- slash marks, two (//), 77
- “sliding windows” technique, 167–168
- SMART domain collection, 183
- Smith and Waterman (SSEARCH), 232
- soft science, 12
- software. *See also individual programs listed by name*
  - described, 412
  - listed, 413
- source, GenBank entry, 74
- speciation, 377
- species information, 179–180
- species tree, 377–379
- specify patterns, 365
- SRS (Sequence Retrieval System), 185, 413
- SSEARCH, Smith and Waterman, 232
- Staden Package, 154
- standard genetic code, table of, 25–26
- star (\*), 292
- stems, 23
- sticky strands, 22–23
- stochastic method, Gibbs sampler, 298
- strands, extended, 330

Strasbourg ClustalW server, 300  
 structural bioinformatics, 15  
 Structural Classification Of Proteins (SCOP), 127  
 structural similarity, 268  
 structure prediction, 269  
 substitution matrix, 223, 257, 286  
 summaries, PubMed, 31–32  
 Swbic resource locator, 414  
 Swiss EMBnet, 160  
 Swiss Institute of Bioinformatics (SIB), 105  
 Swiss-Model server, 127  
 Swiss-Prot database  
   accession number, 111–112  
   described, 412  
   domain, 120–121  
   gathering known collection of sequences from, 280–281  
   synonyms, Swiss-Prot, 113

## • T •

T, IUPAC code, 19  
 T (thymine), 19  
 tandem domains, 252  
 tandem repeats  
   described, 142  
   dot plot, 250–252  
 target database, 203  
 taxonomy, Swiss-Prot, 113  
 tblastn, 201  
 tblastx, 217  
 Tcoffee  
   phylogenetic tree, 400  
   server listed, 301  
 Tcoffee multiple sequence alignment  
   ClustalW versus, 291  
   CORE, evaluating quality with, 290  
   described, 301, 413  
   EXPRESSO, combining sequences and structures with, 290  
   tools, 287  
   using, 287–290  
 TEIRESIAS motif-finding method, 302  
 text sequences, 12  
 thermal cycler, 136  
 3'-terminus, 18

3-D protein structure  
   additional structural features, predicting, 334–336  
   computer, folding in, 351  
   described, 329–330  
   guessing, 340–342  
   homology modeling, 351  
   interactions, predicting, 352  
   interactive exploration, 344–349  
   interplay between multiple alignments and structural analysis, 343–344  
   local segments, 330  
   in movement, looking at, 352  
   patterns, identifiable, 178  
   PDB structures, 350–352  
   from primary to, 336–337  
   retrieving and displaying from PDB site, 337–340  
   sample, illustrated, 16  
   secondary structure, predicting, 330–334  
   sequence and structure, interactive analysis, 349–350  
   sequence/PDB structure relationship, interactive exploration, 344–349  
   sequences, analyzing, 14–16  
   similar shapes, finding proteins with, 350  
 threonine, 11  
 threshold value, 246  
 thymine (T), 19  
 TIGR (The Institute for Genome Research)  
   Assembler, 154  
   bacterial genomes, 94–95  
 TIGRFAM domain collection, 183  
 TMHMM  
   described, 168  
   results, interpreting, 173–174  
   running, 171–173  
 top cursor, Dotlet, 247  
 topological domain, 120  
 TRanslation of European Molecular Biology Laboratory (TrEMBL) nucleotide sequences, 106  
 translocation, 109  
 transmembrane segment, protein  
   described, 120  
   predictions, importance of, 168  
   Protoscale results, interpreting, 170–171

Protscale, running, 168–170  
 TMHMM results, interpreting, 173–174  
 TMHMM, running, 171–173  
 Trees software, 413  
 TrEMBL (TRanslation of European Molecular Biology Laboratory) nucleotide sequences, 106  
 tRNAs, finding in genome, 363  
 tryptophan, 11  
 two sequences, comparing. *See* pairwise comparisons  
 type-1 human immunodeficiency virus (HIV-1), 89–92  
 tyrosine, 11

• U •

U (uracil), 21  
 UniProtKB/Swiss-Prot database  
 accession numbers, 111–112  
 Comments, 114–116  
 Cross-References section, 116–118  
 described, 105–106  
 EGF receptor entry, deciphering, 110–111  
 Entry Name, 111  
 entry sections, 110  
 Features section, 119–123  
 final activities and destination for each protein (translocation), 109  
 folds and functions (scaffold sequence signatures), 109–110  
 Keywords, 118–119  
 linking to, 106–107  
 name and origin of protein, 112–114  
 ORFs, 107–108  
 References, 114  
 sequence, 123  
 UniVec matches, single DNA sequence, 133–134  
 Université Libre de Bruxelles, 158  
 University of Massachusetts Medical School, 135, 136  
 unpublished methods, 409  
 unrooted phylogenetic tree, 399  
 uppercase/lowercase, lost in reformatting, 312  
 uracil (U), 21

U.S. Department of Energy (DoE) whole-genome database, 96–97  
 USC pairwise alignment program, 263

• V •

valine, 11  
 vector sequences, removing single DNA sequence, 130–133  
 VERSION, GenBank entry, 74, 81

• W •

Washington University in St. Louis, 363  
 weak patterns, eliminating, 180  
 weak signals, 180  
 Web servers. *See also individual servers listed by name*  
 Web-based BLAST servers  
 characterized and uncharacterized, integrating, 275–276  
 ExpASy server, selecting sequences on, 276–279  
 multiple-alignment methods, Web addresses for, 276  
 Swiss-Prot server, gathering known collection of sequences from, 280–281  
 Web-based servers  
 advantages, 405  
 alignments, 408  
 borderline results, checking, 409  
 E-values, 408  
 fresh data, importance of, 409  
 licensing issues, 410  
 parameters, 407  
 recording sequence-identification numbers, 407  
 results, saving, 407–408  
 security of data, 406  
 software, installing your own, 410  
 unpublished methods, 409  
 version, server, and database version, 406  
 Webcutter tool, 134–135  
 whole-genome database  
 complete bacterial genomes, 92–94  
 complete viral genomes, 89–92  
 described, 88–89

whole-genome database (*continued*)

DoE, 96–97

TIGR bacterial genomes, 94–95

windows, sliding

described, 167–168

dot plot versus, 239–240

words

BLAST, 224

counting in single DNA sequence, 139–140

frequency, computing, 140–141

WU-BLAST, 232

WWW Signal Scan, 158

## • X •

xenAliTwo pairwise alignment program,  
263

xenologs, 377

## • Y •

Y (pyrimidine)

IUPAC code, 19

RNA nucleotide sequence letters, 21

## • Z •

Zhang Lab, 158

Zhang, Michael (MZE developer), 150



