

# Index

---

Note: Page numbers followed by “f”, “t” and “b” refers to figures, tables and boxes respectively.

- A**  
ABI SOLiD, 57  
ABYSS, 160  
Adenosine 5'-phosphosulfate (APS), 56–57  
Affine gap penalty, 141–142  
Algorithm parameters, 149  
AlgPred, 201  
    input sequence, 203f  
    SDAP developments, 201  
Alignment algorithms, 140–143  
Allergenic protein databases, 198–203  
    allergenicity-prediction paradigm, 200  
    allergenicity-prediction servers, 200–203  
    conformational epitopes, 200  
    databases of, 198–199  
    linear epitopes, 200  
    WHO/IUIS allergen nomenclature, 198–199  
Allergenonline, 199  
Allergens, 198  
Allermatch, 199  
ALLPATH, 160  
All-round Retrieval of Sequence and Annotation (ARSA), 82  
Alpha peptide bonds, 17  
Alternative splicing, intron phase effect, 10f  
American Standard Code for Information Interchange (ASCII), 78  
Amide linkages, 185  
Amino acids  
    hydrophobicity/hydrophilicity scores, 188t  
    relative propensity values, 191t  
Ancestral allele, 47  
Antigenicity prediction, 186–189  
Apomorphy, 51  
APS. *See* Adenosine 5'-phosphosulfate (APS)  
Arginine, 16–17  
Argus microscope, 70f  
ARSA. *See* All-round Retrieval of Sequence and Annotation (ARSA)  
*Aspergillus flavus*, 198  
*Aspergillus fumigatus*, 198  
Asymmetrical exon, 9  
Atlas of Protein Sequence and Structure, 73  
Autapomorphy, 51
- B**  
BankIt, 80  
Basic Local Alignment Search Tool (BLAST), 149  
    *Oatp5/Sico1a6* mRNA sequence, 123f  
    pairwise alignment, 107–110  
    *Sico1a6*, analysis of, 152f  
Bayesian phylogenetic analysis, 214–215  
Binding-inactive regions (BIRs), 24  
Binomial nomenclature system, 50  
Bioinformatics, 27–28  
    analysis of, 75  
    *vs.* computational biology, 74–75  
    definition of, 74  
    goals of, 75  
    technical toolbox, 75–76  
Biological macromolecules, 2  
    3D structure, 197–198  
    FirstGlance in Jmol, 197–198  
    protein data bank (PDB) database, 197  
BIRs. *See* Binding-inactive regions (BIRs)  
Bishop–Friday model, 212–213  
Bit score, 149  
BLAST. *See* Basic Local Alignment Search Tool (BLAST)  
Blocks substitution matrices (BLOSUM)  
    amino acids, 147f  
    PAM matrices, 146  
    scoring matrices, 145–146, 147f  
BLOSUM. *See* Blocks substitution matrices (BLOSUM)  
Bond rotation, 185–186  
Bootstrapping, 215  
    principles of, 216f  
    value, 216  
Bowtie 2, 160  
Bridge amplification, 59  
Bruijn graph, 159–160
- C**  
CAAT box, 11, 168  
*Caenorhabditis elegans*, 172  
 $\alpha$ -Carbon, asymmetric, 15f  
CCDS. *See* Consensus coding sequence (CCDS)  
ceRNAs. *See* Competing endogenous RNAs (ceRNAs)  
CGH. *See* Comparative genomic hybridization (CGH)  
Chain-termination principle, 55–56  
Chaperone molecules, 184–185  
ChIP-seq, 24  
Choanoflagellate *Monosiga brevicollis*, 37–38  
Chou-Fasman method, 190  
    amino-acid relative propensity values, 191t  
    GOR prediction tools, online sources, 191t  
Chou-Fasman Secondary-Structure Prediction (CFSSP) link, 190, 192t  
Chromosomal mutations, 30, 42  
    to nucleosome, organization hierarchy, 19f  
    whole-genome duplication, 35  
circRNAs. *See* Circular RNAs (circRNAs)  
Circular RNAs (circRNAs), 14–15  
Cladistics, 50  
Cladograms, 50–51, 211  
Clustal Omega, 140  
ClustalW, 140  
    algorithm, 140  
    clustal programs, 140  
    multiple alignment, 143f, 143t  
CNPs. *See* Copy number polymorphisms (CNPs)  
CNVs. *See* Copy number variations (CNVs)  
Coactivators, 21  
Coding sequence (CDS), 86  
Coding *vs.* noncoding regions  
    nucleotide composition, 163  
Coiled coils, online tools for analysis, 193t  
Comparative genomic hybridization (CGH), 64  
Competing endogenous RNAs (ceRNAs), 14  
Complementary DNA (cDNA), 78  
Computational molecular biology, 74  
Conformational epitope, 200  
Conjugation, 39  
Consensus coding sequence (CCDS), 118–120  
Conserved domain database (CDD), 195  
    domain analysis, 197f  
    home page, 196f  
    search and analysis launch page, 196f  
Contig, defined, 157  
Contiguous clones, 157–159  
Convergent evolution, 134  
Copy number polymorphisms (CNPs), 30–31  
Copy number variations (CNVs), 30–31, 33, 55–56  
Coregulators, 18  
CpG dinucleotides, 5-methylcytosine (5-mC), 22  
CpG sequence, 20  
Cryptogrammic method, 73–74
- D**  
Darwinism  
    biological evolution/basic premises, 28–30  
    evolutionary principles, 29  
    test tube, 29–30  
Databanks, 102–103

- Data flatfile, 85–86  
 Data normalization, 175–176  
 Data retrieval, 101–103  
   DBGET/LinkDB, 102  
   Entrez/GQuery, 102  
   sequence retrieval system, 102–103  
 Dayhoff, Margaret, 73–74  
 Dayhoff model (PAM), 212–213  
 dbfetch, 82, 104f  
 dbSNP, 80–81  
 DDBJ Trace Archive (DTA), 81  
 Deletion-insertion polymorphisms (DIPs), 91–92, 177  
 Dendrograms, 50–51, 211  
 Deoxyadenosine alpha-thio triphosphate (dATP<sub>α</sub>S), 56–57  
 Deoxynucleotide triphosphates (dNTPs), 56–57  
 Deoxyribonucleic acid (DNA), 2  
   binding proteins, 24  
   conformations, 5  
   double helix, 2  
     base-pairing rules, 4  
     genetic information, 5  
     nucleotides, linkage, 3–4  
     single-stranded, 4  
     structural units of, 2–3, 3f  
   genomic technologies  
     optical mapping, 67–71  
     overview, 70–71  
   high-molecular-weight, 70f  
   methylation, 22  
   polymerase, 59  
   sense/antisense strands, 8f  
   sequences, 162  
     library preparation, 58–59  
     mutations, 12  
     overlapping, 71  
     scoring matrix, 144  
     triple helix, base-pairing rules, 4  
 Derived allele, 47  
 Dihedral angles, 185–186  
 DIPs. *See* Deletion-insertion polymorphisms (DIPs)  
 DisProt database home page, 204–205  
   screenshot of, 204f  
 DisProt disorder-prediction launch page, 205–206, 205f  
 Distal promoter, 167–168  
 DNA. *See* Deoxyribonucleic acid (DNA)  
 DNA Databank of Japan (DDBJ), 79–97  
 DNASTAR's Lasergene, 69–70  
 dNTPs. *See* Deoxynucleotide triphosphates (dNTPs)  
 Domains prediction, 193–197  
   TMHMM prediction, 196–197  
   transmembrane (TM) helices, 196–197  
 Doolittle plots, 189f  
 Double-strand breaks (DSBs), 33  
 Double-stranded DNA (dsDNA) fragments, 57–58  
 Downstream promoter element (DPE), 11, 21  
 DPE. *See* Downstream promoter element (DPE)  
 Drop-down information box, 106f, 108f  
*Drosophila* genes, 66  
*Drosophila melanogaster*, 172  
 DSBs. *See* Double-strand breaks (DSBs)  
 DTA. *See* DDBJ Trace Archive (DTA)  
 Duplicated genes, evolutionary  
   fate, 35–36  
 Duplication-degeneration-complementation (DDC) model, 36–37, 37f  
 Dynamic programming, 140–141  
**E**  
 EB-eye (EBI) search, 82  
 Eck, Richard, 73–74  
 EMBL database, 79–97  
   EMBL-Bank, 79–97  
   expression, 80  
   flatfiles, 91  
   sequence flatfile format, 87–91  
   submission to, 81  
   web-based genome browsers, 117  
 EMBOSS Stretcher, 135  
 em-PCR. *See* Emulsion-PCR (em-PCR)  
 Emulsion-PCR (em-PCR), 57–58  
 Encyclopedia of DNA Elements (ENCODE), 20, 24  
 Enhancer-blocking function, 21–22  
 Entrez home page, 103f  
 Epigenetic code, 23  
 Epigenetic modifications, 24  
 Epigenomics database, 101  
*Escherichia coli*, 167  
   O157:H7: TW14359, optical maps  
     *in silico*, 67f  
     whole-chromosome sequences, 68  
   Pribnow box, 167  
   strains, 68  
 Eulerian path, 159  
 Euler-SR, 160  
 European Molecular Biology Laboratory's  
   European Bioinformatics Institute  
   (EMBL-EBI), 193–195  
   InterProScan, 194f  
 EVD. *See* Extreme value distribution (EVD)  
 Evolutionary systematics, 52  
 Evolutionary taxonomy, 52  
 Evolutionary tree. *See* Phylogenetic trees  
 Exonization, 41  
 ExPASy, 186, 193  
 Expect threshold box, 149  
 Expert Protein Analysis System  
   (ExPASy), 97  
 Expressed sequence tags (ESTs), 40, 161  
 Extinction coefficient, 187f  
 Extreme value distribution (EVD), 149  
**F**  
 FAO/WHO Allergenicity Test, 202f  
 FAO/WHO paradigm, 200  
 FAST-All (FASTA), 149  
 Felsenstein model, 212–213  
 File transfer protocol (FTP), 80  
 Fos-Jun heterodimer, 184  
 Founder effect, 46  
 Frameshift mutation, 30  
 FTP. *See* File transfer protocol (FTP)  
**G**  
 Gap penalties, 140–143  
 Gaps, 140–143  
 Gblocks, 212  
 GenBank, 79–97  
   availability the public, 81  
   history, 80  
   sequence flatfile format, 81–91  
   sequence submission, 80–81  
     to DDBJ, 81  
     to ENA/EMBL-Bank, 81  
     to NCBI, 80–81  
 GenBank divisions, three-letter  
   abbreviations, 92f  
 GenBank sequence database, 92  
 Gene conversion, 33  
 Gene database/portal, 82  
 Gene duplication, 20, 34–37  
 Gene-expression  
   microarray, 62  
   technologies, 55  
 Gene Expression Omnibus (GEO) database,  
   101, 150  
 Gene flow, centrifugal, 46  
 Gene frequency, affecting factors  
   genetic drift, 45–46  
   Hardy–Weinberg equilibrium, 41  
   migration, 43, 44b  
   mutation, 42  
   natural selection, 43–45  
   nonrandom mating, 46  
   in population, 41–46  
 Gene fusion, 38  
 Gene-hnRNA-mRNA-protein relationship, 6f  
 GeneInfo Identifier, 81–82  
 Gene pool, 27–28  
 Gene prediction, *ab initio* prediction, 162  
 Gene References Into Functions  
   (GeneRIFs), 116  
 Gene Sorter on mouse genome, 122f  
 Gene targeting, 64, 64f  
 Genetic drift, 45  
 Genetic hitchhiking, 48–49  
 Genetic information, 2  
 Genetic variations, 30–41  
   diversity  
     gene flow/introduction, 34  
     genome evolution, 34–41  
     recombination/generation of, 33  
   exon shuffling, 37–38  
   gene duplication, 34–37  
   gene fusion/fission, 38–39  
   mutation, molecular basis, 30–41  
   noncoding sequences, origin, 40–41  
   2R hypothesis, 34–37  
 Gene trapping, 65, 65f  
 Gene trees, species trees, 217–218  
 Genome analyzer, 58–59  
 Genome annotation, 160–167  
   *ab initio* prediction algorithms, 165  
   gene prediction, 162–167  
   GenomeScan, 165–166  
   homology-based prediction, 166  
 Genome browsers, 120  
   data visualization, 117–127

- ensembl genome browser, 117–120  
 NCBI's map viewer, 124–127  
 UCSC genome browser, 120–124  
 VEGA genome browser, 127  
 web-based, 117  
 GenomeNet, 102  
 Genome organization, 18–25  
 GenomeScan, 165–166  
 Genome searching, using map viewer, 127–130  
 Genome sequencing, 157–159  
   assembly, 160  
   human, 78, 177  
   methods, 158f  
   sequencing to pyrosequencing, 55–56  
   translation initiation site (TIS), 168  
 Genome structure, 18–25  
   Encyclopedia of DNA Elements (ENCODE) project, 24–25  
   functional sequence elements, 21–22  
   enhancers, 21  
   epigenetic changes, dynamics of, 24  
   epigenetic modifications, 22–24  
   histone code, 23  
   insulators, 21–22  
   locus control region (LCR), 21  
   promoters, 21  
   human genome, 19–20  
 Genome-wide association studies (GWAS), 24  
 Genomic context field, 110–113  
 Genomic data, 78  
 Genomic imprinting, 22–23  
 Genomic regions, 114, 115f  
 Genomic technologies  
   advances, 55  
   DNA, optical mapping, 67–71  
   overview, 70–71  
   genome expression  
     genome editing, 64–66  
     genome-wide mutagenesis, 64–66  
     interference, 64–66  
   high-density oligonucleotide-probe-based array, 62–64  
     genome expression, 62–64  
     tiling array, 63–64  
   mutation detection, 56–57  
   next-generation sequencing (NGS)  
     ABI SOLiD, 59–60  
     Illumina Solexa, 58–59  
     platforms, 57–60  
     Roche 454, 57–58  
     technology, 61–62  
   pyrosequencing, 56–57  
   sequencing to pyrosequencing, 55–56  
   SNP genotyping, 56–57  
 Genotype frequencies, 44b  
 GENSCAN home page, 165f  
 GLIMMER (Gene Locator and Interpolated Markov ModelER), 164  
 Global sequence-alignment method, 135–139  
   *vs.* local sequence alignment method, 135–139  
 Glycine, 17  
 N-Glycosylation, 189–190  
 O-Glycosylation, 189–190  
 Gnathostomata, 34  
 GRAVY value, 187f, 189  
 GWAS. *See* Genome-wide association studies (GWAS)  
 Gypsy, 21–22
- H**  
 Hairpin stem-loops, 169–172  
 Hamiltonian path, 159  
 Hamiltonian traversal path, 159–160  
 Hardy-Weinberg equilibrium principle, 41  
 Hasegawa–Kishino–Yano (HKY) model, 212–213  
 HAVANA (Human and Vertebrate Analysis and Annotation) group, 127  
 HDR. *See* Homology-directed repair (HDR)  
 Heat map, 63  
 Hemoglobin gene, 45  
 Heterochromatin barrier function, 21–22  
 Heterogeneous nuclear RNA (hnRNA), 5–6  
 Heuristic method, 149  
 HGT. *See* Horizontal gene transfer (HGT)  
 Hidden Markov models (HMMs), 163–165, 193  
 Hierarchical shotgun sequencing, 157–159  
 High-scoring segment pairs (HSPs), 154  
 High-throughput genome (HTG), 91–92  
 Histone code, 23  
 Histone H3 (H3K4me1), 25  
 Histone modifications, 23f  
 Hitchhiking effect, 48–49  
 H3K4me1, 24  
 HMMs. *See* Hidden Markov models (HMMs)  
 hnRNA. *See* Heterogeneous nuclear RNA (hnRNA)  
 HomoloGene, 91–92  
 Homologous genes, 36–37, 134  
 Homologous recombination, 33  
 Homology arms, 64  
 Homology-based prediction, 166  
 Homology-directed repair (HDR), 65–66  
 Homology modeling, 190–191  
 Homoplasy, 51–52  
 Hoogsteen edge, 4, 13  
 Hoogsteen hydrogen bonds, 4  
 Hopp plot, 189f  
 Horizontal gene transfer (HGT), 34  
 HSPs. *See* High-scoring segment pairs (HSPs)  
 HTG. *See* High-throughput genome (HTG)  
 HTUs. *See* Hypothetical taxonomic units (HTUs)  
 Human genome sequencing, 35  
 Hydrophathy plot, 186–189  
 Hydrophathy scale, 188  
 Hydrophilicity, 186–189  
 Hydrophobicity, 186–189  
*Hypothenemus hampei*, 39  
 Hypothetical taxonomic units (HTUs), 210
- I**  
 IDP Databases, 204f  
 IDPs. *See* Intrinsically disordered proteins (IDPs)
- IgBLAST. *See* Immunoglobulin search (IgBLAST)  
 Illumina Solexa, 57  
   sequencing, principles of, 60f  
 IMM. *See* Interpolated Markov model (IMM)  
 Immunoglobulin search (IgBLAST), 150  
 Inbreeding depression, 46  
 Industrial melanism, 43–44  
 Initiation factor, 167  
 Initiator (Inr) element, 11, 21  
 International HapMap Project, 177  
 International Knockout Mouse Consortium (IKMC), 65  
 International Nucleotide Sequence Database (INSD), 80  
 Interpolated Markov model (IMM), 164  
 InterProScan  
   European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI), 194f  
   graphical display, 194f  
   Slco1a6, domain analysis, 193–195  
 Intrinsically disordered proteins (IDPs), 18, 203–204  
   analysis, 203–206  
   databases, 204–205  
   online tools, 205t  
   prediction, 205–206  
 Intrinsically unstructured proteins (IUPs), 203–204  
 Introns-early theory, 10  
 Introns-late theory, 10  
 Isopeptide bond, 17  
 IUPs. *See* Intrinsically unstructured proteins (IUPs)
- J**  
 Jones-Taylor-Thornton (JTT) model, 212–213  
 JTT model. *See* Jones-Taylor-Thornton (JTT) model  
 Jukes–Cantor (JC) one-parameter model, 212–213
- K**  
*KanMX* gene, 64–65  
 Kazal domain, 195  
 Kazal\_SLC21, 195  
 Kimura's neutral theory, 49  
 Kimura's two-parameter model, 212–213  
 Knocking out genome expression, 66  
 Kozak sequence, 13, 162–163, 168  
 k-tuples (ktup), 154  
 Kyte plots, 189f
- L**  
 LALIGN pairwise comparison, protein sequences, 141f  
 Last common ancestor (LCA), 50–51, 210, 217  
 Last universal common ancestor (LUCA), 28, 210–211  
 Lateral gene transfer, 39  
 LCA. *See* Last common ancestor (LCA)  
 LCRs. *See* Locus control regions (LCRs)  
 Ledley, Robert, 73–74

- Le Gascuel (LG) model, 212–213  
 Library Page, 102–103  
 LINES. *See* Long interspersed nuclear elements (LINES)  
 Linker histone, 18  
 Linnaean hierarchy, 50  
 lncRNAs. *See* Long noncoding RNAs (lncRNAs)  
 Local sequence alignment method, 135–139  
 Locus control regions (LCRs), 21  
 Long interspersed nuclear elements (LINES), 161  
 Long noncoding RNAs (lncRNAs), 14  
 Lowess (locally weighted scatterplot smoothing) method, 175–176  
 Low-occupancy of TRF (LOT) regions, 24  
 LUCA. *See* Last universal common ancestor (LUCA)  
 Luciferin, 56–57  
 Lysine, 16–17
- M**  
 Major facilitator superfamily (MFS) domains, 195  
 MapSolver™, 67  
 Map Viewer  
   home page, 126f  
   mouse chromosome, gene distribution, 128f  
   *Mus musculus* genome, 126f  
   Oatp-5, master map of, 127f  
 Markov chain Monte Carlo (MCMC) process, 214–215  
 Markov models, 163, 164b  
 Mass Submission System (MSS), 81  
 Mate pairs, 157–159  
 Maximum likelihood (ML), 213–214  
 Maximum parsimony (MP) method, 213–214, 215f  
 McCaskill model, 172–173  
 McDonald-Kreitman method tests, 48  
 MCMC process. *See* Markov chain Monte Carlo (MCMC) process  
*Methanocaldococcus jannaschii*, tRNA endonuclease gene, 36–37  
 MFE. *See* Minimal-free-energy (MFE)  
 MFS domains. *See* Major facilitator superfamily (MFS) domains  
 Microarray analysis, 173–176  
   cluster analysis, 176  
   hierarchical clustering, 176  
   image normalization and clustering, 175f  
   image processing, 174–175  
   *k*-means clustering, 176  
   scanning, 174  
   supervised clustering, 176  
   two-color/two-channel, 174  
   unsupervised clustering, 176  
 Microarray Data Manager (MADAM), 176  
 Microarray fabrication, photolithographic synthesis, 63f  
 Microarray image  
   cluster analysis, 176  
   normalization and clustering, 175f  
   TM4 suite, 176
- Microevolution, 45  
 Micro RNA (miRNA), 172  
 Minimal-free-energy (MFE) RNA sequence, 172–173  
 miRNA. *See* Micro RNA (miRNA)  
 Missense point mutation, 30  
 ML. *See* Maximum likelihood (ML)  
 mlst-1 proteins  
   global alignment, 137f  
   pairwise alignment, 136f  
 Model maker (mm), 125–127  
 Molecular Biology Open Software Suite, 69–70  
 Molecular clock hypothesis, 49  
 Molecular evolution, 27–28  
   clock hypothesis, 49  
   genetics analysis, 211  
   phylogenetics, 49–52  
 Monocistronic RNA, 167  
 Monophyly, 217  
 Motifs prediction, 193–197  
 Mouse chromosome 6, 102  
   acrocentric, 114  
   exons mapped to, 124f  
 Mouse Genome Informatics (MGI) group, 77  
 Mouse Oatp5  
   cDNA, sequence of, 106f  
   coding sequence, 105  
   Ensembl for, 118f, 119f  
   GenBank information, 105f  
   Gene database, 111f  
   mRNA sequence, 123f  
   original submission of, 105  
   RefSeq record, 107f  
 Mouse Slco1a6, 111f  
   chromosome 6 graphics page, 115f  
   domain analysis, 193–195  
   on Ensembl, 118f, 119f  
   exon/intron sequence information, 116f  
   genomic context fields, 112f, 114  
   genomic regions, 114  
   MGI pages, 113f  
   mRNA, 107  
   sequence, 123f, 168  
   mRNAs  
     comparison, 109f  
     Oatp-5, 110  
     partial ProtParam analysis, 187f  
     revision history of, 108f  
     sequence, 205–206, 206f  
     transcript ID, 117–118  
     truncated screenshot, 112f  
 MP method. *See* Maximum parsimony (MP) method  
 MrBayes, 214–215  
 mRNA/gene information, 103–116  
   genomic context, 114  
   genomic regions, transcripts, and products, 114  
   nucleotide structure, 8f  
 Multiple sequence alignments, 139–140  
*Mus musculus*, 86, 130f  
   chromosome 6, 114  
   kidney-specific organic anion, 86  
   Slco1a6. *See* Mouse Slco1a6  
 Mutation, molecular basis, 31f
- N**  
 NAHR. *See* Non-allelic homologous recombination (NAHR)  
 National Center for Biotechnology Information (NCBI)  
   BLAST home page, of nucleotide blast, 151f  
   BLAST pairwise alignment, 136f  
   genome home page, 125f  
   home page, partial view, 101f  
   map viewer, 124–127  
   ORF finder, 169f  
   primary sequence database, 91–97  
   Reference Sequences (RefSeq), 167  
   RefSeq database, 97  
   SNP database (dbSNP), 177  
 Natural selection  
   balancing selection, 45  
   Darwin's theory, 29  
   disruptive selection, 44–45  
   stabilizing selection, 44  
 NCBI. *See* National Center for Biotechnology Information (NCBI)  
 N50 contig, 160–161  
 Needleman-Wunsch algorithm, 136f, 138f, 148  
 Neighbor joining (NJ), 213  
 Neofunctionalization, 35–36  
 Neutral theory, molecular evolution, 47–49  
   positive selection, signatures of, 47–48  
   protein sequence, synonymous/nonsynonymous substitutions, 47  
   selective sweep/hitchhiking effect, 48–49  
 Next-generation sequencing (NGS), 57, 61–62  
 NGS. *See* Next-generation sequencing (NGS)  
 NMR spectroscopy. *See* Nuclear magnetic resonance (NMR) spectroscopy  
 Non-allelic homologous recombination (NAHR), 33  
 Nonhomologous end-joining (NHEJ)  
   pathways, 65–66  
   repair pathway, 65–66  
 Nonsynonymous substitution, 12, 47  
 NSSS. *See* Nucleotide Sequence Submission System (NSSS)  
 Nuclear magnetic resonance (NMR) spectroscopy, 197  
 Nucleic-acid-based search, 152  
 Nucleotide database, 82  
 Nucleotide Sequence Submission System (NSSS), 81  
 Nucleotides, IUPAC codes, 180t  
 Numerical taxonomy, 50
- O**  
 Oases, 160  
 OCTOPUS  
   graphical outputs of, 199f  
   transmembrane-helix prediction, online tools, 197t  
 OMIM (Online Mendelian Inheritance in Man), 101  
 Online multiple alignment tools, 143t  
 Open reading frame (ORF), 9, 12, 79–80



- hexamer composition, 162–163  
prediction of, 167–169  
  online tools, 170*t*
- Operational taxonomic units (OTUs), 210
- Optical mapping, 67–71, 69*f*
- ORF. *See* Open reading frame (ORF)
- Orphan genes, 40
- Orthologous genes, 19–20, 34, 134
- OTUs. *See* Operational taxonomic units (OTUs)
- Overlap-layout-consensus (OLC) algorithm, 159–160
- P**
- Paired ends, 157–159
- Pairwise sequence alignment, 139–140  
  online tools, 140*t*  
  rlst-1c proteins, 139*f*
- PAM250 substitution matrix, 145*f*
- Paraphyly, 217
- Partial Splign output, 163*f*
- Pattern-hit-initiated (PHI)-BLAST, 154
- PAUP. *See* Phylogenetic analysis using parsimony (PAUP)
- pax6a/pax6b* genes, 36–37
- PDB (Protein Data Bank) database, 197
- Peptide bond, 185–186, 185*f*
- PeptideCutter, 186
- Peptide plane, 185–186, 185*f*
- Percent accepted mutation, 74
- Phobius, transmembrane-helix prediction, 205*t*
- Phrap, 159
- Phylogenetic analysis using parsimony (PAUP), 211
- Phylogenetics, 27–28  
  analysis of  
    phylogenetic tree, widespread use of, 209–210  
    tools, 211  
  biological classification systems, 50–52  
    cladistics, 50–52  
    cladograms, 50–52  
    phenetics/phenograms, 50  
  classification, 50–51  
  molecular evolution, 49–52  
  phylogenetic tree, 52  
  systematics, 50–51  
  systematics/biological classification, 50
- Phylogenetic trees, 52, 210–211, 216, 216*f*  
  construction of, 213–215, 214*f*  
  Bayesian phylogenetic analysis, 214–215  
  character-based methods, 213–215  
  distance-matrix, 213  
  molecular marker, selection of, 211–212  
  multiple sequence alignment, 212  
  principles of, 211–217  
  reliability, assessment, 215–217  
  sequence data, evolutionary model, 212–213  
  nested clades, 51*f*  
  presentation of, 210*f*  
  topology of, 210  
  widespread use of, 209–210
- Phylogeny, 49–50
- Phylograms, 211
- Picotiter plate (PTP), 58
- Plesiomorphy, 51
- Polycistronic RNA, 167
- Polymerase chain reaction (PCR), 56
- Polypeptide chain *see also* Protein structure  
  amino acids  
    nonstandard, 18  
    protein function, relationship, 16–17
- Polyphyly, 217
- Polypyrimidine tract, 7
- Position-specific scoring matrix (PSSM), 153–154
- Post-translational modification  
  prediction of, 189–190  
  protein, online analysis tools, 190*t*
- Pribnow box, 167
- Primary sequence databases, 79–97  
  NCBI, divisions of, 91–97  
  redundancy, 91  
  Reference Sequence (RefSeq) database, 92–97  
  sequence accession numbers, 91
- Primer walking, 157
- Prokaryotes, gene prediction, 162
- Promoters, prediction, 167–169
- Protease digestibility, prediction, 186
- Protein Information Resource (PIR) database, 73
- Proteins  
  allergenicity prediction, 198–203  
  3D structure, 197–198  
  physicochemical properties of, 186  
  secondary structure, 192*t*  
  sequence, 133–134  
  threading, 191
- Protein structure, 15–18, 183–185. *See also*  
  Polypeptide chain  
    acidic/basic proteins, 17–18  
    amino acids  
      configuration/chirality, 15–16  
      ionic character, 16  
      peptide bonds, linkage, 17  
      protein function, relationship, 16–17  
     $\beta$ -turn, 184  
    four levels of, 17  
    3.6<sub>13</sub>-helix, 183–184  
     $\alpha$ -helix, 183–184  
      coiled coils, 184  
    primary structure, 183  
    quaternary structure, 185  
    secondary structure, 183  
    tertiary structure, 184–185
- ProtParam, 186
- ProtScale, 188, 189*f*
- Proximal promoter, 167
- Pseudogenization, 35–36
- PubMed, 101
- Pulsed-field gel electrophoresis (PFGE) analysis, 68
- Punctuated equilibrium, 29
- Pyrosequencing technique, 55–57
- Q**
- Quaternary structure, 17
- R**
- Ramachandran plot, 185–186, 185*f*  
  Uppsala Ramachandran Server, 186
- Random genetic drift. *See* Genetic drift
- Ratio-intensity (R-I) plot, 175–176
- Rattus norvegicus*, 92
- Readseq program, sequence formats  
  conversion, 79
- Recoding, 169–172
- Reference assembly, 59
- Reference Sequence (RefSeq) database, 92–97
- Reference SNP cluster ID, 177
- RefSeq IDs, 114*t*
- RefSeq nucleotide sequence, 167
- RefSeq protein database, 152–153
- Regulatory elements (RE), 11
- Reinitiation, 168
- RepeatMasker, 161
- Replication slippage, 32*f*
- Restriction fragment length polymorphisms (RFLPs), 68
- Restriction-site mapping, of input sequence, 169
- Retriointrons, 10
- RHYTHM  
  graphical outputs of, 199*f*  
  transmembrane-helix prediction, 205*t*
- Ribosomal hopping, 172
- rlst-1a proteins, pairwise alignment, 136*f*, 137*f*, 138*f*, 139*f*
- rlst-1c proteins, pairwise alignment, 138*f*, 139*f*
- RNA, features, 12–13  
  circular RNAs (circRNAs), 14–15  
  coding *vs.* noncoding, 14–15  
  competing endogenous RNAs (ceRNAs), 14  
  long noncoding RNAs (lncRNAs), 14  
  messenger RNA (mRNA)  
    instability of, 12  
    5'/3'-untranslated regions, 12–13  
  secondary structures, 13
- RNAi. *See* RNA interference (RNAi)
- RNA interference (RNAi), 22
- RNA secondary structure, 171*f*  
  online tools, 173*t*  
  prediction, 169–173  
  online tools, 173*t*  
  web-based programs, 174*f*
- RNA sequencing (RNA-seq) data, 40, 161
- Roche 454, 57  
  454 sequencing, principles of, 58*f*
- S**
- Scaffolds, 157–159
- ScanAlyze, 176
- Scoring sequence alignment  
  scoring matrix/alignment score/statistical significance, 144–149  
    BLOSUM matrix, 145–148  
    PAM matrices, 144–145  
    PET91 matrix, 144–145  
  statistical significance of, 148–149  
  bit score, 149

- Scoring sequence alignment (*Continued*)  
*E*-value, 149  
*P*-value, 148  
*Z*-score, 148–149
- SDAP database, 201  
 FAO/WHO Allergenicity Test, 202*f*  
 home page, 201*f*
- Secondary databases, 97  
 Expert Protein Analysis System (ExPASy), 97  
 NCBI databases, 98–101  
 on nucleic acid/protein sequences, 98  
 publicly available, 98–101, 98*t*  
 Swiss-Prot, 97  
 UniMES, 97  
 UniParc, 97  
 UniProtKB/TrEMBL, 97
- Secondary-structure prediction  
 accuracy of, 193  
 advances in, 190–193  
 Chou–Fasman methods, 190  
 GOR methods, 190, 191*t*  
 protein, online tools for analysis, 192*t*
- Secondary structure, protein, 192*t*
- Selenocysteine, 5
- Self-fertilization, 46
- Sequence alignment, evolutionary basis, 133–134
- Sequence-assembly data, 130, 159–160
- Sequence data formats, 78–79  
 FASTA format, 78–79  
 PHYLIP format, 79
- Sequence determination, hypothetical pyrogram, 56*f*
- Sequence homology, 134–135
- Sequence identity, 134–135  
 twilight zone of, 200
- Sequence information. *See* Bioinformatics, analysis
- Sequence polymorphism, detection, 176–180
- Sequence read archive (SRA), 80–81
- Sequence Retrieval System (SRS), 101  
 home page, 104*f*
- Sequence similarity, 134–135
- Sequencing-by-ligation approach, 59–60
- Sequencing by synthesis principle, 56, 58–59
- Sequin, 80
- Shine-Dalgarno sequence, 162
- Short interspersed nuclear elements (SINEs), 161
- Short tandem repeats (STRs), 91–92
- Shotgun sequencing, 157–159
- Sigma factor, 167
- Silent point mutation, 30
- Single-base nucleotide substitution (SNPs), 177
- Single-molecule real-time (SMRT) sequencing technology, 62
- Single nucleotide polymorphisms (SNPs), 30–31, 55–56, 91–92, 150, 176–177  
 detection of, 176–180
- haplotype, 177  
 ID number, 179*f*  
 International HapMap Project, 177  
 IUPAC codes, for nucleotides, 180*t*  
 mouse *Slco1a6* gene, 178*f*  
 neighbor, 179*f*  
 neighbor SNP, 179*f*  
 rs266211819, graphic view, 179*f*  
 rs266211819 returns, 178*f*, 179*f*  
*Slco1a6* gene, 178*f*  
 ss370364874, 180*f*
- Single nucleotide variation (SNV), 177
- Slco1a6*. *See* Mouse *Slco1a6*
- Slipped strand mispairing. *See* Replication slippage
- Slippery sequence, 169–172
- Smith-Waterman algorithms, 135, 140*t*  
 blast-like alignment tool (BLAT), 154  
 analysis of, 150–152  
*vs.* FASTA, 154  
 protein query sequence, 149–150  
*Slco1a6*, 152*f*  
 typical basic output, 152–154  
 utility, 149–150  
 value cut-off, 152
- blastn, 150  
 database searching with heuristic versions, 149–154  
 megablast, discontinuous, 150  
 pattern-hit-initiated (PHI)-BLAST, 154  
 protein BLAST (blastp), 153–154  
 sequence comparison, 38–39  
 short nucleotide-sequence matches, 150–151  
 NCBI BLAST home page, 151*f*
- SNPs. *See* Single nucleotide polymorphisms (SNPs)
- SOAPdenovo, 160
- SOLiD sequencing, 59–60  
 principles of, 61*f*  
 sequencing library preparation, 60
- Spea multiplicata*, 44–45
- Speciation, 27
- Spidey, 161
- Splice acceptor, 7
- Splign, 161  
 online tool, 162*f*  
 splice-site detecting alignment algorithms, 161
- Staphylococcus aureus*, 167
- Structural Database of Allergenic Proteins (SDAP), 199
- Subfunctionalization, 36–37
- Submitted SNP ID number, 177
- Sulfolobus solfataricus*, 36–37
- Supercontigs, 157–159
- Swiss-Prot database, 97
- Symmetrical exon, 9
- Synapomorphy, 51
- Synonymous substitution, 47
- Syntenic block, 155
- Syntenic anchors, 155
- Systema naturae, 50
- ## T
- TAL effector nuclease (TALEN) technology, 65–66
- TATA box, 11, 21
- TATA-less promoters, 167–168
- Taxonomic categories, 50
- Taxonomy database, 101
- tbl2asn, 80
- The Institute for Genomic Research (TIGR), 176  
 assembler, 159  
 multiexperiment viewer (MeV), 176  
 Spotfinder, 176
- Tiling path, 157–159
- TMHMM, transmembrane-helix prediction, 205*t*
- TM4 suite, 176
- Torsion angle, 185
- Trace archive, 80
- Transcription-factor-binding sites, prediction, 167–169
- Transcription-related factors (TRFs), 24
- Transcriptomics, 78
- Transfer-messenger RNA (tmRNA), 172
- Translational reprogramming, 169–172
- Translation initiation sites, prediction, 167–169
- Transmembrane domains (TMDs), 107
- Transmembrane (TM) helices, 196
- Transmembrane-helix prediction  
 online tools, 197*t*  
 by RHYTHM, OCTOPUS, Phobius, and TMHMM, 205*t*
- Transmission electron microscopy, 62
- Transposable element (TE) domestication, 20
- Transversion, 30–31
- Trap cassette, 65
- Two-base encoding, 60
- Two rounds (2R) hypothesis, 34
- Typical eukaryotic gene structure, 5–12  
 transcribed genes  
 3'-flanking region, 11–12  
 5'-flanking region, 11  
 transcribed region, 7–11  
 alternative splicing, intron phase, 9  
 introns, evolution of, 10–11  
 intron-splicing signals, 7–8
- ## U
- UniGene database, 91–92, 101
- UniProtKB/Swiss-Prot, 201–203
- UniProtKB/TrEMBL, 97
- Universal Protein Resource Knowledgebase (UniProtKB), 97
- University of California Santa Cruz (UCSC)  
 Genome browser, 117  
 home page, partial screenshot, 120*f*  
 mouse  
 gateway, 121*f*  
 for *Slco1a6*, 121*f*
- 5'/3'-Untranslated region (UTR), 86
- Unweighted pair group method  
 with arithmetic mean (UPGMA)  
 tree, 213

**V**

VEGA. *See* Vertebrate genome annotation (VEGA)  
Velvet, 160  
Vertebrate genome annotation (VEGA)  
  genome browser, 127  
  home page, 128*f*  
VisiGene image browser, 124, 125*f*

**W**

Watson-Crick edge, 13  
Web-Based FASTA servers, 154*t*  
Webin, 81  
Whelan and Goldman (WAG) model, 212–213  
Whole.-genome duplication, 36–37  
Whole-genome shotgun (WGS) sequencing, 157–159  
Whole-genome tiling arrays, 64  
Woods plot, 189*f*

**Z**

Zero-mode waveguide (ZMW), 62  
Zinc-finger nuclease (ZFN), 65–66  
Zippers, online tools for analysis, 193*t*  
Zn-finger DNA-binding domains, 65–66  
Zn-finger nuclease, gene/genome manipulation, 66*f*  
Zwitterions, 16