GLOSSARY

An extensive glossary of genetic terms can be found at the Web site for the National Human Genome Research Institute (*http://www.nhgri.nih.gov/DIR/VIP/Glossary/*). The entries in this glossary provide a brief written definition of the term; the user can also listen to an informative explanation of the term using RealAudio.

algorithm Any sequence of actions (e.g., computational steps) that perform a particular task.

analogous In phylogenetics, characters that have descended in a convergent fashion from unrelated ancestors.

browser Program used to access sites on the World Wide Web. By using hypertext markup language (HTML), browsers are capable of representing a Web page the same way regardless of computer platform.

candidate gene A gene that is implicated in the causation of a gene. Candidate genes lie in a region that has been identified through genetic mapping. The protein product of a candidate gene may implicate the candidate gene as being the actual disease gene being sought.

cDNA library A collection of double-stranded DNA sequences that are generated by copying mRNA molecules. Because these sequences are derived from mRNAs, they contain only protein-coding DNA.

characters and character states In phylogenetics, characters are homologous features in different organisms. The exact condition of that feature in a particular individual is the character state. As an example, the character ‘‘hair color’’ can have the character states ‘‘gold,’’ ‘‘red,’’ and ‘‘yellow.’’ In molecular biology, the character states can be one of the four nucleotides (A, C, T, or G) or one of the 20 amino acids. Please note that some authors define character to mean the character state as defined here.

client A computer, or the software running on a computer, that interacts with another computer at a remote site (server). Note the difference between client and user.

contig Short for “contiguous.” Refers to a contiguous set of overlapping DNA sequences.

cytogenetic map The representation of a chromosome on staining and examination by microscopy. Visually distinct light and dark bands give each chromosome a unique morphological appearance and allow for the visual tracking of cytogenetic abnormalities such as deletions or inversions.

descriptor Information about a sequence or set of sequences whose scope depends on its placement in a record. It is placed on a set of sequences to reduce the need to save multiple redundant copies of information.

domain name Refers to one of the levels of organization of the Internet and used to both classify and identify host machines. Top-level domain names usually indicate the type of site or the country in which the host is located.

download The act of transferring a file from a remote host to a local machine via FTP.

E-mail Electronic mail. Refers to messages that can be composed on the computer and transmitted via the Internet to a remote location within seconds. [*Ant*: snail mail, postal mail.]

EST Expressed Sequence Tags. These are usually short (300–500 bp) single reads from mRNA (cDNA) that are usually produced in large numbers. They represent a snapshot of what is expressed in a given tissue or at a given developmental stage. They represent tags (some coding, others not) of expression for a given cDNA library.

exon Within a gene, a region that codes for part of the gene’s protein product; the “**ex**pressed regi**on**” of a gene.

FAQ Frequently asked questions. Exactly what it sounds like: a compiled list of questions and answers intended for new users of any computer-based resource, such as mailing lists or newsgroups.

feature Annotation on a specific location on a given sequence.

firewall A computer separating a company or organization’s internal network from the public part, if any, of the same network. Intended to prevent unauthorized access to private computer systems.

FTP File transfer protocol. The method by which files are transferred between hosts.

genetic map Gives the relative positions of known genes and or markers. Markers must have two or more alleles that can be readily distinguished.

genome All of the DNA found within each of the cells of an organism. Eukaryotic genomes can be subdivided into their nuclear genome (chromosomes found within the nucleus) and their mitochondrial genome.

Gopher A document delivery system allowing the retrieval and display of text-based files.

TEAM FLY
Team-Fly®

GSS Genome Survey Sequences. This DDBJ/EMBL/GenBank division is similar in nature to the EST division, except that its sequences are genomic in origin, rather than being cDNA (mRNA). The GSS division contains (but will not be limited to) the following types of data: random ''single pass read'' genome survey sequences, single pass reads from cosmid/BAC/YAC ends (these could be chromosome specific, but need not be), exon-trapped genomic sequences, and Alu PCR sequences.

GUI Graphical user interface. Refers to software front ends that rely on pictures and icons to direct the interaction of users with the application.

heuristic algorithm An economical strategy for deriving a solution to a problem for which an exact solution is computationally impractical or intractible. Consequently, a heuristic approach is not guaranteed to find the optimal or ''true'' solution.

homologs/homologous In phylogenetics, particular features in different individuals that are genetically descended from the same feature in a common ancestor are termed homologous. In molecular biology, homologous is often used simply to mean similar, regardless of genetic relationship.

homoplasy Similarity that has evolved independently and is not indicative of common phylogenetic origin.

host Any computer on the Internet that can be addressed directly through a unique IP address.

HTG/HTGS High Throughput Genome Sequences. Various genome sequencing centers worldwide are performing large-scale, systematic sequencing of human and other genomes of interest. The databases have deemed it beneficial to put the unfinished sequences resulting from such sequencing efforts in a separate division. HTG sequence entries undergo a maturation process. In Phase 0, the entry contains a single-to-few pass read of a single clone. In Phase 1, the entry contains unfinished sequence, which may be unordered, contain unoriented contigs, or a large number of gaps. In Phase 2, the entry still contains unfinished sequence but is ordered, with oriented contigs that may or may not contain gaps. In Phase 3, the entry contains finished sequence, with no gaps; at this point, the entry is moved into the appropriate primary GenBank division.

HTML Hypertext markup language. The standard, text-based language used to specify the format of World Wide Web documents. HTML files are translated and rendered through the use of Web browsers.

haplotype Sets of alleles that are usually inherited together.

hyperlink A graphic or text within a World Wide Web document that can be selected using a mouse. Clicking on a hyperlink transports the user to another part of the same Web page or to another Web page, regardless of location.

hypertext Within a Web page, text that is differentiated either by color or by underlining, which functions as a hyperlink.

indel Acronym for ''**in**sertion or **del**etion.'' Applied to length-variable regions of a multiple alignment when it is not specified whether sequence length differences have been created by insertions or deletions.

Internet A system of linked computer networks used for the transmission of files and messages between hosts.

IP address The unique, numeric address of a computer host on the Internet.

Intranet A computer network internal to a company or organization. Intranets are often not connected to the Internet or are protected by a firewall.

Java A programming language developed by Sun Microsystems that allows small programs (applets) to be run on any computer. Java applets are typically invoked when a user clicks on a hyperlink.

LAN Local Area Network. A network that connects computers in a small, defined area, such as the offices in a single wing or a group of buildings.

LOD score For "**l**og **od**ds," a statistical estimate of the linkage between two loci on the same chromosome.

molecular clock The hypothesis that nucleotide or amino acid substitutions occur at more or less a fixed rate over evolutionary time, like the slow ticking of a clock. It has been proposed that given a calibration date and a constant molecular clock, the amount of sequence divergence can be used to calculate the time that has elapsed since two molecules diverged.

mutation studies In Sequin, a set of sequences for the same gene in the same species, perhaps the same individual, in which several different induced mutations are isolated and sequenced.

oligo For oligonucleotide, a short, single-stranded DNA or RNA. Most often used as probes for the detection of complementary DNA or RNA.

orthologs/orthologous Homologous sequences are said to be orthologous when they are direct descendants of a sequence in the common ancestor, i.e., without having undergone a gene duplication event. See also homologs and paralogs.

PAM matrix PAM (percent accepted mutation) and BLOSUM (blocks substitution matrix) are matrices that define scores for each of the 210 possible amino acid substitutions. The scores are based on empirical substitution frequencies observed in alignments of database sequences and in general reflect similar physicochemical properties (e.g., a substitiution of leucine for isoleucine, two amino acids of similar hydrophobicity and size, will score higher than a substitution of leucine for glutamate.)

paralogs/paralogous Homologous sequences in two organisms A and B that are descendants of two different copies of a sequence created by a duplication event in the genome of the common ancestor. See also homologous and orthologs.

pedigree A tree representation of a family (cohort) showing the relationships between members and the pattern of inheritance of a given trait.

phylogenetic studies In Sequin, a set of sequences for the same gene in individuals of different species. The presumption is that the individuals cannot interbreed. Sequin does not allow a single organism name but expects the organism to be encoded in the definition line. It does, however, present a control for setting the proper genetic code.

physical map A genome map showing the exact location of genes and markers. The highest-resolution physical map is the DNA sequence itself.

platform Properly, the operating system running software on a computer, e.g., UNIX or Windows 95. More often used to refer to the type of computer, such as a Macintosh or PC-compatible.

polymorphism A gene or locus that can have one or more alleles or haplotypes.

population studies In Sequin, a set of sequences for the same gene in individuals of a single species. The presumption is that the individuals can interbreed. Sequin allows entry of a single organism name, although it would expect that some distinguishing source information, such as strain, clone, or isolate, be entered for each sequence.

positional cloning Relies on the identification of a gene through pedigree analysis, genetic and physical mapping, and mutation analysis. Does not require extensive knowledge of the biochemistry of the disease to determine the gene responsible for the disease. [*Ant*: functional cloning.]

protein name In a sequence record, the preferred field for a protein feature.

protein description In a sequence record, used if the protein name is not known.

server A computer that processes requests issued from remote locations by client machines.

site An individual column of residues in an amino acid or nucleotide alignment. The residues at a site are presumed to be homologous.

spam Postings to newsgroups or mail broadcast to a large number of E-mail accounts that usually are wholly irrelevant or not of interest to the recipients. Analogous to postal junk mail.

STS Sequenced Tagged Sites. An operationally unique sequence that identifies the combination of primer pairs used in a PCR assay, generating a reagent that maps to a single position within the genome. STS sequences are usually on the order of 200–500 bases in length. dbSTS is a division of GenBank devoted to STS sequences; it is intended to facilitate cross-comparison of STSs with sequences in other divisions for the purpose of correlating map positions of anonymous sequences with known genes.

Telnet An Internet protocol or application that allows users to connect to computers at remote locations and use these computers as if they were physically sitting at that computer.

URL Uniform resource locator. Used within Web browsers, URLs specify both the type of site being accessed (FTP, Gopher, or Web) and the address of the Web site.

user The person using client-server or other types of software.

World Wide Web A document delivery system capable of handling various types of non-text-based media.

INDEX

AACompIdent, 254–255
AACompSim, 254–255
ABI PRISM linkage mapping, 131
Abstract Syntax Notation 1 (ASN.1), 23, 158. *See also* ASN.1 entries
Abstract view, 160
Accession.version combined identifier, 30
Accession numbers, 28–29, 52–53, 57
  on protein sequences, 30–31
ACeDB database, 79, 141
ADAM 10 gene, 294
ADIT service, 88–89
ADSN.1 data model, 23
Advanced Research Projects Agency (ARPA), 2
Advanced structure modeling, 103
Affine gap penalties, 197
Aldolase, 384
Alias names, 9
Aligned sets of sequences, entering on Sequin, 73
Alignment gaps, 38–39, 189
Alignment methods, 329
  optimal, 193–195
Alignment parameter estimation, 331
Alignment quality, assessing, 218–219
Alignments
  classes of, 39
  collections of, 227–228
  data representations of, 39–40
  spliced, 209–210, 211
  statistical significance of, 198
Alignment visualization tools, 218
Align sequences, multiplying, 217–222
Allelic variants list, 183
ALSCRIPT alignment tool, 218
ALSCRIPT program, 222–223
Alu-warning entries, 206
AMAS output, 226
America Online (AOL), 6
Amino acid composition, PROPSEARCH query results based on, 256
Amino acids, 254
  "constellations" of, 255
  substitution rates between, 338–339
Amino acid sequences, 57
Among-site substitution rate heterogeneity, models of, 337–338
AMPS, 221
Analysis of Multiple Aligned Sequences (AMAS), 223–225. *See also* AMAS output
Analysis tools, for protein multiple sequence alignments, 222–227
"Anchored" clusters, 288
Animal gene maps, 119
Animal Genome Database, 130
Animation feature, 102
Annotation/editing functions, in Sequin, 76
Annotation page, 73
Annotations
  entering in Sequin, 71
  manual, 77
  within records, 54–55
Annotators, 79
Anonymous FTP, 11
Archival databases, 47
Arithmetic, in Perl, 418–419
Arithmetic operators, 418
ARPANET, 2
ArrayDB, 404, 405
Arrays
  assigning to lists, 444
  in Perl language, 441–443
Artificial sequences, 67
ASN.1 data description language, 77, 95. *See also* Abstract Syntax Notation 1 (ASN.1)
ASN.1 file, 49
asn2ff program, 49
Assembly
  global, 306–307
  preparing readings for, 308–310
Assembly/finishing methods, 303–322
Assembly software, 306
Assignment shortcut operators, 439
Asynchronous DSL (ADSL), 5
Atlas of Protein Sequences and Structures, 46
Atoms, 85–86
  coordinate data for, 84
Author affiliations, 25–26
Author names
  in databases, 25–26
  hyperlinked, 160
Autodiscrete gamma correction, 338

B. burgdorferi, 381
BAC, 144. *See also* Bacterial artificial chromosome (BAC)
BAC clones, 129, 134, 144, 145, 236, 238, 240, 241
Bacillus subtilis, databases dedicated to, 366
Background frequencies, 195

BAC libraries, 135

Bacterial artificial chromosome (BAC), 114. *See also* BAC entries

Banbury Cross Web site, 247

Banerjee-Basu, Sharmila, 253

BankIt system, 70

Barton, Geoffrey J., 215

Base call accuracy estimates, 305–306

Bases, substitution rates between, 335–337

Basic Local Alignment Search Tool (BLAST), 41–42, 156, 156, 202–204. *See also* BLAST entries

Baxevanis, Andreas D., 1, 155, 233, 253

“Benchmark” sequences, 247

β-pleated sheet, 264

β-strands, 264

β-turn, 264

Bibliographic databases, 25

BIND database, 90

Bioinformatics:
  efforts, NCBI support for, 24
  Internet and, 1–17

Bioinformatics Research Center, 140

Biological analysis, using Perl language to facilitate, 413–449

Biological annotation, 65

Biological databases
  citation formats of, 26
  information retrieval from, 155–184, 155–184

Biological sequences (BIOSEQs), 31–34

Biomolecular structures, 83. *See also* Structures

Bioperl, 449

Biopolymers, 84–85

BIOSCI newsgroups, 9

Bioseqs, classes of, 32–33

Bioseq-sets, 34–35

BioSource (biological source), 40, 55, 56

BioSource descriptors, 40

Bipolymer structure data, visualizing, 97

BITNET, 2
  BLAST, 20, 203
  BLASTN program, 202, 286
  BLASTN search, 51, 67
  BLASTP program, 202
  BLASTP search, 208, 209, 257
  BLAST searches, 135, 144, 365. *See also* Basic Local Alignment Search Tool (BLAST)
  BLAST search service, 41
  BLAST2, 355
  BLAST Web interface, 200
  BLASTX program, 202

Blocks, conditional, 427–430

BLOCKS database, 195, 261–262

BLOSUM matrices, 335, 338, 340

BLOSUM substitution matrices, 195–197

BLOSUM30 matrix, 196

BLOSUM62 matrix, 196, 197, 208

BLOSUM90 matrix, 196

BMP format, 14

Bond feature, 38

Bonding approach, explicit, 86–87

Bonfield, James K., 303

Books link, 160–165

Boolean operators, 160

Bootstrapping, 347–348

Bovine Genome Database, 139–140

Branch, in phylogenetics, 324

Branch-swapping algorithms, 345

Brinkman, Fiona S. L., 323

Brookhaven Protein Data Bank, 91

Browsers, 13, 14. *See also* Web browsers

Bulk E-mail, 8

Burge, Chris, 240

Buried residues, prediction of, 225–227

Cancer Chromosome Aberration Project (CCAP), 129

Cancer Genome Anatomy Project (CGAP), 285, 296–297, 399

Case-sensitive mode, 439

cDNA, 51

cDNA arrays, 397

cDNA clones, 204

cDNA data, 140

cDNA library, 284

cDNA sequencing, 394, 283283

CDS feature, 56–57

CDS intervals, 69, 71

Cedars-Sinai Medical Center, 130

Center for Biological Sequence Analysis, 272

Center for Inherited Disease Research (CIDR), 131

Center for Medical Genetics (Marshfield Medical Research Foundation), 131

centiMorgan (cM) coordinates, 145

Centirays (cR), 117

Centre d’Etude du Polymorphisme Humain (CEPH), 116

CEPH/Généthon Group, 134

CEPH/Généthon YAC project, 135

CEPH MegaYAC library, 134

CEPH pedigrees, 130

CEPH YAC map, 118

CGAP xProfiler, 297. *See also* xProfiler

Character-based tree-building methods, 340, 343–345

Character data, randomized, 347

Character state, 329

Character-state weight matrices, 335

Character weight matrix, 336

Chemical graphs, 85, 87, 94, 97

Chemistry rules approach, 85

Chime program, 101

Chimeric clones, 286

CHLC markers, 130

Chomp function, 421

Chromosomal features map, 180

Chromosomal mapping data. *See* Cytogenetic resources

Chromosome-specific genomic data, 141

Chromosome-specific linkage maps, 130

Chromosome Web pages, 140

Citation formats, 26, 160

Citations, accuracy of, 68

Clade, 324

Cladistic analysis, 323–324, 340

Claverie, Jean-Michel, 247

Client-server applications, 9

Client-server Entrez implementation, 158

Clipboard hyperlink, 172

Clone, defining map position from, 145–146

Clone mapping, 118, 129
CLUSTAL guide tree, 332–333
CLUSTAL W, 12, 217, 221, 227, 329
Cluster analysis, 407
Clustering
  EST, 288–293
  hierarchical, 407–409
Clusters of Orthologous Groups (COGs), 361, 363, 371, 368
  multidomain proteins in, 379–380
Cn3D:
  database viewer, 95
  plug-in, 172
  structure, 106
  structure viewer, 101–102, 174
Coaxial cable, 4
Coding Region (CDS) feature, 37. *See also* CDS entries; Coding sequence (CDS)
Coding regions, incomplete, 76
Coding sequence (CDS), 46, 47, 49, 67, 68–69
Coding sequence region features, 34–35
COGNITOR program, 368, 369, 371, 378
COG phylogenetic search tool, 370, 377
COILS algorithm, 269–270
Comparative genome analysis, 359–390
Comparative genomics, 188
  application of, 382–385
  databases for, 360–365
Comparative Mapping by Annotation and Sequence Similarity (COMPASS), 139
Comparative maps, 119, 121
Comparative predictive methods, 235
Comparative resources, 138–140
Composition, protein identity based on, 254–257
"Compressions," 304
CompuServe, 6, 14
Computational tools, for expression analysis, 399–407
Compute pI/MW tool, 257–258
Computer-aided design (CAD), 97
Computer-aided design (CAD) software, 102
Computer-generated sequences, 67
CON "contig" division, 21
Conditional blocks, 427–430
CON division, 51
  record, example of, 63
Confidence values, use of, 305–306
CONSENSE program, 350
"Consensus approach," 277
"Conservative substitutions," 195
Conserved Domain Database (CDD), 262–263, 373
Conserved gene strings, analysis of, 378–382
Consistency Display, 316
Content-based predictive methods, 235
CONTIG:
  Comparator, 312–313, 314
  Editor, 316–318
  Join Editor, 312, 319
  Selector, 311–312
Contigs, 307
"Continuous gamma," 338
Cooperative Human Linkage Center (CHLC), 130
Coordinate data, 84
"Core substructures," 157
Coriell Cell Repositories, 129
Correlated disorder, 99
Correlation coefficient, 246
CrazyQuant, 397
CRI-MAP program, 131
Cross-referencing, 57
Crystallographic correlated disorder, 100
Cubby service, 165
Curated databases, 47
Curators, 79
Cytogenetic banding, 120
Cytogenetic localizations, 145
Cytogenetic maps, 115
Cytogenetic resources, 128–130

DALI algorithm, 106, 275–277
Data analysis. *See* Phylogenetic analysis
Database accession number, 66
Database cross-reference, 57
Database formats, 47–49
Database hits, statistically significant, 386
Database identifier numbers, 120
Databases. *See also* Biological databases; Medical databases; Structure databases
  archival versus curated, 47
  error propagation and incomplete information in, 385–386
  organism-specific, 365–366
  primary and secondary, 47
  protein-only submissions to, 69
  submission process to, 66–67
  submitting DNA sequences to, 65–81
Database searches, 89
  false positives and false negatives in, 386
Database searching artifacts, 204–208
Database similarity searching, 198–200
Database structure viewers, 100–103
Data description languages, 94
Data dictionaries, 86
Data discovery, 24
Data error, 120
Data files, editing, 95
Data model
  consequences of, 77–79
  using, 19–20
Data repositories, 122–127
Data retrieval, 24
Data transfer rates, 4
dbEST database, 126, 288
dbSNP database, 127, 175
dbSTS database, 126
dbSTS Web site, 136
DDBJ, 49. *See also* DNA Data Bank of Japan (DDBJ)
  submitting sequence data to, 80
DDBJ records, 28
DeArray software, 397
Decision making, in Perl language, 424–427
Def (definition) line, 52
Defined regions, determining and ordering contents of, 143–145
Delta Bioseq, 34
Descriptors, 34, 40, 54, 55, 77
Dideoxy nucleotides, 304

Differential genome display, use of phylogenetic patterns for, 372–373
Digital Differential Display (DDD), 296–297, 399
Digital subscriber lines (DSL), 2, 5
Disassemble Readings function, 319
Discontinuous alignments, 39–40
Discrete gamma approximation, 338
Distance-based methods, 341
Distance-based tree-building methods, 340
Distance-based tree-building procedures, evaluating, 342–343
Distance measures, 121
Distributed document delivery systems (DDDSs), 13
DNA, coding region on, 20
DNA-centered report, 20
"DNA-centered" view, 41
DNA clones, in genomic maps, 114–115
DNA coordinate system, 41
DNA Data Bank of Japan (DDBJ), 46, 360. *See also* DDBJ entries
DNADIST program, 350
DNA features, 20
DNA fingerprinting, 114–115
DNA markers, in genomic maps, 113
DNAML program, 350
DNA probes, 396
DNA/RNA sequences, submission to databases, 67–69
DNA sequences, 66, 69, 144. *See also* Genomic DNA; Sequences
  databases, 46
  defining map position from, 145–146
  homology, 119
  mapping and, 135–136
  multiple protein alignment from, 222
  predictive methods using, 233–251
  submitting to databases, 65–81
DNA sequence tracts, 145
Domain fusions, examination of, 373–378
Domain names, 3
Domain name servers, 2
Dot matrix path graph, 192
Dot matrix representations, 191–192
Dot matrix sequence comparison, 191
DRAWGRAM program, 349
DRAWTREE program, 349
Drosophila projects, 180
Drug targets, 372–373
Dynamic programming, 193

EBI BioCatalog, 15
EcoGene, 365
"Edge effect," 198
Editing functions, on Sequin, 76
eGenome project, 120, 129, 137, 144, 145
eGenome Web site, 136
Electronic data submission, 67
  citing, 27
Electronic mail (E-mail), 7–10
Electronic mapping, 140
Electronic PCR (e-PCR), 113
E-mail addresses, 7, 54
E-mail attachments, 10–12
E-mail servers, 8
EMBL, 49, 50. *See also* European Molecular Biology Laboratory (EMBL)
  submitting sequence data to, 80
EMBL flatfile format, 61–63
EMBL records, 28
EMBL sequence database, 21
Emory University, 142
Enolase, 384–385
Entrez Boolean search statements, 162
Entrez discovery pathway, 159–172
Entrez Genomes division, 21, 31, 35, 42, 112, 123, 144, 360
Entrez Genomes Web site, 144
Entrez map view, 177
Entrez Map Viewer, 175
Entrez queries, 160
  entries resulting from, 170
  text-based, 161
Entrez searches, implementations for, 158–159
Entrez sequence retrieval program, 42
Entrez system, 20, 46, 92, 156–172. *See also* Network Entrez
  formulating a search against, 167
  History feature of, 169
  Limits feature of, 168
  relationships in, 159
  structures accessible through, 173
Enzyme Commission, 382–383
ENZYME database, 382, 383
e-PCR, 123, 125, 145
e-PCR Web interface, 144
Equivalent Bioseqs, 35
Error propagation, in databases, 385–386
Escherichia coli, 359
  databases dedicated to, 365
est2gen program, 210
EST clustering, 288–293. *See also* Expressed sequence tags (ESTs)
EST database (dbEST), 113. *See also* dbEST database
est genome program, 210
EST records, bulk-submission protocol for, 66
EST sequences, 204, 222
Ethernet, 4
Eudora, 8
EUROFAN project, 366
European Bioinformatics Institute (EBI), 46, 132
European Molecular Biology Laboratory (EMBL), 46. *See also* EMBL entries
European Nuclear Research Council (CERN), 13
E values, 198, 386
Evolution, point accepted mutation (PAM) model of, 195
Evolutionary alignment, 217
Exons, 234, 236, 238, 241, 246
  optimal and suboptimal, 240
ExPaSy server, 254, 258
ExPaSy Web site, 382
ExPdb, 275
Experiment files, 307, 308, 311
Explicit bonding approach, 86–87
Explicit sequence, 89
Expressed sequence tag (EST) data, limitations of, 286–288

Expressed Sequence Tags (ESTs), 51, 112, 283–299, 395–396. *See also* EST entries
accessing, 285–286
creation of, 117–118
defined, 284–285
Entrez view of, 287
gene discovery and, 294
gene expression and, 296
sequence polymorphisms and, 296
use in gene prediction, 295
Expression analysis, computational tools for, 399–407
Expression databases, sources for, 402–404
Expression platforms, measurements reported by, 396–399
Extreme value distribution, 198

FASTA:
file format, 41, 48, 49, 73, 92, 264, 436
files, 52, 440
program, 200–202
search, output of, 199
views, 21
FastDNAml program, 353
Feature coordinates, 76
Feature propagation, 76
Feature table, 55–58
documentation for, 69
Fetch, 12
FGENEH, 236–238
FGENES, 238
Filehandles, 422–424, 434
FileMaker Pro, 405, 406
File Transfer Protocol (FTP), 10–12
Fingerprinting, 118, 135, 145
Fitch-Margoliash (FM) method, 341, 342
FITCH program, 349
Flanking markers, 142, 143, 145
Flatfile format, 48
Fluorescence in situ hybridization (FISH), 115, 129
Fluorescent intensity measurements, 398
FlyBase database, 180
FlyBase query, genes view from, 181
Folding classes, 263–269
FORTRAN, 94
FSSP database, 277
Fully qualified domain name (FQDN), 2
Functional information, transfer of, 368

GaIT protein, 209
Galperin, Michael Y., 359
Gamma distribution models, 338
gap4 commands, 312
gap4 database, 307
gap4 program, 308, 311, 319
experiment types in, 319–321
Gap-opening penalty, 197
Gapped alignment programs, 42
Gap penalties, 195–197, 202
Garnier-Gibrat-Robson (GOR) method, 268
GDB BLAST, 123
GDB genomic catalogue, 137
GDB Mapview program, 121
GDB Web site, 136
Gel readings, 304
GenBank, 20, 144, 155. *See also* GenBank sequence database
accession numbers, 177
division code, 51
flatfile (GBFF), 20, 21, 49–58
format of, 59–61
format, 21
formatted records, 22–23
GenPept database, 178
information releases from, 49
map, 125
patented sequences in, 26
records, 28, 47, 56
citations for, 54
release schedule for, 49
sequence database, 45–63
format of, 41
submitting sequence data to, 80
Gene (domain) fusions, examination of, 373–378
Gene-based comparative maps, 119
Genebridge4 (GB4), 132, 294
GeneCards, 120, 177, 404
Gene discovery, ESTs and, 294
Gene evolution, 327
Gene expression, assessing levels of, 296
Gene Expression Database (GXD), 127
Gene Expression Omnibus (GEO), 400
Gene features, 36, 41, 57–58
Gene-finding strategies, 235
GeneID, 245, 248
Gene identification, 233–234
Gene indices, 293
Gene lists, 122
GeneMachine, 249, 250
GeneMap '99, 118, 125, 133
Gene neighborhoods, 378
Gene nomenclature, 68
GeneParser, 245, 248
Gene prediction, in genomic DNA, 295
GeneQuiz, 367
General Seq-id, 31
Gene Recognition and Analysis Internet Link (GRAIL), 235–236
Generic Top Level Domain Memorandum of Understanding (gTDL-MOU), 3
Generic top-level domains (gTDLs), 3
Genes view, from a FlyBase query, 181
Genetic codes, 68–69
Genetic linkage (GL), 120
Genetic linkage (GL) maps, 115–116
Genetic linkage map resources, 130–131
Genetic Location Database (LDB), 121, 129
Genetics Institute, University of Bari, Italy, 129
GenInfo (gi) numbers, 29–30, 48
Genome analysis, 366–382. *See also* Large-scale genome analysis
comparative, 359–390
Genome annotation, common problems in, 385–387
Genome centers, 79
Genome Channel Web tool, 136
Genome comparison, for predicting protein functions, 367–382
Genome context, 367
as a source of errors, 386–387

Genome Database (GDB), 112, 122–123, 144, 145, 175
  as a map repository, 143
  cytogenetic information on, 129
Genome projects, 111
GenomeScan, 241
  versus GENSCAN, 244
Genome sequencing, 360–366
Genome Sequencing Center, 284
Genome Survey Sequences (GSSs), 51
Genome Therapeutics Corporation, 141
Genomic Biology page, 127
Genomic catalogues, 137
Genomic cataloguing, 120, 121, 136–138
Genomic DNA, gene prediction in, 295
Genomic mapping, 111–149
  complexities of, 120–122
Genomic maps, 112
  elements of, 113–115
  types of, 115–119
Genomic records, syntax for, 52
Genomic regions, using mapping resources for defining, 142–143
Genomic research, community-based approach to, 141
Genomics, comparative, 360–365
Genomic Segments, 122
Genomic sequence tracts, 144
Genotator, 249
GenPept format, 41, 46, 49
genQuest program, 236
GENSCAN, 240–241, 248, 249
  output, 242, 243
  Web site, 241
Généthon markers, 130–131
gi (GenInfo) identifier, 53, 57
gi (GenInfo) numbers, 29–30, 48
Giemsa, 115
Global assembly, 306–307
Global sequence alignment, 188–189
Glucose-6-phosphate isomerase, 383
Glyceraldehyde-3-phosphate dehydrogenase, 384
Glycolysis, 382, 383–385
  in *H. pylori*, 385
GM99 Web site, 136
Gopher, 13
GRAIL-EXP, 236
GRAIL2, 236, 239, 248
Grand average of hydrophobicity (GRAVY), 258
Graphical format, in Sequin, 75
Graphical user interface (GUI), 11–12
Graphical view, 74
Graphical viewers, 158
Graphics, presentation, 102–103
Graphs, 40
GRASP, 276
Gribskov collection, 261
Gribskov method of protein analysis, 260
G3 RH (Stanford Generation 3) panel, 132
Guide trees, 219, 329

H. pylori, glycolysis in, 385
Hard link concept, 158
Hard masking, 206
Hash data type, 445–446
Hashes, 441
Headers, 50–55
Heterogeneity, 337
Hidden Markov models (HMMs), 208, 246
Hierarchical alignment software, 221
Hierarchical clustering, 407–409
Hierarchical dendrograms, 407, 408
Hierarchical methods, for automatic multiple alignment, 219–220
High-resolution linkage maps, 131
"High-scoring segment pairs" (HSPs), 156
High-throughput genome sequences (HTGS), 34, 51
History hyperlink, 165
HMMERV software, 208
HMMgene, 246, 249
Hogue, Christopher W. V., 83
HomoloGene, 138, 175, 289
Homologous sequences, 188, 327
Homologs, 289, 327
Homology, 188
Homology maps, 138
Homology model building (threading), 274–275
Homo sapiens Genome View page, 125
HotBot search engine, 15
Human Chromosome 1 Web site, 141
Human gene map, 294–295
Human genome, "working draft" of, 199, 233
Human genome draft sequence, 112
Human Genome Map Viewer, 124–125
Human Genome Organization (HUGO), 140
Human Genome Project, 1, 22, 233
Human markers, 145
Human physical mapping, 134
Hybridization, 397, 398
Hyperlinks, 13
Hypertext markup language (HTML), 14
Hypertext transfer protocol (HTTP), 13

I.M.A.G.E. Consortium, 285
Imagene, 367
Imperial College, 141
Implicit sequence fragments, 90
Implicit sequences, 89, 90
indels, 329
INFOBIOGEN, 129
Informative sites, 329
Input
  combining loops with, 432–433
  in Perl language, 420–421, 433–435
In silico mapping, 139
Institute for Genome Research, The (TIGR), 135, 144, 360. *See also* TIGR Gene Indices
Integrated information retrieval, 156–172
Integrated maps, 119, 121, 136–138
Integrated services digital network (ISDN), 4
International Nucleotide Sequence Database Collaboration, 28, 30, 46, 49, 57
International Standards Organization (ISO), 23
Internet
  bioinformatics and, 1–17
  connecting to, 4–7
  fundamentals of, 2–4
  versus Intranet, 14
  wireless connection to, 5–6

TEAMFLY

Team-Fly®

Internet-accessible phylogenetic software, 354–356
Internet content providers, 6–7
Internet Explorer, 10, 14
Internet Protocol (IP), 2
Internet resources, 16–17, 58–59, 106, 146–148, 183–184, 212, 228, 250–251, 277–278, 298, 321, 356–357, 387–389, 410, 449
Internet service providers (ISPs), 6–7
Internet Software Consortium (ISC), 3
Internet sources, 80–81
Introns, 234
Invariants model, 338
IP addresses, 2
IPW protein, 248

JalView application, 218, 221, 227
Java-based viewer, 101
JNET program, 226, 227
Join function, 444–445
Joint Genome Institute (JGI), 141
Journal articles, 26
JPEG format, 14
JPEG graphics, 355
JPred, 227, 269
JPred server, 222
Judge, David P., 303
"Junk mail," 8

Kans, Jonathan A., 19, 65
Karlin, Sam, 240
Karsch-Mizrachi, Ilene, 45
Keys functions, 446
Key terms, weighted, 157–158
Keywords, 53
Kishino-Hasegawa test, 353
Koonin, Eugene V., 359
"Kringle domain," 190, 192
ktup parameter, 200
Kyoto Encyclopedia of Genes and Genomes (KEGG), 361, 364, 381, 383, 405

Lalign program, 194
Landsman, David, 283
Large-scale gene expression, technologies for, 394–399
Large-scale genome analysis, 393–410
Lawrence Berkeley National Laboratory-University of California, San Francisco, Resource for Molecular Cytogenetics, 129
LDB Web site, 136
Legal features, table of, 38
Leipe, Detlef D., 323
Leucine codons, 339
Levin homolog method, 268
"Likelihood of the tree," 344
Likelihood ratio tests, 348
Linear discriminant analysis, 238
LINKAGE, 116
Linkage maps, 116
LinkOut link, 165
Links, 89
Lists, in Perl language, 444
Loansome Doc service, 165
Local alignments, 191
  optimal and suboptimal, 194
Local area network (LAN), 7
Local Seq-id, 31
Locus ID, 174
LocusLink, 43, 65, 120, 125, 126, 129, 135, 138, 145, 172–178, 291
  as a query interface, 172–174, 175
LocusLink report view, 174, 176
LOCUS name, 21, 28, 50
Lod (log of the odds) score, 116
Log-det transformation, 336
Log-odds approach, 195
"Long branch attraction," 344
LOOK tool, 275
Loops
  combining with input, 432–433
  in Perl language, 430–432
Loop variable, 443
Low-complexity regions (LCRs), 206–207
Lupas COILS method, 270
Lycos, 15
Lynx, 14

M. jannaschii, 383–384
MACAW program, 329
MACCLADE program, 353–354
MacPerl, 415
MacroMolecular Chemical Interchange Format, 94–95
"Macroscopic" software tools, 102
MacStripe, 270
MAGPIE, 367
MALIGN program, 329
Mammalian genomes, 394
MAP, 116
Map Bioseq, 34
MAPMAKER, 116
  MAP-O-MAT Web server, 116
  MAP-O-MAT Web site, 131
Mapping. *See also* Genomic mapping
  interplay with sequencing, 112–113
  sequence-first approach to, 136
Mapping data, MGD, 127
Mapping groups, 120
Mapping projects, 127–142
Mapping resources, uses for, 142–146
Map repositories, 143
Maps, 113
Maps within a Region search, 122–123, 124
Mapview display, 123
Map Viewer, 124–125, 175
Mapviewer utility, 144
Marker names, catalogue of, 137
Markers. *See also* Flanking markers
  genotyped, 130
  ordering, 121
  ordering/labeling error for, 120
  recombination between, 115–116
Marker/sequence links, 135
Marshfield Medical Research Foundation, 131
Masking, 206
Mass spectrometric (MS) techniques, 257
Master sequence, 39
Mathematical models, 19–20
Matise, Tara C., 111
MATLAB software, 408

MaxHom algorithm, 265, 267
Maximum likelihood (ML):
  method, 344
  tree building, 335
Maximum parsimony (MP), 334, 335, 343–344
Max Planck Institute for Molecular Genetics (MPIMG), 129, 141
MAXTREES, 352–353
McKusick, Victor, 180
Medical databases, 181–183
MEDLINE:
  database, 25, 26, 27, 92, 156
  identifier, 54
  MEDLARS layout, 160
  unique identifier (MUID), 27
MEGA program, 341, 342–343
Meiotic maps, 115–116
Meltzer, Paul S., 393
MeSH (medical subject heading) terms, 160
Metabolic pathways, reconstruction of, 382–385
Metacharacters, 436–437
Meta-search engines, 15–16
METREE program, 342
Microarray hybridization, 394, 404
Microarray production, informatics aspects of, 395–396
Microarray results, display of, 406
Microarrays, 297
Microbial Genome Database (MBGD), 361–365
Microsatellites, 114
Microsatellite sequences, 207
Microsoft Exchange, 8, 10
Microsoft Internet Explorer. *See* Internet Explorer
Mindspring, 6
"Minimal genome," 366
Minimum evolution (ME) method, 342
MIPS database, 178
Mitochondrial genome, 142
mmCIF (MacroMolecular Chemical Interchange Format):
  dictionary, 94
  file format, 94–95
  file retrieval, 89
MMDB file format, 94, 95. *See also* Molecular Modeling Database (MMDB)
MMDB standard residue dictionary, 95
MMDB viewer, 101–102
Modeling software, 87
Modules, Perl, 449
Molecular Biology of the Cell (Alberts), 165, 166
Molecular Modeling Database (MMDB), 86, 90–93. *See also* MMDB entries
Molecular populations, 97
Molecular structure data, three-dimensional, 84
Molecular Weight Search (MOWSE), 257
Molecule (mol) type, 51
Molecule information (MolInfo), 40–41
MolInfo descriptor, 40–41
MOLPHY shareware package, 353, 354
MolScript programs, 101
Monomorphic markers, 113
Monte Carlo test of significance, 218–219
month database, 203
MoST program, 367
Mouse Genome Database (MGD), 112, 127, 138, 143, 289
Mouse Genome Initiative (MGI) Database, 127
Mouse Genome Sequence (MGS) project, 127
Mouse markers, 145
mRNA, 51, 58, 69, 284
mRNA alignments, 210
mRNA definition line, 52
mRNA feature, 36
mRNA sequence, 146
MTIDK scoring matrix, 270
MTK scoring matrix, 270
MULPRED, 269
Multidimensional scaling (MDS), 408
Multidimensional scaling plot, 409
MultiMap, 116, 117
Multiple alignment programs, 218
Multiple alignments, 187. *See also* Protein multiple sequence alignments
  collections of, 227–228
  using PSI-BLAST, 222
Multiple
  protein alignment, from DNA sequences, 222
Multiple representation styles, for structural information, 95–97
Multiple sequence alignment, 216
  prediction from, 225–227
Multiple sequence studies, 73
Multiple viewers, 75
Multipoint linkage analysis, 117
Mutation frequencies, 195
Mutation studies, 69
Mycoplasma genitalium databases, 366
MZEF (Michael Zhang's Exon Finder), 238–240, 248, 249

Naming conventions (schemes), 307
Nara Institute of Technology, 365
National Cancer Institute, 129
National Center for Biotechnology Information (NCBI), 2, 20, 112, 155. *See also* NCBI entries
  core data elements used by, 24
  molecular modeling database at, 91–93
  structure query from, 93
National Center for Biotechnology Information (NCBI) data model, 19–43
  examples of, 20–23
  using, 41–43
National Human Genome Research Institute, 141
National Institutes of Health (NIH), 2, 7. *See also* GenBank
NCBI Citation Matching Service, 26. *See also* National Center for Biotechnology Information (NCBI)
NCBI:
  data repository, 123–127
  Desktop, 78–79
  genomic catalogue, 137
  Map Viewer, 135
  Software Toolkit, 21
  Structure division, 92
  toolkit, 95
NDB Protein Finder, 94
Needleman-Wunsch algorithm, 193
Neighborhood words, 202

Neighboring concept, 156–158
Neighbor-joining (NJ) algorithm, 341–342
Neighbor-joining trees, 221
NEIGHBOR program, 349, 355
Netscape Communicator, 14
Netscape Messenger, 8
Netscape Navigator, 10
Network Entrez, 158
Neural networks, 264
Newsgroups, 9–10
NewsWatcher for the Macintosh, 9–10
NIT gene, 248
NMR models, 97–99
NNI swapping, 352
nnpredict algorithm, 264–265
Node, 324, 324
Nomenclature
  challenges of, 120
  difficulties with, 145
"Noncoding RNA genes," 248
Nonglobular regions, 273–274
Nonhierarchical methods, for automatic multiple alignment, 221–222
Nonhuman resources, 130
Nonorthologous gene displacement, 372
Nonparametric models, 338
nr database, 203
N-scores, 261
Nuclear magnetic resonance (NMR), 83
Nucleic Acids Database, 89
Nucleic acid sequence databases, 45
Nucleotide databases, 46, 68–69
Nucleotide page, 72, 73
Nucleotide/protein sets, 34–35
Nucleotide records, 49
  RefSeq, 177
Nucleotide sequence databases, for use with BLAST, 204
Nuc-prot set, 20, 34–35
Number data types, 417
Numeric comparison operators, 426

Objects, Perl, 449
Online Mendelian Inheritance in Man (OMIM), 129, 181, 404
  allelic variants obtained through, 182
  searches using, 183
Operators, 418
  assignment shortcut, 439
  numeric comparison, 426
Operons, analysis of, 378–382
Optical mapping, 118
ORF Finder, 69
Organismal context, as a source of errors, 386–387
Organism qualifier, 56
Organism-specific databases, 365–366
Orthologous genes, 119
Orthologs, 327, 328, 361, 368, 381, 382
  identifying, 327
Ortholog search server, 355
Ostell, James M., 19
Ouellette, B. F. Francis, 45, 65
Outgroup rooting, 346
Output, in Perl language, 420–421, 433–435
OWL database, 257, 269
Oxford Grid, 138, 139

PAC, 144
PAC clones, 134
PAC domain, 378
Pairwise alignment, 187
Pairwise sequence comparison, 337
PAM distances, 339
PAM matrices, 335, 338
PAM250 scoring matrix, 196
Paralogs, 327, 328, 361, 368. *See also* Orthologs
  identifying, 327
Parametric bootstrap, 348
Parsimony, 339, 340. *See also* PAUP software
Parsimony programs, 350
PAS domain, 378
Patents, as bibliographic entities, 26–27
Path graphs, 192–193
  dot matrix, 192
Pattern matching, in Perl language, 436–439
Pattern-matching facility, 447
Patterns, 260
  extracting, 440–441
Pauling, Linus, 263
PAUP (phylogenetic analysis using parsimony), 339, 341, 348
PAUP software, 352–353
PCR-based analysis, 114. *See also* Polymerase chain reaction (PCR) primers
PCR-based markers, 113
PDB chain identifiers, 90. *See also* Protein Data Bank (PDB)
PDBeast project, 91
PDB sequences
  BLAST against, 92
  file format, 94
  file retrieval, 89
  file viewers, 90
  ID codes, 89
  parsing software, 94
  query and reporting, 87–89
  validating, 90–91
PEDANT Web resource, 360–361, 362, 367
PeptideMass, 258
Peptides, signal, 272–273
Perl language
  example of using, 446–449
  pattern matching in, 436–439
  use in facilitating biological analysis, 413–449
Perl modules, 249
Perl scripts, 414, 415, 416
Permutation tail probability test (PTP), 347
Permutation tests, 347
Pfam database, 208, 260, 263, 373, 378
Pfam server, 227
PHD method, 265, 277
PHDtopology method, 272
Phenylalanine codons, 339
Phosphofructokinase, 383–384
Phosphoglycerate kinase, 384
Phosphoglycerate mutase, 384
Phrap engine, 307
Phrapview, 308–310
Phred-style confidence values, 305

PHYLIP (phylogeny inference package) software, 221, 339, 349–351, 348
bootstrap analysis with, 350–351
PhyloBLAST software, 355
PHYLODENDRON, 354
Phylogenetic analysis, 323–357
alignment modification for, 330–331
steps in, 327–328
Phylogenetic data model, 325–327
building, 329–333
Phylogenetic data set, extraction of, 333–334
Phylogenetic patterns (profiles), 368–372
complementary, 374–376
use for differential genome display, 372–373
Phylogenetic sequence data, 329
Phylogenetic software, 348–354
Internet-accessible, 354–356
Phylogenetic studies, 35, 69, 73
Phylogenetic trees, 325, 326. *See also* Trees
evaluating, 346–348
rooting, 346
Physical maps, 118
Physical properties, sequence-based, 257–259
PileUp, 329
PIR, 28. *See also* Protein Information Resource (PIR)
PIR format, 269
Point accepted mutation (PAM) model of evolution, 195
polyA signals, 240
Polymerase chain reaction (PCR) primers, 50. *See also* PCR entries
Polymorphic markers, 113
in genomic maps, 113–114
Polymorphisms, sequence, 296
P1-artificial chromosomes, 114
Population studies, 35, 69, 73
Portals, 6
Position-specific Iterated BLAST (PSI-BLAST), 208–209. *See also* PSI-BLAST entries
Position-specific scoring matrices (PSSMs), 208–209
Position-specific scoring table (PSST), 260
PostScript files, 227
PostScript ribbon diagrams, 101
PostScript viewer, 355
POV-Ray software, 103
PowerBLAST, 35, 249
PREDATOR, 269
PREDATOR algorithm, 267–268
Predictions, 254
Predictive methods
effectiveness of, 246–248
using DNA sequences, 233–251
using protein sequences, 253–278
PredictProtein, 265–267, 269
pregap4 program, 308, 309
Presentation graphics, 102–103
Primary accession number, 53
Primary databases, 47
Primer, 304
Print function, 420
Print resolution, 103
PRINTS database, 261
“Probabilistic model,” 240
PROCRUSTES, 241–245, 249
ProDom, 373
Profiles, 260
ProfileScan, 260–261
ProPack, 329, 333
PROPSEARCH, 255–257
PROSITE, 260, 261
PROTDIST program, 339, 349, 350, 355
Protein accession number, 57
Protein architecture, 83
Protein-based comparative maps, 119
“Protein-centered” view, 20
Protein-coding DNA sequences, 340
Protein context, as a source of errors, 386–387
Protein Data Bank (PDB), 47, 84, 87–91, 253. *See also* PDB entries
Protein databases, 46, 68
Protein evolution, 327
Protein features, 37
Protein Finder, 94
“Protein-folding” problem, 274
Protein functions, genome comparison for predicting, 367–382
Protein identity, based on composition, 254–257
Protein Information Resource (PIR), 47. *See also* PIR entries
Protein multiple sequence alignments, 215–230. *See also* Multiple alignments
analysis tools for, 222–227
defined, 216
from DNA sequences, 222
Protein neighbor, 171
Protein-only submissions, 69
Protein page, 73
Proteins
assigning functions to, 215–216
modular nature of, 190–193
Protein secondary structure, 215
Protein Sequence Analysis server, 268
Protein sequence databases, 45
for use with BLAST, 203
Protein sequences, 66
accession numbers of, 30–31, 53
predictive methods using, 253–278
submitting, 71
Protein structure, images of, 95, 96
ProtEST server, 222
ProtParam, 258
PROTPARS program, 350
Pseudogenes, 146, 210
PSI-BLAST, 259, 268. *See also* Position-specific Iterated BLAST (PSI-BLAST)
multiple alignment by, 222
PSIPRED program, 226, 269
Pub-equiv, 27
Publications, importance of, 24–27
Public databases, 65, 399–407
Public microarray data, 400
PubMed, 27, 156, 175
neighbors to an entry found in, 164
PubMed Central, 27
PubMed identifier (PMID), 27, 54
PubMed link, 165
PubMed record, example of, 163

PUZZLE program, 348, 353
Pyruvate kinase, 384–385

Quadratic discriminant analysis, 238
Quantifiers, 20, 436, 438
Quartet puzzling, 345
"Query-by-position" tools, 137
Query engines, 87–88
Query sequences, 199, 204
Query vector, 255
Quick Gene Search, 127, 128
QUICKMAP application, 135

R. prowazekii, 381
Radiation hybrid (RH) maps, 117
Radiation Hybrid Database (RHdb), 117, 132, 132
Radiation Hybrid Information web site, 117
Radiation Hybrid Mapping Information Web site, 132
Radiation hybrid map resources, 131–134
Radio histogram, 404
Randomized character data, 347
Randomized trees, 346–347
RasMol, 96, 101
RasMol-based viewers, 90, 101
Rat Genome Database, 131, 137–138
Raw Bioseq, 34
Read pairs, 305
READSEQ tool, 49
Recombination events, 115–116
Records, date made public, 51
References, Perl, 449
Reference Sequence (RefSeq) project, 176–177, 360
RefSeq accessions, 43
RefSeq identifiers, 31
Regional map resources, 140–142
Regional physical maps, 135
Region feature, 38
Regular expressions, 436
Related Articles hyperlink, 160
Related Sequences link, 172
RepeatMasker program, 206, 249
Repetitive sequences, 146
Reports, human-readable, 20
Representation, of structural information, 95–100
Research Collaboratory for Structural Bioinformatics (RCSB), 87–91
  database services of, 87
Research Collaboratory for Structural Biology, 84
Residue dictionaries, 85, 86, 95
Residue identification, 223
Restriction fragment length polymorphism (RFLP), 114
RH Consortium, 132
RHdb, 137
RH framework, 137
RHMAPPER program, 117
RHMAP program, 117
Rhodes, Gale, 103
Ribulose-1,5-biphosphate carboxylase (RUBISCO), 73
RIKEN Genomic Sciences Center, 141
RNA editing, 36
RNA feature, 36, 58
RNAi technique, 413–414
Robison, Keith, 15
RPS-BLAST, 263
rRNA, 51, 58, 338
rRNA genes, genomic sequencing of, 67
rRNA sequences, submission of, 73

S. cerevisiae, 393
  databases dedicated to, 366
Saccharomyces Genome Database (SGD), 178, 366
SacchDB Locus view, 179
SAGE database, 293. *See also* Serial analysis of gene expression (SAGE)
SAGE libraries, 399
SAGEmap database, 297, 399, 400
SAGE tags, 394
Sakura system, 70
SAM program, 329
Sanger Center, 141, 321
Sanger Center Web server, 238
Sanger dideoxy sequencing technique, 304
Satellite sequences, 207
Scaling plot, multidimensional, 409
ScanAlyze, 397
SCF files, 307, 308
Schuler, Gregory D., 187
Scientific notation, 417
SCOP system, 104, 225
Scoring matrices, position-specific, 208–209
Scripts, operation of, 416
SEALS package, 367
Search engines, 15
SearchFields interface, 88
Search forms, customized, 122
SearchLite system, 87–88
secE genes, 371
Secondary accession number, 53
Secondary databases, 47
Secondary structure, 263–269
Secondary structure "mask," 221
Secondary structure predictions, 225–227, 266
Segmented Bioseq, 34
"Segmented sequence," 21
Segment records, 21
Segments, 38
Seg program, 206, 207, 208, 274, 386
"Seg-set," 21
Sensitivity value, 246, 247
Seq-annot, 35–40, 42
SEQBOOT program, 350
Seq-descr, 40–41
Seq-feat (sequence feature), 35–38
Seq-graphs, 40
Seq-locs (sequence locations), 35, 36
SEQRES chemical graph, 94
SEQRES keyword, 89, 90
Sequence-alignment programs, 144. *See also* BLAST
Sequence alignments (Seq-aligns), 38, 187–198. *See also* Protein multiple sequence alignments
  classes of, 39
  data representations of, 39–40
  evolutionary basis of, 188–190

Sequence alignments (*Continued*)
  optimal methods of, 193–195
  statistical significance of, 198
Sequence assembly data, file formats for, 307–308
Sequence assembly/finishing methods, 303–322
Sequence-based comparative maps, 119
Sequence comparison, dot matrix, 191
Sequence contigs, 144
Sequence databases, 46, 178–181. *See also* National Center for Biotechnology Information (NCBI)
  growth of, 66
Sequenced Tag Sites (STSs), 51
Sequence editor, 76
Sequence files, finding the length of, 435–436
Sequence Format form, 72, 73
Sequence history information, 54–55
Sequence Identifiers (Seq-ids), 28
Sequence length, 50
Sequence masking, 206
Sequence Neighbors link, 172
Sequence polymorphisms, ESTs and, 296
Sequence records
  biological annotation of, 65
  viewing on Sequin, 74–75
Sequences. *See also* Coding sequences; DNA sequences; Submissions
  accuracy of, 67–68
  annotating, 35–40
  chemical connectivity and, 85
  collections of, 34–35
  describing, 40–41
  joining, 56
  nature of, 67
  organism derived from, 68
  PDB (Protein Data Bank), 90–91
  physical properties based on, 257–259
  release by scientists, 66
  from structure records, 89–90
Sequence similarity, 188
Sequence-tagged clones (STCs), 115
Sequence-tagged sites (STSs), 112, 113
Sequence tracts, 112, 113
Sequencing, interplay with mapping, 112–113
Sequin, 66
  entering sequences on, 72–73
  navigation in, 77
  sequence analysis in, 78–79
  submissions with, 70–77
  Suggest Intervals function in, 69, 71, 73
Sequin editor, 73
Sequin software, 42
Sequin submission tool, 42
Sequin validator, 75–76
Serial analysis of gene expression (SAGE), 297, 395. *See also* SAGE entries
Servers, 8
Short sequence length polymorphisms (SSLPs), 114
Short tandem repeats (STRs), 114
SignalP, 273
  peptides, 272–273
  server, 448, 449
  site, 413
Signatures, 260
Significance filters, 406
sim4 program, 210
SIM algorithm, 194
Similarity, defined, 188
Simple Modular Architecture Research Tool (SMART), 263, 381. *See also* SMART program
Single-chromosome map resources, 140–142
Single nucleotide polymorphism (SNP), 114, 296
Single nucleotide sequences, entering on Sequin, 72–73
Single-resource maps, 121
Site-based predictive methods, 235
Site feature, 38
Site residues, 188
Sites, 329
Skewness test, 346–347
SMART program, 373, 378, 386. *See also* Simple Modular Architecture Research Tool (SMART)
Smith-Waterman algorithm, 193
Smith-Waterman alignment search, 200
Soft masking, 206
Software tools, 43
Solovyev, Victor, 236
Solvent accessibility, 227
Somatic cell hybrid maps, 118
SOPMA method, 268
Source feature, 55–56
Source line, 53–54
Source organisms, information about, 40
Southeastern Regional Genetics Group (SERGG), 129
"Spam," 8
Specificity value, 246, 247
Spliced alignments, 209–210, 211
Split function, 444–445
Sputnik, 249
STACK resource, 293
Staden, Roger, 303
Staden Package, 305, 307
Stanford Generation 3 (G3 RH) panel, 132
Stanford Human Genome Center, 132
Star decomposition, 342
STAR syntax, 94
Statistical Analysis of Protein Sequences (SAPS), 259
Statistical significance, of alignments, 198
Stein, Lincoln D., 413
Steroid fingers, 261, 262
String comparison operators, 426, 427
String interpolation, 419
Strings, 417
STR polymorphisms (STRPs), 114
Structural alignment, 216–217
Structural completeness, 87, 87
Structural information, visualizing, 95–100
Structure data, three-dimensional, 98–99
Structure database records, 83
Structure databases, 83–109
Structure Explorer, 89
Structure file formats, 94–95
Structure modeling, advanced, 103
Structure Neighbors links, 172
Structure queries, 172

Structure records
  free text query of, 92
  sequences from, 89–90
Structures, 83–91. *See also* Secondary structure; Tertiary structure
  mathematical optimization of, 333
  specialized, 269–274
Structure similarity searching, 103–106
Structure Summary pages, 88, 92, 104
  for MMDB structure records, 92
STS content mapping, 118, 134–135
Subalignments, 223–225
Subintervals, 245
Subject sequence, 199
Submission citations, 27
Submission process, 66
  streamlined, 79
Submissions
  electronic, 67
  for EST records, 66
  protein-only, 69
  Sequin, 70–77
  starting new, 72
  World Wide Web, 70
Submitting Authors form, 72
Subroutines, Perl, 449
Substitution matrix, 195
Substitution models, 335–340
  choosing, 339–340
Substitution rate heterogeneity, 337–338
Substitution rate matrix, 337
Substitution rates
  between amino acids, 338–339
  between bases, 335–337
Substitution scores, 195–197
Subtilist Web site, 366
Subtractive hybridization, 285
Subunits, 372
SWALL protein sequence database, 222
SWISS-MODEL program, 275
SwissPDBViewer, 103, 104
SWISS-PROT database, 28, 46, 47, 54, 68, 69, 253, 254, 257, 258, 265
SWISS-PROT format, 49
SWISS-PROT ID, 261
Synthetic sequences, 67

Table of legal features, 38
"Tag value" pairs, 94
Target control, 77
Target frequencies, 195
TATA box, 234
Taxon, 324
TBLASTN program, 202, 286, 371
TBLASTN search, 205
TBLASTX program, 203, 204, 286
TBR swapping, 352
Tcl language, 321
Technology, large-scale gene expression, 394–399
Template Display, 313–316
Tertiary structure, 274–277
TGREASE, 258–259
Thermodynamic calculations, 275
Threading, 274–275
3DBAtlas, 92
Three-dimensional molecular structure data, 84
Three-dimensional structures, animating, 102
3D viewers, 102
TIGR Gene Indices, 293. *See also* Institute for Genome Research, The (TIGR)
TIGR Orthologous Gene Alignment (TOGA) database, 293
Time-resolved fluorescence spectroscopy, 99
Tissue-type plasminogen activator (PLAT) protein, 190, 192
TMbase database, 271
TMpred method, 271
TNG (The Next Generation) panel, 132
TOPITS method, 277
T-PTP test, 347
Trace Display, 316, 317
Trace files, 308
TraceTuner, 306
Transcript Map of the Human Genome, 132. *See also* GeneMap '99
Transcript maps, 117–118
Transmembrane regions, 271
Transmission Control Protocol (TCP), 2
TreeAlign program, 329
Tree-building, 335, 352
  phylogenetic, 325, 326
Tree-building methods, 340–346
  character-based, 343–345
  comparison of, 345
  distance-based, 342–343
TreeDraw, 354
Tree-drawing programs, 354
Tree interpretation, 327
TREE-PUZZLE program, 353
Trees. *See also* Phylogenetic trees
  randomized, 346–347
  searching for, 345–346
Tree search options, 352
TreeTool, 354
TREMBL, 46
Triosephosphate isomerase, 384
TRIPLES database, 366
tRNA feature, 36, 51, 58
Truth, in Perl language, 430
Two-domain proteins, 373

UCLA server, 275
UDB Web site, 136
Unified Database (UDB), 129
Uniform resource locators (URLs), 13
UniGene, 120, 125, 126, 129, 143, 144, 145, 175, 288–293
  queries to, 291
UniGene clusters, 138, 289, 399
UniGene EST clusters, 137
University College London, 141
University of California, Irvine, 141
University of Colorado, 141
University of Texas, San Antonio, chromosome 3 site, 141
University of Toronto chromosome 7 Web site, 141
University of Washington High-Throughput Sequencing Center (UWHTSC), 135, 144
University of Wisconsin, Madison, 365

UNIX:
  environment, 11–12
  system, 8, 236, 353, 414, 416, 422, 434
  workstation, 275
UnStuffIt utility, 414
Unweighted pair group method with arithmetic mean (UPGMA) algorithm, 341, 343
Update process, 77
Update Sequence functions, 76
uRNA, 51
Urokinase-type plasminogen activator (PLAU) protein, 192

Validation, of sequence submission, 75–76
Variable interpolation, 419–420
Variable number of tandem repeat (VNTR) units, 114
Variables, Perl, 417–418
VecScreen database, 67
Vector Alignment Search Tool (VAST), 104–106, 156–157, 275
Version line, 53
Viewer3D, 102
Viewers. *See* Database structure viewers
Virtual Bioseq, 32
"Virtual bonds," 97
Virtual desktop, 103
Virtual libraries, 15
"Virtual mapping," 140
"Virtual Northern," 399, 401, 402
Virtual reality modeling language (VRML), 97
"Virtual subtraction," 399
Visualization software, 86
Visualization tools, 101
VRML file format, 102
v-*sis* viral oncogene, 198

"Warning sequences," 206
Washington University Genome Sequencing Center (WUGSC), 135, 141, 145
Web browsers, 159. *See also* Browsers
Web crawling, 15
WebIN system, 70
Web interface (WEBACE), 141
WebMol, 101
WebMol viewer, 90
Web pages, 13
WEBPHYLIP software, 355
Web sites, 13
  mapping, 120
Weighted key terms, 157–158
Weighted parsimony, 335
WHAT-IF tool, 275
Wheelan, Sarah J., 19
While loop, 430
White, Peter S., 111
Whitehead Center/MIT Center for Genome Research, 132
Whitehead Institute, 119
Whole-genome sequencing projects, 118
WICGR:
  Human Physical Mapping Project Home Page, 134
  mouse mapping project, 137–138
  mouse YAC mapping project, 135
  physical map, 134
YAC data, 144
Windows Notepad, 415
WinZip, 414
WIT (What Is There?) database, 365, 384
Wolfsberg, Tyra G., 283
Word-based methods, 200
Word "hits," 200, 202
Workbenches, 249
World Wide Web, 13–16
  access to, 6
  database similarity search on, 201
  finding information on, 14–16
  submissions on, 70
World Wide Web Entrez implementation, 159
WormPep, 413
WormPep file, 448
WWW Virtual Library, 15

X chromosome resources, 141–142
Xenologs, 327
XGRAIL application, 236, 237, 249
xProfiler tool, 399, 403
X-ray crystallography, 97
X-ray structures, 99

YAC clones, 144
YAC/STS map, 141
Yeast artificial chromosome (YAC) libraries, 114
Yeast Protein Database (YPD), 178, 366

Zebrafish Information Database, 289
Z-scores, 219, 220