Bioinformatics Tools and Databases for Genomics Research

14

14.1 Introduction

Bioinformatics involves the development of statistical tools and techniques and computer software for acquisition, storage, analysis, and visualization of biological information. The term "bioinformatics" has been derived by the fusion of the terms "biology," "information technology," and "statistics." The discipline of bioinformatics has the following three main activities: (1) development of new algorithms and statistical techniques for the assessment of relationships among enormous biological datasets, (2) use of these tools and techniques for analyzing and interpreting the huge biological datasets, and (3) development of databases for an efficient storage and management of the huge amounts of information, and fast search, retrieval and/or analysis of the desired data. Bioinformatics evolved because new tools and techniques were necessary to handle the enormous amino acid and nucleotide sequence data being generated. During the early 1960s, the National Biomedical Research Foundation compiled the first comprehensive collection of amino acid sequences. The European Molecular Biology Laboratory (EMBL) organized their collection of data on nucleotide sequences in 1980; the European Bioinformatics Institute (EBI), Hinxton, UK, now maintains this nucleotide sequence database. The National Center for Biotechnology Information (NCBI), USA, was created during the early 1980s. Sometime later, the DNA Data Bank of Japan (DDBJ) was established. In 1984, the National Biomedical Research Foundation established the Protein Information Resource (PIR), which identifies and interprets the data on amino acid sequences.

14.2 Representation of Nucleotide and Amino Acid Sequences

The amino acid and nucleotide sequences are reduced to digital data. This is greatly facilitated by the use of single-letter codes for the amino acids and the organic bases (Tables 14.1 and 14.2). It may be noted that in RNA sequences the symbol U is used in the place of T. Even those amino acid/base positions that exhibit ambiguity can be adequately represented by single-letter codes. In case of DNA, the sequences of the two complementary strands of a DNA duplex are represented by the symbols for complementary bases, which can be deduced either manually (for short sequences) or by using a computer software. In databases, the nucleotide sequences are listed from the 5′ end (at the extreme left of the written sequence) to the 3′ end of a single strand. The representations of amino acid sequences of protein molecules begin at their N-termini and proceed to their C-termini.

B.D. Singh and A.K. Singh, *Marker-Assisted Plant Breeding: Principles and Practices*,
DOI 10.1007/978-81-322-2316-0_14, © Author(s) 2015

Table 14.1 Single-letter codes for different bases found in nucleotide sequences

Symbol	Meaning	Logic for the symbol	Symbol for the complementary base
A	Adenine	***A***denine	T
C	Cytosine	***C***ytosine	G
G	Guanine	***G***uanine	C
T	Thiamine	***T***hiamine	A
R	G or A	Pu***r***ine	Y
Y	C or T	P***y***rimidine	R
M	A or C	A***m***ino group (bases having)	K
K	G or T	***K***eto group (bases having)	M
S	G or C	***S***trong base pairing	S[a]
W	A or T	***W***eak base pairing	W[a]
H	A, C, or T	Not G[b]	D
B	C, G, or T	Not A[b]	V
V	A, C, or G	Not U[b]	B
D	A, G, or T	Not C[b]	H
N	A, C, G, or T	***N***ucleotide	N

The codes are based on the recommendations of International Union of Pure and Applied Chemistry (IUPAC)
[a]The same symbol is used for the base on the complementary strand since G pairs with C (symbol S denotes both G and C), while A pairs with T (W denotes both)
[b]Not G, letter H comes immediately after letter G in the alphabet; not A, letter B is the next letter to A; not U, letter V follows letter U (denotes T in DNA); not C, letter D occurs just after C

Table 14.2 Single-letter symbols for different amino acids in protein sequences

Symbol	Amino acid	Three letter code
A	Alanine	Ala
B	Asparagine or aspartic acid	Asx
C	Cystine	Cys
D	Aspartic acid	Asp
E	Glutamic acid	Glu
F	Phenylalanine	Phe
G	Glycine	Gly
H	Histidine	His
I	Isoleucine	Ile
K	Lysine	Lys
L	Leucine	Leu
M	Methionine	Met
N	Asparagine	Asn
P	Proline	Pro
Q	Glutamine	Gln
R	Arginine	Arg
S	Serine	Ser
T	Threonine	Thr
V	Valine	Val
W	Tryptophan	Trp
Y	Tyrosine	Tyr
Z	Glutamine or glutamic acid	Glx
X	Any amino acid	Xaa

The codes are based on the recommendations of International Union of Pure and Applied Chemistry (IUPAC)

14.3 Bioinformatics Tools

The genome-sequencing projects triggered the development of such high-throughput technologies that generated sequence data at an unprecedented rapid pace. This necessitated the development of computer programs capable of acquiring, analyzing, classifying, and storing very large volumes of data and retrieving the desired data from the stored data. As a result, the computer hardware capabilities had to be greatly enhanced, new statistical techniques needed to be developed, appropriate computer programs were designed, and suitable data storage and management systems were also implemented. The various computer programs used for the acquisition and analysis of data and detection of associations and patterns as well as to achieve other specific objectives are often referred to as *bioinformatics tools* or simply as *tools*. A large number of different tools is available for achieving a variety of objectives. Some of the tools used for marker discovery and development, gene prediction, association analyses, data storage and management, etc. are briefly described in the following sections.

14.3.1 AutoSNP

Nucleotide sequences of expressed sequence tags (ESTs) can be analyzed to discover single-nucleotide polymorphisms (SNPs). Such SNPs are of great biological significance since they are based on the exons of expressed genes. The AutoSNP computer program carries out automated analysis of EST sequence data and identifies SNPs as well as insertion/deletion (InDel) variations present in them. It aligns the EST sequences and distinguishes between predicted SNPs and sequencing errors on the basis of the redundancy criterion. A putative SNP will be present in multiple reads, while a sequencing error would occur in one or two reads. For each candidate SNP, redundancy score and co-segregation score are estimated. The redundancy score of a predicted SNP locus is the frequency of polymorphism at this locus. The co-segregation score is the likelihood that the predicted SNP will be transmitted together with the other SNPs present in its vicinity in the EST sequence. The AutoSNP output includes the predicted SNPs and InDels along with their redundancy and co-segregation scores. Most of the SNPs and InDels predicted in maize using the AutoSNP tool were validated as true SNPs and InDels. The AutoSNP program is available to research workers free of cost on request to the authors of the program (email: dave.edwards@nre.vic.gov.au; Barker et al. 2003).

The SNPServer (http://hornbill.cspp.latrobe.edu.au/snpdiscovery.html) is a Web interface for using AutoSNP, BLAST, and CAP3 programs for SNP discovery in real time (Savage et al. 2005). BLAST identifies related EST sequences, CAP3 aligns and clusters these sequences, while AutoSNP analyzes the alignments to identify SNPs and InDels. The results from this SNP discovery pipeline, the source of the EST data, as well as their annotation are stored in autoSNPdb. This database can be accessed for free at http://autosnpdb.qfab.org.au/. The database has SNP data on rice, barley, and *Brassica* spp. AutoSNPdb allows identification of SNPs and InDels in specified genes or genes related to specific traits, and between genes of specified pairs/groups of plant varieties. A user-friendly GUI (graphical user interface) enables easy visualization of the SNPs in the database (Duran et al. 2009). Another tool, QualitySNPng, uses a haplotype-based strategy to enable the visualization and detection of SNPs from NGS data, and it does not require a fully sequenced reference genome (http://www.bioinformatics.nl/QualitySNPng/).

14.3.2 SNP2CAPS

The CAPS markers are valuable cost-effective tools for analysis of SNP and InDel polymorphisms in laboratories that are not highly equipped. This is particularly true when common restriction enzymes are used to analyze the CAPS markers. It is quite difficult to manually convert SNP markers into CAPS markers. The computer program dCAPS Finder 2.0 can be used to design PCR primers with such mismatches that either create a restriction site at the selected SNP locus or remove a site existing at the locus. This facilitates the conversion of SNPs to CAPS markers, but designing of such primers successfully is not a simple issue. The program SNP2CAPS (for SNP-to-CAPS) screens multiple aligned sequences for polymorphic restriction sites, analyzes these sites, and identifies such sites that are the most likely candidates for CAPS marker development. This generic program also evaluates the restriction enzymes for their suitability for CAPS analysis in the submitted sequences and selects those enzymes that show at least one restriction site polymorphism in each of the aligned sequences (Thiel et al. 2004).

When polymorphism at a restriction site (Fig. 14.1a) results from an authentic SNP, the restriction site will have an unambiguous sequence, i.e., it will consist of A, T, C, and G only (Fig. 14.1b). In some cases, however, there may be ambiguity (symbol N) at one or two positions in addition to the SNP polymorphism (Fig. 14.1c); however, this is unlikely to affect the CAPS development. Therefore, the above cases would be good candidates for CAPS development. But restriction site polymorphisms may arise merely due to an ambiguous sequence, in which case N would be present within the

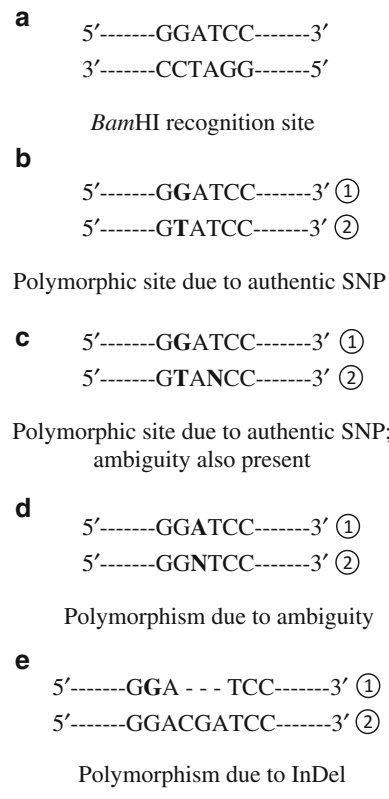

Fig. 14.1 Polymorphism in *Bam*HI recognition site. (**a**) A normal recognition site. (**b**) Polymorphism produced by SNP. (**c**) Polymorphism generated by SNP, but sequence ambiguity (N) is also present. (**d**) Polymorphism due to sequence ambiguity (N). (**e**) Polymorphism caused by InDel within the recognition site. The polymorphisms depicted in (**b**), (**c**), and (**e**) are good candidates for CAPS marker development. 1 and 2 are sequences of the same strand from two different individuals

restriction site sequence in some of the aligned sequences (Fig. 14.1d). Such restriction site polymorphisms are not suitable for CAPS development. In addition, insertion (or deletion) of one or more nucleotides into (or from) the restriction site will also generate restriction site polymorphism (Fig. 14.1e); these would be useful for CAPS development. Thus, the SNP2CAPS program analyzes the submitted sequences and identifies the recognition site polymorphisms suitable for CAPS marker development. The input for SNP2CAPS program is the multiple sequence alignment of the target sequences from different accessions. This input file may be in modified FASTA, ClustalW, MSF, MEME, etc. formats. In addition, it needs an input of restriction enzyme data, which can be downloaded from REBASE (the restriction enzyme database; http://rebase.neb.com/). A high proportion (90 %) of multiple aligned sequences of barley ESTs contained SNPs and InDels; they also had one or more restriction sites that were polymorphic. Further, over 30 % of these polymorphic restriction sites were for ten common restriction enzymes. SNP2CAPS offers a command line as well as a GUI. The SNP2CAPS is freely available from the website http://pgrc.ipk-gatersleben.de/snp2caps/.

14.3.3 TASSEL

The results from association analyses are often confounded by factors like selection, population structure, and family relationships, which may lead to incorrect marker–trait associations. The GLM and MLM approaches were developed to minimize the effects of population structure and/or family relationships on the findings from association studies. The GLM and MLM methods have been implemented in the software TASSEL (*T*rait *A*nalysis by a*SS*ociation, *E*volution and *L*inkage). The GLM method uses a structured association analysis based on a Q matrix to minimize the probability of false associations. The *Q matrix* reflects population structure and is computed by using the STRUCTURE program (Sect. 14.3.4) or by the principal components analysis (PCA) method. The MLM method uses in its model the kinship (K) matrix as well as the Q matrix in an effort to further reduce the risk of discovering false-positive associations. The estimates of K matrix representing the average relatedness between pairs of individuals/lines can be obtained from pedigree information or from genotype data for a large number of unlinked markers covering the whole genome of the organism. TASSEL carries out F-tests and permutation tests and estimates model effect means. When the trait in question does not have normally distributed residual error, some transformation function may be used to generate roughly normal error terms, or a permutation test may be used to generate distribution-independent p-values.

The TASSEL program can handle datasets from plant, animal, and human populations. It

enables the estimation of linkage disequilibrium (LD) as D' and as r^2 and allows graphical visualization of these estimates. Other features of this program include analysis of InDels, diversity analysis, execution of PCA, and imputation of missing data. This package includes several tools for extraction and visualization of data like sequence alignment viewer, neighbor-joining cladogram construction, and many data graphing functions. It has many data management functions and a data browser that provides an interface to relational databases. This software is in Java and is compatible with Windows, Mac, and Linux operating systems. The TASSEL executables, user manual, etc. are available for free from http://www.maizegenetics.net/tassel (Bradbury et al. 2007).

14.3.4 STRUCTURE

The *STRUCTURE* software (ver. 2.3.4 in 2012) is capable of detecting the presence of two or more homogeneous groups within a single population (Pritchard et al. 2000a). A *homogeneous group* is a group of individuals that is at Hardy–Weinberg equilibrium for all of the several random markers. This program implements a Bayesian (Markov chain Monte Carlo) algorithm for model-based clustering of individuals genotyped for several unlinked markers. It can use data from most genetic markers, including SSRs, SNPs, and AFLPs. It attempts to find out the number of homogeneous groups most likely to be present in the given population. The investigator should aim to find the smallest number of groups that accounts for the major structure in the population marker data. It also generates estimates of Q, which depict the likelihood that an individual belongs to a particular cluster. An individual may get assigned to two or more groups if its Q values indicate it to share the genetic properties of these groups. The accuracy of such assignments depends on several factors, including the numbers of individuals genotyped in the sample, groups present in the sample, the marker loci scored; the amount of admixture in the population; and the extent of allele frequency differences among the groups in the sample. This program has been used for detecting genetic structures in the sampled populations, assigning the individuals to different groups of the sample, population admixture, hybridization analysis, etc. Most studies show that STRUCTURE efficiently assigns different individuals to the populations of their origin, particularly when the population has two to four well-differentiated homogeneous groups. After starting the STRUCTURE program in a random configuration, it is run for, typically, 10,000–100,000 steps in simulation and then for another 10,000–100,000 or more steps to get accurate estimates of Q. The program is run several times, each time assuming a different number of groups, ranging from one to ten, in the dataset.

The executables of the program are compatible with Mac, Windows, Linux, or Sun. The computational part of the program, written in C, has a Java front end, which provides several helpful features. The data file should be a text file, and the missing data should be indicated by a number, often −9, which is not used anywhere else in the data file. The STRUCTURE software is free and is available at http://pritch.bsd.uchicago.edu/software/structure2_1.html.

14.3.5 Microarray Software

Microarray technology is a sophisticated precision experimental tool for studying genome-wide gene expression patterns and levels. It generates large quantities of data that require well-designed, user-friendly software for acquisition, analysis, storage, and management. The *TM4 software* is a suite of the following four tools: (1) A *M*icro*A*rray *DA*ta *M*anager (MADAM) tool guides the user through the microarray procedure beginning from RNA isolation to the analysis of data. It also facilitates the entry of data in the database and provides a platform for launching other data entry and management tools. (2) The TIGR Spotfinder tool rapidly and reproducibly analyzes the microarray images as well as quantifies the levels of gene expression. (3) The *MI*croarray *D*ata *A*nalysis *S*ystem (MIDAS) normalizes and filters the data generated by the Spotfinder tool. Finally, (4) the *M*ulti*E*xperiment

*V*iewer (MeV) tool analyzes the gene expression data files and displays the gene expression and annotation information obtained from the microarray experiments. Analysis modules implemented in MeV include the following: PCA, clustering (hierarchical and k-means), self-organizing maps and trees, etc. Bootstrapping and jackknifing procedures are used to generate consensus clusters. In addition, TM4 has a MySQL-based database for storage of the relevant data. This database conforms to the *M*inimal *I*nformation *A*bout a *M*icroarray *E*xperiment (MIAME) standards. TM4 was developed for spotted two-color microarrays, but it can be easily modified for single-color microarray formats. TM4 can be used for a wide variety of biological systems, including plant, animal, and microbial species. It is an extensible, open-source software suite available for free to research workers (http://www.tigr.org/software). The MADAM, MIDAS, and MeV tools can run on Windows, Mac OS X, Linux, and Unix platforms, but the TIGR Spotfinder runs only on Windows (Saeed et al. 2003).

14.3.6 A C. Elegans Database (AceDB)

The *AceDB* database management system was originally designed to handle the data generated from the *Caenorhabditis elegans* genome project. It has many powerful tools for handling genomic and bioinformatics data, which have now been made much more flexible. As a result, they are now used for management of genomic databases of many organisms, including plants. The AceDB system can handle diverse data types, including those pertaining to maps (both genetic and physical maps) and DNA sequences, and it can be easily modified to handle new types of data. AceDB has a full GUI and uses plain text input files, which greatly facilitates the management and distribution of genomic data. It is easy to modify and extend by simple text editing of a single file; this makes AceDB an ideal research tool. The AceDB system can be fully operated by a single biologist. The AceDB database management system is still being used for developing biological databases. Precompiled executables of AceDB for UNIX, Windows, and Macintosh environments along with the relevant documentation are available at the website http://www.acedb.org/.

14.3.7 MAPMAN

Whole-genome gene expression analyses using microarrays and metabolite profiling based on mass spectrometry generate huge amounts of data covering several parameters. The chief limitation in proper exploitation of this data is their proper analysis and interpretation. MAPMAN tool displays the large datasets in form of diagrams that depict the concerned metabolic pathways or other cellular functions and processes; this facilitates the interpretation of these datasets. This tool has two modules, viz., the SCAVENGER and the IMAGEANNOTATOR modules (Thimm et al. 2004). The *SCAVENGER module* collects data on gene expression and metabolite levels and classifies them into hierarchical groups termed as "Bins" and "subBins." A *Bin* corresponds to a specific area of metabolism, e.g., photosynthesis. A Bin can be further divided into *subBins*, e.g., "light reactions," "photorespiration," and "Calvin cycle" in the case of "photosynthesis." The different Bins and subBins are given specific numerical codes reflecting their hierarchical relationships. A specific SCAVENGER module is designed for each type of data: separate modules are used for gene expression and metabolite data. The TRANSCRIPT-SCAVENGER module handles data from gene expression arrays; it sorts the genes into Bins and subBins on the basis of their function deduced from gene annotation information. The assignment of the data to Bins and subBins involves automatic recruitment as well as manual correction. The guiding principles for the assignment are as follows: (1) as many genes as possible, including those with "supposed" annotation, should be assigned to specific Bins, (2) Bin structure should be modified, if needed, to accommodate the relevant data, and (3) as far as possible, each gene should be placed into a single Bin and

subBin. The METABOLITE-SCAVENGER classifies the metabolites into different groups on the basis of either their structures or the pathways in which they occur. The IMAGEANNOTATOR organizes the data groupings generated by the SCAVENGER and displays them as diagrams.

The modular structure of MAPMAN permits editing of the existing data categories, addition of new categories, and the development of SCAVENGER modules for new types of data. MAPMAN needs to be further developed for correction of deficiencies and inclusion of additional applications. For example, the SCAVENGER modules need to be developed for automatically updating the annotation and terminology by error-free acquisition of the GOC (*G*ene *O*ntology *C*onsortium) and other relevant releases. Modules are also needed for the removal of unnecessary redundancies, display of absolute levels of gene expression, or metabolite accumulation. In addition, modules capable of statistical analyses of the data also need to be developed. The IMAGEANNOTATOR module and the instructions for its use are freely available from the website http://gabi.rzpd.de/projects/MapMan/. The SCAVENGER modules can be obtained on request without any charge.

14.3.8 GenScan

The *GenScan* program (Burge and Karlin 1997) predicts complete gene structure, including introns, exons, and the exon–intron boundaries, promoter sites, and poly-A signals in genome sequences of many different types of organisms. The gene structure model used by GenScan is a "probabilistic model" developed for human genes. This model includes the description of the signals for transcription, translation, and splicing and the features related to the lengths and the base compositions of exons, introns, and the intergenic regions. It searches the query sequence for the features of this model, and the stretches of the sequence matching the descriptions of exons, promoters, etc. are identified, and a probability is assigned to each identified stretch. The identified "optimal exons" match the model with the highest probability ($P > 0.99$) and are considered to represent actual exons. GenScan also predicts "suboptimal exons" having acceptable probability levels ($P = 0.50–0.99$) of representing a true exon. Exons predicted with <0.50 probability are discarded as unreliable. This program is capable of predicting multiple genes as well as partial genes located in a given nucleotide sequence. The users can examine the "optimal" and the "suboptimal" sets of predictions to identify nonstandard gene structures like alternatively spliced genes. GenScan can accept and analyze nucleotide sequences of up to one million base pairs in length. It can analyze the sequence of either one or both the stands of a DNA duplex and make consistent prediction of groups of genes. GenScan has high accuracy but is sensitive to exon length. GenScan is by far the most comprehensive and sophisticated gene prediction tool available for free. The GenScan server can be accessed at http://genes.mit.edu/GENSCAN.html. Some other tools designed for gene prediction are FGENESH/FGENES, HMM Gene, GENE ID, GENE PARSER, etc. (Table 12.2).

14.3.9 ClustalW

A multiple sequence alignment is perhaps the most useful investigative procedure in bioinformatics; it is often said that it could be helpful in almost any situation. It helps in prediction of protein structure and function, and is the basis for phylogenetic analyses. The *Clustal* family of programs is perhaps the most extensively used for alignment of multiple sequences. There are two types of Clustal (ver. 2) programs: (1) ClustalW (has a command-line user interface) and (2) ClustalX (has a GUI) (Thompson et al. 1997; Chenna et al. 2003; Larkin et al. 2007). ClustalW is easy to use and is the most frequently used multiple sequence alignment tool. ClustalW uses a progressive method of alignment, in which the sequences are first compared in pairs for similarity. Each similar pair of sequences is then treated as a single

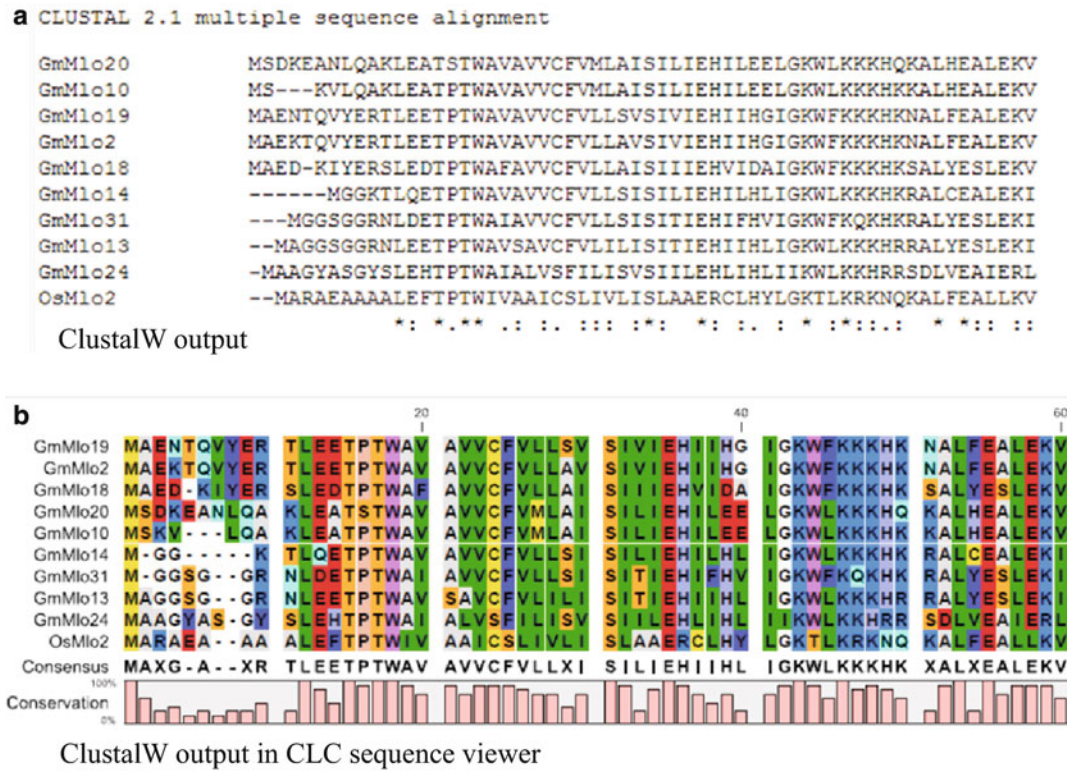

Fig. 14.2 Multiple sequence alignment of nine members of soybean *Mlo* (*GmMlo*) gene family and one rice *Mlo* (*OsMlo2*) gene. (**a**) The output from ClustalW program. (**b**) The ClustalW output visualized using CLC Sequence Viewer (Courtesy, Reena Deshmukh, Varanasi)

sequence, and the sequences so obtained are again compared two-by-two and aligned in pairs. This procedure is repeated till all the sequences are aligned together. A researcher can align the sequences using the default setting, but occasionally one may like to customize the setting to best suit one's needs. The main parameters that can be customized are the substitution matrix and the penalties for gap opening and gap extension.

Clustal programs offer several options for input/output formats, including Clustal, PHYLIP (output only), and FASTA (input only) formats. However, it is most convenient to use the FASTA format for ClustalW input sequences. However, judging the quality of a sequence alignment is essentially an educated guesswork. The bottom row of the ClustalW output of multiple sequence alignment contains stars (*), colons (:), and dots (.) (Fig. 14.2a). A star below a column indicates a fully conserved or an invariant amino acid residue, a colon (:) denotes that all the residues in the column have roughly the same size and hydrophobicity, a dot (.) signifies that the different amino acid residues in the column are either similar in size or hydrophobicity, while lack of a symbol indicates that the residues in the column differ both in size and hydrophobicity. A simple criterion of a sequence block with a good alignment is as follows: it is a gap-free continuous stretch of 10–30 amino acids having 1–3 stars, 5–7 colons, and a few dots scattered in the block. The aligned sequences can also be viewed through the CLC Sequence Viewer (6.8.1.); it uses a color code, in which the same color is used to depict amino acids having similar size and hydrophobicity. In addition, the level of conservation in each column is presented as a bar diagram; fully conserved columns are represented by a bar of full (100 %) height (Fig. 14.2b).

Clustal Omega is the latest addition to the Clustal family. This high-capacity program aligns hundreds of thousands of sequences in only a few hours. It can use multiple processors, and the quality of alignment is superior to those of the earlier versions. However, at present, Clustal Omega has a command-line interface and handles only protein sequences. It may be pointed out that it is preferable to work with protein sequences than nucleotide sequences. Precompiled executables and the source code of the programs (ver. 2.1) for Windows, Linux, and Mac OS X systems are available from www.clustal.org. EBI no longer maintains the ClustalW program, but it has Clustal Omega. ClustalW ver. 2.1 can be used at DDBJ (http://clustalw.ddbj.nig.ac.jp/) and the ClustalW servers (http://www.ch.embnet.org/software/ClustalW.html).

T-Coffee (igs-server.cnrs-mrs.fr/Tcoffee), MEME (http://meme.sdsc.edu/meme/website/intro.html; uses the expectation maximization method), HMMER (http://hmmer.janelia.org/; for protein sequence analysis), MUSCLE (http://www.drive5.com/muscle/; http://www.ebi.ac.uk/tools/muscle/index.html; aligns both DNA and protein sequences), MAP (http://genome.cs.mtu.edu/map.html), and COBALT tool of NCBI (for protein sequence alignments) are some of the other programs that can be used for multiple sequence alignment.

14.4 Bioinformatics Databases

A *database* is a systematized collection of vast amounts of information on a specific topic, e.g., nucleotide sequence, protein sequence, etc., in an electronic environment. The organization of databases is such that it allows regular updating of data and easy search and retrieval of the desired information. There are three types of databases, namely, flat-file, relational, and hierarchical databases. The *flat-file database* is the earliest and the simplest database type, is usually used for storing small amounts of data, and may consist of one or more files. It is easy to set up, but the storage methods are rather complex. In *relational databases*, the data are organized in form of tables. Further, the columns in these tables are indexed according to common features. These databases are constructed using the SQL (*S*tructured *Q*uery *L*anguage) programming language. This type of database has a well-defined design and architecture that minimizes redundancy of stored data, but its setting up is intellectually demanding. It supports very fast data search and can answer complex questions. In the *hierarchical databases*, the data are organized in a hierarchical (ordered tree) structure, and there are two or more levels of data organization (e.g., MAPMAN, Sect. 14.3.7). Construction, operation, and modification of these databases are simple, and data search and retrieval is fast. But they need more space, consume more time, and the same record may need to be stored at two or more places.

Special computer software programs, called database management systems (DBMS), are used to organize, search, and access the data. These programs not only contain raw data records, they also have operational instructions to help identify interconnections in the data records. The DBMS can be either relational or object oriented, the former being the most commonly used. For example, MySQL is a full-fledged, open-source relational DBMS, which has a three-tier architecture, viz., user interface, business logic, and data storage tiers. The biological databases are generally concerned with DNA and protein sequence data storage and management. The primary databases on nucleotide sequences are GenBank, EMBL, DDBJ, and GSDB (Genome Sequence Databases), while Swiss-Prot, TrEMBL, PIR, and MIPS (Martinsried Institute of Protein) are examples of primary databases on protein sequences. Further, there are secondary databases representing some specific sections of the primary databases, and in case of proteins there are composite and structure databases as well. In addition, there are several specialized databases devoted to specific organisms (Table 14.3).

Table 14.3 A selected list of the various "omics" resources and tools for major crop species/groups of crop species (whole-genome sequences, EST sequences, proteomics, transcriptomics, metabolomics, long-insert library, and HTG)

Crop species	Link to "omics" resource/tool
Apple	www.rosaceae.org
Arabidopsis[a]	https://www.arabidopsis.org/
Banana	http://www.musagenomics.org/index.php
Brassica spp. (*Brassica* ASTRA)	http://hornbill.cspp.latrobe.edu.au
Cassava	www.phytozome.net
Cotton	www.cottonmarker.org
GrainGenes[b]	http://wheat.pw.usda.gov/
Gramene[b]	http://www.gramene.org/
Grape	www.vitaceae.org
Legumes	http://www.comparative-legumes.org/
Lotus	http://www.kazusa.or.jp/lotus/
Maize	www.maizegdb.org; http://www.maizegenome.org/
Medicago	http://www.medicago.org/genome/)
Poplar	http://genome.jgi-psf.org/Poptr1/
Potato	www.potatogenome.net
Rape seed	www.brassica.info
RGP[b]	http://rgp.dna.affrc.go.jp/
Rice	http://rice.plantbilogy.msu.edu
Rosaceae	(GDR; http://www.genome.clemson.edu/gdr/)
Sorghum	www.phytozome.net/sorghum
Soybean	http://soybase.org
Tobacco	http://www.intl-pag.org/13/abstracts/PAG13_P027.html)
Tomato	http://solgenomics.net/;
	http://sgn.cornell.edu/help/about/tomatosequencing.html
Wheat	http://www.wheatgenome.org
Metabolic Network[c]	http://www.plantcyc.org/
Protein annotations[c]	http://salad.dna.affrc.go.jp/salad/en/

A comprehensive list of bioinformatics tools/software is available at http://seqanswers.com/wiki/Special:BrowseData/Bioinformatics-application

[a]The model dicot plant species

[b]These databases incorporate analytical, visualization, and interrogation tools

[c]Databases of plant metabolic etwork (PMN) and Surveyed conserved motif ALignment diagram and the Associating Dendrogram (SALAD)

14.4.1 GenBank

GenBank (http://www.ncbi.nih.gov/Genbank) is the main primary genomic DNA sequence database held at NCBI, USA, since 1992. The NCBI staff builds this database from sequences obtained through submissions from various laboratories, exchanges with EMBL and DDBJ nucleotide sequence databases, and patented sequence data from the US Patent and Trademark Office. The sequences stored in GenBank are divided into organismal and functional categories. The nucleotide sequences from specific organisms are stored in the organismal category, while those representing specific functions, such as high-throughput genome (HTG), expressed sequence tags (ESTs), etc. are stored in the functional category irrespective of their source organism. The translations of nucleotide sequences stored in GenBank are stored in the PIR database. The GenBank participates in the International Nucleotide Sequence Database Collaboration (INSDC) along with the EMBL and the DDBJ nucleotide sequence databases. GenBank is interconnected with the EMBL and the DDBJ databases, and these three databases exchange sequence information every day.

14.4.2 Phytozome

There are three different comparative genomic databases on green plants, namely, GreenPhylDB, Plaza, and Phytozome. These databases aim to support studies on genomics-based plant evolution and to facilitate the application of the information on functional genomics obtained from model plants for the improvement of crop plants. The Phytozome database is accessible at the website http://www.phytozome.net. It provides comparative data on genomes and gene families and the tools for their analysis. This database provides complete information on the pattern of changes in the nucleotide sequence and structure of each gene during evolution as well as the evolutionary history of plant gene families and their genomic organization. The large-scale gene family phylogenetic trees are constructed using distance-based methods or, sometimes, distance-plus-character-based methods. It also contains sequences and functional annotations of complete genomes of several (25 in 2012) plant species. The Phytozome Web portal has several widely used tools for search, identification, and evaluation of gene families. It provides information on genomic context of plant genes, gene homologues, and paralogues, RNA transcripts from the given genes, alternatively spliced RNA transcripts and the resulting peptide sequences, and functions of gene families. It also permits putative functions to be assigned to new DNA sequences. It allows access to complete genome sequences available in the database. Retrieval of genes and gene families can be done by either using keyword or through a search based on sequence similarity. Genome-centric views of the genomes present in Phytozome are available through Gbrowse and can be accessed from the Phytozome homepage. The BioMart tool of the database allows customized construction of sequences and annotations of genes and/or gene families (Goodstein et al. 2012).

14.4.3 European Molecular Biology Laboratory Nucleotide Sequence Database

The *EMBL Nucleotide Sequence Database* is maintained by EBI, UK, which is an outstation of the EMBL, Germany. This database is a part of the INSD Collaboration and can be accessed at http://www.ebi.ac.uk/embl. The nucleotide sequence collection of the database is comprehensive and includes all such annotation that is available from public sources. This database serves as the primary source of nucleotide sequences for Europe. The nucleotide sequence data and third party annotation (TPA) are generally submitted by the Webin tool, while sequence alignment data are submitted through Webin-Align. The nucleotide sequence data generated by large-scale genome-sequencing projects and those available from the European Patent Office can be submitted using automated procedures. The INSD Collaboration assigns an accession number to each sequence submitted to the database. Only the accession number of the nucleotide sequence submitted to the database needs to be cited in the publication, and the sequence is considered as part of this publication. The above is a mandatory requirement for publication of sequence information in most journals. The TPA entries carry the prefix "TPA" to distinguish them from primary data. The data are grouped into divisions on the basis of either the methodology used for their generation or the taxonomic status of the source organism; in addition, there are some specialized divisions as well. The other genomic databases held at EBI are Ensembl (a database of genome annotation) and Genome Reviews (has the curated complete genome sequences stored in the EMBL Nucleotide Sequence Database). The daily releases of the database contain new submissions and updated sequence data, while every 3 months the entire database is released. Access to the sequence data is available via FTP, several WWW interfaces, and e-mail. The FTP Server provides access to the daily releases, periodic updates, and the collective files having all types of data. Dbfetch (database fetch) tool is used for the retrieval of sequence data of up to 50 entries via http. The SOAP (*S*imple *O*bject *A*ccess *P*rotocol) tool, on the other hand, enables communication with other systems. The Web-based Sequence Retrieval System (SRS) can access nucleotide sequence data available in other databases at EBI. In addition, tools like FASTA and BLAST are also available (Kanz et al. 2005).

14.4.4 Swiss-Prot

Swiss-Prot (established in 1986; http://www.ebi.ac.uk/) database contains curated and fully annotated protein sequences. Each protein sequence entry consists of core data and the annotation information. The core data comprise sequence and taxonomic data and citation information. The annotation information includes function, domains and sites, secondary and quaternary structures, posttranslational modifications, etc. Effort is made to keep the level of redundancy to the minimum and to afford a high level of cross-referencing to other databases. The Swiss-Prot format closely follows the format of the EMBL nucleotide sequence database. The database *TrEMBL* (*Tr*anslation from *EMBL*) is a supplement to Swiss-Prot and contains unannotated protein sequences. The protein sequences included in TrEMBL are obtained by translating each coding sequence (CDS) available in the EMBL nucleotide sequence database, but the CDSs contained in Swiss-Prot are excluded from this database. The TrEMBL entries are recorded in a Swiss-Prot-like format (Bairoch and Apweiler 1996). The Swiss-Prot and TrEMBL databases are now maintained as two sections of the UniProt Knowledgebase (Sect. 14.4.5).

14.4.5 UniProt Knowledgebase (UniProtKB)

The *UniProt Knowledgebase (UniProtKB*, http://www.uniprot.org) is an expertly curated central database of protein knowledge. This database was developed by the UniProt Consortium and is updated by them every 4 months. The chief objective of UniProtKB is to unify the available information on protein sequence and function and to provide the same to the users. The Consortium comprises workers from Swiss Institute of Bioinformatics (SIB), EBI, UK, and PIR. The UniProtKB is extensively cross-referenced, and all data are freely available to scientists. This database has two main sections: (1) the UniProtKB/Swiss-Prot section (it comprises manually curated records) and (2) UniProtKB/TrEMBL section (it contains automatically curated, annotated, and classified records). The UniProtKB has the following databases. (1) The UniProt Archive (UniParc) database serves as an all-inclusive protein sequence repository. (2) In the UniProt Reference Clusters (UniRef) database, the protein sequences are grouped into different clusters on the basis of sequence identity; this facilitates the search for closely related proteins. (3) Finally, the UniProt Metagenomic and Environmental Sequences (UniMES) database has been designed to serve the emerging area of metagenomics. In addition, UniProtKB contributes gene ontology annotations to the GOC. *Ontology* is a controlled terminology that is used by researchers working with databases of different taxa. For example, gene ontology (GO) pertains to the description of gene product features from several species. Plant ontology (PO) relates to the terminology needed for describing the anatomical and developmental features of different plant species. Similarly, trait ontology (TO) lists the details of evaluation procedure and the environments, in which a specific trait of a given species was assayed. The use of ontologies allows databases to share information with other databases (Magrane and UniProt Consortium 2011; UniProt Consortium 2013).

The main access point to the UniProtKB is the website. The homepage of this website provides a quick introduction to the UniProtKB via a website tour. It also provides tools for a variety of purposes, including querying, data analysis, documentation, Google-like full-text searches, database identifier (ID) mapping, etc. The ID mapping tool converts UniProt IDs to corresponding IDs of several other databases. It also has the BLAST tool for searching similar sequences and ClustalW for aligning multiple sequences. The full-text search does not require information about the data being searched; in addition it is quick and easy. But more complex enquiries can be processed using a field-based text search. The BioMart data management system enables processing of multiple biological queries by accessing the UniProtKB, InterPro, Ensembl, and PRIDE resources. The search results are presented as a table, which can be reorganized by the users.

14.4.6 Gramene

Gramene (http://www.gramene.org) grew out of the RiceGenes project to encourage analysis of functions that are conserved across species as well as those that are specific to individual species. This database began in 2002 as a resource for the rice community and as a collection of comparative mapping studies in grasses. Gramene was organized around the rice genetic, physical, and sequence-based maps. Further, a set of corresponding anchor genetic markers was used to develop the comparative maps of several grass species such as rice, wheat, barley, maize, etc. Gramene has now become a comparative genomic resource for important plant species like *Arabidopsis*, *Brachypodium*, poplar, etc. (Ware et al. 2002; Youens-Clark et al. 2011). It contains data on a variety of relevant topics, including metabolic pathways, QTLs, genes, proteins, genetic diversity, etc. The data stored in the database are integrated with genetic, physical, bin, etc. maps as well as genome browsers. Gramene carries out alignments on the whole genome as well as gene-to-gene basis. It also predicts orthologous and paralogous relationships by constructing gene trees, and implements a synteny analysis to confirm homology.

The main navigation bar on each page of Gramene provides links to various databases, search tools, submission forms, etc. The genetic map can be accessed for a linkage group of a species or for maps having specific features like molecular markers. With the help of the integrated map of rice, one can find the location in the rice genome sequence that corresponds to the given position in the maize, wheat or barley genetic map. Gramene has a curated database containing all publicly available mutants of rice, molecular markers as well as proteins; it also includes descriptions of the variants for physiological or morphological characters, etc. Gene symbol, gene name or Gramene accession number can be used to search for the gene of interest. The search result presents an all-inclusive summary of all the data related to the specified phenotypic variant. The rice genome browser of Gramene displays a variety of information, including gene predictions, genetic markers, coding sequences, etc., and information about the protein encoded by the predicted gene.

Gramene uses ontologies for describing proteins, genes, alleles, and phenotypes, which allows information sharing with other databases. This also permits a gene affecting a developmental stage or an organ in one species to be used for predicting the gene involved in a similar function in another species. Gramene provides several Web services like a Distributed Annotation Server and useful tools like BLAST. Updates of the database are released regularly, while major additions are released twice every year. All Gramene databases and software are free; the downloads can be made via ftp://ftp.gramene.org.

14.4.7 GrainGenes

GrainGenes database is devoted to genetic and genomic information on the following cereal crops and their wild relatives: wheat, barley, oat, and rye (http://wheat.pw.usda.gov). This international database serves as data storehouse as well as information center. It contains curated data on genetic as well as physical maps, probes utilized for constructing the maps, ESTs, EST-derived simple sequence repeats, oligogenes, and QTLs. It contains the contact details of the researchers working with these crop species. It also stores data on sequences, genetic resources, pathology, literature references (on genetics and genomics), and provides links to other databases. The genetic and physical maps available at http://wheat.pw.usda.gov/ggpages/maps are summarized and interactive, and include links to the information on the concerned mapping study. It permits comparison among maps from different populations. QTL data for the relevant species are gathered from various sources, and they are referenced to similar QTLs for rice (in Gramene) and maize (in MaizeGDB). GrainGenes contributes to the development of trait ontology in collaboration

with other databases. It has a database of genotypes (alleles of molecular markers) and phenotypes for wheat and barley varieties (Carollo et al. 2005).

14.4.8 MaizeGDB

The *Maize Genetics and Genomics Database* (*MaizeGDB*) serves the maize research community as the chief source of data on genetics and genomics of maize (http://www.maizegdb.org). MaizeGDB stores data on DNA sequences, genetic studies, QTL experiments, gene products, relevant literature references, and the list of persons/organizations involved in maize research. The genetic data include allelic diversity, maps, primers/probes used for mapping, metabolic pathways, and phenotypic image data. MaizeGDB serves as a portal for the maize genome-sequencing information. Maize sequences available at GenBank are downloaded, curated, analyzed, and assembled into contigs at PlantGDB (Plant Genome Database; http://www.plantgdb.org). However, they are stored at MaizeGDB, which makes them accessible to the users. Cytological map images have been added to the database, and plant ontology terms are being associated with the database records. A researcher can enter the database by typing the term of his/her interest, e.g., *adh1*, into the search field. Appropriate links can be followed to obtain other information related to the search item. The genome browser enables data visualization and displays the data within their chromosomal context. In addition, the MaizeGDB Web service permits the use of data analysis tools like BLAST and GeneSeqer. MaizeGDB homepage offers a set of video tutorials to facilitate effective use of the database (Lawrence et al. 2005).

14.4.9 RiceGeneThresher

RiceGeneThresher (http://rice.kps.ku.ac.th) is a public domain rice genomics resource. The MySQL-based database has integrated information on genetics, genetic markers, genome annotation, ESTs, metabolic pathways etc. It also contains information on protein–protein interaction predictions, and genes that respond to various stresses. RiceGeneThresher is fast and flexible, and has tools for whole-genome mining for QTLs governing the specified traits. The data from studies on inheritance, molecular biology, and various "omics" approaches are analyzed to find the most promising candidate genes within a genomic/QTL region, and to infer their functions. The search for candidate genes may use as query either the relevant DNA marker or base sequence of the target region of the genome. Alternatively, the gene/locus name or gene annotation keywords like gene locus ID may be used as query. The retrieved information is displayed both as graphical and standard Web pages. It includes physical locations of the candidate genes, and nucleotide sequences of the complete genes (including their upstream regions) and their coding regions as well as the amino acid sequences of the proteins produced by their translation (Thongjuea et al. 2009).

14.4.10 Microarray Databases (ArrayExpress and Gene Expression Omnibus)

In addition to in-house databases, there are centralized databases like ArrayExpress and Gene Expression Omnibus (GEO) for gene expression data obtained from microarray experiments. *ArrayExpress* (http://www.ebi.ac.uk/arrayexpress) database is maintained at EBI, UK, and is accessible to the public (Brazma et al. 2003; Parkinson et al. 2005). This database is capable of structured storage of well-annotated data generated by any microarray platform. The annotation used by ArrayExpress conforms to MIAME standard. It is also able to exchange data in Microarray Gene Expression Markup Language (MAGE-ML) format. The other available online facilities include MIAMExpress (the data submission tool), the interface for searching the database, and the analysis tool Expression

Profiler. The query may relate to the experimenter, the laboratory where the work was done, the organism studied, the type of gene expression experiment and/or the type of microarray used in the study. The submissions can be of the following three types: arrays, experiments, and protocols.

The *Gene Expression Omnibus (*http://www.ncbi.nlm.nih.gov/geo) is a public domain storehouse for data on gene expression as well as genomic hybridization generated by high-throughput platforms (Edgar et al. 2002). Heterogeneous datasets can be easily deposited into this database where they are safely stored and can be readily retrieved when required. The objective of GEO is to complement in-house databases for gene expression and to serve as a tertiary central data distribution center. The Platforms, Samples, and Series modules of the GEO function as its central data entities. The *Platform* comprises a list of probes, which determine the molecules detected in the study. The *Sample* describes the group of molecules that was investigated in the study. This description is related to a single platform that was employed to produce the data on abundance of various molecules. The *Series* organizes the samples into meaningful datasets that constitute the experiment.

14.4.11 HarvEST

HarvEST software was initially developed for visualization of EST databases. At present, it supports several activities, including identification of SNPs, designing of genotyping platforms, comparative genomics, association of physical maps with the concerned genetic maps, and designing of microarrays. This software is available for banana/plantain, barley, *Brachypodium*, cassava, citrus, coffee, cowpea, soybean, rice, and wheat; the programs for barley and cowpea are the most complete. HarvEST has databases of crop species-specific EST sequences that have been generally trimmed free of vector and assembled using the CAP3 program, except in the cases of ESTs of rice, soybean, and wheat. The ACE file viewer allows the examination of sequence alignment and identification of SNPs. HarvEST offers a variety of assemblies, synteny, and sequence alignment analyses, archived information on BLAST hits, Boolean and other searches, etc.; these applications do not require Internet connectivity. The query may specify the genes involved in the trait of interest like tolerance to a stress, a developmental stage, or a specific tissue. The search output is either displayed on the screen, or it can be exported as a summary table/sequence file. HarvEST provides links to other sequence databases and supports online searches. The HarvEST software operates in the Windows environment and can be downloaded free for academic use from www.harvest-web.org.

14.5 Sources of Multiple Databases and Tools

There are several bioinformatics resources that provide multiple databases and database search and data analysis tools. It may be pointed out that bioinformatics is developing very rapidly, and software programs that are the best in their respective fields today may become less preferred or even obsolete tomorrow. Therefore, it may often be quite helpful to consult one of the bioinformatics resource locators listed in Table 14.4 to find out the Web locations of the tools that are currently the most appropriate for the needs of a research worker.

14.5.1 National Center for Biotechnology Information

The *National Center for Biotechnology Information* is located in Maryland, USA (www.ncbi.nlm.nih.gov/). The NCBI is an organization of the United States National Library of Medicine (NLM), which itself is a part of the National Institutes of Health (NIH). The NCBI serves as the primary provider of information relevant to biotechnology and biomedicine. The NCBI

Table 14.4 A list of selected bioinformatics resource locators

Resource locator	Web address	Resources related to
ArrayExpress	www.ebi.ac.uk/microarray/	DNA chips
ExPASy	www.expasy.ch	Servers dedicated to proteins; the home of Swiss-Prot
Pasteur	http://bioweb.pasteur.fr/intro-uk.html	Many online tools; miscellaneous links
Phylip	http://evolution.genetics.washington/edu/	Everything related to phylogeny
RNA World	www.imb-jena.de/RNA.html	Links related to RNA
Swbic	www.swbic.org/	Miscellaneous links
NCBI primers	www.ncbi.nlm.nih.gov/education	Very good introductory material on many subjects
Coffee Corner	www.ncbi.nlm.nih.gov	Online protein news
Bio-informer	www.ebi.ac.uk/Information/News	The EBI online news

maintains a series of databases (total number, 66) covering relevant literature, health, organisms, genomes, nucleotide sequences, genes, proteins, chemicals, and pathways (Table 14.5). The major databases maintained at NCBI are GenBank (Sect. 14.4.1) and PubMed (bibliographic database for biomedical literature). Other databases include the Gene, Genome, Epigenomics, Gene Expression Omnibus (Sect. 14.4.10), RefSeq, Structure, Database of Short Genetic Variation (dbSNP), TAXONOMY, etc. The NCBI also provides a variety of tools (58 tools; Table 14.6) for database search (some of the tools) and/or data analysis (most of the tools). The Entrez search engine of NCBI is its main system for text search and retrieval. The other tools include 1,000 Genomes Browser, BLAST, CDTree, Cn3D, Genetic Codes, Open Reading Frame Finder (ORF Finder), SNP Database Specialized Search Tools, TAXONOMY BROWSER, etc. The NCBI Handbook provides a description of the databases and some of the tools like Entrez and BLAST. It also contains information on the manner in which the databases work and the approaches for their utilization. In addition, each NCBI resource (databases and tools) has online help documentation to assist the user in their proper utilization. The NCBI research group conducts studies on the relevance of sequencing errors for database search, develops new algorithms for database search and multiple sequence alignment, constructs nonredundant sequence databases, builds mathematical models for the estimation of statistical significance of sequence similarity, and develops vector models for text retrieval.

14.5.2 Kyoto Encyclopedia of Genes and Genomes

The *Kyoto Encyclopedia of Genes and Genomes* (*KEGG*) is an integrated bioinformatics storehouse in the public domain (http://www.genome.ad.jp/kegg/). It aims to deduce from the genome information an understanding of the higher-order biological functions and their relevance to cells/organisms (Kanehisa et al. 2004). It integrates the current information on genes, proteins, enzymes, reactions, biochemical compounds, and molecular interaction networks. KEGG maintains a suite of databases, namely, PATHWAY, GENES, SSDB, KO, COMPOUND, GLYCAN, REACTION, and enzyme databases. These databases signify graph objects belonging to the following three categories: (1) protein network, (2) gene universe, and (3) chemical universe (Table 14.7). Subsequently, a resource on glycome informatics was developed; this resource integrates protein network, genomic, as well as chemical information (http://www.genome.jp/kegg/glycan/). The GLYCAN database (for carbohydrate structures) along with two useful tools is a part of this resource. In addition, this resource has glycan-related pathways and a map depicting all the possible structural variations in the carbohydrates of the biological world (Hashimoto et al. 2006).

The KEGG databases are organized in a hierarchical manner. These databases interact with numerous external databases. For example, the GENES database selects its entries semiautomatically from different sources, which include

Table 14.5 A list of the genomic resource related databases (except those devoted exclusively to animals) available at the NCBI website (www.ncbi.nlm.nih.gov/)

Database	Type of data stored
BioProject (former Genome Project)	Studies on genomics, functional genomics and genetics, and links to the concerned datasets
BioSample	Description of the biological materials used for experimental assays
Bookshelf	Selected freely downloadable biomedical books covering molecular biology, biochemistry, cell biology, genetics, etc.
CloneDB (former Clone Registry)	Clones and libraries, including sequence data, map positions, etc.
Computational Resources from NCBI's Structure Group	Databases and tools used for the study of structures of macromolecules, conserved domains, classification of protein, etc.
Conserved Domain Database (CDD)	Protein sequence alignments and profiles of the domains that have been conserved during evolution
Database of Expressed Sequence Tags (dbEST)	A division of GenBank comprising short single-pass reads of cDNA sequences
Database of Genome Survey Sequences (dbGSS)	Short single-pass reads of genomic DNA; a division of GenBank
Database of Genomic Structure Variation (dbVar)	Large changes in genome structure, including large InDels, translocations, and inversions
Database of Genotypes and Phenotypes (dbGaP)	Results from studies designed to elucidate the relationship between genotype and phenotype, includes findings from genome-wide association studies (GWAS)
Database of Short Genetic Variation (dbSNP)	Information on single-nucleotide variations, microsatellites, and small-scale InDels
Epigenomics	Richly annotated epigenomics datasets
GenBank	Annotated collection of all publicly available DNA sequences
Gene	Database of genes, especially from completely sequenced genomes
Gene Expression Omnibus (GEO) Database	A repository of MIAME-compliant gene expression data
Gene Expression Omnibus (GEO) Datasets	Curated datasets on gene expression and molecular abundance; derived from the GEO database
Gene Expression Omnibus (GEO) Profiles	Expression and molecular abundance profiles of individual genes; derived from the GEO database
Genome	Sequence and map data for the whole genomes of over 1,000 organisms
HomoloGene	A gene homology tool; identifies putative orthologs; curated orthologs from a variety of sources included via the Gene database
Journals in NCBI Databases	Journals referenced in NCBI database records, including PubMed abstracts
NCBI Glossary	Description of NCBI tools, acronyms, data representation formats, and bioinformatics terms
NCBI Handbook	Description of the various features of NCBI databases and software
NCBI Help Manual	Details of many NCBI resources, including the Entrez system, Gene, SNP, LinkOut, etc.
NCBI Website Search	Static NCBI Web pages, documentation, and online tools
Nucleotide Database	Nucleotide sequences from several sources, including GenBank, RefSeq, TPA, and PDB (Protein Data Bank)
PopSet	Related DNA sequences originating from comparative studies
Protein Clusters	Related protein sequences (clusters) consisting of Reference Sequence proteins
Protein Database	Protein sequences from a variety of sources, including GenPept, RefSeq, Swiss-Prot, PIR, and PDB
Reference Sequence (RefSeq)	Curated, nonredundant DNA, RNA, and protein sequences
Retrovirus Resources	Designed to support research on retroviruses; includes a genotyping tool
Sequence Read Archive (SRA)	A collection of sequence data from the next-generation sequencing platforms

(continued)

Table 14.5 (continued)

Database	Type of data stored
Structure (Molecular Modeling Database)	Macromolecular 3-D structures derived from PDB; also has tools for their visualization and comparative analysis
TAXONOMY	Names and phylogenetic lineages of more than 160,000 organisms
Third Party Annotation (TPA) Database	Sequences and annotations built from the existing primary sequence data in GenBank
Trace Archive	DNA sequence chromatograms (traces), base calls, and quality estimates for single-pass reads
Transcriptome Shotgun Assembly (TSA) Database Computationally assembled sequence database derived from NGS technologies Unigene	Sets of RNA transcripts computationally identified to be encoded by the same gene/pseudogene; also has information on protein similarities and gene expression
Unigene Library browser	Expressed sequence tag (EST) libraries arranged on the basis of organism, tissue type, and developmental stage
UniSTS	Sequence-tagged sites (STSs) derived from STS-based maps and other experiments
Viral Genomes	Biology of viruses, links to viral genome sequences in Entrez genome, etc.
Virus Variation	Sequence sets of selected viruses; also tools for their analyses

Table 14.6 A list of selected tools available at the NCBI website (total tools 58; www.ncbi.nlm.nih.gov/)

Tool	Application
1,000 Genome Browser	An interactive graphical viewer; this tool allows variant calls, genotype calls, aligned sequence reads, and other applications
Amino Acid Explorer	It uses characteristics of amino acids to predict changes in protein sequences due to mutations and functions of amino acid residues in conserved domains
Assembly Archive	This tool links raw sequence with the sequences available in GenBank, EMBL, and DDBJ; it allows viewing of the multiple sequence alignments
Blast Link (Blink)	Blink displays the results of BLAST search of a protein sequence against the protein sequence database at NCBI
Blast Microbial Genomes	BLAST search for similar sequences present in the selected complete eukaryotic/prokaryotic genomes
Blast RefSeqGene	BLAST search of the genomic sequences in the RefSeqGene/LRG set
Basic Local Alignment Search Tool (BLAST)	Searches for regions of local similarity between new nucleotide or protein sequences and those present in the sequence databases
Batch Entrez	Retrieves records from many Entrez[a] databases; results can be saved in various formats
CDTree	Classification of protein sequences; analysis of their evolutionary relationships; creation and updating of protein domain alignments
COBALT	A protein multiple sequence alignment tool; it uses RPS-BLAST, BLASTP, and PHI-BLAST
Cn3D	Visualization of 3-D structures of proteins from NCBI's Entrez retrieval service; displays structure, sequence, and alignment
Concise Microbial Protein BLAST	A specialized BLAST search of database consisting of all proteins from complete microbial genomes
Conserved Domain Architecture Retrieval Tool (CDART)	Displays the functional domains of a protein sequence and lists proteins with similar domain architectures
Conserved Domain Search Service	Identifies conserved domains present in a protein sequence
Digital Differential Display (DDD)	Compares EST profiles to identify genes with significantly different expression levels
Electronic PCR (e-PCR)	Identifies sequence-tagged sites (STSs) present in DNA sequences
Gene Expression Omnibus (GEO) BLAST	Aligns a nucleotide or protein sequence with GenBank sequences included in the GEO database

(continued)

Table 14.6 (continued)

Tool	Application
Gene Plot	Pair-wise comparison of two prokaryotic genomes; displays pairs of protein homologues
Genetic Codes	Provides genetic codes for organisms in the TAXONOMY database
Genome BLAST	Similarity search for query nucleotide or protein sequences in the genomic sequence databases using BLAST tool
Genome ProtMap	Maps each protein from a COG or a VOG back to the genome
Remap Tool	Allows projection of annotation data and other features from one genomic assembly to another or to RefSeqGene sequences
Genome Workbench	Permits viewing and analysis of sequence data
Map Viewer	For special browsing of maps and assembled sequences for a subset of organisms
OSIRIS	Assessment of multiplex short tandem repeat (STR) DNA profiles
Open Reading Frame Finder (ORF Finder)	Detection of all ORFs in the submitted sequence or in a sequence present in a database
Primer-BLAST	Designs PCR primers for a template sequence
ProSplign	Aligns proteins to genomic nucleotide sequences
Related Structures	Finds 3D structures from the Molecular Modeling Database (MMDB) for sequences similar to the query protein sequence
SNP database Specialized Search Tools	Search of the SNP database by genotype, method, population, etc.
Sequence Viewer	Generates a graphical display of a nucleotide or protein sequence and its annotated features; this display is configurable
Splign	Computes cDNA-to-genomic sequence alignments
Tax Plot	Compares genomes on the basis of the encoded protein sequences
TAXONOMY BROWSER	Searches taxonomy tree using partial taxonomic names, common names, etc.
Taxonomic Common Tree	Generation of a taxonomic tree for the selected group of organisms
VecScreen	Quick identification of sequences of vector origin
Vector Alignment Search Tools (VAST)	Identifies similar 3-D structures of proteins
Viral Genotyping Tools	Identifies the genotype of a viral sequence

[a]Entrez: NCBI's primary text search and retrieval system; integrates the PubMed database with 39 other databases

the sequences submitted to the GenBank, RefSeq, and EMBL nucleotide sequence databases, as well as organism-specific databases accessible to the public. These entries are re-annotated by assigning K numbers after the genes are classified in the KO groupings. It is intended that the KEGG would be self-reliant in connecting genome information to cellular functions. It is hoped that KEGG will eventually be able to enable in silico analysis of various biological systems and to create computer representations of cells and organisms. The GenomeNet website (http://www.genome.ad.jp/kegg/) enables easy access to KEGG. The table of contents page of KEGG can be accessed at http://www.genome.ad.jp/kegg/kegg2.html. This page permits access to the databases of KEGG. Academic users can utilize the SOAP server (at http://www.genome.ad.jp/kegg/soap/) to gain computerized access to KEGG.

14.5.3 Molecular Biology Database Collection

The Nucleic Acids Research launched the *Molecular Biology Database Collection (MBDC)* in the year 2000. The MBDC is a centralized online compilation of molecular biology databases (http://www.oup.co.uk/nar/Volume_28/Issue_

Table 14.7 The KEGG databases and their subject matter

Graph object (subject domain)	Major databases	Information stored	Information source
Gene universe	GENES	Information about individual genes	Submissions to GenBank, RefSeq, EMBL databases, and other publicly available databases
	SSDB	Sequence similarity database with ortholog and paralog clusters	GENES database
	KO	Classification of functions of genes in the SSDB database	SSDB ortholog clusters
Chemical universe	COMPOUND	Chemical structures of metabolic and some pharmaceutical and environmental compounds	Manually entered and verified
	GLYCAN	Carbohydrate structures; links to complex lipid and carbohydrate metabolism pathways	CarbBank database; direct entries
	REACTION	Formulas for enzyme catalyzed reactions	
	ENZYME	Enzyme nomenclature; links to the various KEGG databases	The enzyme nomenclature website[a]
Protein network (most unique data object)	PATHWAY	Network of gene products, including protein–protein interactions, metabolic networks, and gene regulatory networks	A collection of manually drawn diagrams called the KEGG Reference Pathway Diagrams (maps)

Based on Kanehisa et al. (2004)
[a]http://www.chem.qmul.ac.uk/iubmb/enzyme/

01/html/gkd115_gml.html). The aim of this compilation is to make these databases more accessible to the scientific community by helping them select the databases best suited to their needs and access them using the links provided. The emphasis of MBDC is to include databases with new value added by data curation, annotation, connections to new data, or inclusion of some other innovative features. The database list, classified on the basis of the information content of the databases, contains minimum redundancy, and the links to the databases are regularly updated (Baxevanis 2000).

14.5.4 Architecture for Metabolomics (ArMet)

ArMet (*Architecture for Metabolomics*; http://www.armet.org) is a data model designed to provide a standard format for describing metabolomics experiments and representing the data obtained from them (Jenkins et al. 2004). It covers the description of the biological source materials, the experimental details (sample collection, preparation, and analysis), and the results obtained. ArMet has the following nine packages: (1) Admin, (2) Biological source, (3) Growth, (4) Collection, (5) Sample handling, (6) Sample preparation, (7) Analysis-specific sample preparation, (8) Instrumental analysis, and (9) Metabolome estimate. These packages are applicable to a wide range of experiments. The core dataset that provides the lowest common denominator for comparison of different datasets is also described. ArMet is not a publicly available data repository; it merely serves as a basis for designing data storage facilities and appropriate data handling tools. ArMet uses controlled vocabularies to allow correct interconnection with other databases. It enables uniform recording of experimental details, ensures completeness and internal consistency of datasets, affords dependable exchange of data, allows a meaningful comparison of data from a range of techniques, supports designing of new experiments, and promotes standard operating procedures. It is designed to handle greenhouse-/phytotron-grown *A. thaliana*, field-grown potatoes, harvest and storage of these

materials, preparation of the materials for analysis, and the analysis itself. In addition, it can handle preparation and storage of peak lists as well. The results of analyses are represented in the Universal Modeling Language (UML) 1, which allows clear data specification and dissemination. The ArMet is built using Oracle with a Microsoft Office front end. It allows designing of sub-packages for new analytical and experimental techniques.

14.5.5 Database Search and Analysis Tools

The amount of data stored in each database is so enormous that it is a daunting task to find the information of interest from a given database. In addition, the data extracted from a database have to be suitably analyzed for the detection of hidden patterns, associations, and other features of interest and for their proper interpretation. These tasks are facilitated by computer programs that have been designed for either searching and retrieving the data from various databases or for searching, retrieving, as well as analyzing the retrieved data. The programs designed for search and, often, analyses of the retrieved data are generally provided by the websites hosting the databases.

14.5.5.1 Entrez

Entrez (introduced in 1991) is one of the most popular search engines at the NCBI, USA, and is accessible at www.ncbi.nlm.nih.gov/entrez/. It is a highly versatile and adaptable text-based search and retrieval system. It can search all major NCBI databases, including PubMed, databases of nucleotide and amino acid sequences, TAXONOMY, Swiss-Prot, etc. Entrez collects data from several sources and retrieves indexes and properly organizes the biomedical information. The data from different sources may have different formats and designs, but they all are integrated into a uniform information model and retrieval system. Entrez has nine nodes [published articles, nucleotide sequences, protein sequences, taxonomy, structure, genomes, Online Mendelian Inheritance in Man (OMIM), PopSet, and books]. Each of these nodes is, in fact, an assemblage of all such data that have been grouped and indexed together; this collection of data is referred to as an *Entrez database*. Each object in an Entrez database has a unique ID number and represents, as far as possible, a stable, objective observation of data. Entrez offers a variety of search criteria for a large number of information types, e.g., all possible citations from a given author that deal with a given subject, standard names for given genes, a given nucleotide/protein sequence in the databases, etc. Entrez also helps deduce relationships between different types of data by linking with the selected nodes and carrying out the necessary computations. The associations detected in this way may be helpful in planning of future experiments as well as facilitate the interpretation of existing information.

14.5.5.2 EBI Search

The *EBI Search*, also known as EB-eye, is a text-based search engine accessible at the website http://www.ebi.ac.uk/ebisearch/advancedsearch.ebi. EB-eye enables easy and consistent access, via a network of cross-references, to the databases hosted at EMBL-EBI. These databases cover nucleotide and protein sequences, structures, gene expression data, reaction maps and pathway models, literature pertaining to the biomedical sciences, as well as the intellectual property relevant to these disciplines. The EBI Search indexes the molecular data and other information contained in these databases and organizes the resources in a hierarchical manner to facilitate search. One can access the EBI Search through the Web or through an interface of the SOAP Web service. The Web page showing the search results gives a summary of hits, i.e., matches, for each query category/domain, the actual list of hits, and other related data, including alternative views. The summary page contains information about the gene, its expression pattern, the encoded protein, the protein structure, and the relevant literature. The summary page can be exported or printed as a

report. The user can look for orthologues of a given gene in another species.

14.5.5.3 BLAST

The *BLAST* (*Basic Local Alignment Search Tool*; Altschul et al. 1990) is the most popular data-mining tool developed ever. The BLAST is a family of user-friendly sequence similarity search tools for the identification of database sequences homologous to the query or submitted nucleotide or amino acid (protein) sequences. This allows prediction of the functions of the submitted sequences and helps in the modeling of 3-D structures of the concerned protein sequences. The BLAST algorithm does not simultaneously use the entire query sequence for similarity search. Instead, it divides the query sequence into several pieces of 11 nucleotides or three amino acids each and uses one piece at a time for similarity search. This is why BLAST is called *local alignment search tool*. The above strategy facilitates a much faster search of database sequences homologous to the query sequence. It may be pointed out that BLAST is much faster and more accurate in using protein sequences than nucleotide sequences. BLAST can be used to find genes in a genome, predict function and/or 3-D structure of a protein, and find members of a gene family. *It is often said that BLAST tool can do almost anything*.

There are five main types of the BLAST (ver. 2.0) tool: BLASTp, BLASTn, BLASTx, tBLASTn, and tBLASTx. The *BLASTp* program is used for comparing the submitted protein sequence against a protein database. The two most popular BLASTp online services are available at the NCBI server (www.ncbi.nlm.nih.gov/BLAST) and the Swiss EMBnet-ExPASy server, the latter offering more options to the users. *BLASTn* compares the query nucleotide sequence with a nucleotide sequence database. The *BLASTx* tool translates the submitted nucleotide sequence into a sequence of amino acids and compares this sequence with the sequences listed in a protein database. This tool is helpful in the correction of sequencing errors and may find a better sequenced corresponding DNA segment deposited in the database. The tool *tBLASTn* uses a protein sequence as query to search a nucleotide sequence database by translating the latter into protein sequences. Finally, *tBLASTx* translates the submitted nucleotide sequence as well as the nucleotide database sequence into protein sequences and searches for homology between the two. In addition to the above, there are several specialized BLAST programs like PSI-BLAST, PHI-BLAST, etc.

Thus, similarity search for a query protein sequence can be done using either BLASTp or tBLASTn programs. BLASTp is the most suitable for indicating the function of a protein, while tBLASTn is the best for searching new genes encoding similar proteins. Similarly, a nucleotide sequence can be used by BLASTn, BLASTx, and tBLASTx tools. BLASTn identifies similar DNA sequences irrespective of the query sequence being a coding or noncoding DNA. But BLASTx analyzes a coding query sequence and identifies similar proteins in the database. tBLASTx, on the other hand, discovers proteins and ESTs encoded by nucleotide sequences comparable to that submitted as the query sequence. When nucleotide sequences are used as query, it is advisable to restrict the search to a subset of the database since *BLAST search using DNA sequences is much slower than that based on protein sequences*.

PSI-BLAST (*P*osition-*S*pecific *I*terated *BLAST*) is used to identify all the members of a very large gene family, which cannot be accomplished by the simple BLAST programs. The first round of PSI-BLAST search is a simple BLASTp search using BLOSUM62 substitution matrix. After this search, the PSI-BLAST program develops a revised substitution matrix on the basis of alignments of the search results with the query sequence. The revised matrix is used for the next round of BLAST search, after which the matrix is again updated; this process is repeated several times. In each repeat of the search (iteration), PSI-BLAST identifies genes that are more distantly related to the query sequence than those detected in the previous rounds of the search. In this way, all such genes are identified that show conservation of amino

acid residues at some positions in the amino acid sequences of the proteins encoded by them. Obviously these positions may be expected to be involved in the cellular functions performed by these proteins so that the amino acid residues at these positions have been conserved in all the members of the gene family.

Suppose a researcher has generated a nucleotide (DNA/RNA) or protein sequence and wishes to identify homologous sequences present in the relevant database. The selected BLAST tool carries out the homology search as follows:

1. The submitted sequence is broken down into small pieces or "words" of 11 consecutive nucleotides or three amino acids each. These "words" are the units for base-per-base (or amino acid-per-amino acid) comparison with the database sequences. It should be noted that *repetitive sequences are masked by the default setting of the BLAST program.*
2. Each match is awarded a positive score, while each mismatch is penalized by a negative score. These scores are awarded according to a substitution-scoring matrix. The gaps introduced in either sequence for making the alignment are also suitably penalized. The amino acid substitution matrices are more reliable than nucleotide substitution matrices. There are 100 or so different amino acid substitution matrices, of which the matrices of the PAM (*P*oint *A*ccepted *M*utation) and BLOSUM (*BLO*cks *SU*bstitution *M*atrices) series are the most popular. *BLOSUM62 is the default substitution matrix for the BLAST program.*
3. The sequence alignment is assigned an overall score, which is the summation of the scores for each of its amino acids/nucleotides. The top-scoring alignments are ranked according to set criteria, which distinguish between a similarity due to ancestral relationship (homologous sequences) and that due to random chance (similar sequences).
4. Discovered homologies or matches are further examined using information accessible through ENTREZ and other search tools.

The general steps followed for BLAST search are as follows (Fig. 14.3). (1) The first step is to specify the parameters of BLAST search. The default parameters are optimal and well tested. But the user may modify them if he/she has some specific reason for doing so. (2) The sequence, for which homology search is to be made, is submitted into the "input sequence" box of BLAST interface. This sequence has to be in the FASTA format. This format is used for representation of nucleotide and amino acid sequences used in alignment/database scanning programs. The first line of FASTA format starts with >; it is the "definition line" that includes a unique identifier. The sequence to be submitted is in single-letter code and begins in the next line. FASTA is the default format for most of the sequence analysis software. FASTA is also a similarity search tool that is much slower than BLAST, but may work better than BLAST with satellite sequences. In contrast, the sequence part of the FASTA format is called *RAW format*; it is used in some sequence analysis software. (3) The appropriate BLAST program is selected depending on the type of similarity search to be

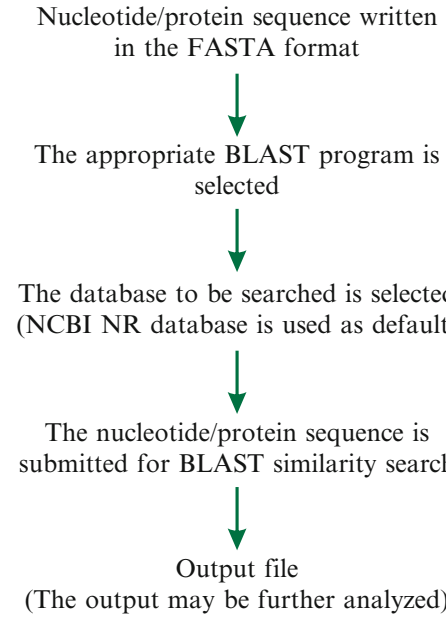

Fig. 14.3 A simplified schematic representation of the steps during sequence similarity search by BLAST. The default settings for BLAST are usually the optimum, but they can be specified by the user. The BLAST program works best with protein sequences

made. In general, it is preferable to work with a protein sequence than with a nucleotide sequence unless it is noncoding DNA/RNA. (4) The appropriate database, from which homologous sequences are to be searched, is specified. The default database used by BLAST is the NR database [nonredundant entries from GenBank, EMBL, DDBJ, and PDB (Protein Data Bank) databases] of NCBI. (5) The query sequence is now submitted to the BLAST server. (6) The results of BLAST search can be obtained either by e-mail or seen at the BLAST interface.

The so-called "traditional" BLAST report has been designed for easy readability. It has three major sections: (1) the header (it lists the BLAST program used, the sequence submitted as query, and the name of the database searched by BLAST), (2) description of each database sequence matching the query sequence, and (3) alignments with the query sequence for each matched database sequence. In addition, the bit score and *E*-values (expectation values) are also provided for each match. This report displays, by default, up to 500 sequences that matched the query sequence. But one can change this number in the case of advanced BLAST page. *As a general rule, when 25 % of the amino acid residues of any two proteins are identical, they are considered as homologous. Similarly, two DNA sequences are regarded as homologous if 70 % of their nucleotides are identical*. It may be pointed out that the above criteria do not work well when the query sequence is less than 100 nucleotides or amino acids in length. If an alignment has a small number of gaps and a few segments with high similarity, it is considered to be a good alignment. But an alignment having some identical amino acid/nucleotide residues distributed here and there over the entire sequence is not regarded as a good alignment. This criterion is useful particularly when the frequency of identical amino acids is around 25 %.

The bit score indicates the degree of similarity and depends on the alignment of similar or identical residues and the gaps, if any, needed for aligning the similar/identical residues in the query and the identified sequences. Therefore, as the bit score increases, the quality of alignment of the submitted sequence to the identified sequence also improves. The *E*-value indicates the likelihood that the detected similarity of the query to the sequence identified from the database is merely due to chance. Therefore, the match between the two sequences increases as the *E*-value decreases. For example, when an identified sequence is the same as the submitted sequence, the *E*-value will be zero. Therefore, to be certain of the homology between the query and the identified sequences, the *E*-value should be lower than 0.0001. The *E*-value depends on the size of the searched database and the system of scoring used for the search. The database sequences identified by BLAST can be considered as homologous to the query sequence only when the two are similar in the same region or, at least, in overlapping regions.

14.5.5.4 Entrez Gene

Entrez Gene, the successor to *Locus Link* program, handles queries concerning various loci. It differs from Locus Link with respect to the following two important features: (1) it has greater scope and (2) it is integrated with the Entrez search and retrieval system. Entrez Gene provides greater access than Locus Link to the genomes that are represented by the Reference Sequences of NCBI. It provides information about genes, including their official names, and allows search for genes homologous to a given gene and to obtain information about them. For example, one can easily obtain information about mouse genes, or genes of several other organisms, that are homologous to a given human gene.

14.5.5.5 Open Reading Frame Finder (ORF Finder)

The *ORF Finder* (*O*pen *R*eading *F*rame *Finder*) tool carries out graphical analysis to identify ORFs present in the submitted nucleotide sequence. Alternatively, it can analyze the nucleotide sequences retrieved from a database to identify the ORFs contained in them. The ORF Finder can detect all the ORFs that equal or exceed specified minimum size. It identifies ORFs using either the standard or an alternative

set of genetic codes from 16 different sets of genetic codes. Several different formats can be used for saving the amino acid sequences deduced from the identified ORFs. Further, these sequences can be used for BLAST searches of the sequence databases. The completeness and the accuracy of the sequence submissions are likely to be improved by the use of ORF Finder tool. The Sequin software package for sequence submission has the ORF Finder tool as a part of the package.

14.5.5.6 Search Tools for SNP Database (dbSNP)

A variety of tools are available on the left side bar of the homepage of SNP database (dbSNP) for searching the dbSNP. These programs allow SNP search by SNP genotype, SNP discovery method, the population in which the SNP was discovered, the researcher who submitted the SNP data, and marker and sequence similarity criteria. Entrez SNP is the main tool for searching dbSNP since it is a part of the Entrez search system. The search for SNPs can be based on qualifiers (aliases) or a specific search field. The Entrez SNP site lists combinations of 25 distinct search fields that can be used for this purpose. The "between markers positional query" is used when a researcher wishes to find SNPs located within any genomic region that is delineated by two STS markers. In addition, the NCBI Map Viewer tool offers other map-based queries.

14.5.5.7 Genome Remapping Service

The *Remap* tool of NCBI is used to project annotation data from one genome sequence assembly to another genomic assembly or to sequence assemblies of the RefSeqGene. It can also be used to transfer the locations of various genomic features across different genomic assemblies. The remapping is made possible through a base-by-base analysis of the concerned nucleotide sequences. The users have the option to either specify the stringency of remapping or to use the default settings. This tool displays the summary of remapping results on the Web page. However, the complete results, including the annotation and the remapped features, have to be downloaded and viewed by using the Genome Workbench graphical viewer of NCBI.

14.5.5.8 Primer-BLAST and Electronic PCR (e-PCR)

The tool *Primer-BLAST* designs pairs of PCR primers with the help of the Primer3 program. These primer pairs are designed for the amplification of the given template nucleotide sequences. The designed primers are used in an in silico PCR reaction using the given sequence as template, and the potential products are determined. These products are automatically used for a BLAST search against the databases specified by the user to assess the specificity of the designed PCR primers to the intended target sequences. The *Electronic PCR (e-PCR)*, on the other hand, is a computational program, which is able to identify STSs within the given nucleotide sequences. This tool searches the nucleotide sequences for potential STSs by using the PCR primers for similarity search. It then assesses the identified sequences matching the PCR primers for their correct order, orientation, and spacing to determine if they can serve as primers for generating known STSs in the given nucleotide sequence.

14.5.5.9 COBALT

The *COBALT* tool of NCBI carries out multiple alignments of protein sequences. The BLAST tools, viz., RPS-BLAST, BLASTp, and PHI-BLAST, are used for sequence similarity searches of the Conserved Domains Database (CDD) and the protein motif database. The protein domains are searched in the CDD by the RPS-BLAST. The tool PHI-BLAST (*P*attern *H*it *I*terated *BLAST*), on the other hand, carries out iterative searches for such sequences that have the pattern stipulated by the user. Prior to each new round of search, the PHI-BLAST program revises the substitution score matrix via PSI-BLAST. The COBALT tool uses the search results to discover a group of pair-wise constraints that are used for the alignment of the multiple protein sequences.

14.5.5.10 Splign and ProSplign

The *Splign* tool is a computer program that carries out cDNA-to-genomic sequence alignments. Similarly, *ProSplign* program aligns protein sequences to genomic nucleotide sequences. Both Splign and ProSplign programs use a variant of the Needleman–Wunsch global alignment algorithm. This algorithm enables these programs to specifically take into account introns and the splice site sequences. As a result, they are able to accurately determine the splice sites in genomic sequences and are tolerant to sequencing errors.

14.5.5.11 PredictProtein

The *PredictProtein server* provides, perhaps, the most comprehensive analysis of protein structure. This server carries out multiple sequence alignments, predicts secondary structures of proteins, detects functional motifs listed in PROSITE, predicts composition-bias regions, finds putative domain structures, and achieves fold recognition by prediction-based threading. It predicts transmembrane helix location and topology, globular as well as coiled-coil regions of proteins, the regions that switch structures, and the sites having cysteine bonds. In addition, it makes use of a collection of methods and databases to predict the presence of signal peptides and the locations of their cleavage sites, glycosylation sites, and cleavage sites for certain proteases, the presence of N-terminal chloroplast transit peptide and its probable cleavage site, and the three-dimensional structures (3-D structures) of protein molecules. It is capable of evaluation of secondary structure prediction accuracy and an automatic evaluation of prediction methods. It is also able to detect remote homologues of the submitted protein sequences.

The query protein sequence may be submitted to the PredictProtein server at cubic.bioc.columbia.edu/predictprotein/. But it may often be easier and faster to access the server at one of the mirror sites, e.g., www.sdsc.edu/predictprotein/, www.embl-heidelberg.de/predictprotein/, and www.cmbi.kun.nl/bioinf/predictprotein/. Alternatively, a request may be submitted to the META-PP server of the PredictProtein website; this server allows the query sequence to be submitted to many servers at once. The META-PP server automatically collects the results from these analyses and provides it to the user. However, it may often be faster and less confusing to access the relevant servers directly rather than using the META-PP server link to them. One may find the most suitable server by checking out EVA (*EV*aluation of *A*utomatic protein structure prediction), the secondary structure server monitoring system, at the website http://cubic.bioc.columbia.edu/eva/.

14.5.5.12 Cn3D and CDTree

Cn3D is a stand-alone tool for viewing 3-D structures of protein sequences obtained from the Entrez search and retrieval service of NCBI. In addition to the 3-D structure, it displays the protein sequence as well as the alignment; it also has powerful annotation and alignment editing features. This program runs on Windows, Macintosh, and UNIX systems. In addition, its configuration can be altered to make it capable of receiving data from the popular Web browsers. *CDTree* is a stand-alone software program that analyzes the amino acid sequences of proteins to determine their evolutionary relationships and also classifies them. It is capable of importing the existing records and hierarchies from the CDD as well as analyzing and updating them. The users can utilize this program for the construction of their own CDTrees and for creating and updating protein domain alignments. This tool is integrated with Entrez-CDD and the Cn3D application.

14.5.5.13 ScanProsite

ScanProsite (www.expasy.ch/tools/scanprosite/) compares the submitted protein sequence with the patterns and profiles listed in the PROSITE database. *PROSITE* database maintained at the ExPASy site (www.expasy.ch) has a collection of functional sites and short sequence patterns or motifs found in many proteins and shown to be associated with some biological property of the proteins. It also contains domain profiles, which

describe every position of an entire protein family. The entries in PROSITE are generally linked to Swiss-Prot and other relevant databases. The PROSITE file includes the sequence entries of the relevant databases that share the matched sequence motif of interest. The characterized motifs are well documented to minimize redundancy. The ScanProsite search result contains information on the sequence and the name of the detected pattern, its likely biological function, name of and hyperlink to the 3-D structure of the pattern (if available), and the list of segments of the submitted protein sequence having this pattern.

14.5.5.14 TAXONOMY BROWSER

The *TAXONOMY BROWSER* search tool provides taxonomic information about various species. The TAXONOMY database of NCBI has information (including scientific and common names) about all organisms, for which some sequence information is known; it includes over 79,000 species. The TAXONOMY server provides genetic information and the taxonomic relationships of the species in question. TAXONOMY has links with other servers of NCBI, e.g., Structure and PubMed.

14.5.6 Genamics SoftwareSeek

The *Genamics SoftwareSeek* is a database of both free and commercial software programs used in molecular biology and biochemistry. This database can be accessed at the website http://genamics.com/software/. This website serves as repository of the listed tools as well. This database has over 1,300 entries, which are growing rapidly. The various tools in the database are classified into 24 different categories, including Biochemistry (101 tools), Chemistry (99 tools), DNA sequence analysis (242 tools), Genetics (206 tools), Genome analysis (95 tools), Image analysis (49 tools), Molecular modeling (177 tools), PCR (15 tools), Analysis of phylogenetic relationships (90 tools), Identification of proteins (48 tools), Analyses of protein sequences (182 tools) and protein structures (96 tools), Prediction of protein (72 tools), and RNA structures (19 tools), and Alignment (165 tools) and presentation of sequences (62 tools). The numbers listed within parenthesis indicate the numbers of tools available under the respective categories. Many of the listed tools can be downloaded from this website. These tools can operate on Windows, MS-DOS, Mac, UNIX, and Linux platforms. This website also has online tools that run through an Internet browser.

14.5.7 Sequence Manipulation Suite

The *Sequence Manipulation Suite* (www.bioinformatics.org/sms2/) is a collection of Web-based computer programs designed for analysis, formatting, and preparation of textual representations of both DNA and protein sequences. The tools available at this Web portal can be used free of charge. The ver. 2 of this portal has a total of 62 tools, which are grouped into the following four categories: (1) format conversion, (2) sequence analysis, (3) sequence figures, and (4) random sequences (Table 14.8). The "format conversion tools" are the second largest in number, and they convert DNA and protein sequences written in one format into another format; they also allow some other types of sequence manipulations. For example, the "EMBL to FASTA" tool converts an EMBL DNA sequence file into the FASTA format. The "sequence analysis" tools form the largest category; these tools analyze the submitted sequences and extract the desired information. For example, the "ORF Finder" tool identifies ORFs in DNA sequences, while the "Reverse Translate" tool converts the submitted protein sequence into the most likely nondegenerate coding DNA sequence on the basis of a codon usage table. The tools in the "sequence figures" category prepare textual representation of sequences. For example, the "Restriction Map" tool identifies and depicts the positions of restriction enzyme cut sites in the submitted DNA sequence. Similarly, the "Translation" tool uses the submitted DNA sequence to prepare a textual map of its translated protein sequence. The tools in the

Table 14.8 A list of the web-based tools available at the sequence manipulation suite ver. 2

Application	Tools
Format conversion	Combine FASTA, EMBL to FASTA, EMBL Feature Extractor, EMBL Trans Extractor, Filter DNA, Filter Protein, GenBank to FASTA, GenBank Feature Extractor, GenBank Trans Extractor, One to Three, Range Extractor DNA, Range Extractor Protein, Reverse Complement, Split Codons, Split FASTA, Three to One, Window Extractor DNA, Window Extractor Protein
Sequence analysis	Codon Plot, Codon Usage, CpG Islands, DNA Molecular Weight, DNA Pattern Find, DNA Stats, Fuzzy Search DNA, Fuzzy Search Protein, Ident and Sim, Multi Rev Trans, Mutate for Digest, ORF Finder, Pairwise Align Codons, Pairwise Align DNA, Pairwise Align Protein, PCR Primer Stats, PCR Products, Protein GRAVY, Protein Isoelectric Point, Protein Molecular Weight, Protein Pattern Find, Protein Stats, Restriction Digest, Restriction Summary, Reverse Translate, Translate
Sequence figures	Color Align Conservation, Color Align Properties, Group DNA, Group Protein, Primer Map, Restriction Map, Translation Map
Random sequences	Mutate DNA, Mutate Protein, Random Coding DNA, Random DNA Sequence, Random DNA Regions, Random Protein Sequence, Random Protein Regions, Sample DNA, Sample Protein, Shuffle DNA, Shuffle Protein
Miscellaneous	IUPAC codes, Genetic codes, Browser compatibility, Mirror this site, Use this site off-line, About this site, Acknowledgments, Reference

www.bioinformatics.org/sms2/

"random sequence" category either generate entirely random sequences or random sequences from a given sample sequence or introduce mutations in the submitted sequence. Each tool has a window for submission of the DNA or protein sequence, and the results are returned as a new page. The output of each program is in the form of HTML commands, which is converted into a standard Web page by the Web browser. One can print, save, or edit the output with the help of either an HTML or a text editor.

14.5.8 PHYLIP

The *PHYLIP* (*PHYL*ogeny *I*nference *P*ackage) is free and comprises a collection of programs designed to construct evolutionary or phylogenetic trees from several types of data. The tree construction methods implemented by PHYLIP include the distance matrix, the parsimony, and the maximum likelihood methods. PHYLIP also carries out bootstrapping. It can use data on discrete characters, distance matrices, nucleotide and protein sequences, frequencies of genes, restriction sites, and DNA fragments. The PHYLIP package comprises 25 different programs for phylogenetic analyses. For example, *protpars* and *dnapars* determine phylogenies of protein and DNA sequences, respectively, by the parsimony method. The tools *proml*, *dnaml*, and *restml* use the maximum likelihood method; they are designed for phylogenetic analyses using data on protein and DNA sequences and the presence/absence of restriction sites, respectively. The *neighbor* package implements Neighbor-Joining and UPGMA methods for phylogenetic tree construction. The package *contml* draws phylogenetic inferences from data on quantitative traits and gene frequencies by the maximum likelihood method. The *drawgram* and the *drawtree* tools are used for drawing rooted and unrooted trees, respectively. The *consense* tool uses the majority rule for drawing consensus trees, while *retree* package rearranges trees, including conversion between rooted and unrooted trees.

The programs in the PHYLIP package are menu driven. The users can select the parameters for the phylogenetic analyses from among the available options. The input data is in the form of a text file prepared in flat ASCII or Text Only format by a word processor or a text editor. In case of nucleotide and protein sequence data, the input should be a high quality multiple sequence alignment of the concerned sequences. The multiple alignment may be done by a suitable program like ClustalW, which can write the output data files in the PHYLIP format. PHYLIP output files have names like outfile and outtree; the outtree files have trees written in a format used

by a number of major phylogeny packages. The source code and the precompiled executables for all the 25 programs of the PHYLIP package for Windows, Mac OS, Mac OS X, and Linux operating systems are available at bioweb.pasteur.fr/intro-uk.html.

Questions

1. "BLAST is the most popular data-mining tool developed ever." Discuss this statement giving appropriate justification for your arguments.
2. "NCBI provides a variety of useful bioinformatics tools." Examine this observation in the light of existing evidence.
3. What is a database? Briefly describe the various types of databases, and summarize the salient features of one nucleotide sequence and one protein sequence database in the public domain.
4. Briefly describe couple of tools used for molecular marker development.
5. Discuss the various applications of Clustal and PHYLIP software packages.
6. "The TASSEL and STRUCTURE programs are primarily relevant for association studies, but TASSEL software has some other applications as well." Evaluate this statement in the light of available relevant information.
7. Briefly describe one bioinformatics resource for each of the following: understanding of the higher-order biological functions, the various molecular biology databases, and tools for sequence manipulation.