Plant Molecular Biology **48:** 59–74, 2002.
© 2002 *Kluwer Academic Publishers. Printed in the Netherlands.*

Surviving in a sea of data: a survey of plant genome data resources and issues in building data management systems

Leonore Reiser*, Lukas A. Mueller and Seung Yon Rhee
*Carnegie Institution, Department of Plant Biology, 260 Panama Street, Stanford, CA 94305, USA (*author for correspondence; e-mail lreiser@acoma.stanford.edu)*

Key words: controlled vocabulary, databases, data management, genomics, information systems, nomenclature

Abstract

Exponential growth of data, largely from whole-genome analyses, has changed the way biologists think about and handle data. Optimal use of these data requires effective methods to analyze and manage these data sets. Computers, software and the World Wide Web are now integral components of biological discovery. Understanding how information is obtained, processed and annotated in public databases allows researchers to effectively organize, analyze and export their own data into these databases. In this review we focus largely on two areas related to management of genomic data. We cite examples of resources available in the public domain and describe some of the software for data management systems currently available for plant research. In addition, we discuss a few concepts of data management from the perspective of an individual or group that wishes to provide data to the public databases, to use the information in the public databases more efficiently, or to develop a database to manage large data sets internally or for public access. These concepts include data descriptions, exchange format, curation, attribution, and database implementation.

Introduction

Biological research during the past decade has generated an exponential increase of data. For example, the number of sequences in GenBank increased from 4864 490 in 1999 to 10 106 023 in 2000, totaling 11 101 066 288 bp (http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html). In addition, exploration-driven methods (e.g. genome sequencing, gene expression profiling) create large data sets that often exist with little biological context, and much of them are published electronically without peer review.

In order to derive meaning from these large data sets, tools are required to analyze and identify patterns in the data, and allow data to be put into a biological context. For the tools to be developed and refined, data must be easily accessible and amenable to analysis. The analyzed data must be fed back into the loop to allow the data to be re-analyzed, refined, verified, unexplored areas to be identified, and new hypotheses to be built. The development and maintenance of systems and procedures that allow the manipulation of data in the above processes can be defined as data management. Good data management practices are fundamental to generators and users of genomic data, as well as those who are concerned with the development of resources for public access (Kaminski, 2000; Stevens *et al.*, 2001).

This paper is divided into two parts. First, we describe different types of data management systems and tools. In the second part, we present issues relevant to the development of data management systems, such as nomenclature, controlled vocabulary, data exchange formats, curation, attribution, conceptual data modeling, and physical database implementation.

Resources and tools for data management

In a recent survey, biologists were asked to assess the required tasks needed to support the utilization and analysis of data (Stevens *et al.*, 2001). Of primary

importance was the ability to retrieve sequences and to perform similarity searches. In addition, the desire for new, more sophisticated visualization tools and increased interoperation between databases was expressed. However, the authors note that many of these desired features are currently available, suggesting that researchers do not have sufficient information on what resources are available. Concurrent with the increase in data generated, there has been a rise in the number of data analysis and management systems and tools, with variable longevity. Rather than provide an exhaustive list of currently available resources, we describe general types of data resources and tools and their functionalities using a few examples.

Biological data resources can be divided into four major types: databases of single data types, databases of multiple data types, data mining tools to search the different databases, and lab management tools for data management in laboratories. Some selected examples of these types of databases are listed in Table 1. This list is by no means exhaustive; more extensive listings have been compiled (Baxevanis, 1998; Baxevanis, 2001; Cartinhour, 1997) and descriptions of these resources are presented in the January annual databases issue of Nucleic Acids Research (e.g.http://nar.oupjournals.org/content/vol29/issue1/). Other descriptions of computational methods and software for sequence analysis have been reviewed elsewhere (Rhee and Flanders, 2000; Rhee, 2000).

Public databases

There are two main types of public databases for submitting, storing, and accessing biological data. There are databases for single data types and for multiple data types. Generally, single-data-type databases contain information about many different organisms and the multiple-data-type databases contain information about a single organism. The advantage of a single-data-type database is the easy access to enormous amounts of data, and the ability to compare and analyze these data across species. Key examples of databases for single data types are listed in Table 1. In contrast, databases for single organisms incorporate diverse data types at a single site for researchers who, for example, would like to design experiments that address the function and interactions of many genes in a given organism. For these applications, it is essential that all the information known for the organism be accessible in an easy, unambiguous, and intuitive way. Of particular value in such databases is the presence of associations between data such as gene expression and mutant phenotype data. This allows the identification of new correlations, which can serve as the basis for future experimentation.

In this section, we describe a few examples of databases and resources for single plant species, *Arabidopsis* and maize, and a few databases where information for multiple plant organisms is presented. Table 1 contains a more comprehensive list of examples of these databases and resources for plant genome data and lists some of the features of each resource along with the URLs. We encourage you to explore these pages to discover the full range of information available from each database.

The Arabidopsis Information Resource

The Arabidopsis Information Resource (TAIR; www.arabidopsis.org) is an information management system providing information about *Arabidopsis* and people working on *Arabidopsis*. It is developed and maintained at Carnegie Institution, Department of Plant Biology and National Center for Genome Resources (NCGR). TAIR's database currently includes data such as clones, genes, sequences, genetic markers, polymorphisms, stocks, *Arabidopsis* researchers, and *Arabidopsis* publications (Huala *et al.*, 2001). The information is accessible in a number of ways: (1) general or advanced text search forms for searching the database, (2) sequence analysis programs such as BLAST, FASTA, or PatternMatching (useful for short sequences such as a protein domains, motifs, or SAGE tags), (3) graphical browse and search programs such as MapViewer and SeqViewer, (4) static Web pages with information relevant to the *Arabidopsis* research community such as information about the progress of functional genomics projects, history and progress of the genome sequencing and annotation projects, and job postings, and (5) large data sets that can be downloaded from the FTP site.

Some of the new features available at TAIR are illustrated in Figure 1. Registered community members who are logged in and affiliated to a lab can order stocks available from the Arabidopsis stock centers (ABRC or NASC) online (Figure 1A; see also ABRC section below). A new graphical genome browser and search tool, SeqViewer, allows searching of the sequenced genome via sequence- or text-based queries (Figure 1B). Data types such as genes, clones, polymorphisms, and markers can be searched and viewed graphically at different zoom levels ranging from the whole chromosome to 10 kb. The graphical display

of each set of data is hyperlinked to a detailed information page generated from the database. In addition, a nucleotide pop-window displays 10 kb of sequence with these data types highlighted on the sequence.

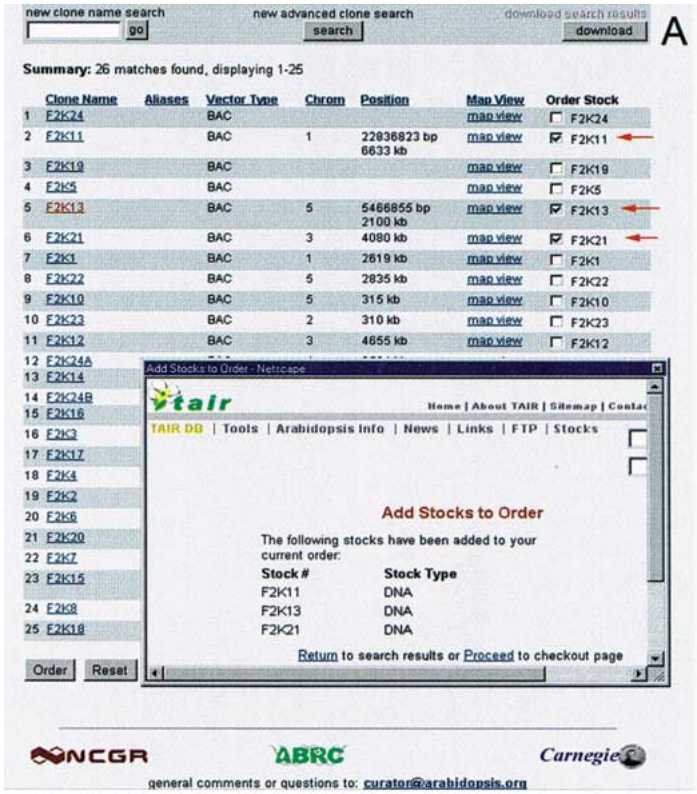
TAIR also provides access to data not in the database. Frequently requested data sets such as FASTA-formatted sequence files, mappings between microarray probes and the annotated genome, and expressed sequence tag (EST) matches to the genome sequence are among some of the data sets available from TAIR's FTP site. Individuals interested in finding out what genes are being studied by groups funded through the Arabidopsis 2010 Initiative (functional genomic characterization) can do so by searching the Functional Genomics section. In addition, Cereon has made their collection of over 56 000 polymorphisms identified between Landsberg *erecta* and Columbia accessions available through TAIR for registered academic researchers. Other information resources include lists of annotated gene families defined by researchers studying the gene families (and linked to their Web sites) and links to external resources and databases.

Data in TAIR are obtained from many sources and curated by a team of Ph.D. level biologists. Data sources include: large-scale genome sequencing projects like the Arabidopsis Genome Initiative (AGI), genome annotation data from TIGR and MIPS (see ATH1 and MATDB below), individual submissions from research groups, as well as from the literature and public web sites. Submission of new data, updates or suggestions for corrections is encouraged and can be sent to the curators (curator@arabidopsis.org). Corrections for gene models should be sent to annotation@arabidopsis.org where they will be evaluated by curators from TAIR and TIGR and, when appropriate, incorporated into the re-annotation of the *Arabidopsis* genome sequence.

Arabidopsis *genome annotation database (ATH1)*
ATH1 (http://www.tigr.org/tdb/ath1/htmls/ath1.html) is developed and maintained at the Institute for Genome Research (TIGR) and contains genome annotation data for *Arabidopsis* from the AGI. The ATH1 user interfaces include BLAST, text searching and graphical browsing of the annotated *Arabidopsis* genome. Detail pages for annotated genes show predicted open reading frames, similar sequences and links to nucleotide and amino acid sequences. TIGR's annotation also includes assignments to functional classes from MIPS and TIGR that can be searched and browsed. Pseudomolecule (non-redundant chromosome sequence), BAC, and gene sequences are can be downloaded from their FTP site. TIGR is currently re-annotating the genome using their computational analysis pipeline processes and association of function and process gene ontology terms (see Controlled vocabulary for more description on gene ontology).

MIPS Arabidopsis thaliana *database (MATDB)*
MATDB is another database containing *Arabidopsis* genome annotation and is developed and maintained at the Munich Information Center for Protein Sequences (MIPS; http://mips.gsf.de/proj/thal/db/index.html). Similar to ATH1, annotations in MATDB can be accessed via BLAST (and FASTA), browsed/viewed graphically or searched via text queries. MATDB has extensive automated gene annotations performed using their PEDANT analysis pipeline (Frishman *et al.*, 2001) and includes matches to INTERPRO domains (Apweiler *et al.*, 2001), SCOP domains (Lo Conte *et al.*, 2000), MIPS functional classes, similar sequences, Pfam (Bateman *et al.*, 2000), and PROSITE (Hofmann *et al.*, 1999) domains. MATDB also describes the rules used by the *Arabidopsis* genome annotation databases for associating a unique gene code (e.g., At2g03400) to each gene on the completed genome sequence. In addition, researchers can submit information about annotation errors at the MATDB site using simple Web forms.

Arabidopsis Genome Resource
The Arabidopsis Genome Resource (AGR; http://ukcrop.net/agr/) integrates the Arabidopsis Genome Initiative (AGI) sequence data with the physical and genetic maps of *Arabidopsis* to provide the necessary components for the study of gene function and the identification of crop plant orthologues of *Arabidopsis* genes. AGR is developed and maintained at the University of Nottingham and maintains the public recombinant inbred (genetic) maps for *Arabidopsis* and integrates this information with physical map data. Sequence homology information is maintained with respect to the public databases (SwissProt, trEMBL, and dbEST, EMBL) by means of BLAST searches. AGR also provides tools to view interactive displays of map and sequence data. For biologists seeking to identify mutations in target gene(s) AGR maintains a database of flanking sequences from insertional mutants that are searchable using BLAST and linked to germplasm requests. Researchers can automate this process using Insert Watch (http://nasc.nott.ac.uk/insertwatch/), which automatically processes BLAST queries against


new clone name search
go
new advanced clone search
search
download search results
download
A
Summary: 26 matches found, displaying 1-25
Clone Name | Aliases | Vector Type | Chrom | Position | Map View | Order Stock
1 F2K24 BAC map view F2K24
2 F2K11 BAC 1 22836823 bp 6633 kb map view F2K11
3 F2K19 BAC map view F2K19
4 F2K5 BAC map view F2K5
5 F2K13 BAC 5 5466855 bp 2100 kb map view F2K13
6 F2K21 BAC 3 4080 kb map view F2K21
7 F2K1 BAC 1 2619 kb map view F2K1
8 F2K22 BAC 5 2835 kb map view F2K22
9 F2K10 BAC 5 315 kb map view F2K10
10 F2K23 BAC 2 310 kb map view F2K23
11 F2K12 BAC 3 4655 kb map view F2K12
12 F2K24A
13 F2K14
14 F2K24B
15 F2K16
16 F2K3
17 F2K17
18 F2K4
19 F2K2
20 F2K6
21 F2K20
22 F2K7
23 F2K15
24 F2K8
25 F2K18
Order
Reset
Add Stocks to Order - Netscape
tair
Home | About TAIR | Sitemap | Contac
TAIR DB | Tools | Arabidopsis Info | News | Links | FTP | Stocks
Add Stocks to Order
The following stocks have been added to your current order:
Stock # Stock Type
F2K11 DNA
F2K13 DNA
F2K21 DNA
Return to search results or Proceed to checkout page
NCGR
ABRC
Carnegie
general comments or questions to: curator@arabidopsis.org


B
tair
SeqViewer
Click on chromosome to see closeup view
Your query : *snp* has 408 match(es)
1
2
3
4
5
Closeup View Options:
Zoom:
1 Mb
Band:
Markers
Polymorphisms
Gene Models
Transcripts
Clones
Whole Genome View Options:
Find any by Name:
Search
Locate Short Nucleotide Sequence(15-150) On Chromosome:
Paste a query sequence or multifasta formated sequences (maximum 4) into this box.
Submit
Reset
Please be patient. Replies directly to your browser may take a minute or two.

Figure 1. Linking of ABRC DNA and seed stock ordering is to search results summary pages is shown in panel A. Stocks are selected by checking a box next to the clone entry (red arrows). Clicking on the order button adds your selection to the current order and displays your current order in a new window. Orders can also be placed from the detail pages. The home page for TAIR's SeqViewer is shown in panel B. Search interfaces are highlighted with yellow boxes. The upper box is for text searches for genes, genetic markers and clones. The lower box is an input form to BLAST up to 4 sequences between 15–150 bp against the genome. Each of the five chromosome are shown in green at the top of the page and hits to the genome sequence are indicated by red bars. In panel B, a search was performed for all SNPs (408) currently placed on the genome sequence. The zoom level options (1 Mb to 10 kb) and objects to be displayed such as markers and genes are selected (blue box).

the insertion data set for hits in a gene of interest and provides email notification when a match is found. The BLAST search is run periodically as new insertion flanking sequences are added to the database.

Arabidopsis Biological Resource Center (ABRC) and Nottingham Arabidopsis Stock Center (NASC)

The ABRC (http://www.biosci.ohio-state.edu/~plantbio/Facilities/abrc/ABRCHOME.HTM; Scholl *et al.*, 2000) and NASC (http://nasc.nott.ac.uk/) are public stock centers, located at Ohio State University and Nottingham University, respectively, where primarily *Arabidopsis* stocks are maintained and distributed. Seed stocks include different ecotypes and accessions of *A. thaliana*, single and multiple mutant lines, pools of insertionally mutagenized lines and stocks of other *Arabidopsis* subspecies and relatives. ABRC and NASC have overlapping but not identical sets of stocks. Mutant, mapping and wild-type accessions are generally shared, but NASC maintains a number of insertionally mutagenized lines that are not available through the ABRC such as the SLAT (http://nasc.nott.ac.uk/info/slat_info1.html) lines from the John Innes Center. ABRC maintains and distributes DNA stocks whereas NASC does not. DNA stocks from ABRC include genomic DNA from mutagenized lines, clones, and clone libraries. ABRC also allows researchers to trace the order history associated with each stock. Currently, the stock information searching and ordering capacity from ABRC's database has been integrated into TAIR. Orders are processed and shipped from the stock center facility at Ohio State University. NASC is integrated with AGR and information about the stocks can be obtained from both the NASC and AGR web sites.

MaizeDB

MaizeDB (http://www.agron.missouri.edu/index.html) is an integrated compilation of genetic resources developed by the Maize Genetics Cooperative. Data are supplied by the stock center, independent databases, literature, and individual researchers. The data are curated and stored in a database that can be accessed by browsing or via text queries. The database contains information about genetic and physical maps, genes, stocks, colleagues, publications, phenotypes, agronomic traits, images, and QTL data. The data are linked to references, and some have links to records in other central repositories of: sequences (GenBank, SwissProt), references (Medline), germplasm (GRIN), other species-specific genome data (*Arabidopsis*, yeast, *Escherichia coli*, RiceGenes, GrainGenes). In addition to the database, MaizeDB also hosts a Web site that includes links to individual maize project Web sites.

Zea mays database (ZmDB)

Zea mays DataBase (ZmDB; http://zmdb.iastate.edu; Gai *et al.*, 2000) is a data repository and an analysis tool for sequence, expression and phenotype data for maize developed and maintained at Iowa State University. The source of most of the data in ZmDB is a collaborative project of maize gene discovery, focused on EST and insertion site flank sequencing and phenotypic analysis using a transposon tagging strategy .The database can be browsed and queried using a number of search parameters. ZmDB also provides software tools for sequence analysis such as BLAST as well as two novel gene-prediction programs developed for plants. Materials generated by the project can be ordered via the Web, including EST and genomic DNA clones, seeds of mutant plants and microarrays of amplified EST and genomic DNA.

UK CropNet and ARS Genome Resource

UK CropNet (www.ukcrop.net/; Dicks *et al.*, 2000) and the ARS Genome Resource (www.ars-genome.cornell.edu) are examples of resources where information about multiple plant species can be obtained. These two sites host similar plant databases and allow searching and browsing of the information stored in the hosted databases. The databases hosted by CropNet and the ARS include ones that are locally developed and maintained as well as external databases. The hosted genome databases are in ACEDB format and include *Arabidopsis*, barley,

Table 1. Selected examples of public databases and resources. These databases include those containing information for single or multiple plant species and those in which plants are represented along with many other organisms. The second column indicates which broad class or specific organism is represented. A brief, partial list of the data to be found at each site is shown in the third column followed by the most current web address.

Name	Organism(s)	Types of data	URL; Reference
GenBank	all	DNA and protein sequence, sequence analysis software, extensive links to other data sources	http://www.ncbi.nlm.nih.gov/Genbank/index.html (Benson et al., 2000)
SwissProt/-trEMBL	all	DNA and protein sequence, sequence analysis software, extensive links to other data sources	http://www.expasy.ch/sprot/ (Bairoch and Apweiler, 2000)
CropSeqDB	crop species	sequence database for 178 crop species and *Arabidopsis*	http://ukcrop.net/cropseqdb.html
COGS	only completely sequenced organisms	precomputed phylogenetic profiles of completed genomes	http://www.ncbi.nlm.nih.gov/COG/ (Tatusov et al., 2001)
Protein Data Bank	all	three-dimensional protein information.	http://www.rcsb.org/pdb/ (Berman et al., 2000)
DDBJ	all	nucleotide/protein sequence database	http://www.ddbj.nig.ac.jp
EMBL	all	nucleotide/protein sequence database	http://www.ebi.uk/embl
TAIR	*Arabidopsis*	genes, clones, genetic markers, maps, sequences, community profiles	http://www.arabidopsis.org (Huala et al., 2001)
ABRC	*Arabidopsis*	DNA and seed stocks	http://arabidopsis.org/stocks (Scholl et al., 2000)
NASC	*Arabidopsis*	seed stocks, RI maps	http://nasc.nott.ac.uk/
MATDB	*Arabidopsis*	annotated genes and proteins	http://mips.gsf.de/proj/thal/db/index.html
ATH1	*Arabidopsis*	annotated genes and proteins	http://www.tigr.org/tdb/ath1/htmls/ath1.html
MaizeDB	mize	genes, clones, markers, maps, phenotypes, references, germplasm	http://www.agron.missouri.edu/
ZmDB	maize	sequences, microarrays, phenotypes, germplasm	http://www.zmdb.iastate.edu/
Soybase	soybean	genes, clones, sequences, maps, markers, traits, germplasm in ACEDB	http://129.186.26.94/
AlfaGenes	alfalfa	genes, clones, sequences, maps, markers, germplasm in ACEDB	http://ars-genome.cornell.edu/cgi-bin/WebAce/webace?db=alfagenes
BeanGenes	*Phaseolus* and *Vigna*	genes, clones, sequences, maps, markers, germplasm, traits in ACEDB	http://beangenes.cws.ndsu.nodak.edu/
SolGenes	solanaceous species such as tomato and pepper	genes, clones, sequences, maps, markers, germplasm in ACEDB.	http://ars-genome.cornell.edu/solgenes/admin.html
SorghumDB	*Sorghum bicolor*	genes, clones, sequences, maps, markers, germplasm, metabolism in ACEDB	http://algodon.tamu.edu/sorghumdb.html

ARS Genome Resource	*Arabidopsis*, barley, *Brassica*, forage grasses, millet, and rice.	access to genes, maps, mutations, clones from multiple ACEDB plant databases.	http://ars-genome.cornell.edu/
UK CropNet	*Arabidopsis*, barley, *Brassica*, forage grasses, millet, rice	as above, including genes, maps, mutations, clones from multiple ACEDB plant databases. Also comparative mapping software	http://ukcrop.net/db.html
TIGR Gene Indices	*Arabidopsis*, maize, barley, wheat, ice plant, *Medicago truncatula*, *Sorghum bicolor*, tomato, soybean, rice, other eukaryotes and prokaryotes.	expressed sequences, tentative consensus sequences, orthologues	http://www.tigr.org/tdb/tgi.shtml
Gramene DB	rice and other grasses	genes, maps, references, clones, phenotypes for comparative genome analysis of rice and other grasses	http://www.gramene.org
GrainGenes	wheat, rye, oat, barley and sugarcane	genetic, sequence and phenotype data	http://wheat.pw.usda.gov/

Brassica, forage grasses, millet, and rice. The data include genes, phenotypes, traits, and chromosome maps. Both sites provide multiple database searching using a single query form that allows you to select one or more databases (including some not in ACEDB format). CropNet also has a BLAST server for sequence similarity searching of individual databases. In addition to providing access to the resident databases, CropNet supports the development of tools for viewing genetic maps (Recombination Viewer) and comparative maps (Grid Map, Pairwise Comparative Map (PCM), Comparative Physical and Genetic Map (CPG Map), Genome Map Viewer (GMV); (http://jic-bioinfo.bbsrc.ac.uk/bioinformatics-research).

TIGR Gene Indices

TIGR's Gene Indices (http://www.tigr.org/tdb/tgi.shtml) compile public EST data into a single resource for within- and cross-species comparison. The plant indices include *Arabidopsis*, *Medicago truncatula*, soybean, barley, potato, tomato, ice plant, rice, wheat, maize and *Sorgum bicolor*. Expressed (cDNA) sequences are used to build tentative consensus (TC) sequences for each organism. Gene indices can be searched by sequence, name, tissue type and gene ontology classification (the gene ontology is described below). Data from the gene indices is incorporated into the TIGR Orthologous Gene Alignment (TOGA; http://tigr.org/tdb/toga/toga.html) database which contains the results of pair-wise alignments used to group the TCs into predicted orthologous groups. The index data are available free of charge to researchers from non-profit organizations.

Data mining resources

Because different information is available from many diverse resources, tools to efficiently query and analyze data from more than one source are often helpful. Data mining resources combine multiple database search engines with software to display and analyze the results. A few types of data mining tools are listed in Table 2 and selected examples are described below. They differ from the Web searching

engines such as Excite (http://www.exite.com) and Google (http://www.google.com) in that they connect to different databases and search for the specific information stored in the databases. The retrieved data can then be passed on to different analysis programs.

SRS (formerly Sequence Retrieval System)

SRS is a data integration and analysis tool originally developed at the European Bioinformatics Institute (EBI; http://srs.ebi.ac.uk) and commercialized by LION Bioscience. Accessible through Web and programmatic interfaces, SRS uses indexes to integrate databases and bioinformatics applications. Key functionality includes: linking information from diverse databases, performing cross-database queries and seamless integration of bioinformatics tools such as BLAST. An example of the use of SRS would be, show me all *Arabidopsis* genes that are membrane-bound proteins and have a known 3-D structure and BLAST these against my favorite database. Another would be, show me all human genes, encoding membrane-bound proteins, involved in glucose metabolism, expressed in the liver, and associated with obesity. Academic users have access to the latest version of SRS (currently 6.1) at no charge and it is available from public servers, such as EBI's where over 140 databases and 20 bioinformatics applications are integrated within SRS. There are also links to many other public SRS sites.

GeneQuiz

GeneQuiz (http://jura.ebi.ac.uk:8765/ext-genequiz/) is a system for large-scale biological sequence analysis, that takes a protein sequence through a series of computational modules to predict protein functions. GeneQuiz provides a search interface (http://jura.ebi.ac.uk:8765/gqsrv/submit) for individual proteins of interest as well as analyzed data for a number of different genomes. This fully automated step-wise method searches for similarities in public databases, and offers an inferred function based upon the sum of all the analysis results. The GeneQuiz modules include a database update; a search system; an interpretation module, and a visualization and browsing system. The modules process information from user queries/public databases through a series of similarity and motif searches; the modules are driven by Perl scripts and automatically transform the data into to the proper format and submit it to the next analysis tool. The results are stored in a simple relational database. The results from the database can be visualized via the Web.

Data storage and analysis resources for the lab

There are an increasing number of public and commercial software packages available for storing, manipulating, and analyzing large sets of customized data locally on a desktop computer or remotely via the internet; examples are listed in Table 3. Some of the features included in these tools are: the ability to define a standard analysis routine with defined parameters (a pipeline); data storage or a virtual lab notebook; search and retrieval of data from external resources, and automatic data transformation for processing data through a series of analysis methods. An important feature is the ability to maintain accurate records of the parameters and results from analysis programs to ensure reproducibility. The potential utility of each system should be evaluated according to the needs (and resources) of each laboratory.

Concepts in data management

Data management is used here to describe how data can be retrieved, stored, analyzed, re-formatted, made accessible to others, and exported to other databases. Some basic questions for any project dealing with large, complex data sets are: What are the data types being managed? What are the methods used to obtain the data? How trustworthy are the data? What will be done with the data and what infrastructure and tools are needed to store and analyze the data? What formats must the data fit in order to be used by analysis software? In the following section, we attempt to give an overview on these issues and how data descriptions, nomenclatures, controlled vocabularies, conceptual data modeling and physical database implementation can be applied to address data management issues. The information presented here is intended for both the general user of public database resources, and people starting to produce high-throughput data who may wish to implement databases in their labs or publish their data in public databases.

Describing data

The variety with which data can be described presents a special challenge for data management and retrieval. Different databases often use different descriptions

Table 2. Examples of tools available on the web for performing queries, retrieving and analyzing data from multiple data sources. Programs available via the internet for acquiring and analyzing data from diverse databases are shown. Some of these resources provide results of their in-house genome analysis in addition to allowing users to input their own queries.

Software	Features	URL/Reference
SRS	search and retrieve sequences from multiple sites and analyze data using many different programs	srs.ebi.ac.uk/
ISYS	application for search, retrieval and analysis of genomic data from diverse plant databases	www.ncgr.org/research/isys/ (Siepel *et al.*, 2001)
GeneQuiz	protein function prediction	http://jura.ebi.ac.uk:8765/ext-genequiz/
KEGG	metabolic pathway profiles, pathway predictions, multiple genome search	http://www.genome.ad.jp/kegg/ (Wixon and Kell, 2000)
InterPro	compiled results of protein similarity searches. Pre-queried searches through the web but also available for local installation	http://www.ebi.ac.uk/interpro/ (Apweiler *et al.*, 2001)

of data, which complicates the process of accessing and comparing the data from these sources. This next section describes examples and illustrates problems associated with diverse data descriptions and presents some of the ways that the problems have been resolved.

Nomenclature

Inconsistent nomenclature can lead to loss of information or incorrect data associations. The nomenclature for gene names illustrates this point. Genes are often queried using their names, which generally are chosen to be descriptive of some functional aspect of the gene product, such as biochemical function, mutant phenotype or a protein-protein interaction. For example, the *Arabidopsis* gene *EMB30* has also been referred to in publications as *GNOM* (Busch *et al.*, 1996; Shevell *et al.*, 1994); both gene names are based upon the mutant phenotype. A search of PubMed with the gene name *GNOM* retrieves five references and *EMB30* yields eight. Only one publication was found in both query results. Thus, a researcher who is unaware of the history of a gene would only be able to access the half of the literature that was available based upon the knowledge of only one name. Genes are often referred to by their symbolic names (e.g. *ADH* for alcohol dehydrogenase), but the same symbolic name is used to refer to more than one distinct locus. For example, in *Arabidopsis* the symbol *FDH* has been used for both the *FIDDLEHEAD1* gene product (encoding a β-ketoacyl CoA-synthase; Yephremov *et al.*, 1999) and *FORMATE DEHYDROGENASE* (encoding a formate dehydrogenase; GenBank accession number AB023897) gene product. A search of NCBI databases by means of Entrez with the term *FDH* does not differentiate between the two, which could lead to false associations if genes are grouped based upon name only.

Several approaches have been taken to resolve problems related to gene nomenclature. Standards for gene nomenclature have been developed by representatives of the research community for naming genes for a specific organism. Guidelines for maize (http://www.agron.missouri.edu/maize_nomenclature.html), rice (http://www.shigen.nig.ac.jp/rice/oryzabase/basic/English/Pages/gene_name.html), *Capsicum* (http://genome.cornell.edu/solgenes/admin/nomen.caps.html) and *Arabidopsis* nomenclature (ADDIN-Meinke,1995; http://arabidopsis.org/info/guidelines.html) have been published. For each organism, the guidelines should be consulted when choosing a symbolic gene name. This approach may be practical for researchers working on a model organism with established guidelines, but there are some limitations to the approach. Guidelines are not always followed by members of the research community, the rules are not general enough to accommodate all organisms, and different formats are used for different organisms that complicates cross-species comparisons.

Another approach uses sequence similarity to define gene families that form the basis for a sequential nomenclature based upon a canonical family name (Price *et al.*, 1996). The idea behind this approach is

Table 3. Examples of data management resources for laboratories and database management systems. A listing of some of the programs that provide sequence analysis tools and sometimes data tracking for computational analysis, and software for creating databases. We have included a brief description of their features. Software applications must be installed locally while web-based tools and the respective databases are accessed via a browser. A partial list of features for each software program is provided. * Indicates software that must be purchased but also allows a free trial period for testing.

Software or company	Abbreviated list of features	Free systems	Platforms/operating	URL
Bionavigator	multiple sequence analysis tools, customizable pipelines, data storage (virtual notebook), report generation	no ∗	Web-based	www.bionavigator.com
Biowire Jellyfish	sequence retrieval and analysis, molecular biology tools (e.g. primer design)	yes	application for Macintosh, Windows, Linux, Unix	www.biowire.com/bw_jsp /home_top.jsp
DoubleTwist	sequence retrieval and analysis, molecular biology tools	yes	Macintosh, Windows, Unix, Linux applications and Web-based tools.	www.doubletwist.com
VectorNTI Suite	molecular biology data management, sequence analysis software	no *	Macintosh, Windows application	http://www.informaxinc.com /products/vectornti/vec tor_suite.html
VectorNTI Viewer	sequence visualization, map making, annotation	yes	Macintosh, Windows	http://www.informaxinc.com /products/vectornti/vector _suite.html
GeneSpring	microarray data management and analysis system	no	Java application for Macintosh, Widows, Unix	http://www.sigenetics.com /cgi/SiG.cgi/Products/Gene Spring/index.smf
Genomax	gene sequence analysis database	no*	Unix http://www.informaxinc.c	om/products/genomax/gen _sas.html
LabBook XML Browser 3.0	sequence visualization and annotation, search and retrieve sequence	no	Windows application	http://www.labbook.com/ products/browser.asp
eLabBook	data management tool with integrated literature searching	no	Windows application	http://www.labbook.com/ products/elabbook.asp
BioDiscovery	software for microarray data management analysis and data mining (CloneTracker, ImaGene, GeneSight)	no	Windows application	http://www.biodiscovery. com/
MySQL	user-defined database management system	yes	Unix, Linux, Windows	http://www.mysql.com/do wnloads/mysql-3.23.html
MS Access	user-defined database management system	no	Windows	http://www.microsoft.com /office/access/default.htm
Oracle	user-defined database management system for large databases	no	Unix, Windows	http://www.oracle.com
Informix	user-defined database management system for large databases	no	Unix, Windows, OS2	http://www-4.ibm.com/software/data/ informixwelcome.html
PostGres	user-defined database management system	yes	Unix, Windows	http://www.postgresql.org
Sybase	user-defined database management system for large databases	no	Unix, Windows	http://sybase.com
Filemaker	user-defined database management system	no	Macintosh, Windows	http://www.filemaker.com /index.html

to provide uniform nomenclature that crosses species boundaries to facilitate comparative analysis between the species. Nomenclature is then based upon a functional characteristics of the gene family that are predicted to be shared among its members. An important caveat to this approach is the difficulty in unambiguously identifying orthologues (White *et al.*, 1999). Naming genes based upon function can be misleading if it is not clear that the assigned name is based upon an inference derived from sequence similarity, rather than experimental evidence. Furthermore, unless great care is taken, this leads to error propagation when functions become associated with less related sequences. For example, gene A gets assigned the name *YFG2* based upon similarity to gene *YFG1* and gene B gets assigned the name *YFG3* based upon similarity to *YFG2*. However, *YFG3* and *YFG1* may be quite distantly related and their biological functions may have diverged. An additional complication is that the *YFG1/YFG2* similarity and the *YFG2/YFG3* similarity may lie in different domains of *YFG2*. In this case there may be no functional similarity at all between *YFG1* and *YFG3*.

Working groups of experts in a given field have been formed to resolve issues in nomenclature for specific gene families. For example, researchers who study phytochrome proposed a phytochrome nomenclature system that has been accepted as the standard for phytochromes (Quail *et al.*, 1994). Recently, a similar community approach has been taken to suggest a nomenclature for phototrophins (Briggs *et al.*, 2001). The advantage of this method is that experts in the field have defined the criteria that must be met for the correct assignment of nomenclature. Matching certain functional criteria in addition to sequence similarity is usually required. For example, in addition to the presence of specific domains, evidence for FMN binding, light-activated autophosphorylation, serine/theronine kinase activity and formation of a flavin C(4a) cysteinyl adduct and its dark decay have been suggested as required attributes for members of the phototropin family (Briggs *et al.*, 2001). The rationale for classification should be explicitly defined for each gene family. For sequence-based methods, BLAST scores, multiple sequence alignments and phylogenetic trees should be presented. Individuals or groups of experts who choose to propose changes in nomenclature should publish their proposals in highly visible journals or databases.

Regardless of the method used to generate a name, names of genes and gene products cannot encompass all information about how they function in an organism. Genes and gene products can be classified in a number of different ways such as shared mutant phenotypes, expression patterns, sequence similarity, and biochemical functions. These alternative modes of description, and emphasis on using controlled vocabularies that uniformly identify what genes and gene products in any organism are doing, will allow more efficient ways of identifying gene functions and roles across species.

Controlled vocabulary

Controlled vocabularies are one way to relate diverse types of data or data from different sources by using a set of shared and defined terms. An advantage of using controlled vocabularies as descriptors of gene function is that it takes the semantic 'load' off of gene nomenclature by using as many terms as are needed to describe a particular gene. These vocabularies do not replace the names or detailed, free-text information, but rather provide a common language of specific terms for grouping identical concepts. Being able to group genes according to their biological roles can be useful in identifying patterns from gene expression data (Schenk *et al.*, 2000) or for classifying gene products and making comparisons between genomes (Arabidopsis Genome Initiative, 2000).

Vocabularies of defined terms used for classification can be organized as simple dictionaries of keywords (e.g. SwissProt) or structured to reflect what is known about the relationships between terms (e.g. the Gene Ontology Consortium, see below). Vocabularies organized as hierarchies allow for queries of a parent term such as 'what gene products are involved in energy metabolism?' that return annotations to its children (e.g. photosynthesis). The Medical Subject Headings (MeSH), MIPs and TIGR functional categories are examples of vocabularies with a hierarchical structure. The MIPS vocabulary is based upon the seminal work of Monica Riley (Riley, 1993) to create a controlled vocabulary for *Escherichia coli* gene function. Additional terms and relationships have been added to accommodate other genomes that are being annotated by MIPS including plants (Frishman *et al.*, 2001).

A system for describing gene products in terms of molecular function, biological process and cellular localization has been developed by the Gene Ontology Consortium (www.geneontology.org), and has been adopted by a number of databases. The GO project seeks to provide a set of shared vocab-

ularies that can be used to describe gene products of any species (Ashburner *et al.*, 2000; Gene Ontology Consortium, 2001). Founded by representatives from *Drosophila* (FlyBase), budding yeast (SGD), and mouse (MGD/MGX) databases, it has since expanded to include *Arabidopsis* (TAIR), nematodes (WormBase), and *Dictyostelium* (DictyBase). The GO structure differs from the previously mentioned vocabularies in that GO terms can have more than one parent. This relationship more accurately reflects biology. Moreover, the content and structure of the ontologies is constantly updated to include relevant terms and to accurately reflect the biological relationships between the terms. Gene products are associated (annotated) with as many terms as deemed appropriate. Because the supporting evidence for each annotation can differ, it is important that GO annotations are associated with an evidence code. Thus inferences of function based upon sequence similarity can be easily distinguished from those based on biochemical assay. In addition, references for the annotations are provided so that the relevant information can be traced to the source. The annotations and their associated sequences are a resource for information transfer from one species to another (Ashburner *et al.*, 2000). GO ontologies have been used to annotate the genomes of fly (Adams *et al.*, 2000; http://flybase.bio.indiana.edu/), yeast (http://genome-www.stanford.edu/Saccharomyces/;www.geneontology.org), man (Venter *et al.*, 2001) and mouse (Kawai *et al.*, 2001; http://www.informatics.jax.org/) and efforts are underway for the annotation of *Arabidopsis* (TAIR, TIGR) and rice genomes (GrameneDB, www.gramene.org).

Many other ontologies and controlled vocabularies have been and are being developed to annotate additional types of biological information such as metabolism, gene expression and mutant phenotypes (Baker *et al.*, 1999; Schulze-Kremer, 1998). The EcoCyc and MetaCyc databases rely upon an ontology of metabolism and gene interactions for querying metabolic pathways and regulatory networks (Karp, 2000). The next wave of data will be coming from systematic approaches to define the function of each gene for a given genome such as knockouts for every gene and systematic analysis of gene expression (http://www.arabidopsis.org/workshop1.html). To cope with the flood of data from these projects, uniform vocabularies are needed for associating shared phenotypes, and patterns of gene expression in terms of anatomy and developmental stages (Eppig, 2000). Vocabularies for microarray data are being developed to facilitate queries and classification of experimental methods including biological sources, experimental treatments and environmental conditions (see Finkelstein *et al.*, this issue; http://www.cbil.upenn.edu/Ontology/MGED_ontology.html). TAIR is collaborating with the Arabidopsis Transposon Insertion Service (ATIS; http://www.jic.bbsrc.ac.uk/staff/michael-bevan/atis/index.htm) to develop common vocabularies to describe anatomy and developmental stages for *Arabidopsis*. Trait vocabularies are available for rice (IRRI) and maize (ZmDB). The Missouri Maize Project/MaizeDB is collaborating with Gramene DB and the International Rice Research Institute (IRRI) to develop an ontology of grass anatomy and development. In the future it will be useful to integrate the vocabularies from different plant species into a resource that will permit cross-species queries.

The use of controlled vocabulary shared by many organisms and databases to describe gene functions will greatly enhance the capability of transferring knowledge across different domains. However, in order to extrapolate the information and devise reasonable hypotheses, the source and reliability of the annotations must also be apparent.

Evidence and attribution

Not all data are of the same quality. The ability to rapidly assess the quality of data can avoid the pitfalls of following misleading or incorrect annotations (Karp, 1998). As described earlier, the type of methods used for the basis of an annotation of a gene should be considered one of the most important aspects of genome-wide annotation efforts. Computational methods of annotation should be distinguished from annotation based upon experimental evidence.

For genomic data generated from computational methods, the analysis tools and parameters should be rigorously documented. Currently there are no standards defined for reporting genomic data. As with experimental data produced from the lab bench, enough detail must be provided for anyone to reproduce the results. Methods of analysis should be defined, as results can differ significantly depending upon what software and database were used. For example, a comparison of results from the annotations of the human genome reveals differences in gene annotation that likely reflect the differences in the methodology used by each group (Aach *et al.*, 2001). In the case of data generated via computational methods such as BLAST, the

evidence should include the scores generated by the analysis method and the parameters used for analysis. However, scores and parameters alone are not sufficient. A vastly ignored aspect of computational analysis is the variation in software versions and in the data sources. A BLAST run against GenBank cannot necessarily be compared to a previous run, as the number of nucleotides in database increases constantly. Because the expect value (e-value) of a BLAST hit is calculated based upon the size of the subject data set (Altschul *et al.*,1990), the e-value for a given sequence changes depending upon when the query was run. Similarly, analysis software can change as new versions are made available and the results may vary depending on the version. Therefore, the supporting evidence should also include information about specific software versions and database releases.

With large-scale genome projects, where massive quantities of data are processed at once, automatic recording of analysis methods and parameters is desirable. Some lab management tools, such as BioNavigator, have this report feature built in.

An important part of data management is tracking the source of data. Having the data traceable to the source facilitates the process of data correction and retrieval of details from the source that may not have been published. The problem of error propagation can be minimized if the source of the problem can be rapidly identified and corrections made (Brenner, 1999). In well-curated databases, the source of data and its history are maintained and made publicly accessible. For users of public databases, this ability to trace data back to their sources is critical to the process of data quality assessment.

Curation

In data management resources that provide large quantities of data, expert assessment of data by biologists is important to assure the uniformity and quality of data. Data acquisition, transformation, validation and annotation are all aspects of curation. Depending upon the size and tractability of the data, some aspects of curation can be done automatically. Large sets of data such as sequence or microarray results are amenable to computational methods. Automatic processing of sequences (i.e. pipelines) that takes raw sequence data, identifies coding regions and exports these sequences to multiple similarity searching packages, necessarily performs operations for data transformation (formatting). Validation steps along the way assure that duplications or incorrect data formats are not created. Analysis reports describing methods and parameters can be automatically generated and linked to the data as evidence. Much of this can be done with little or no manual intervention. However, manual curation is essential to assure data verification and annotation quality. The verification and validation steps not only assure the quality of the data but also can find areas in the automated pipeline methods that can be improved to capture more accurate data. In addition, many issues related to data format such as nomenclature cannot be addressed computationally and require informed input. Annotations, whether computer generated or not, are evaluated by experts and these expert decisions should be captured and made transparent to the user. It is essential that manual vs. automatic methods be distinguished as manual methods are typically of higher quality than automated analyses that have not necessarily been evaluated by experts.

Data exchange formats

Large-scale data that are to be made accessible to the community should be well curated, annotated and documented and appropriately formatted for publication. Journals may publish these data as supplemental material as text, tables or figures or the data may be placed in a public repository and referenced for publication. The data format will depend upon the type of publication method.

For publishing data into a public database, the format is usually defined by the developers of the resource. At present, no universally accepted standards for data formats exist for genomics data; most programs and databases define their own formats. Exchanging data thus often requires scripts to be written in a language such as Perl for converting one format to another. This process can be time-consuming but may be greatly simplified by modifying programs available from BioPerl (www.bioperl.org) for common file formats. Fortunately, standards are emerging, which are notably based on XML (www.xml.org) (Achard *et al.*, 2001). XML-based file formats are GAME (http://www.bioxml.org/Projects/game/game0.1.html) for the genomic annotation data and MAML (http://xml.coverpages.org/maml.html) for microarray data. Non-XML-based formats are also used, such as GFF (http://www.sanger.ac.uk/Software/formats/GFF/) or the ASN.1 format (http://asn1.elibel.tm.fr/) used by GenBank. These data exchange formats can

also be used by developers of small databases to transfer information to and from public databases.

Database implementation

As the quantity of data increases for a project, researchers have to become more sophisticated about data management issues. The best way to handle the information may be to develop a database to manage large quantities of data internally or for public use. The following section provides a brief overview of issues related to designing and implementing a database for biological information.

Conceptual organization of data

The first step in database design is to decide what the database will be used for and how users will interact with it. Once the scope of the database is defined, the data to be stored and how these data are associated with one another is defined. This is often done using a conceptual data model. The model is independent of how the information will be stored in the final, physical implementation on the computer. Entities are defined that informally represent concepts from the real world. The conceptual model developed by Paton *et al.* (2000) for genomic information illustrates how conceptual modeling can be applied to biology. Typical biological entities in genomics are genomes, chromosomes, genes, transcripts, promoters and so forth. In their conceptual model, a genome consists of one or many chromosomes, which can contain one or many chromosome fragments; these chromosome fragments can be either transcribed regions or non-transcribed regions, and so forth. A formal language such as Unified Modeling Language (UML) has been developed for specifying both use cases and conceptual data models (Booch *et al.*,1999). For example, the conceptual data model of TAIR has been developed using UML (http://arabidopsis.org/search/schemas.html).

Physical implementation

For the physical implementation of the data model, a database management system (DBMS) has to be selected. A partial listing of commercial and public products is shown in Table 3. Popular choices for implementing small databases in the lab are FileMaker and Microsoft Access. Both are relational database managers (FileMaker since its latest incarnation) and are quite powerful. Microsoft Access requires a Windows operating system whereas FileMaker is available on both Macintosh and PC platforms, and the files are interchangeable. Both FileMaker and Access have straightforward Web-publication capabilities and intuitive graphical user interface-building capabilities. However, large databases, where thousands of accesses per day are expected, may require an industrial-strength relational database management system (RDBMS), such as Oracle, Sybase or Informix These systems are very powerful but can be expensive. A good alternative to these commercial products are the open-source projects (see below) such as mySQL and PostGres which are available free for most applications. Most of the above databases are queried using a vendor-specific implementation of the Structured Query Language (SQL), which are not 100% compatible to each other. Larger systems should also be implemented using a stable operating system such as UNIX or Linux.

Another popular system for implementing genomic databases is ACEDB, which is a proprietary object oriented database solution for genomics projects that includes pre-defined user interfaces for common applications (Walsh *et al.*,1998). Its query language is not SQL-based, but the system is used by many genomic databases (see http://genome.cornell.edu/acedocs/acedb/faq.html). ACEDB software is an open source project, meaning that the code is freely available for those who wish to use it. Code can be modified to suit specific project requirements and then re-deposited back to the shared resource for others to use.

Conclusion

Modern biology has created an information explosion; the areas of whole-genome sequencing, microarray gene expression, proteomics and now functional genomics have produced a prodigious amount of data. Biologists have a need for tools to manage and analyze these large data sets. Many resources for data handling are available and new tools will be developed to handle the ever-increasing data supply. With an understanding of how the information flow can be managed, biologists can effectively utilize the resources at their disposal to develop testable hypothesis and produce experimental results to share with the research community. The association of controlled vocabularies with biological data will facilitate the ability to perform computational methods and queries of the data. Rigorous data management with detailed record keeping and reporting will enhance the overall quality of

bioinformatics research and development of tools that allow more efficient filtering of the data. With the right tools at our hands, we have the capacity to surf the wave rather than drowning in sea of data.

Acknowledgments

We thank Drs David Flanders and Angela Baldo for their contributions to the text and Eva Huala and Chris Somerville for their comments on the manuscript. We also thank the reviewers, whose excellent comments greatly improved the manuscript. TAIR is supported by NSF grant DBI-9978564.

References

Aach, J., Bulyk, M.L., Church, G.M., Comander, J., Derti, A. and Shendure, J. 2001. Computational comparison of two draft sequences of the human genome. Nature 409: 856–859.

Achard, F., Vaysseix, G. and Barillot, E. 2001. XML, bioinformatics and data integration. Bioinformatics 17: 115–125.

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., George, R.A., Lewis, S.E., Richards, S., Ashburner, M., Henderson, S.N., Sutton, G.G., Wortman, J.R., Yandell, M.D., Zhang, Q., Chen, L.X., Brandon, R.C., Rogers, Y.H., Blazej, R.G., Champe, M., Pfeiffer, B.D., Wan, K.H., Doyle, C., Baxter, E.G., Helt, G., Nelson, C.R., Gabor, G.L., Abril, J.F., Agbayani, A., An, H.J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R.M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E.M., Beeson, K.Y., Benos, P.V., Berman, B.P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M.R., Bouck, J. *et al.*, 2000. The genome sequence of *Drosophila melanogaster*. Science 287: 2185–2195.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. 1990. Basic Local Alignment Search Tool. J. Mol. Biol. 215: 403–410.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N.J., Oinn, T.M., Pagni, M. and Servant, F. 2001. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucl. Acids Res. 29: 37–40.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796–815.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genet. 25: 25–29.

Bairoch, A. and Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucl. Acids Res. 28: 45–48.

Baker, P.G., Goble, C.A., Bechhofer, S., Paton, N.W., Stevens, R. and Brass, A. 1999. An ontology for bioinformatics applications. Bioinformatics 15: 510–520.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L. and Sonnhammer, E.L. 2000. The Pfam protein families database. Nucl. Acids Res. 28: 263–266.

Baxevanis, A.D. 1998. Information retrieval from biological databases. Meth. Biochem. Anal. 39: 98–120.

Baxevanis, A.D. 2001. The Molecular Biology Database Collection: an updated compilation of biological database resources. Nucl. Acids Res. 29: 1–10.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. and Wheeler, D.L. 2000. GenBank. Nucl. Acids Res. 28: 15–18.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. 2000. The Protein Data Bank. Nucl. Acids Res. 28: 235–242.

Brenner, S.E. 1999. Errors in genome annotation. Trends Genet. 15: 132–133.

Busch, M., Mayer, U. and Jurgens, G. 1996. Molecular analysis of the *Arabidopsis* pattern formation of gene GNOM: gene structure and intragenic complementation. Mol. Gen. Genet. 250: 681–691.

Cartinhour, S.W. 1997. Public informatics resources for rice and other grasses. Plant Mol. Biol. 35: 241–251.

Dicks, J., Anderson, M., Cardle, L., Cartinhour, S., Couchman, M., Davenport, G., Dickson, J., Gale, M., Marshall, D., May, S., McWilliam, H., O'Malia, A., Ougham, H., Trick, M., Walsh, S. and Waugh, R. 2000. UK CropNet: a collection of databases and bioinformatics resources for crop plant genomics. Nucl. Acids Res. 28: 104–107.

Eppig, J.T. 2000. Algorithms for mutant sorting: the need for phenotype vocabularies. Mamm. Genome 11: 584–589.

Frishman, D., Albermann, K., Hani, J., Heumann, K., Metanomski, A., Zollner, A. and Mewes, H.W. 2001. Functional and structural genomics using PEDANT. Bioinformatics 17: 44–57.

Gai, X., Lal, S., Xing, L., Brendel, V. and Walbot, V. 2000. Gene discovery using the maize genome database ZmDB. Nucl. Acids Res. 28: 94–96.

Gene Ontology Consortium. 2001. Creating the gene ontology resource: design and implementation. Genome Res. 11: 1425–1433.

Hofmann, K., Bucher, P., Falquet, L. and Bairoch, A. 1999. The PROSITE database, its status in 1999. Nucl. Acids Res. 27: 215–219.

Huala, E., Dickerman, A.W., Garcia-Hernandez, M., Weems, D., Reiser, L., LaFond, F., Hanley, D., Kiphart, D., Zhuang, M., Huang, W., Mueller, L.A., Bhattacharyya, D., Bhaya, D., Sobral, B.W., Beavis, W., Meinke, D.W., Town, C.D., Somerville, C. and Rhee, S.Y. 2001. The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. Nucl. Acids Res. 29: 102–105.

Kaminski, N. 2000. Bioinformatics. A user's perspective. Am. J. Respir. Cell Mol. Biol. 23: 705–711.

Karp, P.D. 2000. An ontology for biological function based on molecular interactions. Bioinformatics 16: 269–285.

Karp, P.D. 1998. What we do not know about sequence analysis and sequence databases. Bioinformatics 14: 753–754

Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, A., Fukunishi, Y., Konno, H., Adachi, J., Fukuda, S., Aizawa, K., Izawa, M., Nishi, K., Kiyosawa, H., Kondo, S., Yamanaka, I., Saito, T., Okazaki, Y., Gojobori, T., Bono, H., Kasukawa, T., Saito, R., Kadota, K., Matsuda, H. A., Ashburner, M., Batalov, S., Casavant, T., Fleischmann, W., Gaasterland, T., Gissi, C., King, B., Kochiwa, H., Kuehl, P., Lewis, S., Matsuo, Y., Nikaido, I., Pesole, G., Quackenbush,

J., Schriml, L.M., Staubli, F., Suzuki, R., Tomita, M., Wagner, L., Washio, T., Sakai, K., Okido, T., Furuno, M., Aono, H., Baldarelli, R., Barsh, G., Blake, J., Boffelli, D., Bojunga, N., Carninci, P., de Bonaldo, M.F., Brownstein, M.J., Bult, C., Fletcher, C., Fujita, M., Gariboldi, M., Gustincich, S., Hill, D., Hofmann, M., Hume, D.A., Kamiya, M., Lee, N.H., Lyons, P., Marchionni, L., Mashima, J., Mazzarelli, J., Mombaerts, P., Nordone, P., Ring, B., Ringwald, M., Rodriguez, I., Sakamoto, N., Sasaki, H., Sato, K., Schonbach, C., Seya, T., Shibata, Y., Storch, K.F., Suzuki, H., Toyo-oka, K., Wang, K.H., Weitz, C., Whittaker, C., Wilming, L., Wynshaw-Boris, A., Yoshida, K., Hasegawa, Y., Kawaji, H., Kohtsuki, S. and Hayashizaki, Y. 2001. Functional annotation of a full-length mouse cDNA collection. Nature 409: 685–690.

Lo Conte, L., Ailey, B., Hubbard, T.J., Brenner, S.E., Murzin, A.G. and Chothia, C. 2000. SCOP: a structural classification of proteins database. Nucl. Acids Res. 28: 257–259.

Meinke, D. 1995. Genetic nomenclature guide. *Arabidopsis thaliana*. Trends Genet. (AUTHOR: PLEASE MENTION VOLUME): 22–23.

Paton, N.W., Khan, S.A., Hayes, A., Moussouni, F., Brass, A., Eilbeck, K., Goble, C.A., Hubbard, S.J. and Oliver, S.G. 2000. Conceptual modelling of genomic information. Bioinformatics 16: 548–557.

Price, C., Reardon, E.M. and Lonsdale, D. 1996. A guide to naming sequenced plant genes. Plant Mol. Biol. 30: 225–227

Rhee, S.Y. 2000. Bioinformatic resources, challenges, and opportunities using *Arabidopsis* as a model organism in a post-genomic era. Plant Physiol. 124: 1460–1464.

Rhee, S.Y. and Flanders, D.J. 2000. Web-based bioinformatic tools for *Arabidopsis* researchers. In: Z.Wilson (Ed.) Arabidopsis: A Practical Approach, Oxford University Press, Oxford, pp. 225–265.

Riley, M. 1993. Functions of the gene products of *Escherichia coli*. Microbiol. Rev. 57: 862–952.

Schenk, P.M., Kazan, K., Wilson, I., Anderson, J.P., Richmond, T., Somerville, S.C. and Manners, J.M. 2000. Coordinated plant defense responses in *Arabidopsis* revealed by microarray analysis. Proc. Natl. Acad. Sci. USA 97: 11655–11660.

Scholl, R.L., May, S.T. and Ware, D.H. 2000. Seed and molecular resources for *Arabidopsis*. Plant Physiol. 124: 1477–1480.

Schulze-Kremer, S. 1998. Ontologies for molecular biology. Pac. Symp. Biocomput.: 695–706.

Shevell, D.E., Leu, W.M., Gillmor, C.S., Xia, G., Feldmann, K.A. and Chua, N.H. 1994. EMB30 is essential for normal cell division, cell expansion, and cell adhesion in *Arabidopsis* and encodes a protein that has similarity to Sec7. Cell 77: 1051–1062.

Siepel, A., Farmer, A., Tolopko, A., Zhuang, M., Mendes, P., Beavis, W. and Sobral, B. 2001. ISYS: a decentralized, component-based approach to the integration of heterogeneous bioinformatics resources. Bioinformatics 17: 83–94.

Stevens, R., Goble, C., Baker, P. and Brass, A. 2001. A classification of tasks in bioinformatics. Bioinformatics 17: 180–188.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D. and Koonin, E.V. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucl. Acids Res. 29: 22–28.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. and Zhu, X. 2001. The sequence of the human genome. Science 291: 1304–51.

Walsh, S., Anderson, M. and Cartinhour, S.W. 1998. ACEDB: a database for genome information. Meth. Biochem. Anal. 39: 299–318.

White, J.A., Apweiler, R., Blake, J.A., Eppig, J.T., Maltais, L.J. and Povey, S. 1999. Report of the Second International Nomenclature Workshop. Cambridge, UK, 1–2 May 1999. Genomics 62: 320–323.

Wixon, J. and Kell, D. 2000. The Kyoto encyclopedia of genes and genomes – KEGG. Yeast 17: 48–55.

Yephremov, A., Wisman, E., Huijser, P., Huijser, C., Wellesen, K. and Saedler, H. 1999. Characterization of the FIDDLE-HEAD gene of *Arabidopsis* reveals a link between adhesion response and cell differentiation in the epidermis. Plant Cell 11: 2187–2201.