B

BIOINFORMATIC DATABASES

At some time during the course of any bioinformatics project, a researcher must go to a database that houses biological data. Whether it is a local database that records internal data from that laboratory's experiments or a public database accessed through the Internet, such as NCBI's GenBank (1) or EBI's EMBL (2), researchers use biological databases for multiple reasons.

One of the founding reasons for the fields of bioinformatics and computational biology was the need for management of biological data. In the past several decades, biological disciplines, including molecular biology and biochemistry, have generated massive amounts of data that are difficult to organize for efficient search, query, and analysis. If we trace the histories of both database development and the development of biochemical databases, we see that the biochemical community was quick to embrace databases. For example, E. F. Codd's seminal paper, "A Relational Model of Data for Large Shared Data Banks" (3), published in 1970 is heralded as the beginning of the relational database, whereas the first version of the Protein Data Bank (PDB) was established at Brookhaven National Laboratories in 1972 (4).

Since then, especially after the launching of the human genome sequencing project in 1990, biological databases have proliferated, most embracing the World Wide Web technologies that became available in the 1990s. Now there are hundreds of biological databases, with significant research efforts in both the biological as well as the database communities for managing these data. There are conferences and publications solely dedicated to the topic. For example, Oxford University Press dedicates the first issue of its journal *Nucleic Acids Research* (which is freely available) every year specifically to biological databases. The database issue is supplemented by an online collection of databases that listed 858 databases in 14 categories in 2006 (5), including both new and updated ones.

Biological database research now encompasses many topics, such as biological data management, curation, quality, integration, and mining (6). Biological databases can be classified in many different ways, from the topic they cover, to how heavily annotated they are or which annotation method they employ, to how highly integrated the database is with other databases. Popularly, the first two categories of classification are used most frequently. For example, there are archival nucleic acid data repositories [GenBank, the EMBL Data Library, and the DNA Databank of Japan (7)] as well as protein sequence motif/domain databases, like PROSITE (8), that are derived from primary source data.

Modern biological databases comprise not only data, but also sophisticated query facilities and bioinformatic data analysis tools; hence, the term "bioinformatic databases" is often used. This article presents information on some popular bioinformatic databases available online, including sequence, phylogenetic, structure and pathway, and microarray databases. It highlights features of these databases, discussing their unique characteristics, and focusing on types of data stored and query facilities available in the databases. The article concludes by summarizing important research and development challenges for these databases, namely knowledge discovery, large-scale knowledge integration, and data provenance problems. For further information about these databases and access to all hyperlinks presented in this article, please visit http://www.csam.montclair.edu/~herbert/bioDatabases.html.

SEQUENCE DATABASES

Genome and protein sequence databases represent the most widely used and some of the best established biological databases. These databases serve as repositories for wet-lab results and the primary source for experimental results. Table 1 summarizes these data repositories and gives their respective URLs.

GenBank, EMBL, and the DNA Databank of Japan

The most widely used biological data bank resource on the World Wide Web is the genomic information stored in the U.S.'s National Institutes of Health's GenBank, the European Bioinformatics Institutes' EMBL, and Japan's National Institute of Genetics DNA Databank of Japan (1, 2, 7). Each of these three databases was developed separately, with GenBank and EMBL launching in 1980 (4). Their collaboration started soon after their development, and DDBJ joined the collaboration shortly after its creation in 1986. The three databases, under the direction of the International Nucleotide Sequence Database Collaboration (INSDC), gather, maintain, and share mainly nucleotide data, each catering to the needs of the region in which it is located (4).

The Ensembl Genome Database

The Ensembl database is a repository of stable, automatically annotated human genome sequences. It is available either as an interactive website or downloadable as flat files. Ensembl annotates and predicts new genes, with annotation from the InterPro (9) protein family databases and with additional annotations from databases of genetic disease [OMIM (10)], expression [SAGE (11,12)] and gene family (13). As Ensembl endeavors to be both portable and freely available, software available at Ensembl is based on relational database models (14).

GeneDB Database

GeneDB (15) is a genome database for prokaryotic and eukaryotic organisms. It currently contains data for 37 genomes generated from the Pathogen Sequencing Unit (PSU) at the Welcome Trust Sanger Institute. The GeneDB

Wiley Encyclopedia of Computer Science and Engineering, edited by Benjamin Wah.

Table 1. Summary of Genome and Protein Sequence Databases

Database	URL	Feature
GenBank	http://www.ncbi.nlm.nih.gov/	NIH's archival genetic sequence database
EMBL	http://www.ebi.ac.uk/embl/	EBI's archival genetic sequence database
DDBJ	http://www.ddbj.nig.ac.jp/	NIG's archival genetic sequence database
Ensembl	http://www.ensembl.org/	Database that maintains automatic annotation on selected eukaryotic genomes
GeneDB	http://www.cebitec.uni-bielefeld.de/groups/brf/software/gendb_info/	Database that maintains genomic information about specific species related to pathogens
TAIR	http://www.arabidopsis.org/	Database that maintains genomic information about *Arabidopsis thaliana*
SGD	http://www.yeastgenome.org/	A repository for baker's yeast genome and biological data
dbEST	http://www.ncbi.nlm.nih.gov/dbEST/	Division of GenBank that contains expression tag sequence data
Protein Information Resource (PIR)	http://pir.georgetown.edu/	Repository for nonredundant protein sequences and functional information
Swiss-Prot/TrEMBL	http://www.expasy.org/sprot/	Repository for nonredundant protein sequences and functional information
UniProt	http://www.pir.uniprot.org/	Central repository for PIR, Swiss-Prot, and TrEMBL

database has four key functionalities. First, the database stores and frequently updates sequences and annotations. Second, GeneDB provides a user interface, which can be used for access, visualization, searching, and downloading of the data. Third, the database architecture allows integration of different biological datasets with the sequences. Finally, GeneDB facilitates querying and comparisons between species by using structured vocabularies (15).

The Arabidopsis Information Resource (TAIR)

TAIR (16) is a comprehensive genome database that allows for information retrieval and data analysis pertaining to *Arabidopsis thaliana* (a small annual plant belonging to the mustard family). *Arabidopsis thaliana* has been of great interest to the biological community and is one of the few plants whose genome is completely sequenced (16). Due to the complexity of many plant genomes, *Arabidopsis thaliana* serves as a model for plant genome investigations. The database has been designed to be simple, portable, and efficient. One innovate aspect of the TAIR website is MapViewer (http://www.arabidopsis.org/servlets/mapper). MapViewer is an integrated visualization tool for viewing genetic, physical, and sequence maps for each *Arabidopsis* chromosome. Each component of the map contains a hyperlink to an output page from the database that displays all the information related to this component (16).

SGD: Saccharomyces Genome Database

The Saccharomyces Genome Database (SGD) (17) provides information for the complete *Saccharomyces cerevisiae* (baker's and brewer's yeast) genomic sequence, along with its genes, gene products, and related literature. The database contains several types of data, including DNA sequence, gene-encoded proteins, and the structures and biological functions of any known gene products. It also allows full-text searches of articles concerning *Saccharomyces cerevisiae*. The SGD database is not a primary sequence repository (17), but a collection of DNA and protein sequences from existing databases [GenBank (1), EMBL (2), DDBJ (7), PIR (18), and Swiss-Prot (19)]. It organizes the sequences into datasets to make the data more useful and easily accessible.

dbEST Database

dbEST (20) is a division of GenBank that contains sequence data and other information on short, "single-pass" cDNA sequences, or Expressed Sequence Tags (ESTs), generated from randomly selected library clones (http://www.ncbi.nlm.nih.gov/dbEST/). dbEST contains approximately 36,843,572 entries from a broad spectrum of organisms. Access to dbEST can be obtained through the Web, either from NCBI by anonymous ftp or through Entrez (21). The dbEST nucleotide sequences can be searched using the BLAST sequence search program at the NCBI website. In addition, TBLASTN, a program that takes a query amino acid sequence and compares it with six-frame translations of dbEST DNA sequences can also be useful for finding novel coding sequences. EST sequences are available in the FASTA format from the "/repository/dbEST" directory at ftp.ncbi.nih.gov.

The Protein Information Resource

The Protein Information Resource (PIR) is an integrated public bioinformatics resource that supports genomic and proteomic research and scientific studies. PIR has provided many protein databases and analysis tools to the scientific community, including the PIR-International Protein Sequence Database (PSD) of functionally annotated protein sequences. The PIR-PSD, originally created as the Atlas of Protein Sequence and Structure edited by Margaret Dayhoff, contained protein sequences that were highly annotated with functional, structural, bibliographic, and sequence data (5,18). PIR-PSD is now merged with UniProt Consortium databases (22). PIR offers the

PIRSF protein classification system (23) that classifies proteins, based on full-length sequence similarities and their domain architectures, to reflect their evolutionary relationships. PIR also provides the iProClass database that integrates over 90 databases to create value-added views for protein data (24). In addition, PIR supports a literature mining resource, iProLINK (25), which provides multiple annotated literature datasets to facilitate text mining research in the areas of literature-based database curation, named entity recognition, and protein ontology development.

The Swiss-Prot Database

Swiss-Prot (19) is a protein sequence and knowledge database and serves as a hub for biomolecular information archived in 66 databases (2). It is well known for its minimal redundancy, high quality of annotation, use of standardized nomenclature, and links to specialized databases. Its format is very similar to that of the EMBL Nucleotide Sequence Database (2). As Swiss-Prot is a protein sequence database, its repository contains the amino acid sequence, the protein name and description, taxonomic data, and citation information. If additional information is provided with the data, such as protein structures, diseases associated with the protein or splice isoforms, Swiss-Prot provides a table where these data can be stored. Swiss-Prot also combines all information retrieved from the publications reporting new sequence data, review articles, and comments from enlisted external experts.

TrEMBL: A Supplement to Swiss-Prot

Due to the large number of sequences generated by different genome projects, the Swiss-Prot database faces several challenges related to the processing time required for manual annotation. For this reason, the European Bioinformatics Institute, collaborating with Swiss-Prot, introduced another database, TrEMBL (translation of EMBL nucleotide sequence database). This database consists of computer-annotated entries derived from the translation of all coding sequences in the nucleotide databases. This database is divided into two sections: SP-TrEMBL contains sequences that will eventually be transferred to Swiss-Prot and REM-TrEMBL contains those that will not go into Swiss-Prot, including patent application sequences, fragments of less than eight amino acids, and sequences that have proven not to code for a real protein (19, 26, 27).

UniProt

With protein information spread over multiple data repositories, the efforts from PIR, SIB's Swiss-Prot and EBI's TrEMBL were combined to develop the UniProt Consortium Database to centralize protein resources (22). UniProt is organized into three layers. The UniProt Archive (UniParc) stores the stable, nonredundant, corpus of publicly available protein sequence data. The UniProt Knowledgebase (UniProtKB) consists of accurate protein sequences with functional annotation. Finally, the UniProt Reference Cluster (UniRef) datasets provide nonredundant reference clusters based primarily on UniProtKB. UniProt also offers users multiple tools, including searches against the individual contributing databases, BLAST and multiple sequence alignment, proteomic tools, and bibliographic searches (22).

PHYLOGENETIC DATABASES

With all of the knowledge accumulating in the genomic and proteomic databases, there is a great need for understanding how all these types of data relate to each other. As all biological things have come about through the evolutionary process, the patterns, functions, and processes that they possess are best analyzed in terms of their phylogenetic histories. The same gene can evolve a different timing of its expression, a different tissue where it is expressed, or even gain a whole new function along one phylogenetic branch as compared with another. These changes along a branch affect the biology of all descendant species, thereby leaving phylogenetic patterns in everything we see. A detailed mapping between biological data and phylogenetic histories must be accomplished so that the full potential of the data accumulation activities can be realized. Otherwise it will be impossible to understand why certain drugs work in some species but not others, or how we can design therapies against evolving disease agents such as HIV and influenza.

The need to query data using sets of evolutionary related taxa, rather than on single species, has brought up the need to create databases that can serve as repositories of phylogenetic trees, generated by a variety of methods. Phylogeny and phylogenetic trees give a picture of the evolutionary history among species, individuals, or genes. Therefore, there are at least two distinct goals of a phylogenetic database: archival storage and analysis (28). Table 2 summarizes these repositories.

Many of the aforementioned data repositories offer functionalities for browsing phylogenetic and taxonomic information. NCBI offers users the Taxonomy Databases (1, 13), which organize the data maintained in its repositories from the species perspective and allows the user to hierarchically browse data with respect to a Tree of Life organization. NEWT is a taxonomy database (http://www.ebi.ac.uk/newt/) that connects UniProtKB data to the NCBI taxonomy data. For every species, NEWT provides information about the taxon's scientific name, common name and synonym(s), lineage, number of UniProtKB protein entries in the given taxon, and links to each entry.

Tree of Life

The Tree of Life (29) is a phylogenetic repository that aims to provide users with information from a whole-species point of view. The Tree of Life allows users to search for pages about specific species through conventional keyword search mechanisms. Most interestingly, a user can also navigate through the "tree of life" using hierarchical browsing starting at the root organism, popularly referred to as "Life," and traverse the tree until a species of interest is reached. The species web page contains information gathered and edited by recognized experts about the species

Table 2. Summary of Phylogenetic Data Repositories

Database	URL	Feature
NCBI Taxonomy Database	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Taxonomy	Whole-species view of genomic and proteomic data stored in GenBank
Tree of Life	http://tolweb.org/tree/	Species-centric hierarchical browsing database modeling the evolutionary relationships between species
TreeFam	http://www.treefam.org/	Repository for phylogenetic trees based on animal genomes
TreeBASE	http://www.treebase.org/treebase/	Archival peer-reviewed phylogenetic tree repository
SYSTERS	http://systers.molgen.mpg.de/	Protein cluster repository with significant phylogenetic functionalities
PANDIT	http://www.ebi.ac.uk/goldman-srv/pandit/	Protein domains repository with inferred phylogenetic trees

as well as peer-reviewed resources accessible through hyperlinks (29).

TreeFam

TreeFam is a database of phylogenetic trees of animal gene families. The goal of TreeFam is to develop a curated database that provides accurate information about ortholog and paralog assignments and evolutionary histories of various gene families (30). To create and curate the trees and families, TreeFam has gathered sequence data from several protein repositories. It contains protein sequences for human (*Homo sapiens*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), chicken (*Gallus gallus*), pufferfish (*Takifugu rubripes*), zebrafish (*Danio rerio*), and fruitfly (*Drosophila melanogaster*), which were retrieved from Ensembl (14), WormBase (31), SGD (17), GeneDB (15), and TIGR (32). The protein sequences in TreeFam are grouped into families of genes that descended from a single gene in the last common ancestor of all animals, or that first appeared in animals. From the above sources, families and trees are automatically generated and then manually curated based on expert review. To manage these data, TreeFam is divided into two parts. TreeFAM-B consists of the automatically generated trees. It obtains clusters from the PhIGs (33) database and uses BLAST (34), MUSCLE (35), and HMMER (36) and neighbor-joining algorithms (37) to generate the trees. TreeFAM-A contains the manually curated trees, which exploit algorithms similar to the DLI algorithm (DLI: H. Li, unpublished data) and the SDI algorithm (38). TreeFAM contains 11,646 families including about 690 families that have curated phylogenetic trees. Therefore, as more trees get curated, the TreeFam-A database increases, whereas TreeFam-B decreases in size.

TreeBASE

TreeBASE (39) was developed to help harness the explosively high growth in the number of published phylogenetic trees. It is a relational database and contains phylogenetic trees and the data underlying those trees. TreeBASE is available at http://www.treebase.org and allows the user to search the database according to different keywords and to see graphical representations of the trees. The user can also access information such as data matrices, bibliographic information, taxonomic names, character states, algorithms used, and analyses performed. Phylogenetic trees are submitted to TreeBASE by the authors of the papers that describe the trees. For data to be accepted by TreeBASE, the corresponding paper must pass the journal's peer review process (39).

SYSTERS Database

SYSTERS is a protein clustering database based on sequence similarity (40). It can be accessed at http://SYSTERS.molgen.mpg.de/. SYSTERS contains 185,000 disjoint protein families gathered from existing sequence repositories: Swiss-Prot (19), TrEMBL (19) and complete genomes: Ensembl (14), The Arabidopsis Information Resource (16), SGD (17), and GeneDB (15). Two innovative features of this repository are the SYSTERS Table and SYSTERS Tree. The SYSTERS Table for a family cluster contains a variety of information, most notably accession numbers as well as accession numbers for a variety of external databases [IMB (41), MSD (42), ENZYME (43), INTERPRO (9), PROSITE (8), GO (44)]. There can be several redundant entries in the table for one protein sequence. As SYSTERS data rely on external protein databases, there is always an entry name (protein name) and an accession number for each entry but there may not be a gene name. For each family cluster that consists of more than two nonredundant entries, a phylogenetic tree is available. The phylogenetic trees are constructed using the UPGMA (45) method. No more than 200 entries are displayed in a tree; the selection process when a cluster contains more than 200 entries is not clear.

PANDIT (Protein and Associated Nucleotide Domains with Inferred Trees)

PANDIT is a nonredundant repository of multiple sequence alignments and phylogenetic trees. It is available at http://www.ebi.ac.uk/goldman-srv/pandit. The database consists of three portions: protein domain sequence alignments from Pfam Database (46), alignments of nucleotide sequences derived from EMBL Nucleotide Sequence Database (2), and phylogenetic trees inferred from each alignment. Currently PANDIT contains 7738 families of homologous protein sequences with corresponding DNA sequences and phylogenetic trees. All alignments are based

on Pfam-A (47) seed alignments, which are manually curated and, therefore, make PANDIT data high quality and comparable with alignments used to study evolution. Each family contains three alignments: PANDIT-aa contains the exact Pfam-A seed protein sequence alignment; and PANDIT-dna contains the DNA sequences encoding the protein sequences in PANDIT-aa that could be recovered; and PANDIT-aa-restricted contains only those protein sequences for which a DNA sequence has been recovered. The DNA sequences have been retrieved using cross-references to the EMBL Nucleotide Sequence Database from the Swiss-Prot (19) and TrEMBL (19) databases. To ensure accuracy, PANDIT performs a translation of the cross-referenced DNA sequences back to the corresponding protein sequences.

PANDIT database is intended for studying the molecular evolution of protein families. Therefore, phylogenetic trees have been constructed for families of more than two sequences. For each family, five different methods for tree estimation have been used to produce candidate trees. These methods include neighbor-joining (37), BioNJ (48), Weighbor (49), FastME (50), and PHYML (51). Neighbor-joining, BioNJ and Weighbor are methods used to produce phylogenetic tree estimates from a pairwise distance matrix. FastME uses a minimum evolution criterion with local tree-rearrangements to estimate a tree, and Phyml uses maximum likelihood with local tree searching. At the end, the likelihood of each tree from the candidate set is computed and the tree with the highest likelihood is added to the database.

STRUCTURE AND PATHWAY DATABASES

Knowledge of protein structures and of molecular interactions is key to understanding protein functions and complex regulatory mechanisms underlying many biological processes. However, computationally, these datasets are highly complex. The most popular ways to model these datasets are through text, graphs, or images. Text data tend not to have the descriptive power needed to fully model this type of data. Graphical and the image data require complex algorithms that are computationally expensive and not reliably accurate. Therefore, structural and pathway databases become an interesting niche from both the biological and the computational perspectives. Table 3 lists several prominent databases in this field.

The Protein Data Bank

The Protein Data Bank (PDB) is an archive of structural data of biological macromolecules. PDB is maintained by the Research Collaboratory for Structural Bioinformatics (RCSB). It allows the user to view data both in plain text and through a molecular viewer using Jmol. A key goal of the PDB is to make the data as uniform as possible while improving data accessibility and providing advanced querying options (52, 53).

To have complete information regarding the features of macromolecular structures, PDB allows a wide spectrum of queries through data integration. PDB collects and integrates external data from scientists' deposition, Gene Ontology (GO) (54), Enzyme Commission (55), KEGG Pathways (56), and NCBI resources (57). PDB realizes data integration through data loaders written in Java, which extract information from existing databases based on common identification numbers. PDB also allows data extraction at query run time, which means implemented Web services extract information as the query is executing.

The Nucleic Acid Database

Nucleic Acid Database, also curated by RCSB and similar to the PDB and the Cambridge Structural Database (58), is a repository for nucleic acid structures. It gives users access to tools for extracting information from nucleic acid structures and distributes data and software. The data are stored in a relational database that contains tables of primary and derivative information. The primary information includes atomic coordinates, bibliographic references, crystal data, data collection, and other structural descriptions. The derivative information is calculated from the primary information and includes chemical bond lengths, and angles, virtual bond lengths, and other measures according to various algorithms (59, 60). The experimental data in the NDB database have been collected from published literature, as well as from one of the standard crystallographic archive file types (60, 61) and other sources. Primary information has been encoded in ASCII format file (62). Several programs have been developed to convert between different file formats (60, 63, 64, 65).

The Kyoto Encyclopedia of Genes and Genomes

The Kyoto Encyclopedia of Genes and Genomes (KEGG) (56) is the primary resource for the Japanese GenomeNet service that attempts to define the relationships between

Table 3. Summary of Structural and Pathway Databases

Database	URL	Feature
The Protein Data Bank (PDB)	http://www.rcsb.org/pdb/	Protein structure repository that provides tools for analyzing these structures
The Nucleic Acid Database (NDB)	http://ndbserver.rutgers.edu/	Database housing nucleic acid structural information
The Kyoto Encyclopedia of Genes and Genomes (KEGG)	http://www.genome.jp/kegg/	Collection of databases integrating pathway, genomic, proteomic, and ligand data
The BioCyc Database Collection	http://www.biocyc.org/	Collection of over 200 pathway and genomic databases

the functional meanings and utilities of the cell or the organism and its genome information. KEGG contains three databases: PATHWAY, GENES, and LIGAND. The PATHWAY database stores computerized knowledge on molecular interaction networks. The GENES database contains data concerning sequences of genes and proteins generated by the genome projects. The LIGAND database holds information about the chemical compounds and chemical reactions that are relevant to cellular processes. KEGG computerizes the data and knowledge as graph information. The KEGG/PATHWAY database contains reference diagrams for molecular pathways and complexes involving various cellular processes, which can readily be integrated with genomic information (66). It stores data objects called generalized protein interaction networks (67, 68). The PATHWAY database is composed of four levels that can be accessed through the Web browser. The top two levels contain information about metabolism, genetic information processing, environmental information processing, cellular processes, and human diseases. The others relate to the pathway diagram and the ortholog group table, which is a collection of genes and proteins.

The BioCyc Database Collection

The BioCyc Database Collection (69) is a compilation of pathway and genome information for different organisms. Based on the number of reviews and updates, BioCyc databases are organized into several tiers. Tier 1 consists of three databases, EcoCyc (70), which describes *Escherichia coli* K-12; MetaCyc (71), which describes pathways for more than 300 organisms; and the BioCyc Open Compounds Database (69), which contains a collection of chemical compound data from BioCyc databases. Tier 2 contains 12 databases computationally generated by the Pathologic program. These databases have been updated and manually curated to varying degrees. Tier 3 is composed of 191 databases computationally generated by the Pathologic program with no review and updating (69).

The BioCyc website allows scientists to perform certain operations, e.g., to visualize individual metabolic pathways, to view the complete metabolic map of an organism, and to analyze, metabolomics data using the Omics Viewer. The website also provides a spectrum of browsing capabilities such as moving from a display of an enzyme to a display of a reaction that the enzyme catalyzes or to the gene that encodes the enzyme (69).

MICROARRAY AND BOUTIQUE BIOINFORMATIC DATABASES

Both microarray databases and boutique databases offer interesting perspectives on biological data. The microarray databases allow users to retrieve and interact with data from microarray experiments. Boutique databases offer users specialty services concerning a particular aspect of biological data. This section reviews such databases and synopsizes these reviews in Table 4.

The Stanford Microarray Database

The Stanford Microarray Database (SMD) (72) allows researchers to retrieve, analyze, and visualize gene expression data from microarray experiments. The repository also contains literature data and integrates multiple related resources, including SGD (17), YPD and WormPD (73), UniGene (74), dbEST (20), and Swiss-Prot (19).

Due to the large number of experiments and datasets, SMD uses comprehensive interfaces allowing users to efficiently query the database. For each experiment, the database stores the name of the researcher, the source organism of the microarray probe sequences, along with a category and subcategory that describe the biological view of the experiment. The user can create a query using any of these criteria to narrow down the number of experiments.

The Yale Microarray Database

The Yale Microarray Database (YMD) (75) is another repository for gene expression data. It is Web-accessible and enables users to perform several operations, e.g., tracking DNA samples between source plates and arrays and finding common genes/clones across different microarray platforms. Moreover, it allows the user to access the image file server, to enter data, and to get integrated data through linkage of gene expression data to annotation databases for functional analysis (75). YMD provides several means of querying the database. The website contains a query criteria interface (75), which allows the user to perform common queries. The interface also enables the user to choose the format of the output, e.g., which columns to be included and the type of output display (HTML, EXCEL, TEXT, or CLUSTER). Finally, the query output can also be dynamically linked to external annotation databases such as DRAGON (76).

Table 4. Summary of Microarray and Boutique Databases

Database	URL	Feature
The Stanford Microarray Database	http://genome-www5.stanford.edu/	Repository for raw and normalized microarray data
The Yale Microarray Database	http://www.med.yale.edu/microarray/	Repository for raw and normalized microarray data
The Stem Cell Database	http://stemcell.princeton.edu/	Database for human and mice stem cell data
The BrainML Data Server	http://www.neurodatabase.org,	Databases containing information necessary for understanding brain processes

The Stem Cell Database

The Stem Cell Database (SCDb) (77), supported by Princeton University and the University of Pennsylvania, is a unique repository that contains information about hematopoietic stem cells from mice and humans. It is closely associated with the Stromal Cell Database (http://stromalcell.princeton.edu/), also supported by Princeton University and the University of Pennsylvania. Data for this repository are obtained from various peer-reviewed sources, publications, and libraries. Users can query on various aspects of the data, including gene name and other annotations, as well as sequence data (77).

The BrainML Data Server

The BrainML Data Server is a repository containing data that pertain to the understanding of neural coding, information transmission, and brain processes and provides a venue for sharing neuro-physiological data. It acquires, organizes, annotates, archives, delivers, and displays single- and multi-unit neuronal data from mammalian cerebral cortex (78). Users can obtain the actual datasets, provided by several laboratories, all in common format and annotated compatibly. The Web interface provides a tool called QueryTool that allows the user to search by metadata terms submitted by the researchers. Another Tool, Virtual Scilloscope Java Tool, displays time-series and histogram datasets dynamically. The datasets can also be downloaded for analysis.

RESEARCH CHALLENGES AND ISSUES

Although extensive efforts have been made to catalog and store biological and chemical data, there is still a great amount of work to be done. Scientists are figuratively drowning in data. Therefore, there is a strong need for computational tools that allow scientists to slice through the mounds of data to pinpoint information needed for experiments. Moreover, with research methodologies changing from library-based to Web-based, new methods for maintaining the quality of the data are needed. Maintenance and updates on bioinformatic databases require not only automatic tools but in most cases also in the curation process. This process involves manual checks from biologists to ensure that data are valid and accurate before integrating this data into the database. There are two major research challenges in the area of bioinformatic databases: (1) development of software tools that are reliable, scalable, downloadable, platform-independent, user-friendly, high performance, and open source for discovering, extracting, and delivering knowledge from large amounts of text and biomedical data; and (2) development of large-scale ontology-assisted knowledge integration systems. The two issues also give rise to others, such as how we can maintain the quality (79) and the proper provenance of biological data when it is heavily integrated. Some work has been done toward the first issue, as discussed in Ref. 80.

Knowledge Discovery from Data (KDD)

The KDD process, in its most fundamental form, is to extract interesting, nontrivial, implicit, previously unknown, and potentially useful information from data. When applied to bioinformatic databases, KDD refers to diverse activities, including bioinformatic data cleaning and preprocessing, pattern and motif discovery, classification and clustering, biological network modeling, and bioinformatic data visualization, to name a few. An annual KDD Cup is organized as the Data Mining and Knowledge Discovery competition by the ACM Special Interest Group (81, 82). Various KDD tools have been developed to analyze DNA and protein sequences, whole genomes, phylogeny and evolutionary trees, macromolecule structures, and biological pathways. However, many of these tools suffer from inefficiency, low accuracy, and unsatisfactory performance due to factors, including experimental noise, unknown model complexity, visualization difficulties with very high-dimensional data, and the lack of sufficient samples for computational validation. Another problem is that some KDD tools are platform dependent and their availability is limited.

One emerging trend in KDD is to apply machine learning, natural language processing, and statistical techniques to text and biomedical literature mining. The goal is to establish associations between biological objects and publications from literature databases such as MEDLINE, for example, finding all related literature studying the same proteins from different aspects. It has been shown that incorporating information obtained from biomedical literature mining into sequence alignment tools such as BLAST can increase the accuracy of alignment results. This shows an example of combining KDD methods with traditional sequence analysis tools to improve their performance. However, these KDD methods are not yet fully reliable, scalable, or user-friendly, and many of the methods still need to be improved.

Large-Scale Knowledge Integration (LKI)

LKI of heterogeneous, distributed bioinformatic data is supposed to offer users a seamless view of knowledge. However, with a few exceptions, many current bioinformatic systems use hyperlink navigation techniques to integrate World Wide Web repositories. These techniques result in semantically meaningless integrations. Often, websites are not maintained, datasets are poorly curated, or in some cases, the integration has been done improperly. With these concerns, efforts based on current biological data integration that create advanced tools to help deliver knowledge to the bioinformatics community fail or become dataset dependent.

A major research challenge in bioinformatics is integrating and representing knowledge effectively. The informatics community has effectively integrated and visualized data. However, research must be taken to the next phase where knowledge integration and knowledge management becomes a key interest. The informatics community must

work with the biomedical community from the ground up. Effective, structured knowledge bases need to be created that are also relatively easy to use. The computer science community is starting to address this challenge with projects in the areas of the Semantic Web and semantic integration. The bioinformatics community has started to create such knowledge bases with projects like the Gene Ontology (GO) and Stanford's biomedical ontology (http://bioontology.org/) (more are listed under the Open Biological Ontology, http://obo.sourceforge.net/). Ontologies and meta-data are only the beginning. It is well known in the computer science community that meta-data management can be a tricky, complicated process. Attempting this in the biomedical realm is downright difficult. Researchers currently must wield complicated ontologies to classify even more complex data. Extensive research is needed into how to develop better ontologies as well as to manipulate them more effectively.

Ontology also assumes that there is a general consensus within the bioinformatics field as to the format and structure of the data, with mechanisms for minimizing synonyms and homonyms. This is not true for many types of data. For example, many plant species have binomial names identical to animal species. Many genes have been given different names when found in one species or one tissue as compared with another. In almost every area of medicine as well as biology, researchers can identify contentious nomenclature issues. This standardized naming problem has serious consequences.

KDD and LKI are not separate; rather, they interact with each other closely. For example, as mentioned, one area of KDD is extracting knowledge from peer-reviewed journal articles for clinical use. However, due to the variety of ways to specify biological objects such as species, regulatory pathways, and gene names, KDD tools have difficulty extracting knowledge from these articles. These articles often represent a majority of data the scientific community have concerning the various biological objects. Due to the lack of standardized representations, one can only employ information retrieval algorithms and give the user a confidence level to the knowledge extracted. Great amounts of knowledge are lost because we cannot exploit a standardized knowledge base while examining peer-reviewed literature. As another example, the GO contains a graph structure that illustrates the relationship among molecular functions attributed to genes. If this structure can be combined with KDD processes such as clustering and classification algorithms, one can produce more biologically meaningful clusters or classification outcomes. These examples illustrate the importance of combining KDD and LKI, which is a challenging problem in the field.

Data Provenance

As demonstrated by the above databases as well as the previous issues, there are large quantities of data interchanging between tens if not hundreds of databases regularly. Furthermore, scientists are revolutionizing how research is done by relying more and more on the biological databases and less and less on original journal articles. Thus, the issue of preserving how the data are obtained becomes a paramount concern (83). The field of data provenance investigates how to maintain meta-data describing the history of a data item within the database. With databases cross-listing each other's entries, and with data mining and knowledge discovery algorithms generating new information based on data published in these databases, the issue of data provenance becomes more and more significant.

BIBLIOGRAPHY

1. D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp, D. L. Wheeler, GenBank, *Nuc. Acids Res.*, **28**: 15–18, 2000.
2. G. Cochrane, P. Aldebert, N. Althorpe, M. Andersson, W. Baker, A. Baldwin, et al., EMBL Nucleotide Sequence Database: developments in 2005, *Nuc. Acids Res.*, **34**(1): D10–D15, 2006.
3. E. F. Codd, A relational model of data for large shared data banks, *CACM*, **13**(6): 377–387, 1970.
4. A. M. Lesk, *Database Annotation in Molecular Biology*. West Sussex, England: John Wiley & Sons, 2005.
5. M. Y. Galperin, The molecular biology database collection: 2006 update, *Nuc. Acids Res.*, **34**: D3–D5, 2006.
6. J. T. L. Wang, C. H. Wu, and P. P. Wang, *Computational Biology and Genome Informatics*, Singapore: World Scientific Publishing, 2003.
7. K. Okubo, H. Sugawara, T. Gojobori, and Y. Tateno, DDBJ in preparation for overview of research activities behind data submissions *Nuc. Acids Res.*, **34**(1): D6–D9, 2006.
8. N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, E. DeCastro, P. S. Langendijk-Genevaux, M. Pagni, C. J. A. Sigrist. The PROSITE database. *Nuc. Acids Res.*, **34**(1): D227–D230, 2006.
9. N. J. Mulder, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, et al., InterPro, progress and status in 2005. *Nuc. Acids Res.*, **33**: D201–205, 2006.
10. S. E. Antonarakis and V. A. McKusick, OMIM passes the 1,000-disease-gene mark, *Nature Genet.*, **25**: 11, 2000.
11. V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler, Serial analysis of gene expression. *Science*, **270**, 484–487, 1995.
12. D. L. Wheeler, D. M. Church, A. E. Lash, D. D. Leipe, T. L. Madden, J. U. Pontius, G. D. Schuler, L. M. Schriml, T. A. Tatusova, L. Wagner, and B. A. Rapp, Database resources of the National Center for Biotechnology Information, *Nuc. Acids Res.*, **29**: 11–16, 2001, Updated article: *Nuc. Acids Res.*, **30**: 13–16, 2002.
13. A. J. Enright, I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis, Protein interaction maps for complete genomes based on gene fusion events, *Nature*, **402**, 86–90, 1999.
14. T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, et al., The Ensembl genome database project. *Nuc. Acids Res.*, **30**, 38–41, 2002.
15. C. Hertz-Fowler, C. S. Peacock, V. Wood, M. Aslett, A. Kerhornou, P. Mooney, et al., GeneDB: A resource for prokaryotic and eukaryotic organisms, *Nuc. Acids Res.*, **32**: D339–D343, 2004.
16. D. W. Meinke, J. M. Cherry, C. Dean, S. D. Rounsley, and M. Koornneef, Arabidopsis thaliana: A model plant for genome analysis, *Science*, **282**: 679–682, 1998.
17. J. M. Cherry, C. Adler, C. Ball, S. A. Chervitz, S. S. Dwight, E. T. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, S. Weng, and D. Botstein, SGD: Saccharomyces Genome Database, *Nuc. Acids Res.*, **26**: 73–79, 1998.

18. C. H. Wu, L. S. Yeh, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Z. Hu, R. S. Ledley, P. Kourtesis, B. E. Suzek, C. R. Vinayaka, J. Zhang, W. C. Barker, The protein information resource, *Nuc. Acids Res.*, **31**: 345–347, 2003.
19. C. O'Donovan, M. J. Martin, A. Gattiker, E. Gasteiger, A. Bairoch, and R. Apweiler, High-quality protein knowledge resource: SWISSPROT and TrEMBL. *Brief. Bioinform.*, **3**: 275–284, 2002.
20. M. S. Boguski, T. M. Lowe, and C. M. Tolstoshev, dbEST — database for expressed sequence tags, *Nature Genet.*, **4**: 332–333, 1993.
21. G. D. Schuler, J. A. Epstein, H. Ohkawa, and J. A. Kans, Entrez: Molecular biology database and retrieval system, *Methods Enzymol.*, **266**: 141–162, 1996.
22. C. H. Wu, R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, et al., The Universal Protein Resource (UniProt): AN expanding universe of protein information. *Nuc. Acids Res.*, **34**(1): D187–191, 2006.
23. C. H. Wu, A. Nikolskaya, H. Huang, L. S. Yeh, D. A. Natale, C. R. Vinayaka, et al., PIRSF: Family classification system at the Protein Information Resource. *Nuc. Acids Res.*, **32**: D112–114, 2004.
24. C. H. Wu, H. Huang, A. Nikolskaya, Z. Z. Hu, and W. C. Barker, The iProClass integrated database for protein functional analysis. *Comput Biol Chem.*, **28**: 87–96, 2004.
25. Z. Z. Hu, I. Mani, V. Hermoso, H. Liu, C. H. Wu, iProLINK: An integrated protein resource for literature mining. *Comput Biol Chem.*, **28**: 409–416, 2004.
26. E. Gasteiger, E. Jung, and A. Bairoch, SWISS-PROT: Connecting biomolecular knowledge via a protein database, *Curr. Issues Mol. Biol.*, **3**: 47–55, 2001.
27. T. Etzold, and P. Argos, SRS—an indexing and retrieval tool for flat file data libraries. *Comput. Appl. Biosci.*, **9**: 49–57, 2003.
28. J. T. L. Wang, M. J. Zaki, H. T. T. Toivonen, and D. Shasha (eds), *Data mining in Bioinformatics*, London, UK: Springer, 2005.
29. D. R. Maddison, and K.-S. Schulz (eds.). The Tree of Life Web Project. Available: http://tolweb.org. Last accessed July 26, 2006.
30. W. M. Fitch, Distinguishing homologous from analogous proteins, *Syst. Zool.*, **19**: 99–113, 1970.
31. N. Chen, T. W. Harris, I. Antoshechkin, C. Bastiani, T. Bieri, D. Blasiar, et al., WormBase: A comprehensive data resource for Caenorhabditis biology and genomics, *Nuc. Acids Res.*, **33**: D383–D389, 2005.
32. B. J. Haas, J. R. Wortaman, C. M. Ronning, L. I. Hannick, R. K. Smith Jr., et al., Complete reannotation of the Arabidopsis genome: Methods, tools, protocols and the final release. *BMC Biol.*, **3**:7, 2005.
33. P. Dehal, and J. L. Boore, Two rounds of whole genome duplication in the ancestral vertebrate, *PLoS Biol.*, **3**: e314, 2005.
34. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, *Nuc. Acids Res.*, **25**: 3389–3402, 1997.
35. R. C. Edgar, MUSCLE: A multiple sequence alignment method with reduced time and space complexity, *BMC Bioinformatics*, **5**:113, 2004.
36. S. R. Eddy, Profile hidden Markov models. *Bioinformatics*, **14**: 755–763, 1998.
37. N. Saitou and M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Mol. Biol. Evol.*, **4**: 406–425, 1987.
38. C. M. Zmasek and S. R. Eddy, A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics*, **17**: 821–828, 2001.
39. M. J. Sanderson, M. J. Donoghue, W. H. Piel, and T. Eriksson, TreeBASE: A prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life, *Am. J. Bot.*, **81**(6): 163 1994.
40. T. Meinel, A. Krause, H. Luz, M. Vingron, and E. Staub, The SYSTERS Protein Family Database in 2005. *Nuc. Acids Res.*, **33**: D226–D229, 2005.
41. J. Reichert, J. Suhnel, The IMB jena image library of biological macromolecules: 2002 update, *Nuc. Acids Res.*, **30**: 253–254, 2002.
42. H. Boutzelakis, D. Dimitropoulos, J. Fillon, A. Golovin, K. Henrick, A. Hussain, et al., E-MSD: The Eurepoean Bioinformatics Institute Macromolecular Structure Database. *Nuc. Acids Res.*, **31**: 458–462, 2003.
43. A. Bairoch, The ENZYME database in 2000. *Nuc. Acids Res.*, **28**: 304–305, 2000.
44. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, et al., Gene ontology: Tool for the unification of biology, The Gene Ontology Consortium, *Nature Genetics*, **25**: 25–29, 2000.
45. R. C. Dubes and A. K. Jain. *Algorithms for Clustering Data*, Englewood Cliffs, NJ: Prentice Hall, 1988.
46. A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, E. L. L. Sonnhammer, The Pfam protein families database, *Nuc. Acids Res.*, **30**: 276–280, 2002.
47. E. L. L. Sonnhammer, S. R. Eddy, and R. Durbin, Pfam: A comprehensive database of protein domain families based on seed alignments, *Proteins: Struct. Funct. Gene.*, **28**: 405–420, 1998.
48. O. Gascuel, BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data, *Mol. Biol. Evol.*, **14**: 685–695, 1997.
49. W. J. Bruno, N. D. Socci, and A. L. Halpern, Weighted neighbor joining: A likelihood-based approach to distance-based phylogeny reconstruction, *Mol. Biol. Evol.*, **17**: 189–197, 2000.
50. R. Desper and O. Gascuel, Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle, *J. Comput. Biol.*, **9**: 687–705, 2002.
51. S. Guindon and O. Gascuel, A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood, *Syst. Biol.*, **52**: 696–704, 2003.
52. T. N. Bhat, P. Bourne, Z. Feng, G. Gilliland, S. Jain, V. Ravichandran, et al., The PDB data uniformity project. *Nuc. Acids Res.*, **29**, 214–218, 2001.
53. N. Deshpande, K. J. Addess, W. F. Bluhm, J. C. Merino-Ott, W. Townsend-Merino, Q. Zhang, et al., The RCSB Protein Data Bank: A redesigned query system and relational database based on the mmCIF schema, *Nuc. Acids Res.*, **33**: D233–D237, 2005.
54. The Gene Ontology Consortium, Gene Ontology: Tool for the unification of biology, *Nature Genetics*, **25**: 25–29, 2000.
55. G. P. Moss (2006, March 16). Enzyme Nomenclature: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes by the Reactions they Catalyse, Available: http://www.chem.qmul.ac.uk/iubmb/enzyme/. Accessed: July 27, 2006.
56. M. Kanehisa and S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nuc. Acids Res.*, **28**: 27–30, 2000.

57. D. L. Wheeler, D. M. Church, R. Edgar, S. Federhen, W. Helmberg, T. L. Madden, et al., Database resources of the National Center for Biotechnology Information: update, *Nuc. Acids Res.*, **32**: D35–D40, 2004.
58. F. H. Allen, S. Bellard, M. D. Brice, B. A. Cartwright, A. Doubleday, H. Higgs, et al., The Cambridge crystallographic data centre: Computer-based search, retrieval, analysis and display of information. *Acta Cryst.*, **35**: 2331–2339, 1979.
59. M. S. Babcock and W. K. Olson, A new program for the analysis of nucleic acid structure: implications for nucleic acid structure interpretation, *Computation of Biomolecular Structures: Achievements, Problems, and Perspectives*, Heidelberg: Springer-Verlag, 1992.
60. K. Grzeskowiak, K. Yanagi, G. G. Prive, and R. E. Dickerson, The structure of B-helical C-G-A-T-C-G-A-T-C-G, and comparison with C-C-A-A-C-G-T-T-G-G: the effect ofbase pair reversal. *J. Bio. Chem.*, **266**: 8861–8883, 1991.
61. R. Lavery and H. Sklenar, The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids, *J. Biomol. Struct. Dynam.***6**: 63–91, 655–667, 1988.
62. H. M. Berman, A. Gelbin, J. Westbrook, and T. Demeny. *The Nucleic Acid Database File Format*. New Brunswick, NJ: Rutgers University, 1991.
63. S.-H Hsieh. *Ndbfilter. A Suite of Translator Programs for Nucleic Acid Database Crystallographic Archive File Format*. New Brunswick, NJ: Rutgers University, 1992.
64. J. Westbrook, T. Demeny, and S.-H. Hsieh. *Ndbquery. A Simplified User Interface to the Nucleic Acid Database*. New Brunswick, NJ: Rutgers University, 1992.
65. A. R. Srinivasan and W. K. Olson, Yeast tRNAPhC conformation wheels: A novel probe of the monoclinic and orthorhombic models. *Nuc. Acid Res.*, **8**: 2307–2329, 1980.
66. M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya, The KEGG databases at GenomeNet. *Nuc. Acid Res.*, **30**: 42–46, 2002.
67. M. Kanehisa, *Post-genome Informatics*. Oxford, UK: Oxford University Press, 2000.
68. M. Kanehisa, Pathway databases and higher order function. *Adv. Protein Chem.*, **54**: 381–408, 2000.
69. P. D. Karp, C. A. Ouzounis, C. Moore-Kochlacs, L. Goldovsky, P. Kaipa, D. Ahren, S. Tsoka, N. Darzentas, V. Kunin, and N. Lopez-Bigas, Expansion of the BioCyc collection of pathway/genome databases to 160 genomes, *Nuc. Acids Res.*, **19**: 6083–6089, 2005.
70. R. Caspi, H. Foerster, C. A. Fulcher, R. Hopkinson, J. Ingraham, P. Kaipa, M. Krummenacker, S. Paley, J. Pick, S. Y. R., C. Tissier, P. Zhang and P. D. Karp, MetaCyc: A multiorganism database of metabolic pathways and enzymes, *Nuc. Acids Res.*, **34**: D511–D516, 2006.
71. P. Romero, J. Wagg, M. L. Green, D. Kaiser, M. Krummenacker, and P. D. Karp, Computational prediction of human metabolic pathways from the complete human genome, *Genome Biology*, **6**: 1–17, 2004.
72. C. A. Ball, I. A. Awad, J. Demeter, J. Gollub, J. M. Hebert, T. Hernandez-Boussard, H. Jin, J. C. Matese, M. Nitzberg, F. Wymore, Z. K. Zachariah, P. O. Brown, G. Sherlock, The Stanford Microarray Database accommodates additional microarray platforms and data formats, *Nuc. Acids Res*, **33**: D580–582, 2005.
73. M. C. Costanzo, J. D. Hogan, M. E. Cusick, B. P. Davis, A. M. Fancher, P. E. Hodges, et al., The Yeast Proteome Database (YPD) and *Caenorhabditis elegans* Proteome Database (WormPD): Comprehensive resources for the organization and comparison of model organism protein information. *Nuc. Acids Res.*, **28**: 73–76, 2000.
74. G. D. Schuler, Pieces of the puzzle: Expressed sequence tags and the catalog of human genes, *J. Mol. Med.*, **75**: 694–698, 1997.
75. K. H. Cheung, K. White, J. Hager, M. Gerstein, V. Reinke, K. Nelson, et al., YMD: A microarray database for large-scale gene expression analysis. *Proc. of the American Medical Informatics Association 2002 Annual Symposium*, San Antonio, Texas, November 9–11, 2002, pp. 140–144.
76. C. M. Bouton and J. Pevsner, DRAGON: Database Referencing of Array Genes Online. *Bioinformatics*, **16**(11): 1038–1039, 2000.
77. I. R. Lemischka, K. A. Moore, and C. Stoeckert. (2005) SCDb: The Stem Cell Database, Available: http://stemcell.princeton.edu/. Accessed: July 28, 2006.
78. D. Gardner, M. Abato, K. H. Knuth, R. DeBellis, and S. M Erde, *Philosophical Transactions of the Royal Society B: Biological Sciences*. **356**: 1229–1247, 2001.
79. K. G. Herbert, N. H. Gehani, W. H. Piel, J. T. L. Wang, and C. H. Wu, BIO-AJAX: An Extensible Framework for Biological Data Cleaning, *ACM SIGMOD Record*, **33**: 51–57, 2004.
80. G. Chang, M. Haley, J. A. M. McHugh, J. T. L. Wang, *Mining the World Wide Web*, Norwell, MA: 2001.
81. H. Shatkay, N. Chen, and D. Blostein, Integrating image data into biomedical text categorization. *Bioinformatics*, **22**(14): 446–453, 2006.
82. A. S. Yeh, L. Hirschman, and A. A. Morgan, Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *Bioinformatics*, **19** (Suppl 1): i331–339, 2003.
83. P. Buneman, A. Chapman, and J. Cheney, Provenance Management in Curated Databases, *Proc. of ACM SIGMOD International Conference on Management of Data*, Chicago, Illinois June 26–29, 2006.

KATHERINE G. HERBERT
Montclair State University
Montclair, New Jersey

JUNILDA SPIROLLARI
JASON T. L. WANG
New Jersey Institute of Technology
Newark, New Jersey

WILLIAM H. PIEL
Peabody Museum of Natural History, Yale University
New Haven, Connecticut

JOHN WESTBROOK
Protein Data Bank and Rutgers, The State University of New Jersey
Piscataway, New Jersey

WINONA C. BARKER
ZHANG-ZHI HU
CATHY H. WU
Protein Information Resource and Georgetown University Medical Center
Washington, D.C.