Syst. Biol. 43(4):585–589, 1994

Report

Interoperability of Biological Data Bases: A Meeting Report

JUDITH A. BLAKE,[1,4] CAROL J. BULT,[1,5] MICHAEL J. DONOGHUE[2,6]
JULIAN HUMPHRIES,[3] AND CHRIS FIELDS[1,7]

[1]*The Institute for Genomic Research, 932 Clopper Road, Gaithersburg, Maryland 20878, USA*
[2]*Harvard University Herbaria, 22 Divinity Avenue, Cambridge, Massachusetts 02138, USA*
[3]*Section of Ecology and Systematics, Cornell University, Ithaca, New York 14850, USA*

Systematic biologists, no less that any other scientists, have experienced the onslaught of electronic storage and management of information. The entry of biological information into multiple, often specialized electronic data bases represents a major transfer of information to a new and powerful medium. Data bases of DNA and protein sequences, genetic and physical maps, biochemical data, phenotypes and strains, biogeographical data, museum collections information, and other types of data already exist; many others are under development. As the amount of information electronically available has increased, new thinking about distributed information systems has generated a strong interest in implementing software tools and data base structures to enable true data base interoperability. True interoperability differs from standard Internet browsing or "hot-link" access in that information residing in two or more data bases can be exchanged in response to ad hoc requests via standard protocols such as standard query language (SQL) issued from a remote site. Data bases that are interoperable in this way become parts of a federation of data bases (Fasman, 1994; Waterman et al., 1994).

The Workshop on Database Interoperability was held in Gaithersburg, Maryland, 25–27 June 1994, cosponsored by the Department of Energy, The Institute for Genomic Research, and MasPar Computer Corporation. This meeting brought together, for perhaps the first time, members of the computational genomics community and the systematics/collections data base community to discuss issues relating to data base development and interoperability with special attention to data resources developed with a relational SQL-compliant data base. Discussion was focused on the practical issues of establishing cross-data-base query capabilities. The topics addressed are of great practical importance to the systematics community, especially as we struggle with the development of adequate tools for managing the vast storehouse of information that is relevant to systematic research. This meeting report draws attention to developments in data base design and interoperability that may have important practical implications for the work of systematists.

Thirty-three participants representing 17 data bases (Table 1, see acronyms) attended the Workshop. The purpose was (1) to review and at least partially integrate schema and semantics from sequence, citation, specimen, and taxonomic/phylogenetic relational data bases and (2) to discuss mechanisms for inter-data-base queries. Discussion of the technical and semantic issues relevant to establishing some degree of interoperability among such data bases dominated the meeting. The format included brief presentations by individuals representing the different data bases and data resources, round-table discussions, and on-line demonstrations of data bases and software tools.

The development of federations of data bases comes from the recognition that data bases contain a defined set of objects and annotations of primary interest to the curator, as well as related information of secondary interest that could be represented solely through pointers to other data bases containing that information. In a federated system, each data base would focus on representing a particular set of biological information. The data base personnel would curate the information within their area of expertise (or provide tools and support for community data curation). Replication of related information from other data bases may still be necessary for efficiency or other reasons, but pointers between data bases would reduce the need to replicate most of the data (an ideal federation would be fully normalized, with each data type represented in only one data base). As a result, the curatorial staff for a particular data base would direct their efforts towards curating the data that they understand the best.

Discussions at the meeting unveiled a significant difference between collections and genome data bases concerning the replication of data types among data bases. In collection data bases, e.g., the biological collection data bases represented in the SMASCH con-

[4] E-mail: blake@tigr.org.

[5] E-mail: bult@tigr.org.

[6] E-mail: m_donoghue@harvard.edu.

[7] Present address: The National Center for Genome Resources, Santa Fe, New Mexico 87505, USA. E-mail: chris.fields@ncgr.org.

TABLE 1. Participating data bases/data resources at the Workshop on Database Interoperability, Gaithersburg, Maryland, June 1994.

Acronym	Data base/data resource	Affiliation	Representative[a]	Description
CGSC	Coli Genetic Stock Center	Yale University, New Haven, CT	Mary Berlyn	genotypes, pedigrees of strains, details and registry of alleles and other mutations, and mapping data for *Escherichia coli*
HDB, ATCC	Hybridoma Data Base, American Type Culture Collection	American Type Culture Collection (ATCC), Rockville, MD	Lois Blaine, Donna Maglott	data on biological materials distributed by the ATCC
SST; EGAD	Sequences, Sources, Taxa; Expressed Gene Anatomy Database	The Institute for Genomic Research (TIGR), Gaithersburg, MD	Judith Blake, Carol Bult, Chris Fields, Anthony Kerlavage, Owen White	linking data base for molecular sequences and collection/specimen data; gene expression information
RDA	Remote Database Access	National Institute of Standards and Technology (NIST), Gaithersburg, MD	Kevin Brady, Joan Sullivan	software for interoperability of remote SQL-compliant data bases
GSDB	Genome Sequence Database	National Center for Genome Resources, Santa Fe, NM	Michael Cinkosky, Gifford Keene	comprehensive data base for nucleotide sequences and related information
TreeBASE	TreeBASE	Harvard University, Cambridge, MA	Michael Donoghue, Bill Piel	data sets and trees from published phylogenetic analyses
SMASCH	Specimen Management System for California Herbaria	University of California, Berkeley	Tom Duncan	relational data model for the management of California herbaria
MGD	Mouse Genome Database	Jackson Laboratory, Bar Harbor, ME	Janan Eppig, Joel Richardson	comprehensive information on the mouse genome
GDB, CitDB	Genome Data Base, Citations Data Base	Johns Hopkins University, Baltimore, MD	Ken Fasman, Michael Chipperfield	GDB: genes and other genomic landmarks, map location, phenotype and locus data for humans; CitDB: relational bibliographic management data model

TABLE 1. Continued.

Acronym	Data base/data resource	Affiliation	Representative[a]	Description
MUSE	Museum Collections Data Management System	Cornell University, Ithaca, NY	Julian Humphries	software for the management and curation of museum collections
FNA	Flora of North America	Missouri Botanical Garden, St. Louis	Deborah Kama	names and geographic distributions of North American flora
ProLink	ProLink	Boston University, Boston, MA	Kathleen Klose	protein structure, function, and families integrated with chromosomal location and metabolic pathways
Genera	Genera	Johns Hopkins University, Baltimore, MD	Stan Letovsky	software for the integration of relational data bases implemented in Sybase into the World Wide Web
RDP	Ribosomal RNA Database Project	University of Illinois, Urbana	Bonnie Maidak	curated ribosomal RNA sequence data, alignments, phylogenetic trees, and data analysis services
NBII	National Biological Information Infrastructure	National Biological Survey, U.S. Department of the Interior, Washington, DC	Brand Niemann	directory, clearinghouse, and distributive system of data bases to support public access to natural resources data
MaizeDB	Maize Data Base	University of Missouri, Columbia	Mary Polacco	mapped genes, gene products, biochemical pathways, stocks (germplasm), phenotypes, and probes for maize
HICLAS	Hierarchical Classification System	Michigan State University, East Lansing	Sakti Pramanik	data base and X-windows interface for searching and comparing hierarchical taxonomic classification schemes

[a] Meeting participants not representing a specific data base or data resource were Robert Robbins (U.S. Department of Energy), Jay Snoddy (U.S. Department of Energy), Stan Blum, and Winston Hide (MasPar, Inc).

sortium, each data base utilizes essentially the same schema. Each institution records data associated with its specimens, including the herbarium where it is housed, the collection locality, the scientific names applied to it, and the collector(s) of the specimen. The replication of data types is necessary because each institution is recording the same types of information about similar objects or specimens. It is both reasonable and efficient for the data bases describing these objects to be maintained by the institutions responsible for maintaining the objects themselves. This process has been termed the horizontal partitioning of information.

In contrast, genomics data bases seek to represent specialized information that is not elsewhere represented in a curated and accessible form. ProLink, for example, is an integrated data base of protein structure, sequence homolog, and functional pattern information. It has links to data bases of protein sequences, i.e., internal links to PDB (Protein Data Brookhaven) and indirect links to SwissProt, Prosite, OMIM (Online Mendelian Inheritance in Man), and GDB. It does not, unlike the collections data bases, represent many of the same data types as do the sequence data bases or other genomics data bases. Such a data base can be developed and maintained anywhere. This process has been termed vertical partitioning of information. Despite differences, both communities utilize similar concepts in data base design and implementation to manage information.

Throughout the 2-day meeting, major topics of discussion were (1) unique accession keys, in particular the complexity of maintaining unique accession keys and tracking them through different versions of the data bases and merge/splits of the data; 2) semantics, i.e., the meaning or usage of the words used to define elements of the data base; 3) representations of DNA or protein sequences, alignments, and phylogenetic trees; and 4) the need to implement links to diverse, curated taxonomic data bases.

Implementing the concept of pointing to other data bases hinges on the assignment of unique accession numbers to each major object or record in each data base in the federation. Information searched across data bases must come with identifiers unique to the whole community of federated data bases. Accession keys are recorded by other data bases as the only consistent link between the data bases. (The word *accession* has different meanings within the systematics community versus the informatics/data base community. The term *accession key* is used to indicate the unique identifier associated with an object in a data base.) Because one data base's primary key thus becomes a secondary key providing the linking information in another data base, the importance of providing consistent long-term support for accession keys despite the release of updates and new versions of the data base was repeatedly emphasized. One concept for implementing unique identifiers for public data bases is a two-field scheme in which, for the public version, an additional component is added to the accession key; e.g., XX:YY, where XX denotes the data base and serves as a registered code for instructions on how to parse the location-specific key, YY. Such a scheme is already in use by default: GSDB: D00133 denotes a sequence accession key D00133 in the GSDB. This system implies community collaboration on registration or distribution of unique data base identifiers, a challenge the systematics community will need to address.

The ability to exchange information among data bases requires shared ideas on what particular words or entities represent, and such semantic issues were emphasized throughout the meeting. An example from a nucleotide sequence data base is the concept of a gene. In one data base, *gene* might represent only those sequences known to code for proteins. In another data base, *gene* might include sequences that function as RNAs and are not translated into an amino acid sequence. Although it is not reasonable (in practice) to impose standardized semantics on members of a federated system, the concepts represented in the schema need to be clearly defined. It seems reasonable to project that some standardization of semantics will occur as groups work together and communicate information about their schemas. Although the semantic issues are common to all data bases, the emphasis on development of a unified schema will be more apparent among members of the federated system that represent the same entities (the horizontal model of data distributions in the collections data bases above) than among members of the system that only partially share data types (the vertical model of data partitioning represented by the genomics data bases). It seems likely that the semantic details will be worked out a few fields at a time, but the process was clearly stimulated by this meeting.

Of particular interest to systematists are ongoing efforts to develop data bases focused specifically on aligned sets of molecular sequences, phylogenetic trees, and associated information. TreeBASE for example, is a prototype relational data base being designed to manage and explore information on phylogenetic relationships. Its primary function is to store published phylogenetic trees and data matrices and to provide an interactive means of assessing and synthesizing phylogenetic knowledge. SST is also a prototype relational data base being developed to integrate DNA and protein sequence data with specimen, collection, and taxonomic information. Interoperability of these data bases would allow users to quickly find information on the location of voucher specimens for DNA sequences used in a particular phylogenetic analysis, for example, or to retrieve all of the phylogenetic studies involving a particular specimen or sequence. Efforts to incorporate phylogenetic analyses into data bases raise a variety of new issues, including the representation of details on the analyses performed and additional information about trees, such as lengths and other indices.

A major issue confronting all efforts to communicate within and among biological data bases is the standardization of taxonomic classifications. Far too little attention has been devoted to the representation of taxonomic names and their relationships with one another. Consequently, multiple synonymous names may be used for the same entity, even within a single data base. For example, a particular "cow" molecular sequence is identified as coming from *Bos bovis* in one data base, *Bos taurus* in another, and *Bos primigenius taurus* in yet another. This situation confounds attempts to use taxonomic names in complex, cross-data-base queries without the development of synonymy tables for all organisms. Linking specialized taxonomic data bases, which have developed synonymy tables for particular groups, into a federation of data bases that make use of such names may be a robust solution to this problem. However, even with appropriate tools, collaborative efforts will need to be established to "edit" changing concepts of taxonomic entities and classifications. A data base demonstrated at this meeting was HICLAS, a system capable of searching and comparing different taxonomic classification schemes as well as cladograms and phenograms.

Currently, World Wide Web (WWW) is the interface of choice for the biological community, and much of the software development for connecting data bases among the workshop participants is based on this interface (Schatz & Hardin, 1994). However, WWW is designed for hypertext browsing and does not support multi-data-base relational queries. There is a strong need for articulating the requirements necessary for software to support different types of queries. For example, software was demonstrated that used a protocol that simultaneously queries museum collections using MUSE with a WWW client and WWW server scripts to search for taxonomic and geographic data from the participating systems. This multiple-retrieve capability is quite different from the implementation of cross-data-base joins from unique tables. In the latter case, a query invokes a remote data base access protocol and collects portions of the requested data from different data bases and then joins the discrete data items into a response for the user.

To establish true interoperability between distributed data bases, software is needed to identify and retrieve the appropriate information from individual data bases and return the components to the user. Remote data base access (RDA) software is under development by the vendors of relational systems. Of particular interest at the workshop was a demonstration of an RDA software tool that programmers at the National Institute of Standards and Technology are developing. The prototype software splits a complex query into component parts, sends each part to a remote data base to retrieve information, and then combines that information into a standard response for the researcher. The actual implementation of this kind of protocol will herald the era of true relational interoperability.

The Gaithersburg meeting brought together data base developers and information providers from the genomics and systematics communities to discuss common problems and begin the development of a federation of data resources. It seems clear that information will continue to appear in data bases designed for specialized purposes. This is as it should be, inasmuch as a single central data base, meant to encompass all biological information or even large parts of it, would be unable to sustain the standards on content and verification demanded by scientific users. This realization means that it will be essential to develop links between data bases—links that make it possible to assemble information from a variety of sources at once. For these links to be effective, cooperation both within and among scientific communities is essential. As more biological information becomes available in electronic form, the speed and accuracy of retrieval in response to queries across data bases will fundamentally impact the scope and pace of research. Meetings such as this one will hasten the time when we spend less time finding data and more time putting it to good use.

References

FASMAN, K. 1994. Restructing the genome data base: A model for a federation of biological databases. J. Comput. Biol. 1:165–171.

SCHATZ, B. R., AND J. B. HARDIN. 1994. NCSA Mosaic and the World Wide Web: Global hypermedia protocols for the Internet. Science 265:895–901.

WATERMAN, M., E. UBERBACHER, S. SPENGLER, F. R. SMITH, T. SLEZAK, R. ROBBINS, T. MARR, D. T. KINGSBURY, R. GILNA, C. FIELDS, K. FASMAN, D. DAVISON, M. CINKOSKY, P. CARTWRIGHT, E. BRANSCOMB, AND H. BERMAN. 1994. Genome informatics. I. Community databases. J. Comput. Biol. 1:173–190.