OMICS A Journal of Integrative Biology
Volume 7, Number 1, 2003
© Mary Ann Liebert, Inc.

Database Management for Life Science Research: Summary Report of the Workshop on Data Management for Molecular and Cell Biology at the National Library of Medicine, Bethesda, Maryland, February 2–3, 2003

H.V. JAGADISH[1] and FRANK OLKEN[2]

THE CRISIS IN DATA MANAGEMENT FOR BIOLOGICAL SCIENCES

The data deluge

OVER THE PAST 15 YEARS, we have witnessed a dramatic transformation in the practice of molecular biology. What was once a cottage industry marked by scarce, expensive data obtained largely by the manual efforts of small groups of graduate students, post-docs, and a few technicians has become industrialized (routinely and robustly high throughput) and data-rich, marked by factory scale sequencing organizations (such as the Joint Genome Institute, the Whitehead Institute, and the Institute for Genomic Research). Such sequencing factories rely on extensive automation of both sequencing and sample preparation. Commencing with sequencing, such industrialization is being extended to high-throughput proteomics and metabolomics, for example.

While this industrialization of biological research is partly the result of technological improvements in sequencing instrumentation and automated sample preparation, it is also driven by massive increases in public and private investment and dramatic changes in the social organization of molecular biology (e.g., the creation of highly specialized, factory scale organizations for mass genomic sequencing). Such industrialization and the accompanying growth in molecular biology data availability demand similar scale up and specialization in the data management systems that support and exploit this data gathering. To date, the bioinformatics community has largely made do with custom handcrafted data management software or with conventional database management system (DBMS) technology developed for accounting applications.

The industrialization of molecular biology has been largely the province of pharmacological, government, and, to a lesser extent, academic molecular biology research. However, it is clear that we stand at the threshold of clinical application of many of these technologies, for example, as clinical laboratory tests for medical applications. Such clinical applications will entail great increases in the laboratory and data management activities to handle tens or hundreds of millions of assays annually in the United States. Similarly, the approaches and data generation output from ever higher levels of biological complexity will be increasingly data intensive and high throughput.

Instruments, data, and data management systems are complementary goods; in other words, their joint consumption is much more useful than consuming a single commodity at a time. It is trivial to say that data management systems are much more useful if they contain data. Consider also how limited the utility of genomic sequence data would be if we could only publish it in books and manually compare it. The availability of data management software that permits the rapid searching of large genomic sequence databases for similar sequences greatly enhances the utility of such sequence data. Quick sequence comparisons are

[1]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan.
[2]Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, California.

not sufficient by themselves; the fact that many (most) of these sequences have been collected into a few databases (e.g., GenBank) greatly simplifies the comparison task.

Many (if not most) instruments used in molecular biology and chemistry produce spectra, or spectra-like results, for example, infrared spectrographs, gas and/or liquid chromatography, and mass spectrometers. Such instruments are used in conjunction with large community databases of spectra and data management systems that can store and quickly retrieve matching spectra. Without such spectral databases and spectral data retrieval systems, many of these instruments would be much less useful for biology, biochemistry, forensics, and medicine. Again, we see that instruments, database content, and data management software are complementary goods, whose joint consumption has higher utility. The NIST NMR spectral database is but one example of such a database.

We expect that this explosive growth in the volume and diversity of biological and biochemical data will continue into the 21st century—in other words, that 21st century life sciences will be data-rich. Success in the life sciences will hinge critically on the availability of computational and data management tools to analyze, interpret, compare, and manage this abundance of data. Increasingly, much of biology is viewed as an information science, concerned with how cells, organisms, and ecological systems encode and process information in genetics, cellular control, organism development, environmental response, and evolutionary settings.

The need for data management tools

To obtain the full benefit of the massive public investments in generating biological data will require commensurate investments in effective data management systems and judicious choices of how to assemble and manage shared databases. In the past decade there have been debates over what data to collect and in what kind of shared databases the data should be kept, for example, large-scale GenBanks or boutique databases. Substantial investments have been made in building large public databases such as GenBank and Protein Data Bank (PDB). In comparison, discussion of novel database management technology for biological applications and investment in their development have been quite modest.

For the past decade, biologists (and medical informaticists) have relied primarily on commercial relational database technology for this purpose. Acquisition of such commercial off-the-shelf (COTS) software was inexpensive, compared to the costs (and time) required to develop new data management software, and vendors boasted of the universal utility of relational data management technology. A decade of experience suggests that, while very useful, "vanilla" relational database management systems, without incorporating specialized life sciences enhancements, are cumbersome instruments for constructing and managing life sciences databases. Note that most approximate sequence matching, graph queries on biopathways, and three-dimensional shape similarity queries are still being performed outside of relational data management systems. The extensibility features of object relational databases have proven difficult to use, provide limited query optimization, and have thus far been of limited help in extending query languages—for example, to handle graph queries. Object-oriented databases have had limited success in providing efficient or extensible declarative query languages. In this report, we discuss (mostly) proposals for research and development of improved data management technology to better address issues of data management for the life sciences.

Data management tools can interface molecular and cellular data with both image data and physiological data, which will be important to scale across the levels of living systems and particularly to translate the findings of basic biology to health care. Similarly, public health depends on our ability to integrate and query data from very diverse, very fragmented, non-standard, distributed data sources and databases.

In short, to turn the vast amounts of new information being generated through scientific experiments into knowledge that can be applied towards better practice in medicine, agriculture, and environmental science, federal agencies need to encourage a profound, deep partnership between experimental biology and database management.

THE UTILITY OF DATABASE MANAGEMENT SYSTEMS

Data management technology development has been an effort to identify common problems and abstractions concerning the storage and retrieval of data across a broad set of applications, and to construct

general purpose DBMSs to address these common requirements. Such DBMSs are tools that application developers can use to build applications involving data management. DBMSs permit specialization of programming effort and the amortization of DBMS development costs across many applications. The use of DBMSs reduces the cost and time required to develop and maintain data management applications. The existence of a multi-billion dollar per year market for DBMS software suggests that such systems are quite useful to application developers.

DBMSs differ in many details; however, they commonly embody several important ideas:

- The notion of data independence, that is, the independence of logical data model from physical data structures, avoids the need to change applications (queries) when physical data structures are changed
- The use of high-level declarative query languages that make it possible for a non-expert to specify complex queries
- The use of automated query optimization of such declarative query languages
- The use of specialized indices (data structures) and query processing algorithms to efficiently answer queries
- Efficient management of a hierarchy of storage devices (e.g., RAM, disk) via caching, buffering, and page-based indices, to enhance performance and minimize cost
- Transaction management software to permit concurrent processing of update queries and quick recovery from aborted transactions or system/disk crashes
- Declarative specification of integrity constraints and their enforcement, such as referential integrity constraints (i.e., no dead links), providing essential support for data curation and consistency maintenance
- Automated partitioning and usage of parallel computers for large-scale query processing
- Technologies for answering queries that span multiple heterogeneous geographically distributed databases (or heterogeneously formatted datasets)

For small data sets that are analyzed by a single user, it is often possible to side-step database management systems altogether. Indeed, simple home-grown programs, and Perl scripts in particular, have adequately served the needs of many a scientist. However, as the size of the data grows, the complexity of the analysis grows, and the diversity of the sources grows, these home-grown solutions do not scale easily. The value of developing cross-cutting technology for data management becomes more apparent.

REQUIREMENTS OF BIOLOGICAL DATA MANAGEMENT

DBMS researchers and vendors have often advertised that their products have universal applicability. In fact, data management technology development has been shaped by different applications over the past 30 years. Commercial (banking, payroll, and inventory) applications drove the development of relational DBMS, computer-aided design (CAD) applications drove the development of object-oriented databases, management information systems have driven data warehousing and online analytical processing (OLAP) data management technology, and web content and e-commerce technology have driven XML data management systems. Biological applications have their own requirements, which will require further advances in data management technology. These include the following:

1. A great diversity of data types: sequences, graphs, three dimensional structures, images.
2. Unconventional types of queries: similarity queries, for example, sequence similarity, pattern-matching queries, pattern finding queries.
3. Ubiquitous uncertainty (and sometimes even inconsistency) in the data
4. Extensive requirements for data curation (data cleaning and annotation)
5. A need for large-scale data integration (hundreds of databases)
6. A need to support detailed data provenance
7. Extensive requirements for terminology management
8. Support for rapid schema evolution
9. A need to support temporal data

10. A need to provide model management for a variety of mathematical and statistical models of organisms and biological systems

These topics are discussed more extensively in the full technical report. Here, we briefly elaborate on only a few of these points.

Diversity of data types and queries

One of the striking features of biological data is the great diversity of data types: sequences, graphs, three-dimensional structures, scalar and vector field data. The queries posed against these data types are also diverse, and different from common commercial queries, for example, similarity and pattern matching queries. Whereas conventional (accounting) databases are dominated by exact match (equality) and range (inequality) queries, biological applications involve the pervasive use of similarity queries, for example, classic sequence similarity queries, but also including subgraph isomorphism, pattern matching queries (e.g., regular expressions, Hidden Markov models) and pattern identification queries. These topics are discussed further in the full technical report (cited below).

This diversity of data and query types has two implications for data management technology. First, we need to develop specialized indexing and query processing techniques to deal with these specialized data and query types. Second, we need to develop more extensible data management systems. Current DBMSs have object-relational facilities that offer some extensibility features that have been used to support geographic information systems and chemo-informatics systems. Most of the workshop participants believe that current extension facilities are too limited and cumbersome to fully cope with the diversity of biological data and queries.

Data provenance

Questions of data release policies for biological data are properly questions of public policy, not technical discussion. However, it has become increasingly clear that good data management infrastructure for recording and querying data provenance—the origin and processing history of data—is vital if we are to effectively encourage the sharing of biological and biomedical data. Data provenance issues have been largely neglected by the database research community except for a few researchers in statistical data management and data warehousing. This area clearly needs further work to support bioinformatics data sharing. The topic is also of increasing interest to the regulatory community (e.g., the U.S. Food and Drug Administration).

The classic approach to sharing knowledge in the biology community has been to publish journal articles. Authors receive public acclaim and acknowledgment in exchange for publication of their knowledge. Individual articles and authors are acknowledged via bibliographic citations (or sometimes co-authorship), and systems have been developed to record the number of citations that papers have received. We believe that similar mechanisms are needed to acknowledge "publication" of datasets in shared databases, so as to encourage rapid, effective sharing of data. Data management support for tracking data provenance (origins) can provide the analog of citations. Usage tracking software can potentially provide analogs to bibliographic citation counts. Support for automatic tracking and querying of data provenance is fairly undeveloped in current DBMSs.

There are other important motivations for recording and querying data provenance. Knowledge of the source and processing history of data items permits users to place the data in context and helps to assess its reliability. Data provenance histories also facilitate revision of derived data when the base data (or analysis codes) change. DBMS support is needed to facilitate the automated update of provenance information as the database is updated and the automatic propagation of provenance information with query results. Experience in other settings, for example, geographic information systems, indicates that unless metadata (e.g., data provenance) is automatically updated, it is likely to quickly become outdated.

Data integration

Many, if not most, applications of biological and biomedical databases require the ability to access data from many different databases (and datasets). There has been a veritable explosion in specialized biologi-

cal databases, recently tabulated at more than 500. Many researchers regard these specialized databases as extremely valuable, in part due to the very detailed and careful curation of the databases by specialists in particular database domains. However, no matter how good the data management technology for data integration, we do not foresee that it will be practical for data integration to succeed in a world of hundreds of biological databases unless the database providers provide extensive assistance in the form of publicly accessible, machine processable documentation concerning the database schemas, contents, query interfaces, and query languages. Adoption of such current technology by database providers was seen as a pressing issue.

How to encourage the adoption of "best practices" for the design and documentation of biological databases. The current practice of only providing access to most specialized biological databases via web-based forms is not sufficient; query APIs and query languages are needed to facilitate data integration. The provision of suitable data documentation and adoption of standard data exchange formats and query languages, and APIs must be seen as a social obligation of investigators similar to careful description of experimental methods in publication. We note that this documentation will need to be machine processable, that is, encoded as formal ontologies, schemas, and terminologies, not merely natural language texts. The efforts of the Microarray Gene Expression Data Society, to develop standard schemas for micro-array data, represent an instance where significant steps have been taken in this direction.

We note that the structural biology and genomics communities have also resorted to various social sanctions to encourage data sharing, for example, requirements of depositing data in PDB or GenBank prior to acceptance of papers for publication, and requirements for data deposition as a condition of grant renewal and a criterion for funding of new grants. We anticipate that similar activism by federal research program managers and journal editors will continue to be required, both for data deposition and to assure adoption of best practices to facilitate data sharing, such as complete documentation, data exchange encoding, and support for query APIs (e.g., Web Services Description Language) and query languages (e.g., SQL, OQL, XQuery).

RECOMMENDATIONS

Interdisciplinary research

The past decade has seen the rise of bioinformaticists (a.k.a. bioinformaticians), a new group of researchers/developers operating across the disciplines of biology, statistics, computer science, and mathematics. Their interdisciplinary activities now have their own professional society, conferences, and journals.

Orchestrating fruitful interdisciplinary research across biology, bioinformatics, and data management is not easy. Even within the workshop, there was heated debate about the best strategy to accomplish this. Lack of sufficient interaction among biologists, bioinformaticists, and data management researchers can easily lead to attempts to reinvent well-known data management technologies by bioinformaticists, or sterile pursuits of insignificant or misunderstood problems by data management researchers. Also, the time scales of data management research and development are often incompatible with the production requirements of ongoing biological laboratories or public databases. Despite early plans and efforts (e.g., by DOE), the major human genome sequencing centers have generally not been major sources of innovative data management technology. The most intellectually fruitful endeavors have often come from data management or computer science research groups with looser collaborations with biologists. The time required to develop new database technologies often exceeds the time demands of most biologists or bioinformaticists, who must produce biologically relevant data to sustain funding.

Research funding

A sustained program supported across the federal agencies at the frontier between biology and data management technology will allow us to share the database expertise of the information technology (IT) professionals with bioinformaticists and biological experimentalists supported across the federal agencies. There are needs for both research in database management technologies and innovative application of existing database technology to biological problems. Funding agencies will have to set up appropriately staffed review panels charged with suitable review criteria for funding such interdisciplinary work. Adequate fund-

ing for small, medium, and large-scale collaborative research projects, as well as including funding within those collaborative projects to train a new generation of database management experts, will be important.

For fastest progress in the biological sciences, we must encourage both the development of biological database content and improved data management technology for managing this content. We must recognize that these are two complementary but quite different endeavors. At least some of the funding for these two types of activities must be separated, or "fenced." (This is often described as different "colors" of money.) Otherwise, the obvious, pressing needs of delivering today's content will too frequently overshadow technology's promise of a far better tomorrow, one that will require more sophisticated approaches. Most research-driven companies recognize this tension and fund at least some of their research activity from central corporate sources, rather than solely through product divisions. In similar fashion, funding agencies should create a targeted funding program for data management specialists to collaborate with life scientists in developing superior data management technology for life science applications. It is also important to achieve closer integration of database research and production database systems development.

Various participants suggested the following taxonomy of funding for data management activities in the life sciences. It is important that there be adequate support for all of these activities and not just selected ones:

- Basic research in biologically relevant data management technologies: work in database theory and algorithms for life sciences data management, prototype DBMS development to support life sciences needs, and testing of algorithms and DBMSs with life sciences data
- Production database systems development
- Database content assembly and curation
- Database integration
- Ontology development
- Continuing support for the availability, maintenance, growth, and evolution of public databases

It is also valuable to define challenge problems that push the boundaries of data management technology, which, if successful, would enable major advances in biomedical science. Well-specified challenges can help direct data management researchers toward important bioinformatics problems. Creation of test data sets and benchmarks are also worthy endeavors in themselves, and should be supported as appropriate and possible. Much of this work must be done by life scientists. The availability of such test data sets and query benchmarks facilitates the comparison of new approaches to older ones.

Education and training

We expect, in the foreseeable future, that it will become important to have physicians and experimental biologists trained in computational methods, just as training in genetics has now become routine for physicians. Biology is often an exercise in induction (generalization from many instances), whereas computer science is more often a deductive enterprise, because computer algorithms/systems are usually designed, not evolved, artifacts. Solution to a specific biological data management problem is of less interest to a computer scientist than the generalization of this problem to a class of data management problems, all of which can be solved in one fell swoop through an appropriate computational advance—and rightly so, since this paradigm is significantly more cost-effective in the domains to which it is applicable. We note that experimental design and algorithmic design are often similar endeavors.

This dichotomy has significant repercussions not just on how we undertake research activities, but also in how we train scientists. Currently, too few biologists are being trained in the underlying algorithms and statistical theory underpinning various sequence matching techniques. This impedes their ability to effectively select appropriate computational or statistical methods, and slows the adoption of new algorithms and statistical methods. We need opportunities for people at every level to train themselves in the "other discipline," and to work at the interface between data management and biomedical science. This problem is reminiscent of debates on how to integrate statistics training into various academic disciplines. There is also a need for further curriculum development. The funding for such activities has to be ongoing over a substantial period of time and available at a number of institutions. A typical 3-year funding cycle is insufficient to implement the sort of major changes that are required, and specific federal programs that can address this continuing need are thus necessary.

CONCLUSION

The development of high-throughput methods and the establishment of commercial sources for even highly specialized biochemical reagents for research in molecular and cell biology over the past 15 years has brought a huge increase in the volume and diversity of biological and biomedical data. Clinical use of these technologies has already begun, and extensive, even routine, application is imminent. Full, efficient exploitation of these expensive investments in data collection will require complementary investments in data management technology.

To date, most efforts to manage this data have relied on commercial off-the-shelf DBMSs developed for business data. Better data management technology is needed to effectively address specific data management needs of the life sciences. Such needs include support for diverse data types (e.g., sequences, graphs, and three-dimensional structures) and queries (e.g., similarity-based retrieval), data provenance tracking, and integration of numerous autonomous databases. Effective data integration will require substantial assistance from the developers of individual databases (e.g., provision of machine-processable schemas, ontologies, thesauri, query APIs, and query languages) and use of standard data interchange formats.

The development of data management technology for the life sciences will require specific funding for research, development, curriculum development, and education in data management for the life sciences. Wet lab biologists need powerful access to suitably maintained and managed data repositories. Beyond this continuing expectation, the data management technology must itself continue to advance to meet what will be the expectations of experimentalists. Because the development of new database technology requires longer time scales than those customary in experimental biology or data collection, we see a need to separate some of the funding and its review processes for basic research in relevant data management technologies, from the ongoing funding for public biological database development and operation.

FULL REPORT

This document is a preliminary version of the summary report. An updated version is available at the workshop web site: http://www.lbl.gov/~olken/wdmbio/. A more detailed full report is also available at this site. This site also contains the position papers, the original workshop proposal, and attendee lists.

ACKNOWLEDGMENTS

This document is the product of a workshop funded primarily by the National Science Foundation, Computer and Information Science and Engineering Directorate. The National Library of Medicine at NIH provided us with conference facilities and support in kind. The Department of Energy provided support to one of the organizers, via the *Genomes to Life* Program, as part of the Virtual Institute of Microbial Stress and Survival Project under Contract No. DE-AC03-76SF00098. The full report was the work of the writing committee comprised of Russ Altman (Stanford), Susan Davidson (U. Penn.), Barbara Eckman (IBM), Michael Gribskov (SDSC), H.V. Jagadish (U. Michigan), Toni Kazic (U. Missouri), David Maier (Oregon Health Sciences University), Frank Olken (LBNL), Z. Meral Ozsoyoglu (Case Western Reserve Univ.), Louiqa Raschid (U. of Maryland), and John C. Wooley (UC San Diego). The many attendees of the workshop contributed white papers and discussions that formed the basis of the report and this summary. Many also contributed comments to the report and this summary.

Address reprint requests to:
Dr. F. Olken
Lawrence Berkeley National Laboratory
Computational Research Division, Building 50B3238
1 Cyclotron Road
Berkeley, CA 94720-8147

E-mail: olken@lbl.gov