Database Management for Life Sciences Research

H. V. Jagadish
University of Michigan
jag@umich.edu

Frank Olken
Lawrence Berkeley National Laboratory
olken@lbl.gov

.

The life sciences provide a rich application domain for data management research, with a broad diversity of problems that can make a significant difference to progress in life sciences research. This article is an extract from the Report of the NSF Workshop on Data Management for Molecular and Cell Biology, edited by H. V. Jagadish and Frank Olken. The workshop was held at the National Library of Medicine, Bethesda, MD, Feb. 2-3, 2003.

The Crisis in Data Management for Biological Sciences

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, Whitehead Institute, 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, metabolomics, etc.

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 DBMS (database management system) 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, e.g., 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 U.S. 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, i.e., their joint consumption is much more useful than consuming a single commodity at a time. It is trivial to see that data management systems are much more useful if they contain data. Consider also what 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 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.

In a similar vein, we note that many instruments used in molecular biology and chemistry produce spectra, or spectra-like results, e.g., infrared spectrographs, gas and/or liquid chromatography, mass spectrometers. Such instruments must be complemented with large community databases of spectra, and data management systems that can store and quickly retrieve matching spectra, to provide greatest value to biology, biochemistry, forensics, and medicine.

We expect that this explosive growth in the volume and diversity of biological and biochemical data will continue into the 21st century. 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 evolution.

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

Database management systems 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, CAD (computer aided design) applications drove the development of object-oriented databases, management information systems have driven data warehousing and OLAP (Online Analytical Processing) data management technology, and web content and e-commerce technology have driven XML data management systems. Biological applications have their own requirements that will require further advances in data management technology. These include:

1. A great diversity of data types: sequences, graphs, three dimensional structures, images, etc.
2. Unconventional types of queries: similarity queries, e.g., sequence similarity, pattern matching queries, pattern finding queries, etc.
3. Ubiquitous uncertainty (and sometimes even inconsistency) in the data
4. Extensive requirements for data curation (data cleaning and annotation)
5. A need to support detailed data provenance
6. A need for large scale data integration (hundreds of databases)
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 [1]. Here we briefly elaborate on only a few of these points.

Diversity of Data Types and Queries

A striking feature of biological data is the great diversity of data types: sequences, graphs, 3D structures, scalar and vector field data, etc. The queries posed against these data types are also diverse, and different from common commercial queries. Whereas conventional databases are dominated by exact match (equality) and range (inequality) queries, biological applications involve the pervasive use of similarity queries, e.g., classic sequence similarity queries, but also including subgraph isomorphism, pattern matching queries (e.g., regular expressions, Hidden Markov models) and pattern identification queries.

Sequences: The availability of sequence data, e.g., DNA, RNA, and amino-acid sequences (proteins), has grown explosively over the past decade with the development of automated sequencing machines and large scale sequencing projects such as the human and mouse genome sequencing projects. Sequences are presently often stored as text strings, but this representation is awkward when we want to annotate sequences, since text strings typically lack addressability at the level of individual letters (nucleotides, or amino acids). Often DNA sequences include not only individual nucleotides, but also gaps, usually with a length (or bounds on length) specification of the gap.

Graphs: Many types of graphs occur in biological data, including a directed (or undirected) labeled graph, a nested graph, and a hyper-graph. Examples include various biopathways (metabolic pathways, signaling pathways, and gene regulatory networks), genetic maps (partial order graphs (i.e., directed acyclic graphs), taxonomies (either trees or DAGs), chemical structure graphs, contact graphs (for 3D protein structure), etc. Graphs are easily stored in existing DBMSs, e.g., relational DBMSs. However, many graph queries, e.g., subgraph isomorphism, subgraph homomorphism, and subgraph homeomorphism are difficult (or impossible) to pose and answer efficiently in existing relational DBMSs, which know nothing of graphs.

High-Dimensional Data: It is not unusual for micro-array experiments of gene expression to involve thousands (or tens of thousands) of genes and hundreds (or thousands) of experimental conditions and samples. Generated datasets are arrays of spot intensities over the Cartesian product of genes and samples (e.g., experimental conditions). Often researchers are interested identifying clusters of genes which exhibit similar (or opposite) patterns of gene regulation. Specialized data structures and clustering algorithms are needed to support nearest neighbor, range searching, and clustering queries in high-dimensional spaces.

Shapes: Three-dimensional molecular (protein, ligand, complex) structure data is another common data type. Such data includes both shape information (e.g., ball and stick models for protein backbones) and (more generally) scalar and vector field data of charge, hydrophobicity, and other chemical properties specified as functions over the volume (or surface) of a molecule or complex.

Temporal Data: Temporal data must frequently be managed when studying the dynamics of biological systems. Examples include cellular response to environmental changes, pathway regulation, dynamics of

gene expression levels, protein structure dynamics, developmental biology, and evolutionary biology.

Temporal data in biological settings can either be *absolute* or *relative*. Absolute time-stamping is common in administrative or long term ecological observational databases -- time is recorded relative to an absolute global temporal coordinates such as UTC date-time. Relative time-stamping records time relative to some event -- e.g., cell division, organism birth, oncogenesis, diagnosis, cold shock, etc. Most implementations of time in the database community have focused on absolute time, whereas relative time is much more commonly used in most biological experiments. In complex settings such as disease progression, there may be multiple important events against which time is reckoned.

Scalar and Vector Fields: Scalar and vector field data is normally thought of primarily in the context of spatio-temporal applications such as computational fluid dynamics, weather, climate, oceanography and combustion modeling. However, a number of participants of the workshop argued that such data is quite important for molecular and cell biology applications. Examples include modeling reactant and charge distribution across the volume of a cell, calcium fluxes across the cell surface or cell volume, reactant or protein fluxes across cell membranes, transport across cellular compartments, clinical response to drugs. Efforts in the visualization, computational fluid dynamics, and geographic information systems communities to deal with vector and scalar field data have focused on the development of fiber bundle or vector bundle data models.

Mathematical Models: Much of modern biological data analysis is concerned with the specification, development, parameter estimation, and testing (statistical or simulation) of various mathematical and statistical models of biological systems and datasets. Thus far the database community has largely been concerned with storing and querying input data sets, estimated parameters sets, and simulation output datasets. Relatively little attention has been paid to systematic methods of representing, storing, and querying the mathematical and statistical models being used. One would like to have declarative specification of mathematical and statistical models, means of recording bindings of model variables to database contents, and some way of recording the statistical analysis method (or simulation method) used.

Constraints: Historically, DBMSs have provided mechanisms to specify and enforce a variety of logical constraints on the contents or allowable updates of the database, e.g., referential integrity constraints. Biological databases require a variety of constraint specifications, both logical rules, and mathematical constraints (e.g., equations or inequalities) as first class data types in a biological data management system, with the ability to store, enforce, and query such constraints. In particular, traditional DBMSs typically have no mechanism to enforce non-local constraints.

Examples of mathematical constraints include various conservation constraints such as mass, momentum and energy conservation. Thus individual chemical reactions in a bio-pathway database must satisfy mass balance for each element. Such constraints are local. In contrast, cycles of reactions in thermodynamic database must satisfy energy conservation constraints. These are non-local (global) constraints. Another example of non-local constraints are the prohibition of cycles in overlap graphs of DNA sequence reads for linear chromosomes, or in the directed graphs of conceptual or biological taxonomies

Patterns: On the many data types described above, one would naturally expect similarity queries. These are clearly important. In addition, much effort has gone into specifying, characterizing, and finding patterns (a.k.a. motifs), e.g. in DNA, RNA, and protein sequences. These patterns are often represented as regular expressions or Hidden Markov Models (HMMs) or other types of grammars. Biologists are interested in collecting, storing, and querying these patterns. Patterns thus need to be considered as first class data types, with support for storage and querying.

A second class of queries consists of pattern matching queries, i.e., queries which find instances of sequences, etc. which match a specific pattern. On strings these queries involve pattern specifications such as regular expressions, Hidden Markov Models, or chart grammars. Graph pattern queries might involve patterns specified by graph grammars, subgraph homomorphism queries, etc.

One will also want to be able query collections of patterns (motifs). One such query would involve finding all patterns which match a sequence (the inverse of the customary query). Alternatively, one might ask for patterns which are similar to a specified pattern. Pattern similarity might be defined either structurally (akin to sequence similarity) or in terms of the overlap in the sequences matched by the two patterns from a specified database.

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, which 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 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 acknowledgement 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 papers 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 wit query results. Experience in other settings, e.g., geographic information systems, indicates that unless metadata (e.g., data provenance) is automatically updated, it is likely to quickly become outdated.

Uncertainty

Biological data has a great deal of inherent uncertainty. Often, when a scientist says "A is a B" they mean "A is probably a B, because there is some (possibly substantial) evidence suggesting that such is the case". For example: A protein sequence may be erroneously recorded in GenBank, because only a partial protein was reported; this error is propagated when another scientist runs a Blast search against sequences in GenBank and reports matches against such an erroneous sequence.

For all of these reasons, it is important to recognize uncertainty (and possible inconsistency) of data recorded in biological databases. Standard database technology provides no support for uncertainty, since business-oriented commercial databases typically contain data that is certain. Individual investigators often resolve uncertainties and inconsistencies by manual inspection and editing of datasets – a process known as manual curation. In large scale database setting, explicit representation of uncertainty and automated tools for curation are needed. Difficulties in scaling up curation activities have been the bane of major public biological databases.

Workflow Management

Large scale molecular biology experiments and data analyses need workflow management systems to assist in orchestrating the work and recording the details of what was done to each sample and/or dataset. Both laboratory WFMS (known as LIMS – laboratory information management systems) and computational workflow management systems (sometimes called scientific workflow management) are needed for large projects. Explicit representation and recording of laboratory and computational protocols is useful in driving automated data analyses and subsequent retrieval of experiments on the basis of protocol features. Detailed records on experiments are useful for process yield analyses and process failure diagnosis. Biological workflows differ from conventional workflows in that pooling and splitting of samples is commonplace.

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 biological databases. Many researchers regard these specialized databases as extremely valuable, in part due to the very detailed and careful curation of the data by specialists in particular 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, query languages, etc. Adoption of such current technology by database providers was seen as a pressing issue.

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 will have to be seen as a social obligation of investigators similar to careful description of experimental methods in publication. The efforts of the Micro-array 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, e.g., 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 (SQL, OQL, XQuery, et al.)

In addition to the short-term needs of machine-readable descriptions, and the constructions of wrappers, there are deeper questions to be addressed with regard to model management. Mapping between different data models or data representations is an integral part of any biological database application. For example, there is often external information or archival data that must be imported to augment local computationally or experimentally derived data. Even within a single project, there can be the need for multiple models or representations for the same kind of information, as it moves through various stages, e.g., data entry, data query, data interchange, and data archiving. With Affymetrix gene expression data, for example, data entry may be what Affymetrix produces, data query may be a relational database with some local model, data interchange may use MAGE-ML, and data archive is what some consortium requires.

Data sources evolve as knowledge changes and as new experimental techniques produce more data and different characterizations of the data. As a result, both the schemas that describe the data as well as applications and queries written specific to the original version of the schema must be updated. This is difficult to accomplish, particularly when the data types and structures are complex and when the analysis involves complex transformations or aggregations. Keeping up with evolution becomes significantly more difficult if there is a fundamental change in our understanding of the meaning or the characterization of the data.

Interdisciplinary Research

The past decade has seen the rise of bioinformaticists (a.k.a. bioinformaticians), a new group of researchers 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.

Recommendations

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 IT (information technology) 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 supporting such interdisciplinary work.

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.

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.

Conclusions

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 fifteen 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 either on commercial off-the-shelf DBMSs developed for business data, or on homegrown systems that are neither flexible nor scalable. 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 (such as sequences, graphs, 3D structures, etc.) and queries (e.g., similarity based retrieval), data provenance tracking, and integration of numerous autonomous databases.

Full Report

This article is an extract from the full report, which is available online from the workshop web site [1]. This site also contains the position papers, the original workshop proposal, attendee lists, etc. Position papers from several attendees, and an interim summary report were published in the *OMICS Journal*[2]; these materials are also accessible via the workshop web site.

Acknowledgements

This document is the product of a workshop funded primarily by the National Science Foundation, Computer and Information Science and Engineering Directorate under grant EIA-0239993. 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.

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.

References

[1] *Workshop on Data Management for Molecular and Cell Biology*, web site. http://www.lbl.gov/~olken/wdmbio/

[2] Data Management for Integrative Biology, Special Issue of the *OMICS Journal*, vol. 7, no. 1, Jul. 2003.