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

Data Management in Molecular and Cell Biology: Vision and Recommendations

BARBARA ECKMAN,[1] JULIA RICE,[2] and PETER SCHWARZ[2]

With the completion, or near completion, of the sequencing of the human genome has come the promise, for the first time, of being able to understand intra- and inter-cellular processes from the molecular level up. A wide range of chemical and biological phenomena are under investigation, including:

- The prediction of the 3D structure of proteins
- The determination and understanding of the dynamics and kinetics of protein folding
- The role of protein-protein interactions
- Ligand (small-molecule) and protein binding
- Biomolecular reaction pathways
- Modeling of intracellular processes and intercellular communication

Similarly, there is a wide range of physical models currently available and continuously being developed that span from considering electrons, atoms and molecules as the system elements and modeling aspects of biomolecular reactions at a highly accurate level using quantum mechanics, to considering gene and protein networks within a cell as the system elements and modeling the kinetics and dynamics of cellular processes using differential equation methods or stochastic techniques.

DATA MODELING ISSUES

The data structures used to represent biological data must reflect the physical model used to explore and understand the biological system, but it must also reflect the manner in which the experimental data is collected and catalogued. The provenance of the data must also be captured, e.g., the application that produced it, the date, and parameters used. Since molecular and cell biology are on the frontier of scientific investigation, physical models of biological systems, experimental techniques, and computational methods are all evolving rapidly, and database schemas must also evolve rapidly to keep pace.

Complex interactions in biological systems are frequently modeled as directed (multi-) graphs. To date, data relevant to cellular or pathway modeling has historically been limited in size, because of the difficulty of obtaining it, and so it has been possible to use specialized data management systems concerned mainly with representation and perhaps less with efficiency. This situation will change in the near future, as high throughput gene expression experiments continue to generate more data that can potentially be used in these models. Traditional database management systems must be extended to handle the directed multi-graph data structures and operations on them naturally and efficiently.

[1]IBM Life Sciences, Somers, New York.
[2]IBM Almaden Research Center, San Jose, California.

DATA MANAGEMENT ISSUES

A variety of resource needs

Different physical models pose a range of requirements in terms of amount of data in, amount of computation per input datum, amount of communication during computation, and amount of data produced. The increasing volumes of experimental data will require techniques for bulk data management where *ad hoc* methods once sufficed. Some, particularly those modelled at the finest level of detail (e.g. the electron level) are very compute intensive and are driven by, and produce, a relatively small amount of data. For these methods, the need for the computation to be close to the data or defined within a traditional database system is more a matter of convenience than requirement. At the other end of the spectrum are less compute intensive models, which in the simplest case can be the testing of a hypothesis by running a database query over a large volume of data, where optimized access to data is key. In the middle of the spectrum, molecular dynamics simulations of protein folding generally require a small amount of data to initiate, but can be considerably compute intensive, and can generate a large volume of molecular trajectory data. This output data must be organized to facilitate efficient analysis, such as comparison of a 3D snapshot of the system with all other 3D snapshots, or the aggregation of data from many snapshots to establish an average or identify a salient event. In these cases it may make sense to incorporate both the analysis routines and the data within a traditional database management system.

An additional feature of the physical models is that the models at each level of biological complexity build on the output of the models from the levels below. Thus, for example, if the rate of a given biomolecular reaction changes (as a result of improvements in a reaction rate model) then the outcome of the simulation of the cell may be affected. Incorporating this hierarchy into the data representation would clearly be an advantage.

Data and application integration

Data integration, both syntactic and semantic, continues to be critical. For example, much data has already been collected around focused topics, for example, protein sequence, 3D structure, and protein-protein interactions. But since the focus of these data sources of necessity has been relatively narrow, the data is not necessarily structured optimally for integrating into a unified representation of a cell and all its processes.

While the development of standards for biological data representation is a desideratum, the fast-paced, ever-evolving nature of biological research suggests that heterogeneity will continue to pose a critical challenge for biological data integration. To handle semantic integration, schema management systems are needed to perform such functions as identifying and mapping between corresponding structures in different database schemas that represent the same or related biological objects, and facilitating queries over different versions of evolving schemas.

Scientific data exploration requires complex, optimized queries to be posed over a wide variety of related data sources. While large data warehouses can solve key integration needs, we argue that it is not practical to require all relevant data to be transformed and loaded into a central warehouse. Therefore, a federated approach to integration will continue to be a necessary complement to warehousing, and as a means of integrating specialized operations/functions such as chemical substructure searches, sequence analysis packages, text mining algorithms, or motif searching.

A second need is application integration. Workflow definition and management is key for high-throughput data generation and analytic activities. In these workflows, the results of one application often form the input to another application, for example, the multiple alignments of clustalw are fed into PAUP for phylogenetic tree reconstruction, and the results of a gene expression normalization algorithm may be input to Spotfire for visualization. Standards for data exchange will be critically important in this effort.

TECHNOLOGIES THAT CAN HELP

Database federation

Database federation supplements data warehousing by providing up-to-the-minute data currency and on-the-fly "drill-down" capability to related non-warehoused data, for example, data that is highly volatile.

SQL/MED is an emerging standard for making data sources available in federated systems (Melton et al., 2002; ISO, 2003).

Part 9 of the ISO SQL standard, Management of External Data (SQL/MED) specifically addresses database integration. The standard specifies the interface between a centralized database engine, referred to as the *federated server*, and other queryable data sources, referred to as *foreign servers* and accessed via software modules called *wrappers*. To an application that connects to the federated server, the information managed by the foreign servers appears as part of a single *federated database*, queryable with standard SQL. By standardizing the interface between the federated server and wrappers for foreign servers, SQL/MED will make it possible to take a large step toward "plug and play" database integration. Owners of information can export their data and its specialized search methods via a SQL/MED compliant wrapper that encapsulates the peculiarities of the system in which it is stored. Those who wish to access the information can plug this wrapper into a SQL/MED compliant federated server, and immediately begin to issue SQL queries that access the new information, possibly combining it with their own information or data from still other sources.

The SQL/MED standard has been designed so that such *federated queries* can take advantage of special-purpose capabilities (e.g., chemical structure similarity matching, sequence alignment, etc.) of the foreign servers, and so that an optimizer at the federated server can devise a plan for executing such queries efficiently. Getting past the basic issues of connectivity and optimization by means of SQL/MED will allow practitioners to focus on the remaining challenge of overcoming structural and semantic mismatches between data sources.

While SQL/MED is part of the ISO SQL standard, and therefore based on the relational data model, SQL/MED compliant wrappers should also interoperate smoothly with federated systems whose data models subsume the relational model, for example, object-relational, complex-relational, and collection type data models. Since the relational data model is arguably still the *de facto* data management standard, our position is that it is currently the best choice for a federated *lingua franca*.

Since hierarchical data structures are ubiquitous in biology, it is desirable to access federated data through a structured view, for example, an XML (Bray et al., 2000) view using a declarative query language like XQuery (Boag et al., 2002). Among other advantages, the XML data model provides ordered lists as well as the sets (bags) of the relational model.

Workflow technologies

Although many complex computations are expressible as queries, others are more conveniently modeled as workflows among cooperating applications. XML, and in particular the Web Services Description Language (WSDL) (Chinnici et al., 2002), facilitate this kind of interaction in several ways. Firstly, XML provides a format for information interchange in which the basic data types are standardized in a manner that is independent of the programming languages in which the applications were developed or the hardware platform on which they are deployed. Moreover, XML supports a rich data model that allows for the interchange of complex, hierarchically-structured objects. XML documents are self-describing, with flexible schemas that can accommodate documents with missing or extra information.

Standards for exchanging information between applications are just a starting point. WSDL takes interoperability to the next level of abstraction by defining a standard for describing services that can be accessed via the web. A WSDL specification describes each operation a server can perform, and the format of XML documents representing the input required and output produced. Tools are becoming available that facilitate the development of web services, requiring little more than a mouse-click to turn an existing or newly-developed application into a web service available to other parts of the enterprise or the external world. Once this vision is fully realized, Web Services will allow computations in support of complex business processes that cross organizational boundaries to be assembled from well-defined building blocks and executed on the web.

Grid technologies

Grid technology facilitates efficient implementation of complex computations by means of secure, controlled access to shared computing and data resources. It is becoming clear that understanding biological sys-

tems will require the orchestration of a wide array of software tools for the integration, mining, analysis and visualization of data from public and private databases. At each step in the process, the computing resources (e.g., processors, software, storage) required by a particular tool must be located and allocated to the task at hand, and the necessary data obtained from a database or the output of another tool. In some cases, work must be partitioned so that it can be carried out in parallel on a battery of computing engines. To date, most of these tasks have been handled in an *ad hoc* manner, but the tremendous increase in both the amount of raw data and the number and kind of tools available to analyze it are making this approach untenable.

Grid computing initiatives, like the Open Grid Services Architecture, aim to solve this problem by defining generic web-based infrastructure to simplify these tasks (Foster et al., 2002a,b). The goal is to *virtualize* the resources necessary for a computation, so that users of the grid need not be aware of their location. By a process known as *provisioning*, the grid infrastructure locates the necessary resources and brings them together to perform the computation. Provisioning may involve some or all of: finding an appropriate hardware platform for the computation, obtaining and installing or configuring the necessary software, shipping required datasets to where they are needed, and the establishment of communication channels between parts of a parallel computation. All these tasks must be managed in a manner that allows the sharing of resources to be carefully controlled, so as to meet strict security and quality-of-service guarantees.

Many of the problems that must be solved to implement the grid are standard distributed computing problems that have been addressed by prior research. But this research has typically assumed an environment of closely-cooperating nodes with more-or-less homogeneous software. The advent of standards for Web Services, however, offers the real possibility that a wide range of computing services will be offered through standardized interfaces while preserving the autonomy of individual providers to choose a particular implementation. Such interoperability will be critical to the successful establishment of a broadly accepted grid infrastructure.

CONCLUSION AND RECOMMENDATIONS

The development of XML standards for representation and exchange of biological data is a critical need, and efforts such as the OMG-LSR (www.omg.org/lsr) and the I3C (www.i3c.org) are crucially important. But because molecular and cell biology are on the frontier of scientific investigation, representations of biological objects and systems may be expected to continue to evolve rapidly, and research is needed into schema evolution and mapping.

We see the need for two parallel architectures for integration of federated data and applications, respectively: wrappers written to the SQL-MED API specification, to permit optimized declarative queries over data sources and analytic algorithms; and web services with WSDL interfaces, to enable easy application integration. We urge funding agencies to support—perhaps even require—funded efforts that produce data sources or algorithms to make them available over the Internet to the community as WSDL web services, SQL-MED wrappers, or, ideally, both.

Finally, as the volume of data and the complexity of data management tasks continues to grow with the increasing complexity of the biological systems and processes that are modeled, it becomes ever more important for tools that are made available to the biological research community to be robust and performant. While funding agencies typically support the prototyping of new tools, they usually do not provide continuing support for maintenance and upgrades. The open source movement works well for standalone scripts and relatively simple components, but is not well suited to components of complex database management systems, for example, optimizers, special indexing schemes. We urge researchers developing new optimization algorithms, indexing methods, etc. to work closely with data management vendors to incorporate their new insights and tools into "industrial strength" data management products, where they may be readily maintained and extended for the good of the community.

REFERENCES

BOAG, S., CHAMBERLIN, D., FERNANDEZ, M., et al. (2002). XQuery 1.0: An XML Query Language. W3C Working Draft. Available: www.w3.org/TR/xquery.

BRAY, T., PAOLI, J., SPERBERG-MCQUEEN, C.M., et al. (2000). Extensible Markup Language (XML) 1.0 (Second Edition). W3C Recommendation. Available: www.w3.org/TR/REC-xml.

CHINNICI, R., GUDGIN, M., MOREAU, J.-J., et al. (2002). Web Services Description Language (WSDL) Version 1.2. W3C Working Draft. Available: www.w3.org/TR/wsdl12.

FOSTER, I., KESSELMAN, C., NICK, J., et al. (2002a). Open Grid Service Infrastructure WG. The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. Available: www.globus.org/research/papers/ogsa.pdf.

FOSTER, I., KESSELMAN, C., NICK, J., et al. (2002b). Grid Services for Distributed System Integration. Available: www.globus.org/research/papers/ieee-cs-2.pdf.

INTERNATIONAL STANDARDS ORGANIZATION. (2003). Information Technology—Database Language—SQL—Part 9: Management of External Data (SQL/MED). FCD (Final Committee Draft) 9075-9-200x (currently under ballot). Available: http://sqlstandards.org/SC32/WG3/Progression_Documents/FCD/4FCD1-14-XML-2002-03.pdf.

MELTON, J., MICHELS, J.-E., JOSIFOVSKI, V., et al. (2002). SQL/MED—A Status Report. Available: www.acm.org/sigmod/record/issues/0209/jimmelton.pdf.

Address reprint requests to:
Dr. B. Eckman
IBM Life Sciences
Route 100
Somers, NY 10589

E-mail: baeckman@us.ibm.com