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

Data Modeling and Data Management for the Biological Enterprise

LOUIQA RASCHID

RECENT TECHNOLOGICAL ADVANCES in computational biology have had a fundamental impact on the vast amount of data that is available to the biological enterprise. To date, technological advances in data modeling and data management have not focused on the special needs and challenges of data management for cellular and molecular biology. These needs as well as the strengths and weaknesses of existing solutions for biological data integration are summarized. Next, we discuss challenges and potential solutions to the emerging data management needs of biologists. If these issues are not addressed satisfactorily, it is unclear if the adoption of data management technology will be of benefit to the biological scientist.

DATA REPRESENTATION AND DATA MANAGEMENT NEEDS FOR THE BIOLOGICAL ENTERPRISE

- Data models appropriate for capturing both complex structure and function
- Ability to model heterogeneity and semi-structured data
- Scalable access to large amounts of data that is often stored at remote sources, is accessible via WANS, and is managed by autonomous servers
- Support for scientific exploration of both metadata and data
- Support for non-traditional processing, for example, pattern matching on complex structures; ranking, filtering and data mining on multi-modal data

STRENGTHS AND WEAKNESSES OF AVAILABLE TECHNOLOGY

Existing approaches for data integration can be broadly classified as follows:

- Scripts in languages such as Perl or Python
- Mediation or Federated DBMS
- Data warehouses
- Peer-to-peer access where groups of peers determine their level of data integration
- Service-based, for example, web services a la WSDL
- Emerging technologies such as Semantic Web
- Workflow-like specification of scientific exploration tasks

We briefly review the advantages and disadvantages of some of the more popular solutions that have been deployed in the biological enterprise and identify some example systems. Scripts written in Perl or

Smith School of Business, University of Maryland, College Park, Maryland.

Python are the most common solution. The drawbacks of these solutions are well known and include the difficulty of maintaining and re-using scripts, especially as the underlying data sources evolve. More important, scripts provide no support for the incorporation of data management and data analysis tools; consequently they are not a scalable robust solution. Data Warehousing (DW) is also a popular solution to create repositories for specific tasks. Its strengths include rapid access to data warehoused from multiple remote data sources; the availability of tools for data cleaning; and support for privacy and security. Weaknesses are that DW technology traditionally supports the relational data model and (R)OLAP and is not suited for the complex structure and semi-structured data of biological sources. There is little support to resolve semantic heterogeneity across sources. Further, DW solutions cannot utilize complex search and query processing services, for example, BLAST or search engines, hosted at remote servers, nor can they explore the increasing number of hyperlinks and annotations that are typically added by data curators, after the records are initially entered into the source, for example, the Nucleotide link on PubMed. Finally, the greatest drawback of DW solutions is that data in the warehouse becomes stale and must be refreshed. Data sources such as GenBank or PubMed are constantly being updated. A more recent solution that has been adopted by the biological enterprise includes a variety of architectures for federated access or mediation. Strengths include the ability to access remote servers accessible over WANS. These solutions are more tolerant of semi-structured data compared to DW since they are built on DBMS platforms that are not always limited to relational data models. A major advantage is that they can exploit complex search and computational services hosted at remote servers. As with DW solutions, federation or mediation also does not provide many tools to process complex or semi-structured data. There is little support to resolve semantic heterogeneity across sources. Finally, such solutions may fail when remote servers are inaccessible.

Several systems have been designed for domain specific integration of biomolecular data. BioKleisli (Davidson et al., 1997) and its extensions K2 (Davidson et al., 2001) and Pizzkell/Kleisli (Wong, 2000) follows a mediation approach and enables queries against integrated data sources. P/FDM (Kemp et al., 1999, 2000) provides support to access specific capabilities of sources such as SRS (Etzold and Argos, 1993). No semantic knowledge is expressed or utilized in either system. TAMBIS is primarily concerned with overcoming semantic heterogeneity through the use of ontologies. It provides an integrated view of data sources but offers no ability to explore and exploit alternate identifiers and alternate links (paths). Garlic and its new extension for life science DiscoveryLink (Haas et al., 2000) encapsulates the access to specialized search capabilities into user-defined functions. While they provide extensive cost based optimization to support efficient and seamless data integration, they too are hindered by the lack of knowledge about source capabilities as well as semantic knowledge about relationships among sources and their contents (Eckman et al., 2001a). The OPM multi-database system is based on the Object Protocol Model (OPM; Chen and Markovitz, 1995) and object views (Chen et al., 1997). While OPM provides the ability to evaluate complex queries, it too does not capture knowledge of semantic equivalences of scientific entity instances, links, and paths (Eckman et al., 2001b). The Sequence Retrieval System SRS (Etzold and Verde, 1997) applies full text indexing and keyword based search techniques that are indeed very powerful. However, it is limited in that SRS was not designed to support semantic equivalences. For example, the SRS interface available at EBI offers the powerful capability of retrieving all sequences from EMBL that contain the keyword apoptosis in their description field (DE). However, an SRS query against both EMBL and MEDLINE no longer offers this powerful capability and is limited to full text search on apoptosis; thus the search on both data sources may return large numbers of irrelevant hits.

We now address the challenges that must be addressed if data management tools are to enhance the biological scientist.

DATA INTEGRATION AND SEMANTIC HETEROGENEITY

Providing integrated access to multiple heterogeneous sources is the first step. The next step is providing tools to integrate data across these sources. In this regard, federated/mediated solutions may be more flexible in providing tools to register new sources and to populate the Mediator Catalog. However, both DW and federated/mediated solutions provide little support to the scientist to address the challenges of se-

mantic heterogeneity. Without such support integrated access will not provide the potential benefits of integration across sources.

Data about scientific entities, for example, genes or proteins, are stored in multiple sources. Each source captures some features (attributes) describing both the structure and function of the scientific entity. Typically, information about a single instance of some scientific entity, for example, the gene TP53, may be found in multiple data sources. While there is overlap of data among sources, typically these sources are not replicas. Instead, each source references instances in other sources, and each source captures some information about the structure or function of the scientific entities. Under these circumstances, solving the challenge of data integration across multiple data sources, successfully, requires the acquisition of metadata about the contents and overlap of contents among these sources in order to correctly identify and completely characterize the structure and function of an instance of a scientific entity across multiple sources. Currently, neither DW nor federated/mediated solutions provide adequate support to address this issue.

SCALABLE PERFORMANCE

A key objective of data integration is seamless and efficient access to remote sources. The first aspect of quality of service is the end-to-end latencies or delays associated with a computational task. The second aspect is the quality of the contents or results.

Once the scientist has accomplished a process of discovery, she or he is able to formulate a complex computational task to be evaluated across multiple remote data sources. Mediation, DW, and workflow technologies are all suited to support a reliable and efficient computational platform for data integration. Further research is needed to support learning the costs of query evaluation in noisy WANs; query evaluation with delayed, bursty or completely unavailable sources; cost based query optimization that can exploit the existence of multiple alternate remote sources and the complex search and query processing services hosted by remote servers. Challenges to this task include the following:

- There is difficulty in predicting costs accurately. Learning and other techniques are needed to construct cost models (Gruser et al., 2000; Nie et al., 2001).
- A variety of optimization approaches are needed, e.g., performance targets; alternate sources; adaptive evaluation strategies (special issue of *IEEE Data Engineering 2001,* edited by Hellerstein et al., 2002a,b).
- In many situations, clients, especially automated clients such as crawlers, can overwhelm the computational capability of a data source. There is a need for servers to be able to advertise their constraints and semi-automated mechanisms to enforce these constraints. An example of server constraints are those published by NCBI for users of their E-Search utilities, which prohibit automated tools from accessing their servers during peak access periods.

Typically, query optimization with multiple alternate data sources makes the assumption that the results are independent of the particular source or query evaluation plan that is chosen. For biological data sources, while there is significant overlap of sources, few of the sources are exact replicas. As an example, the three sequence data sources—GenBank, DDBJ, and EMBL—do not all contain the same data about sequences. There has been some research on query evaluation with incomplete, imprecise or alternate but dissimilar sources, as well as flexible query answering and approximate query answering (Duschka, 1997; Florescu et al., 1997, 1998; Naumann, 2001; Naumann et al., 2000; Workshop on Flexible Query Answering; Yerneni et al., 2000). Issues include the following:

- Imprecise values or missing data or dirty data
- Unavailable sources
- Alternate sources and query evaluation plans with dissimilar semantics, for example, result cardinality may vary or characterization of objects in sources may be different

The challenge is query planning that can exploit domain specific semantics to provide answers that closely match the specific result quality and performance requirements of the biological scientist or application. For example, a scientist who is exploring some hypothesis will very like be interested in reducing access

latencies as she or he explores multiple alternatives. However, for a validation task, a scientist would probably want to explore the results from all the relevant sources, despite the overlap of their content.

SCIENTIFIC EXPLORATION

Traditional data access (based on traditional data management technology) requires the specification of the query (e.g., SQL) or some application program that is to be evaluated against the data. This is a limitation on the process of scientific discovery where the scientist wants the ability to express a workflow of potentially complex operators, each of which may have some domain specific semantics. An example is where the scientist wishes to gather a collection of proteins that has a maximal number of links to certain publications on some disease, and where the proteins are associated in a specific database to other proteins with specific characteristics. Current solutions based on scripts provide little support for this task. Solutions based on DBMS technology would not support the scientist to construct a complex query. Thus, there is a critical need for a biological exploration language that can support the user as she or he browses the metadata, contents and links, and query processing services of multiple sources, and allows the user to express a complex workflow and domain specific semantics, corresponding to her task. Without such support, the scientist will be hindered by the limitations of current query languages.

REFERENCES

CHEN, I.A., KOSKY, A.S., MARKOWITZ, V.M., et al. (1997). Constructing and maintaining scientific database views. Proceedings of the 9th Conference on Scientific and Statistical Database Management.

CHEN, I.A., and MARKOWITZ, V.M. (1995). An overview of the object-protocol model (OPM) and OPM data management tools. Information Systems, **20,** 393–418.

DAVIDSON, S., CABTREE, J., BRUNK, B., et al. (2001). K2/Kleisli and GUS: experiments in integrated access to genomic data sources. IBM Systems Journal, **40,** 512–531.

DAVIDSON, S., OVERTON, C., TANNEN, V., et al. (1997). BioKleisli: a digital library for biomedical researchers. Journal of Digital Libraries.

DUSCHKA, O.M. (1997). Query optimization using local completeness. Proceedings of the AAAI/IAAI 249–255.

ECKMAN, B., LACROIX, Z., and RASCHID, L. (2001). Optimized seamless integration of biomolecular data. Proceedings of the IEEE International Symposium on Bio-Informatics and Biomedical Engineering (BIBE 2001).

ECKMAN, B., KOSKY, A., and LAROCO, L. (2001). Extending traditional query-based integration approaches for functional characterization of post-genomic data. BioInformatics **17,** 587–601.

ETZOLD, T., and ARGOS, P. (1993). SRS, an-indexing and retrieval tool for flat file data libraries. Computer Applications of Biosciences **9,** 49–57.

ETZOLD, T., and VERDE, G.. (1997). Using views for retrieving data from extremely heterogeneous databanks. Proceedings of the Pacific Symp. Biocomput. 134–141.

FLORESCU, D., KOLLER, D., LEVY, A.Y., et al. (1997). Using probabilistic information in data integration. Proceedings of the VLDB-97.

FLORESCU, D., LEVY, A.Y., MENDELZON, O. (1998). Database techniques for the world-wide web: a survey. SIGMOD Record **27,** 59–74.

GRUSER, J.-R., RASCHID, L., VIDAL, M.E., et al. (1999). A wrapper generation toolkit to specify and construct wrappers for web accessible data sources (websources). Journal of Computer Systems, Special Issue on Semantics in the WWW **14,** 83–98.

GRUSER, J.-R., RASCHID, L., ZADOROZHNY, V. et al. (2000). Learning response time for websources using query feedback and application in query optimization. Very Large Data Bases Journal (Special Issue on Databases and the Web) **9,** 18–37.

HAAS, L., KODALI, P., RICE, J., et al. (2000). Integrating life sciences data—with a little garlic. Preceedings of the IEEE International Symposium on Bio-Informatics and Biomedical Engineering (BIBE).

KEMP, G., ROBERTSON, C., and GRAY, P. (1999). Efficient access to biological databases using CORBA. CCP11 Newsletter **3.1.**

NAUMANN, F. (2001). Quality-driven query answering for integrated information systems. Lecture Notes on Computer Science, **2261.**

NAUMANN, F., LESER, U., and FREYTAG, J. (1999). Quality-driven integration of heterogenous information systems. Proceedings of the VLDB.

NIE, Z., KAMBHAMPATI, S., NAMBIAR, U., et al. (2001). Source coverage statistics for data integration. Proceedings of the 3rd International Workshop on Web Information and Data Management (WIDM).

WONG, L. (2000). Kleisli, its exchange format, supporting tools, and an application protein interaction extraction. Proceeding of the IEEE International Symposium on Bio-Informatics and Biomedical Engineering (BIBE).

YERNENI, R., NAUMANN, F., and GARCIA-MOLINA, H. (2000). Maximizing coverage of mediated web queries. Technical report, Stanford University.

ZADOROZHNY, V., RASCHID, L., URHAN, T., et al. (2002). Efficient evaluation of queries in a web query optimizer. Proceedings of the ACM SIGMOD Conference.

ZADOROZHNY, V., and RASCHID, L. (2002). Query optimization to meet performance targets for wide area applications. Proceedings of the International Conference on Distributed Computing Systems.

Address reprint requests to:
Dr. Louiqa Raschid
University of Maryland
4315 Van Munching Hall
College Park, MD 20742-3251

E-mail: louiqa@umiacs.umd.edu