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

Designing Efficient User-Friendly Biological Data Management Systems

ZOÉ LACROIX

BIOINFORMATICS can refer to almost any collaborative effort between biologists or geneticists and computer scientists and thus covers a wide variety of traditional computer science domains including data modeling, data retrieval, data mining, data integration, data managing, data warehousing, data cleaning, ontologies, simulation, parallel computing, agent-based technology, grid computing, and visualization. However, applying each of these domains to biomolecular and biomedical applications raises specific and unexpectedly challenging specific research issues.

The design of a biological data management system relies on the access and exploitation of information related to diseases, disorders, and condition. This information is available at multiple data sources and requires sophisticated tools for its access and analysis. Biological is growing at a rate unseen since the earliest days of the field. Gene sequencing robots, new experimental methodologies and online data collection devices are causing exponential growth in the amount of raw data on the web that is available to the life scientists. Life scientists need to exploit transparently these large datasets with various new applications to analyze, mine, cluster, and visualize this wealth of information. A transparent biological data management system should provide life scientists the ability to access data and applications despite the lack of explicit knowledge about where the data are stored, how they are structured, where the application is running.

Many systems have been developed since the meetings on the Interconnection of Molecular Biology Databases (the first of the series was organized at Stanford University in the San Francisco Bay Area, August 9–12, 1994) and the list of queries of the DOE report on Genome Informatics (Robbins, 1993). Although successful, these systems are often limited and fail to meet all users' needs while the needs and the problems to address became significantly more complex. During their development and usage, existing approaches collected fruitful experience. The analysis of the past experiences should benefit the research community. It is time to take some distance and try to get the perspective provided by the accurate insight into the specific problem being addressed by each system, why the particular architecture was chosen, its strengths, and any weaknesses it may have, to evaluate them and provide an overall summary of these approaches, and their characteristics (advantages and disadvantages).

The diversity of data sources and the multiple of applications often distributed on the Internet raise complex issues related to integration. Traditional integration approaches are typically not addressing the two dimensions of the problem: multi-database systems, mediations and warehouses are data-driven whereas agent architectures (CORBA), Web services, and more recently grids are application-oriented. New approaches integrating both data and applications with the flexibility needed to accommodate life scientists are still needed. The problem is made more complex by the semantic mismatches between scientific resources. Information about scientific objects (e.g., a sequence, a gene) is typically spread over multiple data sources, each providing a different identifier for the object. A biological data management system must integrate them all and reconcile these different identifiers in order to provide life scientists a transparent access to each scientific object. Existing efforts to formalize keys for scientific objects, data formats, and

Department of Mechanical and Aerospace Engineering, Arizona State University, Tempe, Arizona.

schemas, that aim to be shared by the scientific community will be simplifying, thus not solving, the biological integration problem.

The complexity of the needs for data integration also raises the question: is it possible to design a system that would actually meet all users needs? Clearly, life scientists and potential users for such systems do not have the same computer skills, they often do not have the same scientific background either. And if they often share the same objectives, they typically use the same vocabularies but with different semantics! Under these circumstances, when the data themselves should be shared beyond these delimited scientific boundaries, does it make sense to design a system that will meet the expectations of all life scientists? If the underlying implementation may remain the same, it seems that systems should offer various user interfaces to satisfy the multiple users needs of this community.

Systems characteristics can be articulated with respect to two orthogonal dimensions: system and user perspective. The user perspective expresses the ability of the system to meet its users' needs and, therefore, it drives the requirements of the system to be developed or chosen. The system perspective captures the characteristics of the system from the technical point-of-view. Much of this perspective is driven by the user requirements however it reflects only one of many possible implementations satisfying these requirements. While both views are helpful in understanding a system, and there is significant overlap between them, the true success of a system is determined by whether or not its users are satisfied. Thus the user perspective is ultimately the more important. However, life scientists' needs are often difficult to evaluate. Life scientists typically follow a query-driven approach: they design a protocol that aims to collect or analyze data in order to answer a particular scientific question. In contrast, computer scientists have a generic approach and aim to design a system that enables users to answer multiple questions. As a consequence, approaches developed by computer scientists often do not provide life scientists with the expressiveness needed to carry out their immediate needs as illustrated in Figure 1. They typically lack the flexibility needed to exploit the latest versions of sophisticated analysis tools or new data repositories. However, they usually provide a framework for efficient query processing.

To better support life scientists, current research should develop approaches to collect and analyze the needs of the scientific community. Efforts in developing users' survey and evaluation criteria to capture the level of usability, flexibility, scalability, and other characteristics expected by users and to be able to evaluate the overall users' satisfaction could benefit future developments. While the analysis of these needs may result in system specifications that may guarantee a level of users satisfaction, users satisfaction may also increase by developing more efficient systems. Current research should also benefit from the development of appropriate performance models, evaluation matrices, cost models and benchmarks adapted to this context.

Existing systems are often difficult to use by scientists who often rely on information technology assistants to express their queries and design various interfaces significantly limiting the access to the underlying system, but providing users a level of understanding and the feeling of safety needed to answer their query pipeline. In this context, safety may involve various data treatments often hidden to the user such as the evaluation of the query (e.g., where the data are collected from, in what particular order, how they are assembled, and analyzed). Even these programmers may have difficulties in expressing the queries since existing languages are often not adapted to the type of queries. Obviously most of existing approaches are not well adapted to life science. New data models and query languages are needed to capture the complexity

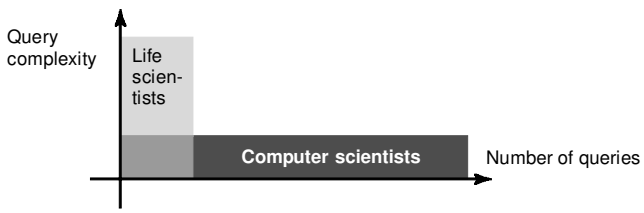


FIG. 1. Query-driven versus generic approaches.

of scientific data representation including temporal, spatial, mathematical, and sequence data, and the expressiveness of their manipulation and analysis.

To conclude, if the research still needs to address the unresolved yet issues on integration of heterogeneous distributed semi-structured data and the various applications that exists to analyze them, more focus on users satisfaction should drive the design and development of future systems. The main research directions can be summarized as follows:

- Design generic systems for specific scientific communities.
- Collect adequate users' requirement from use cases.
- Partner with life scientist to design scientific standards that are approved by the scientific community as well as machine exploitable.
- Design new data models and query languages.
- Design new systems architectures able to integrate both data and applications.

REFERENCE

J. ROBBINS, ed. (1993). U.S. Department of Energy (DOE) Human Genome Project. Report of the Invitational Workshop on Genome Informatics, 26–27 April 1993, Baltimore, Maryland, http://www.ornl.gov/TechResources/Human_Genome/publicat/miscpubs/bioinfo/contents.htm

Address reprint requests to:
Dr. Zoé Lacroix
Arizona State University
P.O. Box 876106
Tempe, AZ 85287-6106

E-mail: zoe.lacroix@asu.edu