Biological Data Management: Research, Practice and Opportunities

Thodoros Topaloglou
Senior Vice President, Scientific Computing
MDS Proteomics, Toronto, Canada
ttopaloglou@mdsp.com

1. Panelists

Susan B. Davidson, Professor of Computer and Information Science, and Bioinformatics, UPenn

H. V. Jagadish, Professor of Electrical Engineering and Computer Science, and Bioinformatics, University of Michigan

Victor M. Markowitz, Chief Informatics Officer Joint Genome Institute and Head of Biological Data Management and Technology Center at Lawrence Berkeley National Laboratory

Evan W. Steeg, Independent Bioinformatics Consultant

Mike Tyers, Senior Investigator, Samuel Lunenfeld Research Institute and Professor of Medical Genetics and Microbiology, University of Toronto

2. Introduction

Biological research and drug development are routinely producing terabytes of data that need to be organized, queried and reduced to useful scientific knowledge. Although data management technology can provide solutions to problems, in practice the data needs of biomedical research are not well served. The goal of this panel is to expose the barriers blocking the effective application of advanced data management technology to biological data.

Management of biological data involves acquisition, modeling, storage, integration, analysis and interpretation of diverse data types including analog signals, digital images, sequences, spreadsheets, taxonomies, structured records and unstructured text data. Existing data management technology is often challenged by the lack of stability, evolving nature, diversity and implicit scientific context that characterize biological data.

3. Biological Data Management

Why is biological data management such an important field? Just in the last 10 years we have witnessed the sequencing of the human genome, the genesis and widespread use of gene arrays, the industrialization of Proteomics and an explosion in biological data available in reference public databases and special purpose information products. These advances have profound influence in biology and drug discovery. Biological research has transformed from a purely experimental to an information-driven discovery science. Drug development is moving away from the hit-and-miss model towards rational and information-driven practices. Today, a biology student can get access to all the elements of a biological system (that is, the complete genome sequence and all the genes and proteins encoded by the genome) from (more than one) archival databases, and do hypothesis-driven science or global analyses.

The volume of available data and the pace by which post genomic era technologies generate data creates both risks and challenges that biological data management has to tackle. Is biological data management up to the task? The answer varies depending on who provides it: the database researcher, the bioscience user, and the bioinformatics practitioner in industry and in research. The answer to this question is also a topic that will be addressed by the panelists. Here we will briefly outline the different perspectives and translate

Proceedings of the 30th VLDB Conference,
Toronto, Canada, 2004

them into a list of provocative statements that the panelists will elaborate on.

3. Biological Data Challenges

What do biological data look like and how are they used in research and drug development? Biological data are broad and diverse. Biology encompasses many domains of knowledge (molecular and cell biology, genetics, structural biology, pharmacology, physiology, etc.) each one of which is concerned with overlapping or complementary entity types, and has it is own terminology and data needs. Furthermore, the variety of experimental procedures yield related but not identical data. New bio-analytical procedures and progress in the science add another dimension of instability to biological data types. The data types vary from sequences, to 3-dim structures, images, graph structures, data tables, semi-structured and unstructured text.

Most of the information that the biological research is interested in is available in public reference databases, specialized private data sources, and the over 12 million articles of the scientific research literature, most of which is accessible on the web. It is estimated that 80% of the biological data are in text form, and the rest resides in databases that range from indexed files, to relational and specialized formats. Biological databases may contain primary data i.e., have their own data entry or submission policy, or secondary data i.e., built by integrating data from primary sources in which case their integrity depends on the constituent sources. As many of these data sources are non-standard and not well documented, accessing, integrating and sharing biological data becomes a challenge and an art.

4. Current Practice

Despite the challenges, scientific users have available a wealth of information, and have built specialized applications to access portions of it. The widespread use of the web and Excel spreadsheets, makes the ad-hoc and unsustainable data access practices, unnoticeable to many. To the bioscientist, database development means production of a dataset and not the construction of system that manages data. The separation of the application from the representation is also not recognized as well as the need for a DBMS. Even when a DBMS is used, standard database design principles are not always followed. Just imagine the gains in productivity if all the loss of time due to sub-optimal means of managing and accessing data, were to be replaced with efficient data management solutions. So, what stands in the way? Education is perhaps the single most significant obstacle. As the biology sounds complicated and distant to the database community, and truly is, similarly data management technology is distant and incomprehensive to the biosciences. Only a small fraction of database professionals has crossed paths with biology compare to financials for instance.

In the biopharmaceutical industry the situation is much better. Due to the size of the data management problem, concerns of productivity and sometimes regulatory requirements, many organizations are taking aggressive steps to establish strong information management practices. Data management technology and database vendors are partners in such endeavors. Oftentimes the state of the art in data management technology may not be able to provide a complete solution and needs to be extended. For example, federated queries against heterogeneous data sources had limited DBMS support a few years ago, support for generic transformations from one representation is still not solved in DBMSs, and workflow management solutions are not there either. Our panelists are invited to describe examples where the biological data management requirements have led to the development of novel solutions that have contributed to the state of the art in database technology.

The scale and scope of the data management problem varies between organizations that generate data and centers that host community databases.

The “data cycle” in an organization that generates data, such as MDS Proteomics or a sequencing laboratory, starts with sample tracking, followed by the laboratory processing that produces vast amount of raw data, followed by analytical processing to translate the signals to measurements, to biological data such as sequence tags or abundance of gene or proteins, and ultimately to conclusions. There are several heterogeneous information systems with distinct data requirements involved in this process. Data and process tracking, data file management, high performance algorithms for data processing, efficient database loading and workflow management are everyday business whether or not optimal technology options exist. Daily production rates are in the order of tens of Gigabytes a fraction of which ultimately makes it into relational databases.

In the community database environments the pressure points are different. The two important functions are data curation / annotation and web publishing. Technology and tools for data curation and annotation has been the focus of bioinformatics research from the early days, and is still an active area of research with a significant data integration and application interoperability challenge as part of it. Fast access to community database data from the web could be served by "off the shelf" technology, however, for historical reasons or due to the specialized data types, specialized indexing schemes and sequence similarity search applications are deployed.

The solutions providers in biological data management vary across the board. Scientific users in academia are supported by bioinformaticians which in some cases, but not always, are linked to the data management community. In research networks such as the NIH, DOE and EMBL, biological data management is well served by groups that have pioneered in the field such as LBL, NCBI, EBI, CBIL and others. In industry, certain companies have built their own centers of excellence and others access the services of specialized vendors. In terms of tools, the field started with its own homegrown solutions such as AceDB and over time migrated to relational technology where the popular choices include MySQL, Oracle and DB2. XML is also popular as a data exchange format and oftentimes is misused as a data management solution but without the management tools support that the database community is active on. Finally, Excel and Perl/CGI contribute to the successes and to the pitfalls in managing biological data as they make users, for a while, independent of professional data management, which ultimately comes in to solve the scalability, performance or integration problems.

5. Open Problems

In recent a workshop organized by NSF[1], scientists from database research and biosciences debated the open problems in biological data management as perceived by both communities and aimed at defining a research agenda for data management technology support of basic and applied research in molecular and cell biology. The participants were asked to contribute an opinion paper based on a list of suggested topics.

We reviewed 27 opinion papers grouped in four categories based on the author's background. The categories include database/comp.science (DB/CS) background with academic (AC) or industry (IN) affiliation, and bioinformatics/biology (BFX/BL) background with academic or industry affiliation. The research institution affiliation was counted as academic. The number under each category indicates the category membership. The list of topics is extracted from the papers themselves and is slightly different than the suggested topics. A topic is given a check if it is mentioned as a challenge, open problem or proposed for future work in a paper. The following table summarizes the importance count of each topic.

Topic	DB/CS		BFX/BL		#
	AC (13)	IN (4)	AC (8)	IN (3)	
hierarchical repr., taxonomies			xx		2
biological data types	xxxx		xxxx		8
heterogenous, complex queries	xx	x	xx		5
schema mgmt, data transformation	x	x			2
data exchange formats	x		x		2
metadata repr. / management	x		xxx		4
ontologies, controlled vocabularies	x	xx	xxxxx	xx	10
interoperability, web services		xx	x	x	4
data quality	x	x	x		3
data provenance	x	xx			3
data modeling / data evolution	x	x	x		3
graph / pathway representation	xxx				3
data integration / semantics	xxxxx	xxx		x	9
workflow management		x	x		2
text mining			xx		2
natural language processing			x		1
understanding scientific requirements	x		xx		3
data life cycle		xx			2
semi-structured data / XML, ANS1	xxx	x			4
usability/ robustness/ performance	x	x			2
visualization	x				1

[1] "Database Management for Life Science Research: Summary Report of the Workshop on Data Management for Molecular an Cell Biology at the National Library of Medicine, Bethesda, Maryland, Feb 2-3, 2003", Jagadish, H.V., and Olken, F. OMICS, Vol. 7, No. 1, 2003.

Ontology development was the top pick with data integration and support for biological data types close behind. The next group of important topics includes support for heterogeneous and complex queries, interoperability and web-services, metadata representation and management, and support for semi-structured data and XML. The low number of industry participants and the exclusion of drug research as an area of focus, leave some open questions that our panelist will have to comment on. For example, one would expect that schema management, workflow management and data provenance are important to technology practitioners while text mining, data quality metrics and visualization are important to the industrial scientist. Nobody besides academic computer scientists selected the graph representation as an important topic, is this really the case?

The fact that ontologies, controlled vocabularies and domain specific terminologies was commented across the board but mostly by the user community, speaks of the importance of domain aware data management. But should the computer scientists embark on developing terminology for biology? Clearly not, but their information management tools should be compatible with such concepts, for instance by allowing attribute domains to be hierarchical terminologies and by providing proper query support for those. Also, repositories of domain terminologies, common schemas components and tools to manage and extend those can shift the paradigm of database development for biosciences as software information repositories impacted software engineering in the past.

Reference to the concepts of semantic web, web services, WSDL, and XML was made mainly by the industrial participants. Clearly these concepts have become popular in industrial bioinformatics and the database community is further ahead, but so far this topic has received little attention in the biological data management literature. Is it an area of promise for follow up research? Good question for our panelists. They are also invited to make the connection with some earlier efforts for creating interoperability and integration infrastructures for bioinformatics, the infamous CORBA/JAVA buzz, which led to some unsatisfying results.

3. Panel Questions

Our panelists are invited to comment on the following questions:

1. Is the current state of biological data management well suited to support the bioinformatics challenges in biological research and drug discovery?
2. Interpret the data of the above table. Do you agree with the top five priorities? Or, what your top five choices would be? Are there any other important topics that should be added?
3. Do the areas above require new or additional research? Or can existing data management tools and methods be adapted for them?
4. Discuss the difference between problems in the above areas and problems in analogous traditional data management areas.
5. Provide examples of biological data management systems that address (some of) the problems in the areas above
6. Comment on adequacy of commercial DBMSs and tools for these problems.
7. Does the database research community need more education with respect to the challenges in biology and drug discovery, and if so, what do you propose