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

Data Management Challenges for Molecular and Cell Biology: An Industry Perspective

VICTOR M. MARKOWITZ

DATA MANAGEMENT for molecular and cell biology involves the traditional areas of data generation and acquisition, data modeling, data integration, and data analysis. In industry, the main focus of the past several years has been the development of methods and technologies supporting high-throughput data generation, especially for DNA sequence and gene expression data (Cambridge Healthtech Institute, 2001). New technology platforms for generating biological data present data management challenges arising from the need to capture, organize, interpret and archive vast amounts of experimental data. Platforms keep evolving with new versions benefiting from technological improvements, such as higher density arrays and better probe selection for microarrays. This evolution raises the additional problem of collecting potentially incompatible data generated using different versions of the same platform, encountered both when these data need to be integrated and analyzed. Further challenges include qualifying the data generated using inherently imprecise tools and techniques and the high complexity of integrating data residing in diverse and poorly correlated repositories.

The data management challenges mentioned above, as well as other data management challenges (Jagadish and Olken, 2003), have been examined in the context of both traditional and scientific database applications. When considering these challenges, it is important to determine whether they require new or additional research, or can be addressed by adapting and/or applying existing data management tools and methods to the biological domain.

The experience gained at Gene Logic in developing data management systems for gene expression data (Gene Logic Products, 2003) suggests that existing data management tools and methods, such as commercial database management systems, data warehousing tools and statistical methods, can be adapted effectively to the biological domain. For example, the development of Gene Logic s gene expression data management system has involved modeling and analyzing microarray data in the context of gene annotations (including sequence data from a variety of sources), pathways, and sample (e.g., morphology, demography, clinical) annotations, and has been carried out using or adapting existing tools. Dealing with data uncertainty or inconsistency for experimental data has required statistical, rather than data management, methods; adapting statistical methods to gene expression data analysis at various levels of granularity has been the subject of intense research and development in recent years (Oz, 2003). The most difficult problems have been encountered in the area of data semantics-properly qualifying data values (e.g., an expression estimated value) and their relationships, especially in the context of continuously changing platforms and evolving biological knowledge. While such problems are encountered across all data management areas, from data generation through data collection and integration to data analysis, the solutions require domain specific knowledge and extensive data definition and curation work, with data management providing only the framework (e.g., controlled vocabularies, ontologies) to address these problems.

In an industry setting, solutions to data management challenges need to be considered in terms of complexity, cost, robustness, performance and other user and product specific requirements. Devising effective solutions for biological data management problems requires thorough understanding of the biological ap-

Data Management Systems, Gene Logic, Inc., Berkeley, California.

plication, the data management field, and the overall context in which the problems are considered. Inadequate understanding of the biological application and of data management technology and practices seem to present more problems than the limitations of existing data management technology in supporting biological data specific structures or queries.

REFERENCES

CAMBRIDGE HEALTHTECH INSTITUTE. (2001). Bioinformatics: getting results in the era of high-throughput genomics. Cambridge Healthtech Institute Report **9.**

GENE LOGIC PRODUCTS. (2003). Available: www.genelogic.com/products.htm.

JAGADISH, H.V., and OLKEN, F. (2003). NSF workshop proposal. Available: http://pueblo.lbl.gov/~olken/wdmbio/wsproposal/1.htm.

OZ. (2003). Available: http://oz.berkeley.edu/users/terry/zarray/Html/index/html.

Address reprint requests to:
Dr. Victor M. Markowitz
Data Management Systems
Gene Logic, Inc.
2001 Center Street
Berkeley, CA 94704

E-mail: markowitz@genelogic.com