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

A Plea for Normalization of Biosciences Information

SHALOM TSUR

THE SCOPE AND REQUIREMENTS of data management systems for commercial purposes and for scientific discovery in the biosciences have rapidly diverged over the last decade. While a detailed comparison would be out of place here, some of the salient features of these systems are worth mentioning. Commercial data management systems are deployed in a relatively stable world, in which the universe of discourse and the basic processes change slowly. Concepts such as "Client," or "Invoice," or "Fulfilling an Order" have an undisputed meaning, have been around for a long time and are unlikely to change in the foreseeable future. Innovation in these systems is driven by the relentless need to "do more for less," to improve efficiency and to extract more performance out of existing organizations by removing as much of the human factor out of the loop as is possible. Results in this world can and are usually measured in well-defined monetary terms. In contrast, data management systems for discovery in the biosciences are deployed in a rapidly changing universe of discourse, which is fuelled by the discovery process itself. The biosciences community is fragmented into subgroups concentrating on different aspects of the discovery process, covering as many as 10 orders of magnitude of size and time, from molecular dynamics to genomics, via proteomics to functional interpretation, to regulatory networks and systems biology. Until recently, each of these subgroups developed their own vocabulary and largely focused on the collection and interpretation of their own data. However, there is a growing consensus that the integration and global sharing of this large variety of information sources will accelerate the overall discovery process and hence, will be of benefit to the community as a whole. Note however that unlike information integration projects the commercial world, no clear and measurable objectives have been set in this respect. An additional dimension in the complexity of integrating biosciences data stems from the desire to combine these with phenotypic data, obtained via the clinical practice, and mine the integrated result for such knowledge as biomarkers for certain diseases (Tsur, 2000).

"DUAL-PURPOSE" TECHNOLOGIES

The information infrastructure, required to support the discovery process in biosciences is complex and utilizes a range of computational technologies, including data management and database systems, schema and information integration, semantic organization and ontologies as well as derivation algorithms such as BLAST (Altschul et al., 1990), data mining methods, data reduction methods and others. The origins of these infrastructure components are varied: some exist as commercial products, some are the results of database systems research in academia but the large majority have been homegrown, in the sense that they were developed in an ad-hoc fashion, in response to the needs dictated by specific research projects. Consequently, the solutions tend to be focused and do not easily lend themselves to more generic applications. By and large, data sources tend to be kept as flat files with a minimum or no descriptive data at all. There is a proliferation of data formats, which makes the exchange of data a complex and costly problem. Because of the inherent complexity and variability of biological data, the adoption of existing commercial so-

Real-Time Enterprise Group, Mountain View, California.

lutions such as relational database technology is problematic. In a commercial world with static or slowly changing schemas, relational solutions work well. In the highly dynamic world of biosciences, the permanent adaptation of schemas to the ever-changing state of knowledge makes the maintenance of relational information sources a burden that consumes more and more resources at the expense of other useful development in support of the discovery process. The present practice is often to maintain information in spreadsheets. While this "solution" offers some added flexibility, it simply shifts the burden—the maintenance of a large and rapidly growing number of spreadsheets becomes the new maintenance problem. The promise of object-oriented database technology in this domain must still be demonstrated. So far, this technology is not widely used and it is not clear whether it can meet the heavy performance requirements that stem from the need to handle the vast volumes of data that are generated.

We have thus a situation where on one hand, technology that is developed by the biosciences community itself cannot be widely deployed and adapted to more generic uses. Even if it could, it would be unrealistic to expect that a research community would be able to develop the technology into fully supported software products and, on its own, offer a level of service that would be expected of a commercial IT vendor. On the other hand, technology developed by IT vendors for commercial purposes, does not meet the requirements of the biosciences community. The notion of *Dual-Purpose Technology* pertains to generic technology that (*i*) was developed for commercial purposes and hence, is maintained, marketed and further developed by its vendors for their business purpose and (*ii*) can be used or adapted to meet the requirements of the biosciences community. Note that in itself, the biosciences community is not yet a sufficient source of potential revenue for the IT vendors to warrant special-purpose development and thus, dual-purpose technology would thus represent the best of both worlds.

The exact requirements of this dual-purpose technology are far from clear at this time. For example, to adapt relational database technology to dual-use would require the utilization of the time-proven commercial methods for query computation and optimization; indexing, data recovery, security and access control but in addition, would require advanced schema capabilities for the seamless integration of meta data, integration with ontologies and support of advanced data models such as graphs and the object-oriented models used in such information integration systems as Kleisli (Chung and Wong, 1999; Wong, 2000), K2 (Crabtree et al., 2003; Davison et al., 2001), TAMBIS (Patton et al., 1999; Stevens et al., 2000), DiscoveryLink (Haas et al., 2001), and others.

STANDARDIZATION OF TECHNOLOGY

To achieve the dual-use objective, it is instructive to look again, at the commercial world and borrow a page from the World Wide Web Consortium (*W3C*) book. This vendor-neutral, non-profit organization plays a major role in promoting standards for interoperability over the web by a well-established process, in which interested parties can contribute and reach consensus with respect to new technologies. Some of its notable success stories are in the promotion of XML as the standard of information exchange. A similar process can be used to promote and set standards for dual technology; either under the direct auspices of W3C, or by an organization that focuses more closely on biosciences such as *I3C,* which at this stage concentrates already on standards for data exchange in the biosciences. The existence of well-established standards would promote the deployment and use of these technologies and hence, would create a significant incentive for vendors to develop these technologies as part of their commercial offerings.

The utilization of already established standards in the commercial domain could be extended beyond those pertaining directly to database technology. For example, achieving application interoperability by the use of web services is a current topic of research and the objective is to rely on XML-based standards. To this end, various proposals are in different stages of the approval process and some are offered as products by vendors: *SOAP* an application-to-application message protocol, *WSDL* a web-services description language, and *UDDI* a yellow pages system for the posting and subscription of web services. Other standards to deal with business process protocols are in advanced stages of research. The biosciences community could immensely benefit from the adoption of these standards for its own needs e.g., the numerous data

sources in existence could be advertised and offered via standard WSDL-based interfaces and as such, would significantly reduce the cost of data exchange. Adding more semantic information could enhance the value of these sources; again, reliance on emerging standards for semantic exchange such as the Resource Description Framework *RDF* and the semantic web would be a major benefit. Another component that could be added via standard services is annotation, curation and provenance information on the data, which is increasingly necessary to build and maintain trust. Lastly, work is ongoing to use XQuery as a tool for information integration in web services (Andrade et al., 2003). The results would be directly applicable to web-services based bioscience data.

SUMMARY

In this short paper the author argues for the adoption of standards in the biosciences domain and following an evolutionary path that has been proved to be so successful in other domains. To mention two examples: the trigger event in the explosive expansion of the world wide web was the adoption of TCP/IP as a standard communication protocol instead of the myriad of incompatible protocols that were used before; the adoption of standards for the trade of financial instruments, such as option contracts, instead of the private and incompatible contracts used by banks before, created an orders of magnitude increase in the volume of business on financial markets. Other examples abound. The opinions offered in this paper are partly based on the experiences of the author as the director of informatics at SurroMed, Inc. and partly on his current research interests. They can best be summarized as a short agenda for research:

- Research issues in the creation of dual-use technology for information management in the biosciences. Create dual-use standards for database management technology and promote these via a standards organization such as W3C.
- Create prototypes to demonstrate the feasibility and utility of these standards.
- Adapt web-services standards to the biosciences domain.
- Use XQuery as a tool for information integration in the biosciences domain.
- Create a test-bed and benchmarks for the comparison of algorithms used in biosciences.

The last point stands alone but seems to be of increasing importance given the proliferation of analysis methods such as BLAST for sequence comparisons, in which the assumptions made in the various implementation of the algorithm and hence the consequences, are not explicit. The use of benchmarks is a long-standing tradition in database performance research.

REFERENCES

ANDRADE, J., DRALUK, V., and TSUR, S. (2001). "XQuery as a tool for liquid data integration—some design considerations. BEA Technical Report, 2001.

ALTSCHUL, S.F., GISH, W., MILLER, W., et al. (1990). Basic local alignment search tool. Journal of Molecular Biology, **215,** 403–410.

CHUNG, S.Y., and WONG, L. (1999). Kleisli: a new tool for data integration in biology. Trends in Biotechnology **17,** 351–355.

CRABTREE, J., HARKER, S., and TANNEN, V. (2003). The information integration system K2. Available: http://db.cis.upenn.edu/K2/K2.doc.

DAVISON, S.B., CRABTREE, J., BRUNK, B.P., et al. (2001). K2/Kleisli and GUS: experiments in integrated access to genomic data sources. IBM Systems Journal **40,** 489–511.

HAAS, L.M., SCHWARTZ, P.M., KODALI, P., et al. (2001). DiscoveryLink: a system for integrated access to life sciences discovery. IBM Systems Journal **40,** 489–511.

PATTON, N.W., STEVENS, R., BAKER, P., et al. (1999). Query processing in the TAMBIS bioinformatics source integration system. Presented at the 11th International Conference on Scientific and Statistical Database Management.

STEVENS, R., BAKER, P., BECHHOFER, S., et al. (2000). TAMBIS: transparent access to multiple bioinformatics information sources. Bioinformatics **16,** 184–186.

TSUR, S. (2000). Data mining in the bioinformatics domain. Presented at the 26th International Conference on Very Large Databases.
WONG, L. (2000). Kleisli, a functional query system. Journal of Functional Programming **10,** 19–56.

Address reprint requests to:
Dr. Shalom Tsur
Real-Time Enterprise Group
1076 El Monte Avenue
Mountain View, CA 94040

E-mail: tsur@pacbell.net