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

Whither Biological Database Research?

TONI KAZIC,[1] ED COE,[2] MARY POLACCO,[2] and CHI-REN SHYU[1]

ABSTRACT

We consider how the landscape of biological databases may evolve in the future, and what research is needed to realize this evolution. We suggest today's dispersal of diverse resources will only increase as the number and size of those resources, driving the need for semantic interoperability even more strongly. Because the complexity of the questions biologists want answered automatically continues to rapidly escalate, we will need to draw upon high-performance computing resources such as the GRID to process complex queries. Finally, we still need data, and our ways of acquiring and curating data must improve by orders of magnitude.

INTRODUCTION

MODERN BIOLOGICAL DATABASES are in an enviable position: everybody wants them and wants them "big"—complete, accurate, current, fast, easy to use, and permanent. The difficulty in meeting these *desiderata* is scale. Today, a database that displayed all these properties would have very large curatorial and programming staffs; complete control over its semantics, usage, and schema, and those of any other databases with which it interacts; and an inexhaustible supply of money. One might trade smaller staffs and less money for narrower domains, less accuracy and currency, more semantic confusion, less automatability, and shorter database lifetimes. None of these trades answers the very real needs of modern biological research, and all result in more work for scientists, curators, technical staff, and program directors.

The blunt fact is that the way we make, use, and maintain databases today simply does not scale by any measure: the quality, quantity, and structure of data; our ability to synthesize, interpret, and model with data; and the human and financial resources needed to produce the databases. Adding 10 times more data can take twice, or 20 times, as many curators, depending on how the data are obtained and vetted and how they relate epistemologically to those already in a database. Incorporating more complex relationships among data can take disproportionate effort, depending on the existing domain model, the nature of the new relationships, and how the database is implemented. Such structural changes in domain models can easily become the topic of irritation or negotiations to all the other databases that use information from the metamorphosing database. Adding 10 times the links can easily take more than 10 times the programming and curatorial staff, depending on the granularity of the cross-references. There is very little automated exchange of information among different databases or between them and computational resources apart from curatorial scripts. One upshot of this is restorage and redistribution of the same data by different databases, in different contexts, and with different algorithmic or informational trimmings.

[1]Department of Computer Engineering and Computer Science, and [2]USDA and Department of Agronomy, University of Missouri, Columbia, Missouri.

Problems of scale also manifest themselves as problems in interacting with the data, worsening as the amount and structural complexity of the data increase. Everyone complains about interfaces, but there is little agreement on what characteristics a good interface has, how many different types of good interfaces there might be for particular sorts of data, or how to measure interface quality. There is no user customization of what one might call fundamental interface media (would you like that in text or pictures or sound?). Interface media themselves haven't strayed far from the illustrated magazines of the Victorian era; the main addition has been very simple and uncontrollable animations. (Clicking links doesn't differ from spreading many things out on the library table and looking among them at once, except to condense the table into an electronic screen and convert a parallel process into a slow serial one.) Adding HTML links produces a geometric increase in the number of clicks, since each link usually references at least two others not already seen. Redistribution of data by different databases makes it very hard for users to discern the origins of the data and which data are duplicated; and even if algorithmically derived data are annotated as such, the annotations do not indicate which algorithm or under what conditions it was run. Some of that information may be in a paper that may be referenced in the database, but even then is often not associated with each entry.

Contemporary armchair economics suggests that many instances of poor scaling are fixed by automation and decentralization. Perhaps these are good things, but implementing them in the context of scientific databases is much, much harder than in business (where it is very nontrivial). The epistemology of modern science is far more complex than (even creative) accounting, fewer sources are electronic, and most data reporting is still by unstructured paper-based text. Moreover, the cost pressures are very stringent. The consensus seems to be that science is best served by making fundamental, and as much other as possible, data and algorithms freely available. Increasingly, those who deal in electronic materials are being held to the same standard as those who distribute biological ones. We strongly agree this *ethos* is best for science, but it does mandate that authors of code and data be productive in an economic sense.

TRENDS AND A FUTURE

Given these jostling realities, what future can one envision for biological databases? Current trends offer some interesting glimpses. By way of example, we suggest the following:

1. Monolithic databases will gradually give way to a set of interoperable databases, sharply reducing restorage of data.
2. The origins and quality of data will be better annotated.
3. Large databases will increase in size.
4. More databases will become large, and many more databases, curating specialized portions or views of the literature, or computations and models, will arise.
5. Some of the literature will exist only in databases.
6. Curators will curate not only their own data, but help others curate related data, and curate not just data but relationships, semantics, nomenclature, ontologies, computations, and models.
7. Every active scientist will become a part-time database depositor and curator, at least for his or her results that don't fit in the paper but must be available for good scholarship or publication.
8. Population of databases will begin with either natural language processing of full-text articles to extract information; or automated submission of structured data; or community deposit and review *via* web sites, all with checking and elaboration by human curators.
9. Interfaces to individual databases will give way to automated portals to many interoperable databases and algorithms, and the portals will synthesize the information and computations into evaluative reviews.
10. Users' robots will continually surf, generating customized notices and interpretations of changes in the data.
11. We will still write and review (and, one hopes, fund) proposals, but their emphases will expand to include the development of new technologies for interoperability, database population, and interaction;

the curation of semantic and epistemological systems; and the addition of different types of data, building domain models that more accurately mimic biological processes, and using those for simulation.

How do we get there? We believe research in the following areas is extremely important in building some version of this future. There are immediate applications to many problems, such as those arising from microarrays and proteomics.

1. **Improved methods for database population and maintenance.** One example is methods to mine semantics, relationships, reasoning, and usage automatically from text, glossaries, and databases: for example, natural language processing to construct, populate, and monitor the usage of ontologies and semantics (Raychaudhuri, 2002; Krauthammer, 2000); and automated reasoning to test the fidelity of code reifying definitions to their natural language versions. Another is distributed curation (Jiang, 2000; Alliance for Cellular Signalling, 2002). While databases can benefit from the community's scientific expertise and help amortize the cost, unconstrained participation is a recipe for boredom, semantic confusion, inadvertant errors, and usage creep (*e.g.*, "gene" becoming "ORF"). A third is direct capture of raw data from instruments followed by automated processing. There has been extensive effort on the second part, but the first requires significant advances in laboratory information management systems (LIMS). These are almost invariably hand-crafted commercial products: what we need are open-source or public domain modules that can plug and play together, raising again the usual problems of standardization of devices and reusable code. Assessment of the accuracy, precision, and sociology of this research is essential.

2. **Semantic interoperability of databases.** Sections of the computer science and commercial communities are heavily invested in research on distribution methods of various kinds (*e.g.*, SOAP and HANDLE; see Cerami, 2002; Corporation for National Research Initiatives, 2002), but as a community we still have far to go in defining the semantics of our data and databases so information can be automatically retrieved with high precision and reliability. In our view, this is by far the hardest part of the interoperability puzzle. How can a database use an ontology to make its data interoperable (Ashburner, 2000)? How can this be done efficiently and cost-effectively, without restructuring the entire database and recurating the data? What granularities of notions are needed? How well do formal languages specify the complex semantics of real biology (Wiederhold, 1999; Kazic, 2000a, 2000b)? How can they be efficiently used to produce interoperability? Can they become biologist-friendly? What other methods could be devised?

3. **Databases.** The community still needs databases: biochemistry; physiology; development; anatomy; phylogeny; morphology; ecologies; phenotypes; diseases of species, landscapes, and ecosystems; laboratory procedures; and computational methods. The results of computations and simulations must be easily available, and specify exactly how the data were obtained—much more complex datatypes than we have heretofore considered. Each of these domains offers an opportunity to discover more efficient methods of data acquisition, curation, and maintenance, both generally and in their particular contexts. Moreover, the databases are crucial as platforms in developing complex queries, simulations, and visualizations—these need large, complicated datasets—and for discovery-based research.

4. **Complex queries.** What kinds of queries are most needed, and how can they execute efficiently over disparate resources? We define a complex query as one that involves operations other than mere retrieval across distributed, independent databases and computational resources. Examples of such operations include feature recognition of images, substructure search among molecular structures, linkage analysis, superposition of folded proteins, and exploration of large networks of reactions (Bayada, 1992; Shyu, 2003; Rost, 2000; Mummaneni, 2002). Clearly, progress will depend on semantic interoperability and the ability to mobilize distant computational resources efficiently. But we also need explicit efforts to identify different types of complex queries and determine how they can be more optimally executed in an inherently transitory environment.

5. **Mobilizing more powerful machinery.** There is little use of high-performance computing and communications by the general biological community, except for certain groups of structural biochemists and computational biologists. Most biologists seem unaware of massively parallel computers, grid computing, terabyte storage, and tools for videoconferencing and remote collaboration (Buyya, 2002). Many database queries are excellent examples of pleasantly parallel calculations, and making and applying simple improvements in distributed technology would improve both access times (especially important for dynami-

cally eliciting information from interoperable databases) and the complexity of queries that can be offered (including those that would benefit from true parallelism; Condor Team, 1986; Globus Team, 2003). Similarly, interactive acquisition and analysis of data by geographically separated users, and storage on remote data farms, are not yet widely used and offer significant opportunities to improve science. But we lack the tools and biological applications for the emerging grid environment of parallel computing and communications, and most of us have not yet joined forces with those developing these new technologies.

6. **Support for complex and combined models.** If one wants to predict the effect of a change in a protein's tertiary structure on drought resistance in maize, one has to combine many different ways of thinking and models that differ in scale and structure (National Science Foundation, 1999). We need to learn how we can combine information and computations from different types of databases and models, not just from the standpoint of interoperability but scientifically. The answers for biology may be very different from those for physics, since biology is so much more strongly data-driven.

7. **Use of multidimensional, multivariate, and multistructure data.** This isn't just visualization, though that's important, unsolved, and badly needed, or dimension reduction (Swayne, 2002); but also asking what characteristics different types of coupled data might share and how to express those characteristics algorithmically. Do big complicated biological datasets have unifying principles that help us grasp and manipulate them? How can we use those principles to support the generation and testing of hypotheses?

8. **Understanding the mathematics and statistics of biology.** What's math got to do with databases? Simple: mathematics and statistics are well-defined techniques for recognizing, expressing, and evaluating consistent patterns in infinite sets of objects. From understanding the deep structure of biology, to determining how to do the microarray experiment, to simulating biological systems, to evaluating how well complex data are mined from databases, all these and more require the closer integration of mathematics and statistics into biology. This integration is highly likely to produce new mathematics and statistics, and it's guaranteed to produce new biology.

9. **Doing this humanely.** No amount of automation will replace human beings, but clumsy and belligerent automation will alienate them and suppress their creativity.

ACKNOWLEDGMENTS

We are grateful to the NIH (T.K.: GM56529; C.R.S.: CA94441) and NSF (M.P. and E.C.: DBI-9872655) for supporting our research.

REFERENCES

ALLIANCE FOR CELLULAR SIGNALLING. (2002). Available: www.afcs.org.

ASHBURNER, M., BALL, C.A., BLAKE, J.A., et al. (2000). Gene ontology: tool for the unification of biology. Nat. Genet. **25,** 25–29.

BAYADA, D.M., SIMPSON, R.W., JOHNSON, A.P., et al. (1992). An algorithm for the multiple common subgraph problem. J. Chem. Info. Comp. Sci. **32,** 680–685.

BUYYA, R., ed. (2002). GRID Computing Info Centre. Available: gridcomputing.com.

CERAMI, E. (2002). *Web Services Essentials.* (O'Reilly and Associates, Inc., Sebastopol, CA).

CONDOR TEAM. (1986). Condor. Available: www.cs.wisc.edu/condor.

CORPORATION FOR NATIONAL RESEARCH INITIATIVES. (2002). HANDLE system introduction. Available: www.handle.net/introduction.html.

GLOBUS TEAM. (2003). The Globus Project. Available: www.globus.org.

JIANG, J., SANGHI, G., KUTIKKAD, G., et al. (2002). The *Agora.* Available: www.the-agora.org.

KAZIC, T. (2000a). Semiotes—a semantics for sharing. Bioinformatics **16,** 1129–1144. Available: www.biochem-info.org/repository/semiotes.ps.

KAZIC, T., JIANG, J., KUTIKKAD, G., et al. (2000b). Glossa Semiotes. Available: www.the-agora.org/glossa/semiote_list.ps.

KRAUTHAMMER, M., RZHETSKY, A., MOROZOV, P., et al. (2000). Using BLAST for identifying gene and protein names in journal articles. Gene **259,** 245–252.

MUMMANENI, A. and KAZIC, T. (2002). The architectural dynamics of cellular biochemistry and molecular biology. In: Callaos, N., ed., *Proceedings of the Sixth World Multiconference on Systemics, Cybernetics and Informatics (SCI 2002).* Available: www.biocheminfo.org/repository/sci2002.ps.

NATIONAL SCIENCE FOUNDATION. (1999). Knowledge and distributed intelligence. Available: www.ehr.nsf.gov/kdi/default.htm.

RAYCHAUDHURI, S., SCHUTZE, H., and ALTMAN, R.B. (2002). Using text analysis to identify functionally coherent gene groups. Genome Res. **12,** 1582–1590.

ROST, B., EYRICH, V., LIU, J., et al. (2000). PredictProtein. Available: http://cubic.bioc.columbiaedu/predictprotein/.

SHYU, C.R., KAK, C., PAVLOPOULOU, A.C., et al. (2002). Using human perceptual categories for content-based retrieval from a medical image database. Comp. Vis. Image Understanding **88,** 119–151.

SWAYNE, D.F., COOK, D., BUJA, A., et al. (2002). ggobi Manual. Available: www.ggobi.org/manual.pdf.

WIEDERHOLD, G., and JANNINK, J. (1999). Composing diverse ontologies. Available: www-db.stanford.edu/SKC/publications/ifip99.html.

Address reprint requests to:
Dr. Toni Kazic
Department of Computer Engineering and Computer Science
143A Engineering Building West
University of Missouri
Columbia, MO 65201

E-mail: toni@athe.cecs.missouri.edu