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

Metadata and Semantics: A Computational Challenge for Molecular Biology

JUDY BAYARD CUSHING

MOLECULAR BIOLOGY APPLICATIONS, like those of other scientific domains, need to store, mine, and view large amounts of specialized quantitative information. The research challenges for data management are many. A brief informal survey of former collaborators indicates that significant progress has been made in shared domain models; tools for sequence production, curation, publication, and visualization; and UI's and API's. That said, many of the research challenges from then, are with us still. Why? Because these are tough problems. To borrow a term from Fred Brooks (who borrowed it from Aristotle), many of the problems we faced 10 years ago were essential to the understanding of the molecular biology and to the wider application of gene and protein sequence data beyond their specific domains. In other words, those problems are problems of semantics. Many of the incidental or accidental problems have been solved, and progress on the essential challenges has been made. If anything, however, today's challenges could be more complex, since the success of the last decade has bred desire for increased functionality and for using valuable sequence information across many disciplines (cell biology, organism biology, paleobotany, ecology, medical research, medical practice). Thus, molecular biology data promise even deeper understanding of life's mysteries, and with that promise comes a need for even better semantics, and the increased challenge.

Public databases such as GenBank, PDB, EMBL, JIPID, SwissProt, etc., make millions of genetic sequences available to molecular biologists (and others), and industry and university laboratories maintain their own private stores of (many more) sequences. In short, there is lots of information, some of it duplicate for everyone, some of it duplicate for some purposes but not for others, that people use for different purposes. Several specific domain-specific information technology problems come to mind: (1) a need for high-level, domain-specialized common interfaces and query languages to exploit heterogeneous databases, (2) the ability for individuals to filter and annotate the sequences, and to share some of those annotations with close colleagues, or even the pubic at large, (3) ability to integrate and track inputs and results from numerous computational biology programs, (4) the need for molecular biology communities, the T4 phage community, to have a more specific view of T4 sequences than the general molecular biology community; (5) the need for non-molecular biologists to have a view of the data that specializes it for their problem domain, though such views must in some way be tied to the "molecular biology" view so that meaningful collaboration can occur. Regarding this fifth problem, with respect to the scientific domain in which I am now working, tree physiologists collaborate with molecular biologists to identify genetic differences among trees that "breathe" or hold water differently. I suggest as a research challenge that we do not yet know enough of how such "new" wildly interdisciplinary collaborations will use data, and that some resources should be set aside to build small experimental prototypes.

A problem not mentioned above, but which is rampant throughout the sciences is how to deal with error (especially as one scales up or down), and missing data. This is not to say that scientists need to work only with data that are "100%" certain and complete, but that they know which data are not so (and by how much). Those working with molecular biology applications need to have a very high degree of trust in, or at least understanding of, the quality and provenance, of the data underlying that information.

Department of Computer Science, The Evergreen State College, Olympia, Washington.

How does all this translate into challenges for computer science research and development? Below are some ideas, but the problem of metadata (and the implications it has for increased understanding of how to process semantics of the data with the data themselves) is perhaps the most intriguing:

1. Improved and extremely flexible metadata facilities for DBMSs—that allow one to document data with more than the metadata required for managing the data in a DBMS. Why can a domain researcher not select what metadata would be visible or tracked across applications, or even extend standard metadata schema as desired? Just as ecologists cannot yet overlay multiple taxonomic information simultaneously over ecological field data, I believe that molecular biologists still lack the ability to organize result items from sequence comparisons into multiple clusters that can be marked, named, annotated and manipulated.
2. Better tracking of derived data and data products, as they pass through filters and sieves.
3. Better computational workbenches, all with a better understanding of how to describe programs and how programs should automatically describe their products.

ACKNOWLEDGMENTS

I gratefully acknowledge the help of many molecular biologists in helping me understand data issues, including but hardly limited to: Tim Hunkapiller, Betty Kutter, Tom Marr, Dave Yee, as well as ideas from ecologists Barbara Bond and Nalini Nadkarni. I also acknowledge funding from the National Science Foundation IRI-9224726 and BIO 9975510, which funded work in molecular biology and ecology databases, respectively.

REFERENCE

BROOKS, F. *No Silver Bullet—Essence and Accident in Software Engineering. Mythical Man Month.* (1995). (Addison Wesley, Boston, MA).

Address reprint requests to:
Dr. Judy Bayard Cushing
Department of Computer Science
The Evergreen State College
2700 Evergreen Parkway
Olympia, WA 98505-0002

E-mail: judyc@evergreen.edu