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

Common Objects: Think Global, Act Local

CHRISTIAN J. STOECKERT, JR.

THE ADAGE "Think global, act local," exhorts individuals to change the world starting with actions within ones local community. The research challenges for data management for molecular and cell biology can be similarly addressed. Local communities have their own cultures, needs, and priorities. This is as true for communities involved in the various knowledge domains of molecular and cell biology (e.g., genomics, structural biology, molecular phylogeny, pharmacology) as it is for geopolitical communities. In the global research community encompassing the many molecular and cell biology knowledge domains we need to communicate information in order to integrate data on a gene's chromosomal location, the structure of the protein that the gene encodes, and the small molecules that bind to that protein. Communication requires a common standard not only at the syntactic level but also at the semantic level. A top down approach of imposing a single standard for all molecular and cell biology knowledge domains is impractical if not impossible to enforce. A bottom up approach of each knowledge domain generating its own standards would take advantage of the work already done in those local communities and best utilize the expertise within those communities. Standards for protein structure have been generated through the efforts of the Research Collaboratory for Structural Bioinformatics (RCSB) and the Protein Data Bank (Westbrook, 2002). Standards for assigning and describing molecular function, biological processes, and cellular component have been established through the Gene Ontology (GO) Consortium (Ashburner, 2000). Microarray standards have been established through the efforts of the Microarray Gene Expression Data (MGED) Society (Brazma, 2001; Spellman, 2002). While these and other standards exist, certainly they are not available for all molecular and cell biology knowledge domains.

The challenges for a bottom-up approach to generating a common standard for molecular and cell biology data management are to first establish standards in the various communities or knowledge domains comprising this broadly-defined area of research and secondly to have these different standards work together. These challenges are related in that the standards that are developed at the community or local level must be compatible at the global level. To think global in this sense is to have local standards efforts (within knowledge domains) look to and learn from existing standards efforts. The existing standards efforts to consider should not be restricted to molecular and cell biology (e.g., GO) but should also include standards effort in the mediums used by molecular and cell biologists, computational biologists and bioinformaticists. Examples include the standards efforts in computer industry specifications by the Object Management Group (OMG, www.omg.com/), in bioinformatics programming by the Open Bioinformatics Foundation (http://open-bio.org/), and in web technologies by the World Wide Web Consortium (W3C, www.w3.org/). The challenge of building local data standards thus includes consciousness raising of other efforts and their relevance.

If the "think global, act local" approach is taken, then ultimately data managers for molecular and cell biology will share common objects. Objects are defined here as instances of some concept (class). Common objects are ones that can be shared and understood between data systems. Despite the popularity of relational data management systems and indexed text files, most computational biology and bioinformatics

Department of Genetics and Center for Bioinformatics, University of Pennsylvania, Philadelphia, Pennsylvania.

systems have an object layer and are capable of generating some form of eXtensible Markup Language (XML, www.w3.org/XML/). Objects provide an abstraction that permits heterogeneity in schema or data representation for individual data systems. Thus data in legacy systems can be mapped to common objects for exchange with other systems such as is the case for the MicroArray Gene Expression (MAGE) object model (Spellman, 2002). A standard representation for objects exists in the Unified Modeling Language (UML, www.omg.org/uml/). Standards for exchanging objects exist such as the Common Object Request Broker Architecture (CORBA) by the OMG (www.omg.org/gettingstarted/corbafaq.htm) and Simple Object Access Protocol/XML protocol (SOAP/XMLP) by the W3C (www.w3.org/2000/xp/Group/). Thus, the means for sharing objects exist; the challenge is to create standard ones. There is certainly overlap between different molecular and cell biology domains and an associated challenge to encouraging the generation of domain standards is to minimize the overlap in standard objects. The Global Open Biological Ontologies (GOBO) is one effort to encourage common data format usage and minimal overlap of representation (www.geneontology.org/doc/gobo.html). GOBO is also requiring open or freely available standards which is a practical necessity for common objects.

Abstraction of molecular and cell biology data to commonly agreed upon objects that can be exchanged though a popular language such as XML is the central point of this paper. Others may feel that the key challenges are the limitations in expressivity of current data types that can be exchanged. Whether new data types or new domain-specific standards in data representation are developed, it should be kept in mind that these local actions should still allow data systems to think globally (i.e., across domains with common objects).

REFERENCES

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

BRAZMA, A., HINGAMP, P., QUACKENBUSH, J., et al. (2001). Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat. Genet. **29,** 365–371.

SPELLMAN, P.T., MILLER, M., STEWART, J., et al. (2002). Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol. **3,** RESEARCH0046.

WESTBROOK, J., FENG, Z., JAIN, S, et al. (2002). The Protein Data Bank: unifying the archive. Nucleic Acids Res. **30,** 245–248.

Address reprint requests to:
Dr. Christian J. Stoeckert, Jr.
Department of Genetics and Center for Bioinformatics
University of Pennsylvania
1415 Blockley Hall
423 Guardian Drive
Philadelphia, PA 19104

E-mail: stoeckrt@pcbi.upenn.edu