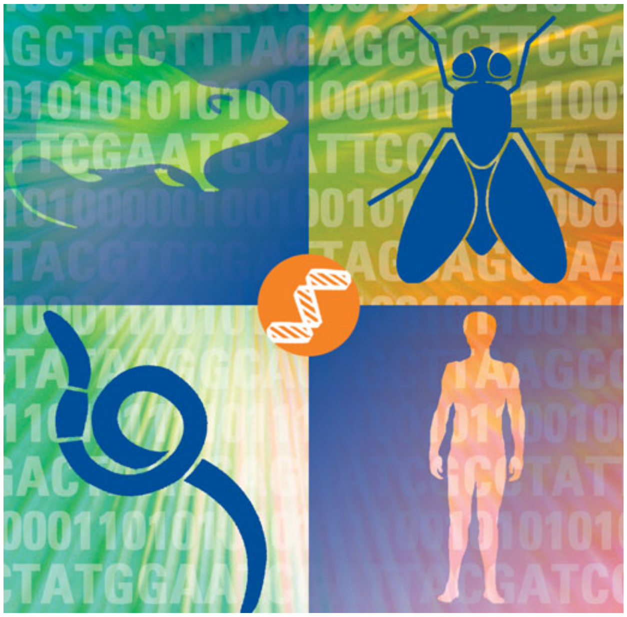
Bioinformatics: embracing genetic ontology

The problem of biological information lies not just in untidy databases but in the language of research.

BY MARK S. LESNEY

Many of the problems involved in dealing with the bioinformatics glut, especially with databases containing incompatible formats, were discussed in this department in April ("A league of IT's own?", p 51). But much of the problem is not in the database systems or the means of presentation, but in the science itself—how scientists do, think about, define, and report their own research. The proliferation of terminologies, taxonomies, genes, proteins, and 'omes, as well as the chaotic fragmentation of specialization, is creating what many see as an ontological crisis in molecular biology that has some of its greatest impact in the bioinformatics arena. In a very real sense, when computer scientists try to address what biologists do when they study and write about biology, they are confronted with the equivalent to the problem of herding cats. To quote Michael Ashburner, joint head of the European Bioinformatics Institute, "Biologists would rather share their toothbrush than share a gene name" (*1*).



Bioinformatics originated primarily as a means of automating huge volumes of sequence information that was relatively homogeneous and defined. The field took what could be considered an expert-system approach to the material—mimicking how an individual expert would analyze the information if she only had the time and patience, and could visit a hundred labs across the country simultaneously. From the start, this involved the need to annotate the databases to identify potential gene families and to perform comparative genomics with the sequences of other species. Predicting and comparing protein products were of course relatively easy functions and were rapidly introduced.

But even in these early stages, ontological cracks began to appear in the well-built bioinformatics structures, and now they are growing into huge stress cracks as the move from genomics to physiomics proceeds apace. Not only gene names, but the names of enzymes, structural proteins, diseases, and physiological pathways enter the babble of voices, differently defined across species and disciplines. Hence the development of the concept of genetic ontology—and the move toward software to deal with it.

Up with ontology

The concept of ontology with regard to genetics vocabularies is a straightforward one. As defined by the Stanford Knowledge Systems Laboratory, "An ontology is an explicit specification of some topic . . . a formal and declarative representation which includes the vocabulary (or names) for referring to terms in that subject area and the logical statements that describe what the terms are, how they are related to each other, and how they can or cannot be related to one another."

Sharing such an ontology, identifying terms and fixing definitions, is one of the first steps in the maturation of a scientific discipline. Consider the chaos in chemistry before the naming of the elements and molecules moved beyond alchemical descriptive forms, which had as much to do with astrology as science. Or, perhaps most pertinent, in biological taxonomy before Linnaeus set up standards and a language for defining genus and species—the familial relationships of organismic life.

The consortium

The Gene Ontology (GO) Consortium evolved in 1998 from the realization of such shared needs and interests among the scientists (referred to as curators) in charge of the three model organism databases: the *Saccharomyces* Genome Database (SGD), FlyBase (the *Drosophila* genome database), and the Mouse Genome Informatics databases (MGD/GXD) (*www.informatics.jax.org/mgihome/GO/ontology.shtml*). According to the position paper written by the scientists involved, "The GO Consortium is to produce a dynamic controlled vocabulary that can be applied to all eukaryotes, even as knowledge of gene and protein roles in cells is accumulating and changing" (*2*). The initial three collaborating genome organizations involved in the GO Consortium, as well as newer corporate and public collaborators, maintain links from the GO site.

The core of the GO Consortium is the maintenance of three evolving databases (described and accessible at *www.geneontology.org*) containing the accepted terminology to be used for *gene product* annotation. These databases are named *molecular*

Advertisers

MDL INFORMATION SYSTEMS, INC.
14600 Catalina Street, San Leandro, CA
94577, 510-895-1313, 510-614-3608,
www.mdl.com

INVITROGEN CORP
1600 Faraday Ave., Carlsbad, CA
92008, 760-603-7200, 760-602-6500,
www.invitrogen.com

IBC USA
One Research Drive Suite 400A,
Westborough, MA 01581, 508-616-5550,
508-616-5533, www.ibcusa.com

GILSON, INC.
3000 W. Beltline Hwy., Middleton, WI
53562-0027, 800-445-7661,
608-831-4451, www.gilson.com

function, *biological process*, and *cellular component*. The GO Consortium defines these terms as follows.

Gene product "is a physical thing", in this case either a protein or an RNA molecule. Because no one believes in the possibility of policing gene- and protein-naming conventions across all of biology, genes and gene products do not reside within the purview of the GO Consortium. They exist as items in separately maintained public and commercial databases to which the GO annotation is applied. In many, if not most, cases databases place GO identifier annotations on the genes rather than on the protein or RNA products, which is somewhat outside of the original intent of the curators, because genes often code for multiple gene products, thereby leading to potentially confusing annotation.

Molecular function "is what something does" either actually or potentially. Typically, this is a specific enzyme activity, of which an individual gene product could have several. For example, a protein could have the potential to be a fairly generalized alcohol dehydrogenase when purified, but in vivo it may only be involved in specific sugar–alcohol reactions.

Biological process "is a biological objective"; this is a more global category than that of simple molecular function, including categories on the order of "cell growth and maintenance" or "membrane transport".

Cellular component "is just that . . . but with the proviso that the component is part of some larger object" (anatomical, e.g., the nucleus or the Golgi apparatus) "or gene product group" (e.g., a ribosome or a heterodimeric protein).

Each database term has a unique GO identification number assigned to it so that there can be no confusion. Multiple GO numbers can be assigned to a single gene product. Key to all of these terms is the requirement within each for annotated documentation of the appropriate literature or authority involved in order to validate the accuracy of the claimed annotation and for cross-checking the evidence behind it (*3*).

The need for GO to constantly evolve has resulted in the need for permanent and open institutions to maintain and update the databases. GO was easiest to implement with the extant nonhuman genome databases, most of which had a single agreed-upon "home". Yet, as pointed out in some of the earliest discussions of the GO concept (*4*), the Human Genome Project is problematic for such controlled implementation of GO annotation because the massive database has no one real home and represents, in many cases, a conflict of proprietary interests.

Voluntary compliance

For the human genome, no one institution or individual organizes and enforces compliance, even at the level of "voluntary" annotation that the other databases have achieved. Yet, despite the voluntary and scattered nature of applying GO standards to the human genome, the major companies and databases still seem to be moving rapidly in that direction.

> **"Biologists would rather share their toothbrush than share a gene name."**

According to the GO Consortium, this is understandable because it is obviously in the best interests of companies and researchers to adopt uniform standards as that is what consumers, including journals and other scientists, are coming to demand and expect. To that end, GO terminology is being incorporated by a growing number of software manufacturers as either an integral component of their database annotations or an added feature to whatever unique annotation treatments or metadata they seek to provide their customers.

Among the growing number of companies committed wholly or in part to working with the GO Consortium and/or in the annotation of their proprietary databases and services are Celera, Compugen, Affymetrix, AstraZeneca, GlaxoSmithKline Pharmaceuticals, and Incyte Genomics.

Moving to markups

Annotation is such a critical issue in bioinformatics that the attempts at standardization of terminologies are widespread. Markup languages, most based on XML, are proliferating to apply these annotations easily to existing databases. Among these are the Genome Annotation Markup Elements, which provides a syntax for exchange of genomic annotation, and the XOL Ontology Exchange Language, an XML-based language for exchanging genetic ontologies.

One effort particularly pertinent to proteomics above and beyond the three GO databases is the Integrated Resource of Protein Domains and Function Sites (InterPro), an XML-based databank database which, among other things, attempts to organize domain and motif annotations (*www.ebi.ac.uk/interpro*).

But in the long run, it is unlikely that any of these approaches will be the final word in annotation. To this end, researchers are attempting to develop methodologies for automatic "re-annotation" of databases, hoping to eliminate the time-consuming step of requiring expert curator creation and modifications of new annotation concepts (*5*). To this end, curators of databases such as SGD are maintaining a library of abstracts referring to the genes and gene products in question, should subsequent rounds of annotation be required. Ultimately, it appears that, no matter the mechanics of annotation, a significant and growing portion of the bioinformatics community has made a commitment to the ontology concept. There is even a Plant Ontology Consortium (*www.bioinformatics.org/sprig/poc.html*) to deal with botanical genomics. So, as an organizing principle for the developing database Tower of Babel, the new ontological approach is here to stay.

References

(1) Pearson, H. *Nature* **2001,** *411,* 631–632.

(2) Ashburner, M.; et al. *Nature* **2000,** *25* (1), 25–29.

(3) The Gene Ontology Consortium. *Genome Res.* **2001,** *11,* 1425–1433.

(4) Gene Ontology annotation and the human genome; www.geneontology.org/minutes/20001210_Banbury.txt.

(5) Raychaudhuri, S.; et al. *Genome Res.* **2002,** *12,* 203–214.

Mark S. Lesney is a senior associate editor of *Modern Drug Discovery*. Send your comments or questions about this article to mdd@acs.org or to the Editorial Office address on page 3. ■