Ontologies for Programs, not People

by Lawrence Hunter

In a recent editorial [1], the wry and erudite Sydney Brenner expressed his disdain for recent efforts to create computational ontologies of molecular biology. In essence, Brenner argued that building a network of names of biological entities is a waste of time. It is the nucleotide sequences or amino acid conformations of these objects, not their names, "that create the processes that produce outcomes for cells, organs and organism," he says. "Very simply, the network we should be interested is not the network of names but the network of the objects themselves."

The honorable Dr. Brenner's editorial misses the point -- several points, actually. First, the essence of the Gene Ontology, of which he is specifically critical, and other knowledge-bases of molecular biology, such as EcoCyc [2] or the Unified Medical Language System [3], is not in the list of names they embody, but in the _relationships_ represented. The names are convenient symbols to which more complex statements can be attached. Without the names, it is impossible to specifically represent relationships such as "activates" or "binds to". Surely that sort of information must be the kind of thing that Brenner means when he says we are interested in the interactions among the objects themselves, rather than their names.

Second, if we are to build useful databases of the interactions that Brenner suggests ought to hold our interest, then there are significant advantages to being able to make statements about classes of genes and gene products together, using the terminology that molecular biologists are accustomed to. For example, representation of the statement "the balance between pro- and anti- apoptotic members of the BCL2 family of genes determines whether apoptosis proceeds" is straightforward if we use an ontology with an appropriate level of abstraction, and painfully difficult if we are limited to expressions of direct interactions between pairs of genes and proteins.

Third, it is important to be clear about to whom the "we" in Dr. Brenner's argument refers. Knowledge-bases are not generally used directly by an end user, but by computer programs in order to accomplish complex inference tasks. Many productive and promising approaches to bioinformatics require a computationally manipulable representation of existing biological understanding -- incomplete and incorrect as it may be -- as a vital prerequisite. For example, inference from gene expression data using Bayesian networks [4] can take advantage of online sources of information about possible probabilistic dependencies among expression levels of various genes. Knowledge-bases built from textbooks, review articles, or even the Oxford Dictionary of Molecular Biology can provide precisely this sort of computationally useful information.

Fourth, if bioinformaticians are to build useful tools for managing the ever-growing onslaught of research publications resulting from high-throughput instrumentation and exacerbated by the collapse of subdisciplinary distinctions, then they must first create computer programs that recognize references to genes, proteins and other biological entities in texts. Automatically linking references to molecular entities and processes in texts (such as Medline abstracts) to the appropriate entries in molecular databases such as Genbank can save enormous amounts of researcher time, and facilitate the kind of

biology that Brenner holds dear. Such a mapping, however, requires the presence of a well represented knowledge-base of molecular biological entities -- perhaps like the Gene Ontology.

Dr. Brenner is, of course, entitled to his opinion about the utility of efforts like the Gene Ontology and the UMLS. Perhaps he doesn't need any of the computational tools for analyzing high-throughput data in light of prior knowledge, or managing the vast scientific literature, either. However, for those of us who use bioinformatics software to advance scientific understanding, broad community efforts at knowledge representation like the Gene Ontology Consortium are invaluable.

Lawrence Hunter, PhD, is the director of the Center for Computational Pharmacology at the University of Colorado School of Medicine, and the founder of the International Society for Computational Biology.

Literature Cited

[1] Brenner, S. "Life Sentences: Ontology Recapitulates Philology" *Genome Biology* 2002 3(4):1006.1-1006.2. Also published as *The Scientist* 16[6]:12, Mar. 18, 2002

[2] Karp, PD "Pathway databases: a case study in computational symbolic theories." *Science* 2001 Sep 14;293(5537):2040-4

[3] Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The Unified Medical Language System: an informatics research collaboration. *Journal of the American Medical Informatics Association* 1998;5(1): 1-11.

[4] Segal E, Taskar B, Gasch A, Friedman N, Koller D. "Rich probabilistic models for gene expression." *Bioinformatics.* 2001;17 Suppl 1:S243-52.