Ontologies for Biologists - A Community Model for the Annotation of Genomic Data

Michael Ashburner
Department of Genetics
University of Cambridge & EMBL - EBI
Hinxton, Cambridge, England
m.ashburner@gen.cam.ac.uk

The representation of biological knowledge in databases is a necessity for modern biomedical research. Historically, there has been very little collaboration or coordination between different database providers. Although many grandiose schemes for the "integration" of biological databases have been proposed over the years, none have been practical to the point of implementation. Yet the need for integration remains, as many biologists, both those at the bench and those who analyse data computationally, wish to integrate data from a diversity of sources. The Gene Ontology Consortium (GOC) began, some seven years ago, to develop a resource that could be used by both the model organism databases (e.g. FlyBase, WormBase, Mouse Genome Database, The Arabidopsis Information Resource) and the large "horizontal" databases (e.g. UniProt, GeneDB, TIGR Gene Index) as a standard for the annotation of gene products. The GOC now maintains several structured controlled vocabularies for the annotation of gene products. The first three of these are used for the annotation of gene products with respect to these domains: their molecular function, their cellular location and the biological processes in which they are involved. This database of nearly 18,000 terms is now used for the annotation of the gene products of all of the major experimental eukaryotes and many prokaryotes.

The philosophy of the GOC is now being extended to cover further domains of biological knowledge. Under the umbrella of "obo" (open biological ontologies) structured controlled vocabularies are now available, or are being developed, for sequence annotation, anatomies and development, cells and tissues, mouse pathology and experimental treatments.

In this talk I will discuss how the concept of ontologies can be used for the intelligent design of database schema, and for the development of common tools for data exchange. I will also discuss some of the major limitations of the current models of data representation used by the GO Consortium, and proposals that will make the design of ontologies for shared use both more flexible and powerful.

URLS:

www.geneontology.org
obo.sourceforge.net