Opinion

Gene Ontology: looking backwards and forwards

Suzanna E Lewis

Address: Department of Molecular and Cell Biology, University of California, 539 Life Sciences Addition, Berkeley, CA 94720-3200, USA.
E-mail: suzi@fruitfly.org

Published: 15 December 2004

Genome ***Biology*** 2004, **6:**103

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/6/1/103

Abstract

The Gene Ontology consortium began six years ago with a group of scientists who decided to connect our data by sharing the same language for describing it. Its most significant achievement lies in uniting many independent biological database efforts into a cooperative force.

Long ago, in the pre-genome era, biological databases had to come to terms with a formidable amount of work. After Crick and Watson elucidated the structure of DNA, the field of molecular biology exploded and an ever-increasing amount of information needed to be carefully managed and organized. This was particularly true after the invention of methods to sequence DNA in the late 1970s [1,2] and, consequently, the initiation of the genome sequencing programs in the late 1980s, all of which led to an even faster acceleration of work in this field. Keeping pace with molecular developments were biological data-management efforts. These first began emerging in the 1960s when Margaret Dayhoff [3] published the *Atlas of Protein Sequence and Structure* [4], which later went online as the Protein Identification Resource (PIR [5]). More than 30 years ago, in the 1970s, the first protein-structure database, Protein Data Bank (PDB [6]), was founded [7] and the Jackson Laboratory developed the first mammalian genetics database [8]. A few years later the first depositories for nucleotide sequences were established - with the EMBL 'Data Library' [9] beginning in 1981 [10] at Heidelberg, Germany and GenBank [11] in 1982 [12] at Los Alamos, New Mexico - followed soon afterwards by the formal establishment of the PIR in 1984 [13] for proteins. By the late 1980s and 1990s biological databases were popping up everywhere: in 1986 SwissProt [14]; in 1989 *Caenorhabditis elegans* AceDB [15]; in 1991 *Arabidopsis* AtDB [16]; in 1992 [17] The Institute for Genomic Research (TIGR) [18]; in 1993 FlyBase [19]; and in 1994 [20], *Saccharomyces* Genome Database (SGD) [21]. These groups all took advantage of concurrent technological advances and pioneered the use of the internet, the worldwide web, and relational database management systems (RDBMSs) and standard query language (SQL), when these technologies first became available during the 1980s and 1990s [22-24]. Thus, many biological databases bloomed, flourished and, until the late 1990s, all of them operated primarily autonomously.

Having many independent genome databases made a large number of researchers very happy but there were shortcomings. The most important research limitation was that the full potential of these isolated datasets could not be realized until they were as integrated as possible. But there is a practical constraint: biological databases are inherently distributed because the specialized biological expertise that is required for data capture is spread around the globe at the sites where the data originate. Whatever the solution to biological integration, it would have to acknowledge that the primary sources of data are distributed investigators.

The community of biological data managers was initially very small and the pioneer database developers largely knew one another. They made many attempts to work together towards an integrated solution, either by facilitating the transfer of knowledge between databases or by merging them. The annual AceDB [15] workshops are one example of these efforts. In the early 1990s these two-week sessions brought together participants working with many organisms, such as pine trees, tomatoes, cows, flies, weeds, worms, and others. Unfortunately, AceDB was dependent upon what became outmoded technology and did not adapt to the web or RDBMSs sufficiently quickly to allow it to survive as a general solution. There were also a number of meetings organized to attempt - ultimately in vain - to design the ultimate biological database schema, such as the Meeting on the Interconnection of Molecular Biology Databases held at Clare College, Cambridge in 1995 [25]. Creating a federated system failed for reasons too numerous to list, but the biggest impediment was getting the many people involved to agree on virtually everything. It would have created a technological behemoth that would be unable to respond to new requirements when they inevitably occurred. Even small-scale collaborations between two databases failed (for example in the case of SGD [21] and the Berkeley Fly Database, a precursor of Flybase [19] - my personal experience). While we decided to share technology, the RDBMS and programming language, this commonality was moot because we did not also share a common focus. SGD had a finished genome while Berkeley was managing expressed sequence tag (EST) and physical mapping data. The central point is that the solution to biological database integration does not lie in particular technologies.

At the same time, an approximate solution to this problem was being demanded by the research communities whom the model organism databases served. These communities increasingly included not just organism-specific researchers, but also pharmaceutical companies, human geneticists, and biologists interested in many organisms, not just one. Another contributing factor was the recent maturation of DNA microarray technology [26,27]. The implication of this development was that functional analysis would be done on a large scale, and the community risked losing the capacity to leverage the power of these new data fully if the data were poorly integrated. For those orchestrating a genome database this was not merely an intellectual exercise: we had to find a solution or risk losing funding. We were highly motivated.

The most fundamental questions for the biologists served by the model organism databases revolve around the genes. What genes are there, what are their mRNA and peptide sequences, where are they in the genome, when are they expressed and how is their activity controlled, in what tissue, organ, and part of the cell are they expressed, what function do they carry out and what role does this play in the organism's biology? Both pragmatically and biologically, then, it made sense for the solution similarly to revolve around the genes. One essential aspect of this, which everyone agreed was necessary, was systematically recording the molecular functions and biological roles of every gene.

One of the first functional classification systems was created in 1993 by Monica Riley for *Escherichia coli* [28]. Building primarily upon this system, Michael Ashburner began assembling what became the forerunner of the Gene Ontology (GO), originally to serve the requirements of FlyBase. Similarly, TIGR created its functional classification system around this time. These early efforts were systematic, in that they were using a well-defined set of concepts for the descriptions, but they were limited because they were not shared between organisms. SGD [21], FlyBase [19], TIGR [18], Mouse Genome Informatics (MGI) [29], and others, all independently realized that we could essentially solve a significant portion of the data-integration issue if a cross-species functional classification system were created. In our ideal world, sequence (nucleic acid or protein), organism, and other specialty biological databases would all agree on how this should be done.

In 1998, it became simply imperative for those responsible for community model organism databases to act, as the number of completely sequenced genomes and large-scale functional experiments was growing. Our correspondence that spring contained many messages such as these: "I'm interested in being involved in defining a vocabulary that is used between the model organism databases. These databases must work together to produce a controlled vocabulary" (personal communication); and "It would be desirable if the whole genome community was using one role/process scheme. It seems to me that your list and the TIGR list are similar enough that generation of a common list is conceivable" (personal communication). In July of that year, Michael Ashburner presented a proposal at the Montreal International conference on Intelligent Systems for Molecular Biology (ISMB) bio-ontologies workshop to use a simple hierarchical controlled vocabulary; his proposal was dismissed by other participants as naïve. But later, in the hotel bar, representatives of FlyBase (me), SGD (Steve Chervitz), and MGI (Judith Blake) embraced the proposal and agreed jointly to apply the same vocabulary to describe the molecular functions and biological roles for every gene in our respective databases. Thus we founded the Gene Ontology Consortium.

Six years have now passed and GO has grown enormously. GO is now clearly defined and a model for numerous other biological ontology projects that aim similarly to achieve structured, standardized vocabularies for describing biological systems. GO is a structured network consisting of defined terms and the relationships between them that describe three attributes of gene products, their Molecular Function, Biological Process and Cellular Component [30]. There are many measures demonstrating its success. At present there are close to 300 articles in PubMed referencing GO. Among large institutional databanks, Swiss-Prot now uses GO for annotating the peptide sequences it maintains. The number of organism groups participating in the GO consortium has grown every quarter-year from the initial three to roughly two dozen. Every conference has talks and posters either referencing or utilizing GO, and within the genome community it has become the accepted standard for functional annotation. While it is impossible in hindsight to pinpoint exactly why it has succeeded, there are certain definite factors involved that are discussed below.

In brief: we already had 'market share'; our careers were such that we could take risks; we were and are practical and experienced engineers; we have always worked at the leading edge of technology; it was in our own self-interest; we had 'domain knowledge'; and we are open. When considering 'market share', a significant advantage that we (those managing biological databases) had, though it is not often considered, is our stewardship of key datasets. The commencement of GO also coincided with the completion of many key genome sequences. Once sequencing is finished, database groups annotate, manage and maintain the sequence. This put us in the right position to succeed because of the influence these data have. The decisions we make in our management of the data have a great deal of downstream effect. Every researcher, whether bench-scientist or informaticist, who utilizes the genomic data of mouse, *Drosophila*, yeast, or other organisms, is influenced by our choices as to how the data are described and organized. In contrast to broad-spectrum archival repositories, these data are annotated by specialists in the biology of a given organism who have a detailed understanding of its idiosyncratic biology. This expertise anchors the captured knowledge in experimental data. As other organism specialists joined - such as the *Arabidopsis* Information Resource (TAIR) [31], which joined soon after the start, as well as microbial and pathogen databases [32] - the impact of GO increased. Given the large established constituency of biologists who use FlyBase, SGD, MGI, and TAIR, it is not surprising that our decision to jointly develop GO was influential.

In addition to holding majority share of these critical research resources, the careers of the people involved are built on successful collaborative efforts. The professionals who are responsible for the biological databases fall roughly into two classes. They are either tenured principal investigators who wish to contribute to their community or PhD-level researchers (both biologists and computer scientists) who have especially chosen a non-academic career track. As individuals, they do not have much to gain by, for example, publishing papers as individuals. Papers are published, of course, about the content of the database or techniques for managing the data, but an individual's personal publication record is not a primary criterion upon which their career is evaluated. Rather, careers are measured by the success of the project and the strength of an individual's contribution to the project's goals. This attitude allowed us to remove both our egos and our concern for individual recognition from the search for a solution to the data-interconnection problem.

Apart from these organizational and social factors, each GO consortium scientist had a successful background in producing large information resources. Everyone had their own institutional knowledge of the requirements for biology and proven experience in engineering management and development. They knew how to decompose a large and complex project into smaller readily measurable milestones, which is an extremely difficult thing to do. Understanding the theoretical requirements of a problem is necessary, but not sufficient. The experience and practical skill to effectively direct the development and implement a solution were also essential.

Complementing our existing skills was our willingness to use new technologies. A key characteristic of the scientists who initiated GO is that they are 'early adopters' of new technologies. There is a definite behavior pattern in this group of exploring technological innovations. We had always sought new strategies to solve our problems: for example, the internet, the worldwide web, RDBMSs, new programming languages (such as Perl and Java), and through to ontologies, all of which we began to work with before the methodologies were mature and well-established. In short, we have a tradition of experimentation. It is not very surprising that scientists are willing to experiment, but this mindset extends to computer science as well and enables us to exploit advances in that field to address the needs of biology. We will take advantage of anything that will help us get the job done.

The GO consortium is inherently collaborative, and collaborations are hard - very hard - because of geography, misunderstandings, and the length of time it takes to get anything resolved and completed. Within the consortium, collaboration is made even more difficult because we must discuss and agree upon mental concepts and definitions in addition to concrete issues such as data syntax and exchange. Still, we actively sought collaboration, because it was in our own self-interest. Our users, upon whose support we depend, were demanding the ability to ask the same query of different genomic databases and to receive comparable answers. Every biological database would gain through cooperation.

One of the most significant contributing factors is our deep knowledge of the domain of biology. No problem can be solved successfully if you do not understand its nuances. The consortium succeeded by utilizing knowledge from many disparate fields: selectively exploiting what has been learned in the field of artificial intelligence and the study of ontologies; constrained by practical engineering considerations and incremental development; all the while bearing in mind the niceties of the biology being represented. Domain knowledge is essential to GO's success, and without it we could not maintain biological fidelity.

Last, and perhaps most important, is that we have always been open. All of the vocabularies, the annotations, and the software tools are available for others to use. Our success is best illustrated by how much they are used [33]. This openness is essential in the scientific environment in which we work. To provide a technology without a willingness to reveal all source code and data is tantamount to throwing away the lab notebook. Providing outside researchers with the ability to completely understand the methods that are used is mandatory for scientific progress. GO is not perfect, but its success is primarily due to revealing everything. The feedback we receive from others is what enables the consortium to improve with age.

Our plan for the future is to build on this base. We are actively seeking ways and building tools to help new biological databases utilize GO and thus extend our data coverage to include more organisms. We will remain pragmatic in our choice of technologies and remain sufficiently flexible to be able to exploit new advances. We will incrementally advance the sophistication of the underlying software architecture, one example of which is shown by our collaboration with Reactome [34], a project generating formal representations of biological pathways. We will seek out domain experts as the biological coverage of the GO extends into new areas, so that biological veracity is maintained. Similarly, we will work with experts to extend the scope of available ontologies to cover other critical areas of biological description, such as anatomies, cell types, and phenotypes, as illustrated by the Open Biological Ontologies [35] project. Finally, we will continue to work cooperatively and remain open as this has been shown to be the most scientifically productive approach.

In summary, GO has succeeded because it is not a technical solution *per se*. Technology is more than just an implementation detail, of course, but it will never be a silver bullet. We want to continue integrating our knowledge forever and technologies are short-lived. So, the solution must be to adopt new technologies as they arise while the primary focus remains on cooperative development of semantic standards: it's about the content, not the container. Perhaps ironically, the impact of shifting the focus away from a technical solution to the biological data integration problem is that we have begun sharing technology. Once the mechanism for a dialog was in place, we discovered many other areas where our interests coincided. There are now organized meetings for professional biological curators to meet and discuss standard methodologies [36]. The Generic Model Organism Database (GMOD) [37] effort makes these common tools available to the community and serves as a forum for a wide spectrum of interests. It is this unforeseen outcome, consolidating the disparate databases into a cooperative community engaged in productive dialogs, that, in my view constitutes the single largest impact and achievement of the Gene Ontology consortium to date.

References

1. Sanger F, Coulson AR: **A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.** *J Mol Biol* 1975, **94:**441-448.
2. Maxam AM, Gilbert W: **A new method for sequencing DNA.** *Proc Natl Acad Sci USA* 1977, **74:**560-564.
3. **Dr Margaret Oakley Dayhoff - Pioneer in Bioinformatics** [http://www.dayhoff.cc/index.html]
4. Dayhoff MO, Eck RV, Chang MA, Sochard MR: *Atlas of Protein Sequence and Structure.* Silver Spring: National Biomedical Research Foundation; 1965.
5. **PIR Protein Information Resource** [http://pir.georgetown.edu/home.shtml]
6. **PDB** [http://www.rcsb.org/pdb/]
7. **PDB Current Holdings** [http://www.rcsb.org/pdb/holdings.html]
8. **Research Milestones at the Jackson Laboratory** [http://www.jax.org/about/milestones.html]
9. **EMBL Nucleotide Sequence Database** [http://www.ebi.ac.uk/embl/index.html]
10. **Brief History of EMBL** [http://www.embl.org/aboutus/generalinfo/history.html]
11. **GenBank** [http://www.ncbi.nlm.nih.gov/Genbank/index.html]
12. **Bioinformatics milestones** [http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/milestones.html]
13. **PIR Mission and History** [http://pir.georgetown.edu/pirwww/aboutpir/history.html]
14. **UniProt/Swiss-Prot** [http://www.ebi.ac.uk/swissprot]
15. **AceDB** [http://www.acedb.org/]
16. **New Directions in Genome Databases at Stanford** [http://weedsworld.arabidopsis.org.uk/Vol3ii/Cherry-Flanders-Petel.WW.html]
17. **The Institute for Genome Research 1992-1999** [http://www.tigr.org/about/history.shtml]
18. **TIGR** [http://www.tigr.org/]
19. **FlyBase** [http://www.flybase.org]
20. **About SGD** [http://www.yeastgenome.org/aboutsgd.shtml]
21. **SGD** [http://www.yeastgenome.org/]
22. **Ted Codd: The Rise of Relational Databases: 1970** [http://www.nap.edu/readingroom/books/far/ch6.html]
23. **The Moschovitis Group: Internet Is Defined Officially as Networks Using TCP/IP** [http://www.historyoftheinternet.com/chap4.html]
24. **A Little History of the World Wide Web** [http://www.w3.org/History.html]
25. **Meeting on Interconnection of Molecular Biology Databases** [http://www.ai.sri.com/~pkarp/mimbd.html]
26. Fodor SP, Rava RP, Huang XC, Pease AC, Holmes CP, Adams CL: **Multiplexed biochemical assays with biological chips.** *Nature* 1993, **364:**555-556.
27. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science* 1995, **270:**467-70.
28. Riley M. **Functions of the gene products of *Escherichia coli*.** *Microbiol Rev* 1993, **57:**862-952.
29. **MGI** [http://www.informatics.jax.org/]
30. **An Introduction to Gene Ontology** [http://www.geneontology.org/GO.doc.html]
31. **TAIR** [http://www.arabidopsis.org/]
32. **Gene DB** [http://www.genedb.org/]
33. **Gene Ontology - A Bibliography** [http://www.geneontology.org/GO.biblio.html]
34. **Reactome** [http://www.reactome.org]
35. **Open Biological Ontologies** [http://obo.sourceforge.net]
36. **Biocurator** [http://tesuque.stanford.edu/biocurator.org/]
37. **Generic Model Organism Database Construction Set** [http://gmod.sourceforge.net/]