Creating the Gene Ontology Resource: Design and Implementation

The Gene Ontology Consortium[2]

The exponential growth in the volume of accessible biological information has generated a confusion of voices surrounding the annotation of molecular information about genes and their products. The Gene Ontology (GO) project seeks to provide a set of structured vocabularies for specific biological domains that can be used to describe gene products in any organism. This work includes building three extensive ontologies to describe molecular function, biological process, and cellular component, and providing a community database resource that supports the use of these ontologies. The GO Consortium was initiated by scientists associated with three model organism databases: SGD, the *Saccharomyces* Genome database; FlyBase, the *Drosophila* genome database; and MGD/GXD, the Mouse Genome Informatics databases. Additional model organism database groups are joining the project. Each of these model organism information systems is annotating genes and gene products using GO vocabulary terms and incorporating these annotations into their respective model organism databases. Each database contributes its annotation files to a shared GO data resource accessible to the public at http://www.geneontology.org/. The GO site can be used by the community both to recover the GO vocabularies and to access the annotated gene product data sets from the model organism databases. The GO Consortium supports the development of the GO database resource and provides tools enabling curators and researchers to query and manipulate the vocabularies. We believe that the shared development of this molecular annotation resource will contribute to the unification of biological information.

As the amount of biological information has grown, it has become increasingly important to describe and classify biological objects in meaningful ways. Many species- and domain-specific databases have strategies to organize and integrate these data, allowing users to sift through ever-increasing volumes of information. Biologists want to be able to use the information stored in disparate databases to ask biologically interesting questions. They want to know, for example, which genes or gene products contribute to the formation and development of an epithelial sheet, or what are the DNA-binding proteins involved in DNA repair but not in DNA replication, or what evidence is there that the mouse *Pax6* gene product is involved in eye morphogenesis. In addition, researchers want to be able to expand such queries to find gene products in different organisms that share characteristics. To support this kind of research, databases must rigorously organize and annotate the biological properties of gene products. Searching for these types of information in the context of examining microarray expression data, sequencing genotypes from a population, or identifying all glycolytic enzymes is difficult, if not impossible, without computational tools and well-defined annotation systems.

The Gene Ontology (GO) Consortium was formed to develop shared, structured vocabularies adequate for the annotation of molecular characteristics across organisms (The Gene Ontology Consortium 2000). The original intent of the group was to construct a set of vocabularies comprising terms that we could share with a common understanding of the meaning of any term used, and that could support cross-database queries. It soon became obvious, however, that the combined set of annotations from the model organism groups would provide a useful resource for the entire scientific community. Therefore, in addition to developing the shared structured vocabularies, the GO project is developing a database resource that provides access not only to the vocabularies, but also to annotation and query applications and to specialized data sets resulting from the use of the vocabularies in the annotation of genes and/or gene products.

An ontology (Gruber 1993, 1995) has two primary pragmatic purposes. The first is to facilitate communication between people and organizations. The second is to improve

[1]Corresponding author.
E-MAIL jblake@informatics.jax.org; (207) 288-6132.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.180801.

[2]Alphabetical list of authors in the Gene Ontology Consortium: Michael Ashburner (FlyBase, http://flybase.bio.indiana.edu), Catherine A. Ball (*Saccharomyces* Genome Database, http://genome-www.stanford.edu), Judith A. Blake (Mouse Genome Database and Gene Expression Database, http://www.informatics.jax.org), Heather Butler (FlyBase, http://flybase.bio.indiana.edu), J. Michael Cherry (*Saccharomyces* Genome Database, http://genome-www.stanford.edu), John Corradi (Mouse Genome Database and Gene Expression Database, http://www.informatics.jax.org), Kara Dolinski (*Saccharomyces* Genome Database, http://genome-www.stanford.edu), Janan T. Eppig (Mouse Genome Database and Gene Expression Database, http://www.informatics.jax.org), Midori Harris (*Saccharomyces* Genome Database, http://genome-www.stanford.edu), David P. Hill (Mouse Genome Database and Gene Expression Database, http://www.informatics.jax.org), Suzanna Lewis (Berkeley *Drosophila* Genome Project, http://www.fruitfly.org), Brad Marshall (Berkeley *Drosophila* Genome Project, http://www.fruitfly.org), Chris Mungall (Berkeley *Drosophila* Genome Project, http://www.fruitfly.org), Leonore Reiser (The *Arabidopsis* Information Resource, http://www.arabidopsis.org), Sue Rhee (The *Arabidopsis* Information Resource, http://www.arabidopsis.org), Joel E. Richardson (Mouse Genome Database and Gene Expression Database, http://www.informatics.jax.org), John Richter (Berkeley *Drosophila* Genome Project, http://www.fruitfly.org), Martin Ringwald (Mouse Genome Database and Gene Expression Database, http://www.informatics.jax.org), Gerald M. Rubin (Berkeley *Drosophila* Genome Project, http://www.fruitfly.org), Gavin Sherlock (*Saccharomyces* Genome Database, http://genome-www.stanford.edu), and J. Yoon (The *Arabidopsis* Information Resource, http://www.arabidopsis.org).

interoperability between systems. We have consciously chosen to begin at the most basic level, by creating and agreeing on shared semantic concepts; that is, by defining the words that are required to describe particular domains of biology. We are aware that this is an incomplete solution, but firmly believe that it is a necessary first step. These common concepts are immediately useful and can be used ultimately as a foundation to describe the domain of biology more fully.

The use of ontological methods to structure biological knowledge is an active area of research and development (e.g., Guarino 1998; Jones and Paton 1999; http://www.cs.utexas.edu/users/mfkb/related.html, http://ontolingua.stanford.edu). Independent of the application of ontological methods to the biological domain, however, researchers were constructing vocabularies to categorize cellular functions (e.g., Riley 1993). These types of classifications have been applied in a manner that supports the ability to search for genes by physiological roles in databases such as EcoCyc and MetaCyc (Karp et al. 2000). Other efforts at comprehensive vocabularies include the medical subject heading (MeSH) vocabularies which have been applied to the scientific literature via MEDLINE (Delozier and Lingle 1992; Lowe and Barnett 1994).

The Gene Ontology Project

The GO Consortium

The GO Consortium was established in 1998 as a collaboration between three model organism databases: FlyBase, the genome database for *Drosophila* (The FlyBase Consortium 1999); the *Saccharomyces* Genome Database (SGD) (Ball et al. 2000); and the integrated Mouse Genome Informatics databases, Mouse Genome Database, MGD (Blake et al. 2000) and Gene Expression Database, GXD (Ringwald et al. 2000), hereafter referred to jointly as MGI. During 2000, two more model organism groups, The *Arabidopsis* Information Resource (TAIR) (Huala et al. 2001), and the *Caenorhabditis elegans* group (http://www.wormbase.org/) joined the GO Consortium.

Goals of the GO project

GO endeavors to develop cross-species biological vocabularies that are used by multiple databases to annotate genes and gene products in a consistent way. Three extensive ontologies are under development, for (1) molecular function, (2) biological process, and (3) cellular component. These particular classifications were chosen because they represent information sets that are common to all living forms and are basic to our annotation of information about genes and gene products. This effort parallels work in the computational biology community to provide tools for implementing biological ontologies (Schulze-Kremer et al. 1998; http://www-smi.stanford.edu/projects/bio-ontology/). Rather than focusing on tools and procedures for implementation of ontologies, our effort primarily focuses on knowledge domain development and the biological annotations that are applied to model organism gene products.

One important feature of the GO project is that the development of the GO vocabularies is independent of the association of particular gene products with GO terms. The Consortium members work together to construct and define the terms in the vocabularies and to specify the relationships between terms. Then, the ontologies are used to annotate gene products in the databases of the Consortium members. Each model organism information resource incorporates the vocabularies into its data query and visualization tools as appropriate.

The goals of the GO project have been carefully defined, as shown in Box 1. We recognize that there exists a biological relationship between a molecular function, the involvement of a series of functions in a biological process, and the unfolding of that process at a given time and space in the cell. It follows that GO could logically be expanded to reflect all cellular operations and states at a given time. However, the GO Consortium members have chosen to initially focus on three precise sets of terms that are of immediate and exceptional utility to the researcher and that span our various organismal domains. Although we anticipate that it may be necessary to expand GO in the future to incorporate more sophisticated biological concepts, the effort described here is an essential start to creating a shared language of biology.

The strength of the GO approach lies in its focus on the specifics of the biological vocabularies and on the establishment of precise, defined relationships between the terms. The structure of the ontology permits the implementation of robust query capabilities far beyond the development of a simple dictionary of terms or keywords. For example, "DNA replication" is represented in GO as a part of "DNA metabolism" and as a part of "DNA replication and cell cycle," which is itself a part of the "cell cycle." The term is also found as a part of the "mitotic S phase." There are, therefore, multiple pathways and terms that can be used to recover information about gene products that have been annotated to the molecular function 'DNA replication" (Fig. 1).

The assignment of a defined term as an attribute of a gene product also allows a subsequent query, via the defined

Box 1. The Goals of the Gene Ontology Consortium

1. To compile a comprehensive structured vocabulary of terms describing different elements of molecular biology that are shared among life forms.
   - -Terms are defined, may have synonyms and are organized into broader and narrower refinements.
   - -Separate vocabularies are used to define separate dimensions of biology.
2. To describe biological objects (in the model organism database of each contributing member) using these terms.
3. To provide tools for querying and manipulating these vocabularies.
   - -To add new vocabularies for additional aspects of biology.
   - -To permit researchers to locate both terms and biological objects either via the Web or in more complex ways.
   - -To allow others to set up satellite databases.
4. To provide tools enabling curators to assign GO terms to biological objects.
   - -Sequence-based methods
   - -Editorial annotations
   - -Microarrays
   - -Protein binding experiments

WHAT GO IS NOT:

1. GO is not a way to unify biological databases. Sharing nomenclature is a step toward unification, but is not, in itself, sufficient.
2. GO is not a dictated standard, mandating nomenclature across databases. Groups participate because of self-interest and cooperate to arrive at a consensus.
3. GO does not define homologies between gene products from different organisms. The use of the GO results in shared annotations for gene products from different organisms, and this may reflect an evolutionary relationship, but the shared annotation is in itself not sufficient for such a determination.

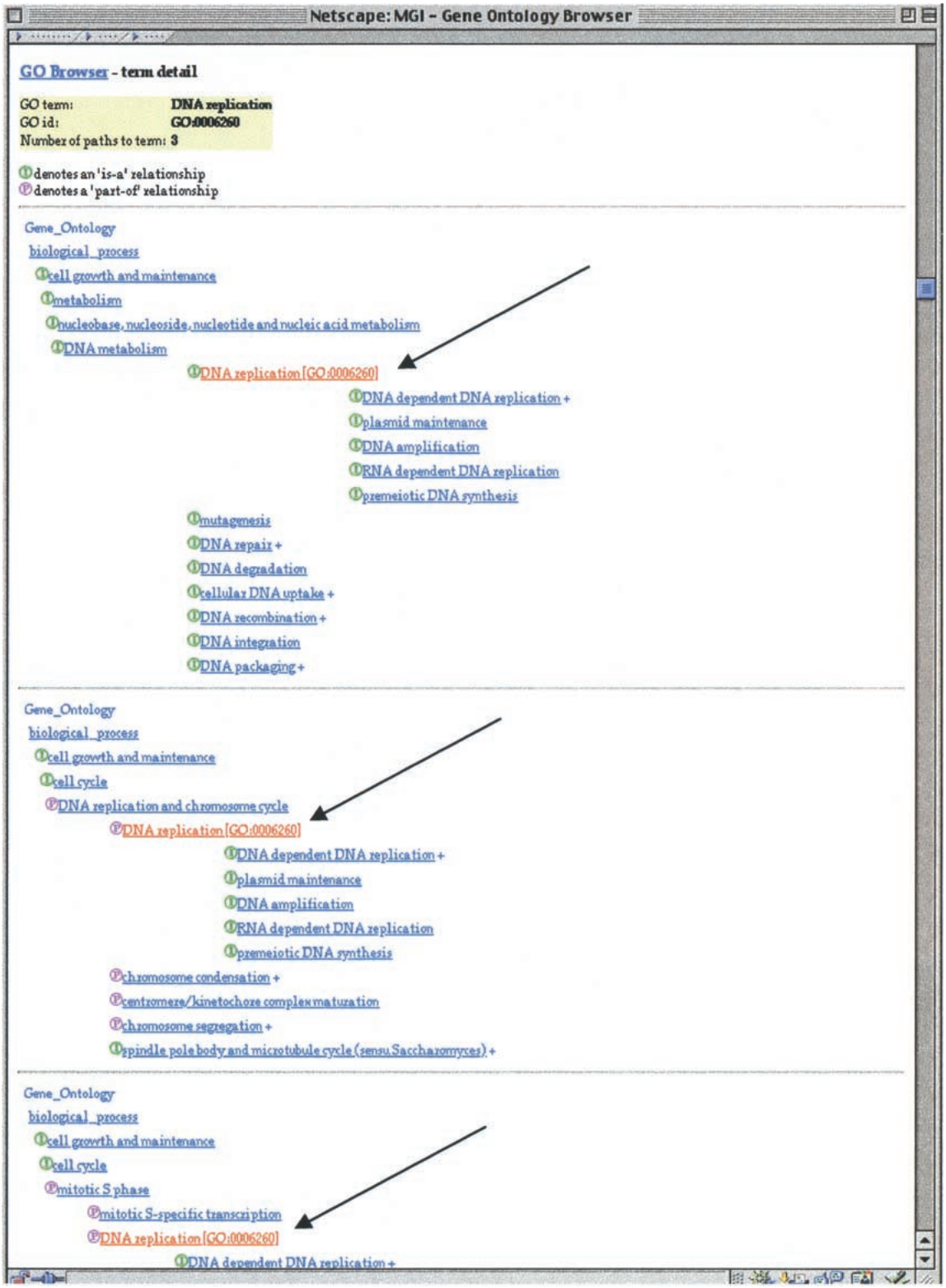

Figure 1 Multiple pathways. This figure illustrates the different pathways that represent the biological process of DNA replication. Viewed in the GO Browser, the relationships between the terms, the GO ID for DNA replication, and information about the number of pathways incorporating this biological process are displayed.

term, to recover all gene products known to share that attribute. For some searches, particularly searches by a precisely defined function, the gene products from yeast and *Drosophila* may show strong structural similarities to the gene product in mouse or other organisms. Some may consider this sufficient information to conclude that these represent orthologs, but the GO project itself will not draw that conclusion, as different evolutionary paths can result in a shared molecular function (e.g., yeast and fly alcohol dehydrogenases). For example, the gene known as "*Car11*, carbonic anhydrase 11, in the mouse, is a member of the carbonic anhydrase gene family, but does not have the enzymatic function of carbonate dehydratase that other members of this family do. Therefore, in the GO annotations, this gene is not associated with the GO molecular function term 'carbonate dehydratase' (GO identification no. GO:0004089). Assertions of homology or evolutionary relatedness between gene products from multiple organisms currently lie outside the scope of the GO project. The intent of GO, rather, is to robustly define information known about each specific gene product, and then to provide the ability to explore the information through searches by, for example, molecular function or cellular location, that recover the gene products known to share the attribute.

The Three Ontologies

The GO Consortium is developing three ontologies: molecular function, biological process, and cellular component, to describe attributes of gene products or gene product groups. Briefly, molecular function describes what a gene product does at the biochemical level. Biological process describes a broad biological objective. Cellular component describes the location of a gene product, within cellular structures and within macromolecular complexes.

The Ontology of Molecular Function

Molecular function is defined as what a gene product does at the biochemical level. It describes only what is done without specifying where or when the event actually occurs or its broader context. Examples of broad functional terms are "enzyme," "transporter," or "ligand." Examples of more specific functional terms are "adenylate cyclase" or "Toll receptor ligand."

There is a potential for semantic confusion between a gene product and its molecular function because very often a gene product is named by its molecular function or at least by one of its molecular functions. Enzymes are obvious examples of this phenomenon.

The Ontology of Biological Process

Biological process refers to a biological objective to which the gene product contributes. A process is accomplished via one or more ordered assemblies of functions. It often involves transformation in the sense that something goes into a process and something different comes out of it. Examples of broad biological process terms are "cell growth and maintenance" or "signal transduction." Examples of more specific terms are "pyrimidine metabolism" or "cAMP biosynthesis."

A biological process is not equivalent to a pathway. Specifically we are not capturing or trying to represent any of the dynamics or dependencies that would be required to describe a pathway in the present implementation. It is understood that a network of relationships connects specific molecular functions to one or more biological processes, but it is beyond our current scope to explicitly develop and represent these interconnections. Instead, we seek to define the molecular function of a gene product as precisely as possible, and, similarly, to note each and any biological process in which a gene product is involved, as described below.

The Ontology of Cellular Component

Cellular component refers to the place in the cell where a gene product is found. These terms reflect our understanding of cell structure in a generic sense.

Cellular component includes terms describing complexes where multiple gene products would be found, such as the "ribosome" or "proteasome." It also includes terms such as "nuclear membrane" or "Golgi apparatus." Thus, the term "cellular component" encompasses a broad concept of "location" as the place in the cell where the gene product is active. For example, cellular component terms can be a"place" such as the nuclear outer membrane (GO: 0005640; synonym: outer envelope), or it can mean a "place" such as the histone deacetylase complex (GO: 0000118).

Ontology Structure and Standards

The Structure of the Ontologies

The ontologies are structured vocabularies in the form of directed acyclic graphs (DAGs) that represent a network in which each term may be a "child" of one or more than one "parent." An example from the molecular function vocabulary is the function term "transmembrane receptor protein-tyrosine kinase" and its relationship to other function terms. It is a subclass both of the parent "transmembrane receptor" and of the parent "protein tyrosine kinase."

Relationships of child to parent can be of the "is a" type or the "part of" type. The "is a" type refers to when a child is an instance of the parent (in an example from the cellular component vocabulary, a mitotic chromosome is a instance of a chromosome). The "part of" type refers to when a child is a component of the parent (e.g., the telomere is a component of a chromosome). Child terms may have more than one parent term and may have a different class of relationship with its different parents. Although it is difficult to manage the different types of relationships within the same ontology, the relationships must be of many varieties if we are to accurately reflect the semantics. The expressive capabilities of a DAG as compared to tree and logic language, and the rules we implement to address how the logic in querying works in respect to the different relationship types, permit this complex representation of relationship.

Each term in the ontology is an accessible object in the GO data resource. Every term has a unique identifier to be used as a database crossreference in the collaborating databases. Each term is (or will be) defined and each definition will cite the source from which its definition was obtained. Query and implementation tools have been developed to exploit the detailed relationships captured in the ontologies themselves. Although each term in an ontology has a relationship with at least one other term, this information is not incorporated in the identifiers because, among many other considerations, the location of the term within an ontology (i.e., its parents and children) may change.

Defining Terms

Definitions for GO terms are being provided as part of the development of the ontologies. We are using, as much as possible, the Oxford Dictionary of Molecular Biology (1997),

with permission and attribution. Other definitions are obtained from standard reference works in biochemistry and molecular biology or sources such as SWISS-PROT (Bairoch and Apweiler 2000). The source of each definition is stored and is available. Definitions are accessible to users both from the Web pages and as part of the data available by searching with the GO Browser (see below).

It is only through the careful attention to the precise definition of a term as it is implemented in GO that scientific curators at multiple sites from multiple research communities can successfully annotate gene products in the context of the GO collaboration. Often the same term can be used with different meaning in different research communities. One of the strengths of the GO Consortium approach to this reality is that we work to provide a definition for each controversial term that works for all the annotation groups. Although nomenclature (i.e., use of terms with specific meanings) wars are common in biology, the need for shared vocabularies has resulted in our drive and commitment to reaching a consensus so that a given term is used by all with a specific meaning.

Other attributes of ontology terms are supported. For example, a term may have one or more synonyms. There may be cross-database references to the Enzyme Nomenclature Database (http://expasy.proteome.org.au/enzyme/) or to other specialized vocabularies such as those for species-specific anatomies.

Standards for GO

Members of the Consortium group may contribute to updates and revisions of GO. The GO editor works together with several scientific curators to analyze and refine aspects of the ontologies that are all works in progress. Additions and changes to the ontologies come from the collaborating databases and from the broader community (see below). Some of the working principles for the development of GO are: (1) all paths must be true; (2) terms should not be species specific, but should represent at least class level coverage; (3) all attributes of GO must be accompanied by appropriate citations; and (4) all annotations of gene products to GO terms must incorporate controlled statements of the type of evidence that supports the relationship, as well as appropriate citations. These rules and guidelines are documented, with examples, at the GO web site.

True Path Rule

The "True Path Rule" is an example of the types of procedures incorporated into the general GO guidelines. The pathway from a child term to its top-level parent(s) must always be true. If a new gene product is found to break this rule, or if species specificity becomes a problem, a restructuring of the hierarchy should occur by adding more nodes and connecting terms that creates a new path to maintain the validity of the upward hierarchy. Consider terms describing chitin metabolism in the biological process ontology. Chitin metabolism is part of cuticle synthesis in the fly, and is also part of cell wall organization in yeast. Figure 2 illustrates how the biological process ontology is constructed for this example.

The chitin example exemplifies the current paradigm concerning the expansion of the GO; that is, to refine and extend the ontology to be semantically correct. As the GO is thus expanded, however, we have considered the concerns as to (1) how we will continue to maintain consistency within the GO structures as the ontology expands to incorporate fine levels of molecular detail, and (2) how we will know when to limit the expansion because the ontologies reflect too much species-level detail.

We are working on solutions to the concern about endless expansion and the maintenance of internal consistency. First, we will continue to develop the GO database and related annotation tools. Computationally, the size of the ontologies isn't really an issue. Curatorially, it will be increasingly difficult to maintain the semantic consistency we desire without software tools that perform consistency checks and controlled updates. In addition, without dedicated curation of various subsections of the GO ontologies, we will lose overall consistency in the project. Thus, we actively refine and extend particular areas within an ontology as we realize the need to do so. For example, in the biological process ontology, we have recently grouped all cellular processes independently of the multicellular processes. This reorganization will allow us, with the help of the molecular biology informatics community, to refine and update this particular section of the ontology.

Species–Specific Considerations

Many molecular functions and biological processes do not exist in all organisms. The set of GO terms, however, is meant to be inclusive and integrity among the terms in the hierarchy must be consistent for all organisms. Our current convention is to include any term (such as "pattern formation") that applies to more than one taxonomic class of organism (e.g., Mammalia and Aves classes of the phylum Chordata). This consideration may change as the utility of the resource across various species is tested. Within the ontologies themselves, there are cases where a word or phrase has different meanings when applied to different organisms. In such cases, the ontologies have one term for each meaning, distinguished from other like terms both by the definition and by the use of the sensu designation. For example, GO:0007322 is the term "mating (sensu *Saccharomyces*)" to distinguish it from "mating (sensu *Caenorhabditis elegans*)." The term "mating (sensu *Saccharomyces*)" is to be used by other yeast species in their annotations when the mating process is comparable to that exemplified by *Saccharomyces cerevisiae*.

Updating the Ontologies

Terms can be added to GO by curators of the participating databases. E-mail notification of the Consortium members alerts participants to the addition or restructuring of any aspect of the ontologies. Adding new terms can be as straightforward as the recognition that an additional term is needed, for instance, for another type of DNA repair enzymatic activity. It can also be as complicated as deciding how "mitotic spindle orientation," a child of "spindle assembly," relates to "establishment of cell polarity," a child of "cytoskeleton organization and biogenesis." Curators carefully evaluate changes to GO, especially those that change "parent" to "child" relationships and thus might have an impact on the current use of the terms by participating databases. One interesting aspect of GO is that because it represents biological knowledge independent of single-gene annotations, as GO continues to be refined and to evolve, the annotation of and knowledge about gene products associated with GO terms will automatically be refined as well. GO welcomes input from users.

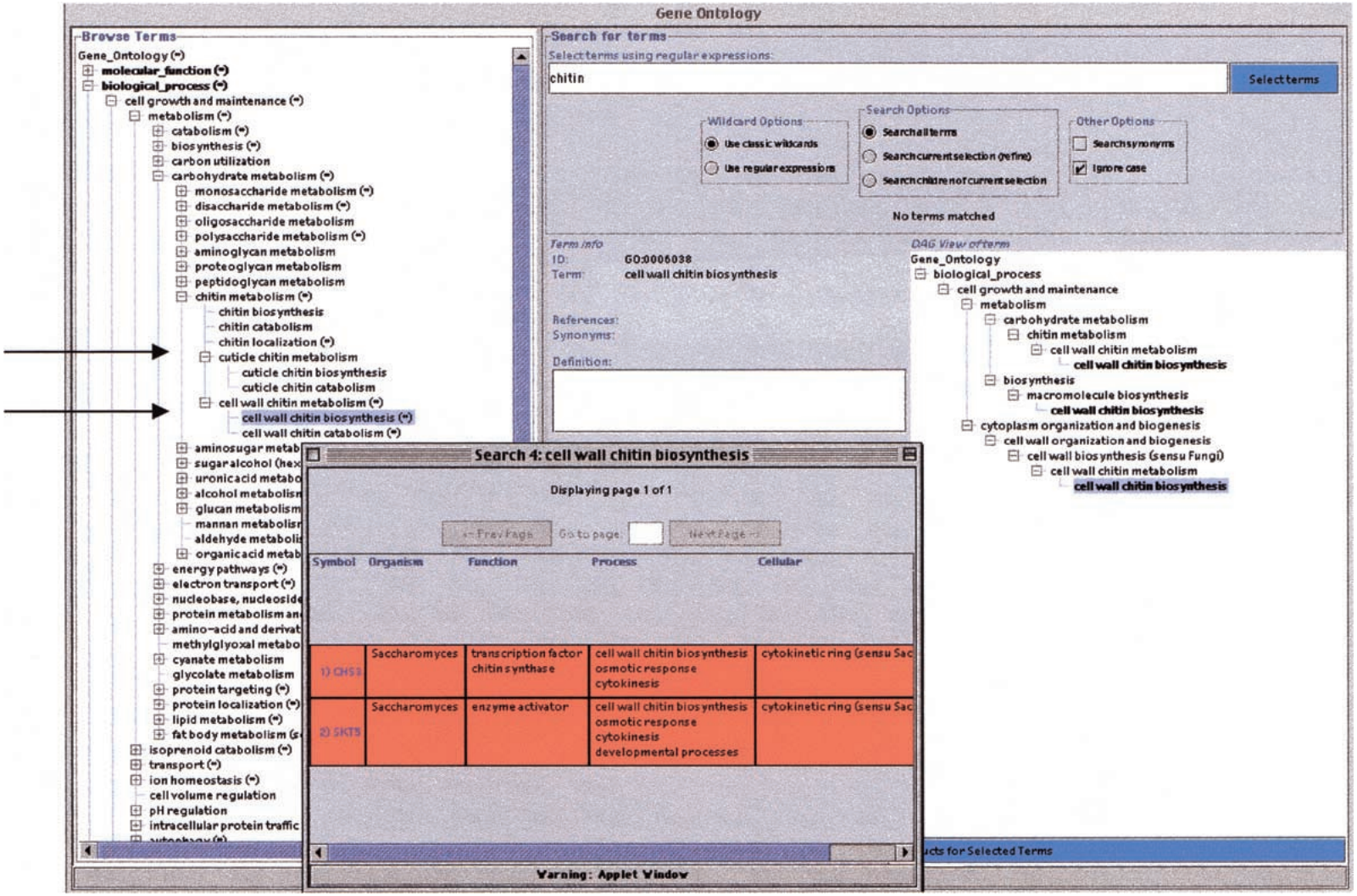

Figure 2 Extension of GO to reflect multiple uses of the term "chitin". The "chitin" example illustrates the development of the GO structure to accurately reflect the understanding of the relationships. Cuticle chitin metabolism is a process independent of cell wall chitin metabolism. This illustration utilizes the BDGP GO Browser, which supports links to gene products annotated to a given GO term. Here, the genes annotated with the term "cell wall chitin biosynthesis" are represented, with links back to the SGD database representations for these genes.

Outstanding Design and Implementation Issues

The current development model will carry the GO project a long way. This practical approach to the development of ontologies can be improved by further definition of the procedures used to create and maintain semantic order among concepts. Such issues are under regular discussion by the GO Consortium. For example, we debate about when a "process" begins and when it ends. Many of these discussions are summarized in the GO Usage Guide which is linked to the GO home page and is found at http://www.geneontology.org/GO.usage.html#description.

A specific example of the importance of definitions in understanding the development of GO is that of term "actin cytoskeleton" GO:0005856. Is the actin cytoskeleton a "type of " cytoskeleton, or is it "part of" the cytoskeleton? The answer depends on the definition of the term "cytoskeleton." As seen in the MGI GO Browser at http://www.informatics.jax.org/go/GO.cgi?id=GO:0005856, the definition we employ in the GO starts "Any of the various filamentous elements within the cytoplasm of eukaryotic cells. . . . . " Thus, the actin cytoskeleton is identified as a "type of" cytoskeleton. It will take time and dedicated effort to sort through the various definitions of common biological terms. This exercise in itself will clarify the vocabulary of biology for all users.

We are using a DAG to represent the terms and their relationships. This is an improvement over a simple hierarchical tree because it allows for a more specific term to be a child of multiple broader terms and thus captures the biological reality. We use the shorthand "term" to refer to the semantic definition of a concept that is embodied in a particular text string. It is not terribly complicated data structure, but it is sufficient to represent a vocabulary and that is what GO currently is. The primary limitation of this approach is that there is no governance over how the terms are applied to ensure that the triad of function, process, and component terms used to describe a single gene product makes sense biologically. At present this is the sole purview of the curator making the assignment and we rely on the dedication to accuracy of the curators. We plan to implement more robust solutions in the future.

The growth of the vocabularies is organic. We recognize that certain assumptions about development and releases (e.g., that early and frequent releases would alienate users and result in the failure of a project) are not necessarily true. Furthermore the addition of extra people brings new insights and eventually leads to a product that is more robust than anything a small group of developers could possibly have achieved. To be successful using this model requires the Consortium to be extremely responsive and attentive to the feedback that we receive from the community. We consciously make an effort to make extensions and modifications immediately when these are pointed out by the community. This responsiveness encourages people to continue to assist us in making improvements.

Annotating Gene Products to GO

The creation of the ontologies and the association of ontology terms with gene products are two independent operations. A gene product is a physical entity: a protein, or a functional RNA. Examples of gene products (by name) are alpha-globin or small ribosomal RNA. Gene products may assemble into entities that function as complexes, or gene product groups. Genes, gene products, gene product precursors, and gene product complexes can each and all be associated with one or more GO terms. Each gene product can be described in this system as having one or more functions, being involved in one or more biological processes, and as occurring in one or more cellular locations. Until a participating database curates as independent objects each different product and each product complex (e.g., from differential splicing or posttranslational modification), the annotation by GO terms uses the "gene" as a surrogate for all its products and their complexes.

It is critically important to distinguish between gene products and attributes of the product such as function that are often incorporated into the gene or gene product name. Genes and gene products are frequently named by their function. In fact, many revisions in nomenclature have occurred as the knowledge of the function of the gene product has developed. For example, the mouse gene originally designated as *c* (albino), based on the single gene inheritance of this coat color phenotype, has been cloned and its function identified, and hence, its symbol was revised to *Tyr* (tyrosinase).

In defining GO terms it has been essential to focus on the terms as a "function" or "process" term rather than on representing the "product" itself. A particular gene product may have one or more molecular functions and therefore will be associated with one or more molecular function terms. For example, the mouse gene *Abca4* is annotated to two terms in the function ontology: (1) ATP-binding cassette transporter (GO:0004009), and (2) phospholipid transfer (GO:0005548). Additionally, some gene products function in multiple enzymatic reactions, as defined by Enzyme Commission (EC) numbers. For example, the product of the mouse gene *P4hb* can function as either an isomerase or a dioxygenase, and is therefore annotated to both (1) protein disulfide isomerase (GO:0003756) and (2) procollagen-proline, 2-oxoglutarate-4-dioxygenase (GO:0004656). These examples illustrate that the GO term describes the chemical reaction carried out by the enzyme and is not a reference to the enzyme molecule itself.

Annotation of a gene product to one ontology is independent of its annotation to other ontologies. For example, the enzyme gamma-glutamyl transpeptidase, encoded by the mouse gene *Ggtp*, is annotated to a molecular function: gamma-glutamyl transferase (GO:0003824) and two biological processes: (1) glutathione metabolism (GO:0006749) and (2) spermatogenesis (GO:0007283), based on the study of the phenotype of mice with a mutation in this gene. Whereas the former biological process can be deduced from molecular function, the latter cannot, illustrating the value of independently annotating gene products using more than one ontology. Another example demonstrating the importance of the independence of ontologies is the annotation of the isoforms of malate dehydrogenase in yeast. The products of the *MDH1*, *MDH2*, and *MDH3* genes all have the same molecular function, malate dehydrogenase (GO:0004470), but localize to different cellular components and act in different biological processes.

Evidence and Citations for Gene Product Annotations

The annotations of gene products to the GO vocabularies are attributed to a source, which may be a literature reference, another database, or a computational analysis. The annotations include not only the source attribution, but also an indication of the evidence on which the annotation is based. A simple controlled vocabulary is used to describe the evidence supporting the attribution, such as "inferred from mutant

phenotype" or "inferred from direct assay." The complete set of "evidence statements" can be viewed at http://www.geneontology.org/GO.evidence.html. Referencing each annotation with both experimental method and citation is intended to help researchers evaluate the reliability of an annotation and is critically important to the future evaluation and use of these annotations. One might have greater confidence in an assignment based on direct experimental evidence than one based solely on a computational method such as sequence similarity. Furthermore, researchers may give some forms of experimental evidence more credence than others; for example, the observation that a mutation of a specific gene leads to a specific phenotype does not automatically mean that the gene product is directly involved in the biological process affected.

Creating a Shared Data Resource

The GO Web Site

As a public community effort, we have endeavored to incorporate suggestions from those using GO and to provide the annotation files and other documentation for the GO project. The GO web site (http://www.geneontology.org) provides downloadable versions of the ontologies, the term definitions, the species-specific gene product annotations, and other information. This site also includes query tools and a database implementation of GO (in MySQL). As the GO Consortium members collaborate to develop GO, other annotation groups are incorporating GO terms and philosophy in the annotation of gene products in other contexts (Adams et al. 2000).

A GO friends mailing list allows those not actively involved in the creation of the ontologies or the annotation of gene products to the ontologies to contribute to and ask questions of this project. The GO vocabularies exist as text files and a single XML file that provides the data for the ontology browsers.

GO Browsers

Three GO browsers are now available for the GO vocabularies, two developed at the Berkeley *Drosophila* Genome Project (BDGP) and the other developed by the MGI group. The BDGP Java-based GO Browser provides a query interface to the ontologies that allows users to use either regular expressions or simple "wild card" characters to query the database. The display includes a full DAG representation of the query results along with its definition and other information about these terms. In addition, it is the only one of the browsers able to then follow the terms to find all the gene products that one of the collaborating database have associated with these GO terms. This GO Browser accesses the nascent MySQL GO database at UC Berkeley. The Browser and documentation are available from http://www.fruitfly.org/annot/go/.

The BDGP HTML-based browser is more experimental. It uses frames to display results; one frame for entering simple queries, a second frame to display full information (including the definition) for a single term that the user has selected, and yet another frame to display the relationship of the query result to all of its parent and child terms. This browser can be accessed at http://www.fruitfly.org/~bradmars/cgi-bin/go.cgi.

The MGI GO Browser allows one either to browse or to search GO terms but not the annotations of gene products to the GO terms. A "Term Detail" page displays relevant information about each term, including its definition, any synonyms, and its relationships to other ontology terms. A "Query Summary" page displays all matches to the GO terms in the ontology category. The MGI GO Browser is also available from the GO site, http://www.geneontology.org/.

The Gene Ontology Database

The ontologies and the gene annotations have been loaded into a relational database for more robust representation and query capabilities. Implemented in both MySQL and Informix, the data model incorporates the relationships between terms and includes versioning of terms, their synonyms, and definitions. The association files of organism-specific gene-product annotations are also part of the database representation. The GO database (http://www.fruitfly.org/annot/go/database) is being built and maintained by the BDGP.

Documentation

A general documentation file to guide users is available (http://www.geneontology.org/GO.doc.html). In particular, bioinformatics systems that want to download GO annotations will find information about file structures and syntax. Additional information about GO, including links to other publications, a bibliography, and other information for users can be reached from the GO home page. Guidelines for the content and style of the ontologies have been assembled into the GO Usage Guide, available at http://www.geneontology.org/GO.usage.html. The controlled vocabulary for evidence supporting annotations is available, along with examples of the kinds of experiments that would fall into each category, at http://www.geneontology.org/GO.evidence.html.

Current Development and Future Plans

The GO Consortium meets on a regular basis approximately four times a year. We expect that other model organism database and annotation groups will participate in the development of GO and will use the GO vocabularies and tools in their annotation work. We will continue to collaborate with interested users on the translation of other vocabularies to GO.

Our current development efforts are focusing on the creation of a GO annotation tool and on the enhancement of the GO database. We may extend the vocabulary set to include such useful sets as "cell types" or "tissues." We expect the vocabulary development to be ongoing as we create a resource that can accommodate changes in our understanding of biology.

Summary

The GO project has united several model organism database groups by providing a shared annotation system for describing some primary aspects of organismal biology. The ontologies are already being used by private and public data providers as a method of annotating and cross-referencing their gene and gene product information. The project has enhanced and promoted the development of robust strategies for presenting and querying across extensive classifications. The GO project has been a seminal collaborative work and has resulted in the development and implementation of important biological ontologies and in the development of a model organism information resource.

ACKNOWLEDGMENTS

The GO Consortium is support by NHGRI grant HG02273. In addition, SGD is supported by a P41, National Resources grant

from National Human Genome Research Institute (NHGRI) grant HG01315; MGD by a P41 from NHGRI grant HG00330; GXD by National Institute of Child Health and Human Development grant HD33745; FlyBase by a P41 from NHGRI grant HG00739, and the Medical Research Council, London; and TAIR by the National Science Foundation grant DBI-9978564. The GO Consortium thanks Ken Fasman of AstraZeneca for his support and enthusiasm. We thank Monica Riley and Gretta Serres of GenProtEC and Michelle Gwinn of TIGR for allowing us to incorporate terms and relationships from their own projects. We thank Oxford University Press for permission to reproduce text from the *Oxford Dictionary of Biochemistry and Molecular Biology* within the GO definitions. We also thank AstraZeneca for financial support. The Stanford group acknowledges a financial gift from Incyte Genomics.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al., 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Bairoch, A. and Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28:** 45–48.

Ball, C.A., Dolinski, K., Dwight, S.S., Harris, M.A., Issel-Tarver, L., Kasarskis, A., Scafe, C.R., Sherlock, F., Binkley, G., Jin, H., et al. 2000. Integrating functional genomic information into the *Saccharomyces* Genome Database. *Nucleic Acids Res.* **28:** 77–80.

Blake, J.A., Eppig, J.T., Richardson, J.E., Davisson, M.T., and the Mouse Genome Database Group. 2000. The Mouse Genome Database (MGD): Expanding genetic and genomic resources for the laboratory mouse. *Nucleic Acids Res.* **28:** 108– 111.

Delozier, E.P. and Lingle, V.A. 1992. MEDLINE and MeSH: Challenges for end users. *Med. Ref. Serv. Q* **11:** 29–46.

Gruber, T.R. 1993. A translation approach to portable ontologies. *Knowl. Acq.* **5:** 199–220.

———. 1995. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum. Computer Stud.* **43:** 907–928.

———. 1998. In *Formal ontology in information systems* (ed. N. Guarino), IOS Press, Washington, DC.

Huala, E., Dickerman, A.W., Garcia-Hernandez, M., Weems, D., Reiser, L., LaFond, F., Hanley, D., Kiphart, D., Zhuang, M., Huang, W., et al. 2001. The Arabidopsis information resource (TAIR): A comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.* **29:** 102–105.

Jones, D.M. and Paton, R.C. 1999. Toward principles for the representation of hierarchical knowledge in formal ontologies. *Data Knowl. Eng.* **31:** 99–113.

Karp, P., Riley, M., Saier, M., Paulsen, I.T., Paley, S M., and Pellegrini-Toole, A. 2000. The EcoCyc and MetaCyc databases. *Nucleic Acids Res.* **28:** 56–59.

Lowe, H.J. and Barnett, G.O. 1994. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *JAMA* **271:** 1103–1108.

Oxford Dictionary of Biochemistry and Molecular Biology. 1997. (eds. A.D. Smith, et al.) Oxford University Press, NY.

Riley, M. 1993. Functions of the gene products of *Escherichia coli*. *Microbiol. Rev.* **57:** 862–952.

Ringwald, M., Eppig, J.T., Kadin, J.A., Richardson, J.E., and the Gene Expression Database Group. 2000. GXD: A gene expression database for the laboratory mouse-current status and recent enhancements. *Nucleic Acids Res.* **28:** 115– 119.

Rison, S.C.G., Hodgman, T.C., and Thornton, J.M. 2000. Comparison of functional annotation schemes for genomes. *Func. Integ. Genom.* **1:** 56–69.

Schulze-Kremer, S., Karp, P.D., Musen, M.A., and Altman, R.B. 1998. Ontologies for molecular biology tutorial. 6th International Conference on Intelligent Systems for Molecular Biology, Montreal, Canada.

The FlyBase Consortium. 1999. The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res.* **27:** 85–88.

The Gene Ontology Consortium. 2000. Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25:** 25–29.

Received January 18, 2001; accepted in revised form May 14, 2001.