The Ontology of the Gene Ontology

Barry Smith, PhD[1,2], Jennifer Williams[3] and Steffen Schulze-Kremer, PhD[4]

[1]Institute for Formal Ontology and Medical Information Science, University of Leipzig
[2]Department of Philosophy, University at Buffalo, NY
[3]Ontology Works, Inc., Odenton MD 21113
[4]Institute of Informatics, Free University of Berlin

Published in Proceedings of AMIA Symposium 2003

The rapidly increasing wealth of genomic data has driven the development of tools to assist in the task of representing and processing information about genes, their products and their functions. One of the most important of these tools is the Gene Ontology (GO)*, which is being developed in tandem with work on a variety of bioinformatics databases. An examination of the structure of* GO*, however, reveals a number of problems, which we believe can be resolved by taking account of certain organizing principles drawn from philosophical ontology. We shall explore the results of applying such principles to* GO *with a view to improving* GO*'s consistency and coherence and thus its future applicability in the automated processing of biological data.*

INTRODUCTION

One of the most important tools for the representation and processing of information about gene products and functions is the Gene Ontology (GO).[1] GO is being developed in tandem with work on a variety of biological databases within the framework of the umbrella project OBO (for: open biological ontologies).[2] It provides a controlled vocabulary for the description of cellular components, molecular functions, and biological processes.

Representatives from a number of groups working on model organism databases, including FlyBase (Drosophila), the Saccharomyces Genome Database (SGD) and the Mouse Genome Database (MGD), initiated the Gene Ontology project in 1998 in order to provide a common reference framework for the associated controlled vocabularies.[3]

As of June 19, 2003 GO contains 1297 component, 5396 function and 7290 process terms. The total number of GO informal term definitions is 11020. Terms are organized in parent-child hierarchies, indicating either that one term is more general than another or that the entity denoted by one term is part of the entity denoted by another.

Database compilers create associations between GO terms and entries for genes or gene products in their databases in order to describe the processes in which the latter are involved. Typically, such associations (or 'annotations') are first of all established electronically and later validated by a process of manual curation which requires the annotator to have expertise both in the biology of the genes and gene products and in the structure and content of GO.

GO AND ONTOLOGY

The Gene Ontology, in spite of its name, is not an ontology as the latter term is commonly used either by information scientists or by philosophers. It is, as the GO Consortium puts it, a 'controlled vocabulary'. Information scientists tend to view ontologies as terminologies with which axioms and definitions are associated, formulated (for example in some Description Logic framework) in ways which make them suitable for supporting software applications. Philosophers generally regard ontologies as theories of the different types of entities (objects, processes, relations) existing in given domains.[4] Where information systems ontologists seek to maximize reasoning efficiency even at the price of simplifications on the representational side, philosophical ontologists argue that both logically rigorous formalization and representational adequacy can bring benefits both for the stability of an ontological framework and for its extendibility in the future.

GO's primary focus is not ontological in either sense. Certainly it uses hierarchies of terms. But its authors have focused neither on software implementations nor on the logical expression of the theory encompassing these terms. Rather their efforts have been directed toward providing a practically useful framework for keeping track of the biological annotations that are applied to gene products.[5]

This means that when faced with the trade-off between (1) formal and ontological coherence, stability and scalability, and (2) the speedy population of GO with biological concepts, preference was given by the GO consortium to the latter. Too little attention was thereby paid to the significance of those ontological or quasi-ontological terms – such as *function*, *part*, *component*, *substance*, *action*, *domain*, *complex* – which were employed in GO's construction.

GO, GONG and Protégé-2000: GO's success in serving the biological community has led some researchers to attempt to expand its utility by using GO in tandem with bioinformatics applications. Some are now seeking to use GO as a basis for replacing manual comparison of the properties of sets of gene products with automatic reasoning. The Gene Ontology Next Generation (GONG) project[6] is attempting to improve GO's suitability for use by computers by rendering GO in a description logic.[7] Another effort applies the Protégé 2000[8] frame-based[9] ontology editor and associated tools to browse and edit GO and to verify certain kinds of consistency within GO.[10]

All of these efforts, however, accept GO as it is, and seek to supplement it with formal reasoning tools. Here, in contrast, we show that the existing architecture of GO harbors problems which stand in the way of such formalizing efforts. We show how, by taking account of certain organizing principles drawn from philosophical ontology, GO's consistency and coherence and thus its future applicability in the automated processing of biological data might be enhanced.

The Need for Expert Knowledge: One problem faced by GO's curators is that "Many molecular functions and biological processes do not exist in all organisms".[5] For example, only certain kinds of organisms have the cellular component known as *glycosome*; only certain kinds of organisms are susceptible to *budding*. Sometimes the needed supplementary information is included in a GO term's definition, but where such information is not included one must rely on manual inspection and expert knowledge to determine under what set of conditions GO's stated relations are applicable (e.g. that *glycosome is part-of cytoplasm* only for Kinetoplastidae). This places severe obstacles in the way of using GO as a basis for computer applications; for the latter do not have access to expert knowledge about how relations stated with perfect generality in GO should have their applications restricted in specific contexts.

GO's TRIPARTITE ARCHITECTURE

GO is divided into three disjoint term hierarchies: the *cellular component*, the *molecular function*, and the *biological process ontologies*.

The Cellular Component Ontology: The vocabulary of GO's cellular component ontology consists of terms such as *flagellum*, *chromosome*, *ferritin*, *extracellular matrix* and *virion*. This ontology is the GO counterpart of anatomy within the medical framework. It is intended to allow biologists to note the physical structure with which a gene or gene product is associated. GO includes in this vocabulary both the extracellular environment of cells and the cells themselves (that is, *cell* is subsumed in GO by *cellular component*). Cellular components are physical and measurable entities. They are, in the terminology of philosophical ontology, *substances* or *independent continuants* (things, or objects). They endure through time.

The Molecular Function (Activity) Ontology: The GO definition of *molecular function* is: "the action characteristic of a gene product." *Molecular function* accordingly subsumes terms describing actions, for example *ice nucleation*, *binding*, or *protein stabilization*, entities which do not *endure* but rather *occur*. Until recently the reading of "function" as meaning *action* was beset by some confusion in virtue of the fact that the molecular function hierarchy includes terms such as *anti-coagulant* (defined as: "a substance that retards or prevents coagulation") and *enzyme* (defined as: "a substance … that catalyzes"), which refer neither to functions nor to actions but rather to substances. This confusion has been remedied to a degree by a policy change effective as of March 1, 2003 whereby "All GO molecular function term names [with the exception of the parent term *molecular function* and of the whole node *binding*] are to be appended with the word 'activity'."[3] Thus the change solves problems with terms such as *structural molecule*, which is defined as meaning: "the action of a molecule that contributes to the structural integrity". As "molecule" is normally used, of course, the term *structural molecule* refers not to an action, but rather to that which *performs* an action. However, because only names have been changed, but not associated definitions, some inconsistencies still remain.

The Biological Process Ontology: A biological process is defined in GO as: "A phenomenon marked by changes that lead to a particular result, mediated by one or more gene products". Biological process terms can be quite specific (*glycolysis*) or very general (*death*). Molecular function and biological process terms are clearly closely interrelated. The process of anti-apoptosis, for example, certainly involves the molecular function now labeled *apoptosis inhibitor activity*. GO's curators attempt to clarify the relationship as follows: "A biological process is accomplished via one or more ordered assemblies of molecular functions."[11] This would suggest that molecular functions are *initiators* of biological processes – "activity", unlike "process", connotes *agency*. But it would suggest also that they stand to such processes in a *part-of* relation. At the same time, however, GO's authors insist that *part-of* holds only between entities within a single hierarchy and never *between* the three GO vocabulary sets, and in general they

have provided too little guidance as to the role of the different sorts of temporal entities within their ontology as also as to the relations between the three term-hierarchies by which GO is constituted. We shall return to these problems below.

ONTOLOGICAL DISTINCTIONS

As GO increases in size and scope, it will, as the GO Consortium accepts, "be increasingly difficult to maintain the semantic consistency we desire without software tools that perform consistency checks and controlled updates".[5] As the Gene Ontology expands, therefore, and thus becomes more useful to researchers, its semantic integrity will at the same time become more difficult to maintain through manual inspection and curation. The addition of each new term will require the curator to understand the entire structure of GO in order to avoid redundancy and to ensure that all appropriate linkages are made with other terms. One method to improve on the current approach would be to make explicit the criteria used for discriminating subclassifications by introducing a decision-tree methodology into the construction of each hierarchy. This would involve the recording of explicit statements as to the basis for given classification choices and thus enable retrospective checking. In addition, it would facilitate the more reliable identification of concepts already in the ontology.[12] We believe, however, that to reap the full benefits from this methodology certain distinctions drawn by philosophical ontologists need to be kept in mind:

Universals vs. particulars: Philosophers distinguish between *universals* (also called kinds, species, types) and *particulars* (also called individuals, exemplars, instances, tokens). Examples of universals are: the species *E. coli*, the function: *to boost insulin production*. Examples of particulars are: the *E. coli* bacterium now existing in this Petri dish, the function of this gene to boost insulin production in these beta cells in your pancreas. GO terms correspond, in philosophical terminology, to universals, that is to entities which are multiply instantiable. Thus the universal corresponding to the term *Cell* is instantiated by every actual cell.

Continuants vs. Occurrents: Orthogonal to the distinction between universals and particulars is that between *continuants* and *occurrents*. Continuants, as the name implies, are entities which continue to exist through time. Organisms, cells, chromosomes are all continuants: they preserve their identity from one moment to the next, even while undergoing a variety of different sorts of changes. The parts of continuants – for example your arms and legs – are also continuants. The principal mark of a continuant is that, if it exists *at* a time, then so also do all its parts.

Occurrents (also called events, processes, activities) are in contrast never such as to exist in full in a single instant of time; rather, they are such as to unfold themselves in their successive phases, in the way in which, for example, a process of viral infection unfolds itself through time. Processes characteristically have a beginning, a middle and an end. Where your *arm* is part of you, your *youth* is part of that process which is your life. Note that part-whole relations never cross the continuant/occurrent divide.

Dependent vs. Independent: Some entities (planets, people, molecules, atoms) have an inherent ability to exist without support from other entities. Others require such support in order to exist: a viral infection is dependent upon a virus and upon the organism which is infected; the function of an organ is dependent upon the existence of the organ.

Continuants and Occurrents in GO: The continuant/occurrent opposition corresponds in first approximation to the distinction between substances (objects, things) and processes. GO's cellular component ontology is in our terms an ontology of substance universals; its molecular function and biological process ontologies are ontologies of function and process universals. But functions, too, are from the perspective of philosophical ontology continuants. For if an object has a given function – which means a *token* function – for a given interval of time, then this token function is present in full at every instant in this interval. It does not unfold itself in phases in the manner of an occurrent. If, however, the token function gets *exercised*, then the token *process* that results does indeed unfold itself in this manner. Each function thus gives rise, when it is exercised, to processes or activities of characteristic types.

It is tempting, in light of this, to conceive the relation between GO's molecular function hierarchy and its biological process hierarchy as one of function to exercise of function. Biological processes would then be accomplished when assemblies of molecular functions are exercised in ways which reflect the assemblies of cellular components which the given functions are the functions of. A further advantage of this solution is that it allows us to do justice to the fact that a function continues to exist even when dormant or is for some other reason not being exercised.

The relabeling step of March 2003makes clear, however, that a solution along these lines is not what GO's authors have in mind. Relabeling 'functions' as 'activities' signifies that GO's authors hold molecular functions to be occurrent rather than continuant entities. But if this is so, then for clarity's sake they

should take the further step of relabeling the *molecular function* hierarchy a *molecular activity* hierarchy.

If we understand this relabeling step correctly, 'function' in fact means *functioning*, it signifies the function as exercised rather than as the potential to be exercised. Thus to assign a function/activity to a gene product is not necessarily to assert that this product possesses at all times the potential to perform the given activity. On the other hand however the GO Introductory Documentation[11] still states that 'Molecular function defines the tasks that a physical gene product (or gene product group) does *or has the potential to do*.' (Emphasis added) We believe that GO here considers 'potential' to indicate the ability to perform a given activity *under appropriate circumstances*. For instance, a transporter protein complex would not be expected to demonstrate transporter activity when the substrate it transports is not present. Thus when all molecular function nodes have been relabeled as activities, the term 'activity' still does not mean 'activity', but rather (something like): potential for activity under certain circumstances.

We believe that the failure to resolve such problems reveals itself in coding errors, for instance in the definitions of terms such as *transporter* (GO: 0005215). Currently (June 19, 2003) this is defined as: "Enables the directed movement of substances (such as macromolecules, small molecules, ions) into, out of, or within a cell." Transporter *activity*, however, would more properly be defined as this directed movement itself, or better still: as the catalysis of this movement.

But now again: how are we to resolve the problem of the relation between the molecular function and the biological process hierarchies? The solution in terms of function and exercise is no longer available, yet given that part-whole relations are not allowed to traverse GO hierarchies, we do not know what solution could replace it.

Dependent and Independent Entities in GO: Where GO's cellular component ontology contains terms denoting independent entities, its function/activity terms denote dependent entities, which means entities which have a necessary reference to the substances in which they inhere. Thus a binding function/activity is dependent on the several molecules (or molecule parts) involved when binding occurs.

The biological process hierarchy, too, encompasses dependent entities. These are occurrents that require support from some substance in order to allow them to occur. Consider for example the term *germination* (GO:0000844), defined as "The physiological and developmental changes by a seed, pollen grain, spore or zygote that occur after release from dormancy". Here it is evident that the process of germination can only be expressed by means of some substance. Another clear example is *viral life cycle* (GO:0016032), defined as: "A set of processes by which a virus reproduces".

GO RELATIONS

The Relation *isa*: Although GO documentation refers to *isa* as meaning *instance of*,[13] the *isa* relation is clearly used in such a way to indicate *is a kind of* or specialization relations between universals (e.g., "a eukaryotic cell is a cell"), rather than instantiation relations between particular entities and their universal kinds. The *isa* relation is distinct also from the relation of part to whole. Confusingly, however, *isa* is sometimes also used with the meaning *part-of*, as in the definition of *lysosomal hydrogen-transporting ATPase V0 domain* (GO:0046610), which treats the V0 and V1 domains as kinds of V-type complexes, rather than as component parts thereof. We believe that such errors derive, again, from a lack of attention to ontological principles.

The *isa* relation in its intended meaning indicates a *necessary* relationship. That is, when we say "*eukaryotic cell isa cell*", we mean that *every* eukaryotic cell is a cell. However, there are cases in GO where *isa* is used to indicate non-necessary subsumption. For example, the term *transport* was defined in GO (as of June 19, 2003) as meaning: "The directed movement of substances (such as macromolecules, small molecules, ions) into, out of, or within a cell." The term *cell motility* as: "Any process involving the controlled movement of a cell." The term *cell growth and/or maintenance* as: "Any process required for the survival and growth of a cell." Now, however, the first two are connected by *isa* to the last. The GO statement: *transport isa cell growth and/or maintenance*, is to be read as indicating that *every* transport process is a cell growth and/or maintenance process, which is however not true of transports such as *viral intracellular protein transport* (GO:0019060).

The Relation *part-of*: The intended meaning of *part-of*, as explained in the GO Usage Guide, is: "*can be a part of*, **not** *is always a part of*". In addition, the *part-of* relation is intended to behave transitively.[13] GO uses *part-of* for representation of parts of both substances and processes (e.g. *activation* is *part-of fertilization*) and of functions/activities. *Part-of* appears also in each of the following kinds of statements:

– *membrane part-of cell*, intended to mean "a membrane is a *part-of* any cell"

– *flagellum part-of cell*, intended to mean "a flagellum is *part-of* some cells"

– *replication fork part-of cell*, intended to mean: "a replication fork is *part-of* the cell (nucleoplasm) only during certain times of the cell cycle"
– *regulation of sleep part-of sleep*, which should be corrected to: "regulation of sleep is co-located with and is causally involved with the sleep process".

We believe that each of these usages should be represented as a different relation, and that the system of relations involved should be explicitly presented. This means also extending the ontology to take explicit account of the notion of time and of the ways in which time is involved for example in determining different *part-of* relations. A revision along these lines will in addition bring much greater adaptability to the purposes of automated reasoning about dynamic aspects of biological phenomena.

CONCLUSION

Benefits of the GO Approach: The GO approach has brought considerable benefits:

1) Work on populating GO could start immediately, without its authors needing to solve some of the intricate problems which face ontologies when formalized as logical theories.

2) Populating GO does not require the completion of complex protocols of formally determined steps but can be done intuitively by the expert biologist.

3) There are few formal constraints standing in the way of easy incorporation of existing controlled vocabularies from the biological domain.

4) The principle of unique identifiers allows GO terms to be used for database annotation without consideration of their place in the GO hierarchy.

Drawbacks of the GO Approach: Focusing on the rapid population of GO has, however, had its drawbacks:

1) It is unclear what kinds of reasoning are permissible on the basis of GO's hierarchies.

2) The rationale of GO's subclassifications is unclear. The reasoning that went into current choices has not been preserved and thus cannot be explained to or re-examined by a third party.

3) No procedures are offered by which GO can be validated.

4) There are insufficient rules for determining how to recognize whether a given concept is or is not present in GO. The use of a mere string search presupposes that all concepts already have a single standardized representation, which is not the case.

A Modest Proposal. Our work here has consisted of no more than an initial analysis of GO terms and organization with respect to some basic ontological distinctions. Our proposal is to use these basic distinctions to determine a revised upper-level ontology for GO in which the relations between the three existing hierarchies and the roles of GO's central ontological notions (including the various existing uses of *part-of*, as also of *process*¸ *action*, *activity*, etc.) would be clearly specified and branches transplanted or divided accordingly. At the same time, definitions should be formulated in a way that records the conditions under which stated relations are applicable.

It is no easy to task to determine how large an intellectual investment this proposal would require (although some necessary steps – for instance the substitution of "activity" for "function" in the name of GO's *molecular function ontology* – would involve no cost at all). We are however convinced that the benefits will significantly outweigh the costs involved. For as ever more biological concepts come to be represented formally, the introduction into GO of a robust ontological architecture along the lines here suggested would:

1) help to avoid coding errors,

2) ensure that computer systems will be able to assume more of the burden of ontology curation,

3) ensure that such systems are better able to use GO as a basis for automated reasoning,

4) facilitate GO's interoperability with other biological databases and terminology systems and with associated ontologies.

REFERENCES

1. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nature Genet. 2000; 25: 25-29.
2. http://obo.sourceforge.net.
3. http://www.geneontology.org.
4. Smith B. Ontology. In L. Floridi, editor, Blackwell companion to philosophy, information and computers, Oxford, 2003.
5. Gene Ontology Consortium. Creating the Gene Ontology resource: Design and implementation. Genome Res. 2001; 11: 1425-1433.
6. http://gong.man.ac.uk/background.html.
7. Wroe CJ, Stevens R, Goble CA. A methodology to migrate the gene ontology to a description logic environment using DAML+OIL. Pacific Symposium on Biocomputing 2003; 8: 624-635.
8. http://protege.stanford.edu.
9. Fikes R and Kehler T. The role of frame-based representations in reasoning. Commun. ACM 1985; 28: 904-920.
10. Yeh I, Karp PD, Noy NF, Altman RB. Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO). Bioinformatics 2003; 19 (2): 241-248.
11. Gene Ontology general documentation. http://www.geneontology.org/doc/GO.doc.html

12. Schulze-Kremer S. Ontologies for molecular biology. Pacific Symposium on Biocomputing 1998; 3: 693-704.
13. Gene Ontology usage guide: http://www.geneontology.org/doc/GO.usage.html.

Acknowledgments: This work was supported by the Wolfgang Paul Program of the Alexander von Humboldt Foundation. Thanks are due also to Bill Andersen, Olivier Bodenreider, Louis Goldberg and Jane Lomax.