

Gene Ontology (GO) for microbe-host interactions and its use in ongoing annotation of three *Pseudomonas syringae* genomes via the Pseudomonas-Plant Interaction (PPI) website

C. W. Collmer,^{1,2*} M. Lindeberg,¹ and A. Collmer¹

¹Cornell University, Ithaca, NY, 14850, U.S.A., and ²Wells College, Aurora, NY, 13026, U.S.A.

NOTE: This chapter will be appearing in Proceedings from the 7th International Conference on *Pseudomonas syringae* Pathovars and Related Pathogens (ICPSRP), held in Agadir, Morocco, November 12-16, 2006. The original publication of this chapter is available at www.springerlink.com.

Genome-scale sequencing of plant pathogens is an increasingly important tool for exploring host-pathogen interactions. While comparative structural analyses of three *Pseudomonas syringae* pathovars that differ in pathogenicity (*P. s. pv. tomato* DC3000, *P. s. pv. phaseolicola* 1448A, and *P. s. pv. syringae* B728a) offer valuable insights, a second type of analysis based on the ability to find and compare gene products with similar function (but possibly different structure) across even diverse pathogens is now possible through genome annotation using the controlled vocabularies of the Gene Ontology Consortium (GO; <http://www.geneontology.org>). As members of the Plant-Associated Microbe Gene Ontology (PAMGO) Interest Group (<http://pamgo.vbi.vt.edu>), we have been developing precisely-defined GO terms that describe the biological processes involved in a microbe's interactions with its host (e.g. GO:0044409, entry into host; GO:0044002, acquisition of nutrients from host). Using these terms for annotating the genes of prokaryotic (or eukaryotic) pathogens (or mutualists) that attack plant (or animal) hosts promises to offer a new mechanism for viewing the diversity of microbe strategies for overcoming common host challenges (e.g. finding all gene products across diverse microbes that are annotated to GO:0044414, suppression of host defenses). However, because genome annotation and analysis must continue over time to match experimentation and discovery, genome projects must provide portals for continually updating genome sequence and annotation data. The Pseudomonas-Plant Interaction (PPI) website (<http://pseudomonas-syringae.org>) exemplifies the "hub model" of web-based information management for small-scale genome projects. This model maximizes access to the most current analytical tools (e.g. the Artemis Genome Viewer) through links to primary off-site resources, while tailoring instructional tutorials, specific information, and organization to the needs of the *P. syringae* community. The website also serves as a portal for submission of updated genome sequence or GO annotation data as it becomes available.

Key words: plant pathogenesis, bacterial pathogen

*Corresponding author: Candace W. Collmer, ccollmer@wells.edu

Full-genome sequencing has the potential to unveil formerly hidden microbial secrets, if the right questions can be formulated and the tools to answer them are available. Thus, important and obvious questions in relationship to plant pathogenesis could be posed once full sequences were available for three *Pseudomonas* species---*P. syringae* pv.

tomato DC3000, a plant pathogen; *P. aeruginosa* PA01, an animal pathogen; and *P. putida* KT2440, a non-pathogen. Equally interesting questions arose with the availability of genomic sequences for three *P. syringae* pathovars that differ in pathogenicity---*P. s. pv. tomato* DC3000, which causes bacterial speck on tomato and *Arabidopsis*; *P. s. pv. phaseolicola* 1448A, the causal agent of halo blight on bean; and *P. s. pv. syringae* B728a, which causes brown spot on bean. To address these questions, genome-scale analyses are essential. However, the tools for addressing genome-wide questions that focus specifically on plant pathogenesis have until recently been rather limited. That limitation has in part derived from the lack of universally understandable terms that can be used to tag and fully annotate genes in microbes that function in pathogenesis. In this paper, we report progress to date in working with the Gene Ontology (GO) Consortium (<http://www.geneontology.org>) to develop precisely-defined GO terms that describe the biological processes involved in a microbe's interactions with its host, including those of pathogenesis. In addition, we describe features of the Pseudomonas-Plant Interaction (PPI) website (<http://pseudomonas-syringae.org>) that allow for the viewing as well as the continuous updating of genome sequence and annotation data as experimentation and discovery continue on three *P. syringae* pathovars.

In the process of genome-wide analysis, it is relatively straightforward to do informative structural comparisons by aligning genes and even whole genomes; such a comparison could show, for example, that a particular gene active in one genome is inactivated by insertion of a transposable element in others. However, the ability to find and compare gene products with similar function (regardless of structural similarity) across genomes has been much harder. Comparisons based on the functions of gene products require not only that functions be assigned to genes in the process of annotation, but also that the vocabulary for describing such functions be universally understandable. Indeed, the annotation of the full genome sequence of *P. syringae* pv. *tomato* DC3000 in collaboration with scientists at The Institute for Genomic Research (TIGR) included the assignment of genes to broad, prokaryotic role categories derived from those of Riley (1993); a comparison of the relative number of genes assigned to these different cellular roles revealed that DC3000 differs significantly from both the animal pathogen *P. aeruginosa* PA01 and the non-pathogen *P. putida* KT2440 in having greater transport capacity for sugars, but less for amino acids (Buell *et al.*, 2003). Recognition of the utility of such comparisons coupled with the need for a much larger number of functional categories had already led to TIGR membership in the Gene Ontology Consortium, and it was TIGR scientists who spear-headed the development of GO terms appropriate for prokaryotic gene functions.

Since 1998, the GO Consortium has been developing three ontologies---Molecular Function, Biological Process, and Cellular Component---each comprised of a hierarchical-like arrangement of terms whose definitions and interrelationships are precisely defined (Gene Ontology Consortium, 2001). As the original developers of GO were associated with the mouse, yeast (*Saccharomyces*), and fly (*Drosophila*) genome projects, the early versions of the ontologies were focused on eukaryotic cells; as mentioned above, terms related to prokaryotes were integrated into the continually developing ontologies as their need for the annotation of prokaryotic genomes became apparent. Terms assigned during annotation to a particular gene from each of the three ontologies tell what the gene's product does (its molecular function), where it does it (its cellular component), and why it does it (as a part of what larger biological process). Each GO term (or phrase) within these ontologies has a unique GO identifier, which once annotated to a gene product in one organism then allows searches for genes with the same identifier (e.g. function or process) in other organisms, regardless of whether the products of these different genes share structural similarity. While the three GO ontologies have been continually expanding since 1998, at the time the *P. syringae* pv. *tomato* DC3000 genome was annotated, there were few GO terms appropriate for annotating gene products implicated in the virulence of plant pathogens.

However, the need for such terms, and for GO annotation in general, as a tool for answering critical questions about plant pathogenesis through cross-genome comparisons is easily apparent. Without GO terms, searching across genomes for gene products with similar functions is difficult unless they also share structural features. For example, while gene names can be searched, they are often based on idiosyncratic phenotypes, which obscure gene function (or address only one of several). Secondly, while one can search databases for specific function terms in Keywords or Comments fields, such searches are doomed to be incomplete because of inconsistent terminology (e.g. "attachment," "adhesion," "prepenetration activity" could be used by different authors to describe similar events). In addition, there is presently no way to systematically add new information on the function of gene products in a manner that supports comparison across genomes. A good example of the problem is found in the genes encoding the type II protein secretion pathway in bacterial plant pathogens. The outer membrane pore protein is known as XcpD in *Pseudomonas*, OutD in *Erwinia*, and XpsD in *Xanthomonas*. Worse, there can be common components involved in type II protein secretion and type IV pilin biogenesis, and one cannot tell from the name designation whether a given protein serves one or both systems. Because both systems are important in pathogenesis, they have been extensively studied in several different organisms. But new information on the

functions of secretion system components is not accessible in any systematic way through current sequence-oriented searches of databases. Information available on a gene product in one organism could prove useful for understanding a comparable one in another if their functional relationships were easily recognizable. For instance, learning experimentally that a gene product is involved in type IV pilin biogenesis could prompt a search for gene products in other organisms known to be involved in the same process; finding that one of these was also annotated for a role in type II protein secretion could suggest a meaningful experimental question relating to the gene product of the original organism.

It was the recognition of the potential utility of such searches for answering critical questions about plant pathogenesis, and of the need for more GO terms related to pathogenesis, that led in 2004 to the formation of the Plant-Associated Microbe Gene Ontology (PAMGO) interest group. As members of that group, we have been working with the GO Consortium to develop precisely-defined GO terms that describe the biological processes involved in a microbe's interactions with its host (Gene Ontology Consortium, 2006). The PAMGO interest group (<http://pamgo.vbi.vt.edu>) comprises scientists from six institutions working on genome projects of both prokaryotic and eukaryotic microbial plant pathogens---bacteria, oomycetes, fungi, and nematodes. Our original goal was to develop higher level biological process terms appropriate for annotating genes in these organisms that had been implicated in their pathogenic interactions with plants. However, we soon realized the broader utility of developing terms that were as general as possible, relevant not only for pathogens of all kingdoms, but also for the whole range of host-microbe interactions (from mutualism through parasitism) and for all hosts, animals as well as plants. Given that microbes of all types (whether prokaryotic or eukaryotic) that approach a plant to initiate a relationship (whether ultimately mutualistic or pathogenic) must all begin by recognizing their host, the utility of a GO term that could be used to find all annotated microbe genes involved in host recognition seemed obvious. In addition, the fact that at least some biological processes involved in pathogenesis are shared by both plant and animal pathogens (e.g. the injection of pathogen-encoded effectors into host cells via a type III secretion system) underscored the desirability of broad GO terms that could serve both communities.

Using such broad GO terms for annotating the genes of prokaryotic (or eukaryotic) pathogens (or mutualists) that interact with plant (or animal) hosts promises to offer a new mechanism for viewing the diversity of microbe strategies for overcoming common host challenges (e.g. finding all gene products across diverse microbes that are annotated to the [PAMGO-developed] term GO:0044414, suppression of host defenses). However, not all

GO terms are broad ones; GO terms are imbedded in the ontologies within “tree” structures (directed acyclic graphs, or DAGs) that show their relationships as “children” of broader terms or “parents” of more specific ones. Thus, while microbial gene products that effect recognition of either a plant or an animal host could all be collected by a search using the broad term “recognition of host,” that broad term could have more specific “children” terms that describe either plant-specific or animal-specific recognition processes appropriate for the annotation of some gene products but not others. In addition, such a hierarchical-like structure allows for the annotation of genes with terms denoting broad functional categories, or to terms with more specificity as knowledge accrues.

PAMGO members began in 2004 with a proposal to the GO Consortium for about thirty higher level (broad) biological process terms describing the range of interactions between microbes and their hosts. This proposal generated intensive discussion before, during, and after its presentation at a GO Content meeting in August of 2004. Following the consideration of three alternative “tree” options at a GO Consortium meeting in September, a final set of thirty-five new terms was submitted to GO in December, 2004, and accepted into the GO Biological Process ontology in January, 2005 (Gene Ontology Consortium, 2006). The process of developing these terms included broad-ranging discussion across the wider GO community about the definitions of terms such as pathogenesis and symbiosis. In the end, a broad definition of the parent term symbiosis was adopted, indicating an interaction between organisms including parasitism, commensalism, and mutualism, and acknowledging that the three are not always discreet categories of interactions but rather occur on a continuum of interaction ranging from parasitism to mutualism. Most of these terms are shown in Figure 1 as children of the parent term GO:0044404, symbiotic interaction between host and other organism.

In addition to providing precise definitions for all terms, as well as the option of annotating a gene product at whatever level of specificity is warranted by the available information (e.g. “binding” versus “protein kinase binding”), annotation using the Gene Ontology also requires the concomitant assignment of an appropriate Evidence Code. The thirteen Evidence Codes indicate the type of data on which each annotation is based, which in turn provides some measure of the relative degree of certainty in the annotation. For instance, an annotation that is “Inferred from Direct Assay” is more reliable than one “Inferred from Electronic Annotation.”

As genome annotation using the new GO terms began in 2005-2006 for genes implicated in the virulence of *Pseudomonas syringae* pathovars, *Erwinia chrysanthemi* 3937, and *Phytophthora* spp., it was immediately obvious that more specific GO terms were needed to capture information available in the published literature. Accordingly,

a three-day jamboree for GO term creation in July, 2006, which involved five PAMGO members and two members of the GO editorial board, produced a draft list of several hundred GO terms that are more specific children to the

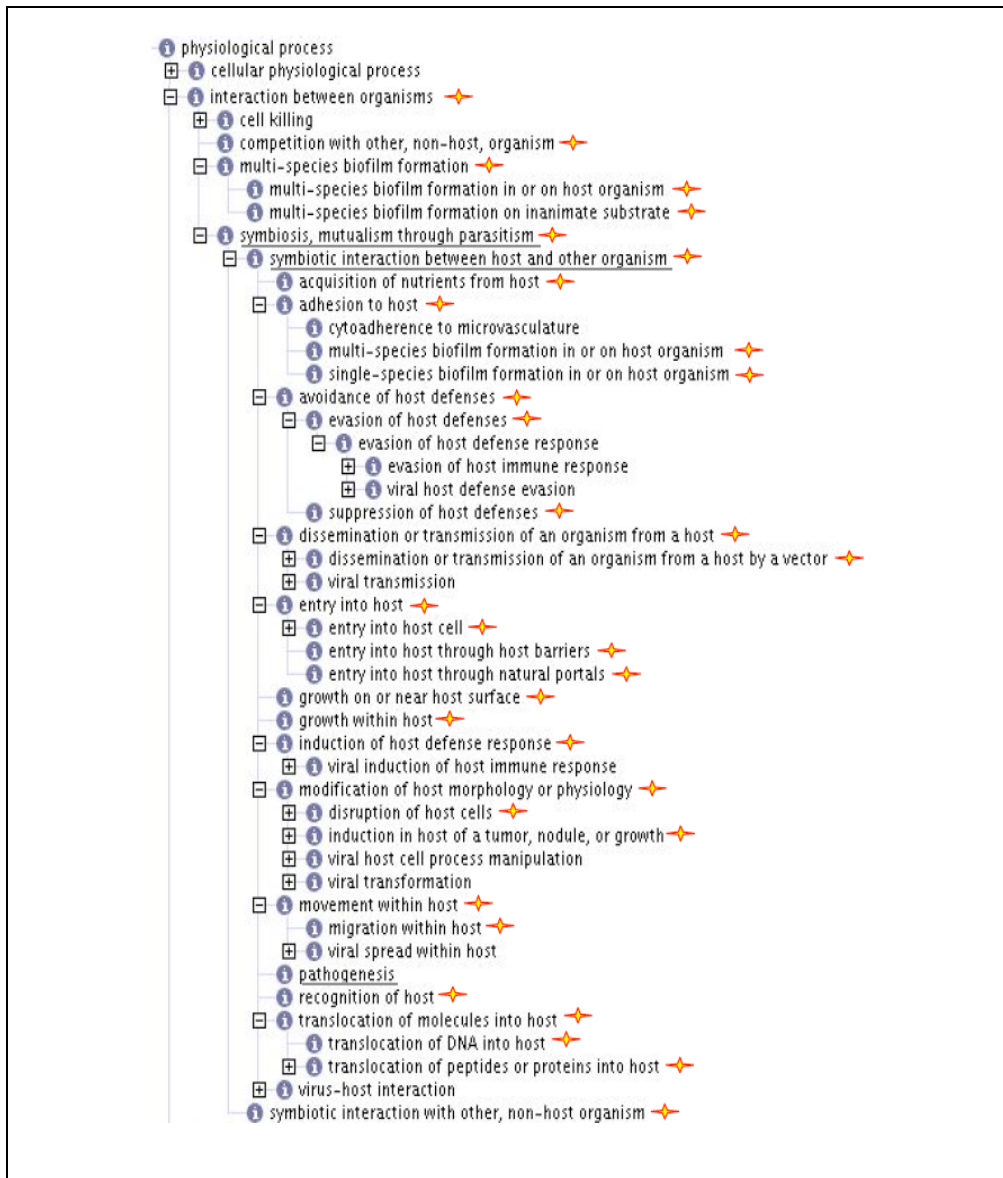


Figure 1. Newly developed GO terms for the Biological Process ontology, showing the parent-child relationships among terms. All terms marked by a star were developed by PAMGO and added to GO in January, 2005. The unmarked terms were already in GO but were integrated as children to newly-developed PAMGO terms when appropriate.

original 35 broader terms. While these newly proposed terms describe many biological processes involved in the interaction between a microbe and its host, a large number are children of GO:0044003, modification of host morphology or physiology. Examples of new, more specific terms include, for example, “induction by other

organism of host defensive cell wall thickening,” “modification of host morphology or physiology via protein secreted by type III secretion system,” and “induction by other organism of host phytoalexin production.” These new terms are still in the process of review but should soon be integrated into the GO and be available for the continuing annotation of the genomes of plant pathogens. And as research continues to provide new insights and understanding, the GO ontologies will continue to grow.

Because genome annotation and analysis must continue over time to match experimentation and discovery, whole genome sequencing projects must provide portals for not only presenting but also continually updating genome sequence and annotation data. To meet this need for the *Pseudomonas syringae* pathovars, a website was developed to provide general resources for navigating and viewing the genome data and for communicating updates to the annotation. Few smaller-scale genome projects have access to the extensive foundational resources that support websites for the major model organisms, characterized as they are by extensive on-site databases and a large support staff with specialized skills. The Pseudomonas-Plant Interaction (PPI) website (<http://pseudomonas-syringae.org>), in contrast, exemplifies the “hub” model of web-based information management, offering a model by which researchers can create and maintain websites to leverage the data from smaller scale projects. The two chief hallmarks of the “hub” model are (i) emphasis on the use of pre-existing, primary databases and analytical tools whenever possible, and (ii) maintenance by personnel familiar with the organism and embedded within the research community. The direct use of primary databases and tools has many advantages, including the fact that this approach reduces the expense of developing and maintaining on-site databases in a personnel-intensive manner, and that by directly accessing primary databases and tool development sites, users are ensured of access to the most up-to-date resources. Furthermore, this approach allows the “hub” model website to be maintained primarily by biological researchers themselves, who can design and adapt the site to meet the evolving needs of the research community while focusing limited time and personnel power on high-priority organism-specific resources.

In its current form, the PPI site is a central hub for genome research on the three *P. syringae* pathovars with a special focus on navigating the existing annotation, viewing the genome, and communicating annotation updates. The GenBank genome annotation records for each strain are used as the primary database, with updates to those annotation records being forwarded there directly. In addition, the recent, intense interest in Hop effector proteins has made development of an on-site Hop database a high priority for the *P. syringae* research community. The Hop database currently maintains a regularly updated list of characterized proteins together with information related to

their nomenclature, phylogeny, and phenotypic characteristics. Users interested in viewing the genome are directed to the Artemis genome viewer and Artemis Comparison Tool developed at Sanger rather than to an on-site genome viewer, with the PPI site focusing on Artemis-related resources tailored to the needs of the *P. syringae* research community. Resources include tutorials for viewing the *P. syringae* genomes in Artemis, and Artemis-readable files that allow researchers to view *hop* genes and other virulence factors of special interest to researchers by location in the genome. The PPI website also serves as a portal for submission of updated genome annotations, such as translational start sites, which come from either personal communication or review of the literature. These are then forwarded to GenBank and the Comprehensive Microbial Resource (CMR) at TIGR. GO annotation data also can be communicated through the PPI website and from there to TIGR, where it becomes part of annotated genome updates sent from TIGR to the GO Consortium databases, ensuring broad accessibility..

ACKNOWLEDGEMENTS

The work of the Plant-Associated Microbe Gene Ontology (PAMGO) interest group has been supported by NSF/USDA award #EF-0523736. Summer work by C.W.C. was supported by NSF award #DBI-0077622.

NOTE: This chapter will be appearing in Proceedings from the 7th International Conference on *Pseudomonas syringae* Pathovars and Related Pathogens (ICPSRP), held in Agadir, Morocco, November 12-16, 2006. The original publication of this chapter is available at www.springerlink.com.

REFERENCES

- Buell, C.R., Joardar, V., Lindeberg, M., Selengut, J., Paulsen, I.T., Gwinn, M.L., Dodson, R.J., Deboy, R.T., Durkin, A.S., Kolonay, J.F., Madupu, R., Daugherty, S., Brinkac, L., Beanan, M.J., Haft, D.H., Nelson, W.C., Davidsen, T., Liu, J., Yuan, Q., Khouri, H., Fedorova, N., Tran, B., Russell, D., Berry, K., Utterback, T., Vanaken, S.E., Feldblyum, T.V., D'Ascenzo, M., Deng, W.-L., Ramos, A.R., Alfano, J.R., Cartinhour, S., Chatterjee, A.K., Delaney, T.P., Lazarowitz, S.G., Martin, G.B., Schneider, D.J., Tang, X., Bender, C.L., White, O., Fraser, C.M., and Collmer, A. (2003) The complete sequence of the Arabidopsis and tomato pathogen *Pseudomonas syringae* pv. *tomato* DC3000. Proc. Natl. Acad. Sci. USA 100 (18): 10181-10186.
- Gene Ontology Consortium. (2001) Creating the Gene Ontology resource: design and implementation. Genome Res. 11 (8): 1425-1433.

Gene Ontology Consortium. (2006) The Gene Ontology (GO) project in 2006. Nucleic Acids Res. 34 (Database Issue): D322-D326.

Riley, M. (1993) Functions of the gene products of *Escherichia coli*. Microbiol. Rev. 57 (4): 862-952.