

Ontology Usage at ZFIN

Doug Howe and Christian Pich

Zebrafish Model Organism Database,
5291 University of Oregon, Oregon, OR 97403, USA
<http://zfin.org>

Abstract. The Zebrafish Model Organism Database (ZFIN) provides a Web resource of zebrafish genomic, genetic, developmental, and phenotypic data. Four different ontologies are currently used to annotate data to the most specific term available facilitating a better comparison between inter-species data. In addition, ontologies are used to help users find and cluster data more quickly without the need of knowing the exact technical name for a term.

1 Introduction

ZFIN is the model organism database for *Danio rerio*, the zebrafish and provides a centralized resource for zebrafish genomic, genetic, phenotypic, and developmental data. The ZFIN database contains highly integrated, manually curated information about genes, gene expression, mutant phenotypes and antibodies[1]. Web-based search interfaces and tools allow viewing and analysis of the data, facilitate the understanding of gene function and regulation, and promote scientific discovery.

One significant obstacle when searching for annotated data is knowing the exact ontology term name to search with. A curator may use one name to annotate a phenotype with a given anatomical structure while a user may know the same structure by a different name. To overcome this problem ontologies are created that use a definitive name for each entity and support the extensive use of synonyms. Such a dictionary is built as a directed acyclic graph (DAG) in which entities have one or more relationships to each other in a noncyclical pattern. DAGs provide a way to structure the various entities being modeled and lend themselves for various reasoning in regards to parent-child relationship questions. Currently, ZFIN uses four different ontologies[3] to annotate gene function, gene expression, and phenotypic information: Gene Ontology (GO)[2], Zebrafish Anatomy Ontology (AO), Entity Quality Ontology (PATO) and the Spatial Ontology. Curators at ZFIN try to annotate to the most specific ontology term available. This can make it hard for end users to find those annotations as they may not know the specific term name or they may wish to query with a more general term. Fortunately, due to the DAG-structure of the ontologies, annotations can be looked up by a higher level term name by performing an ontologically aware search that includes all subterms (children terms) if desired. For example, a researcher is looking for genes that are expressed in the eye.

The smart search returns all expression data that are annotated directly to the term eye and all its child terms as well, e.g. retinal pigmented epithelium. In addition, ZFIN provides auto-completion as the user types a term name into a search entry field, i.e. a list of terms is provided that matches the users input. Synonymous names for all terms are included and help the user locate the best entity match.

2 Current Challenges

Ontologies are useful tools to capture and describe expression, phenotype, and gene functional data. However ontologies are often incomplete and sometimes annotation of one observation can be made in several distinct ways. Consider situations in which a term is not found in a given ontology but the term might be created as a cross-product of terms from two or more independent ontologies. For example, if the term "fin development" was not currently in the GO, it could be emulated as a cross product of the term "fin" from the zebrafish AO and the GO term "development". Currently, ZFIN accomplishes this through a technique of post-composition, annotation-time term composition in which a new entity is emulated as the intersection of two existing ontology terms. Including such post-composed terms in term lookups, annotation displays, download files, and web-based search results then becomes a more intricate problem. If a user queries for phenotypes affecting the anatomical structure "actinotrichium", which is part of the fin, should an annotation with the post-composed term "fin":"development" be included in the results set? If so, how and when is this association accomplished? Even more difficult is following the proper logic when a phenotype annotation uses an existing GO term, like "neural retina development" and a user then makes a query for phenotypes that affect the "retina" by specifying the "retina" term from the zebrafish AO. How should the zebrafish AO term "retina" logically and rigorously return phenotypes that are annotated with the GO term "neural retina development"? There is no direct logical link (see Fig. 1) other than a string match between the zebrafish AO term "retina" and the GO term "neural retina development". One way to remedy the missing links could be to create a separate ontology that contains the relationships between the terms of the two ontologies in question. Work has already begun on this path in the case of linking species specific anatomy ontologies and the biological processes of the Gene Ontology[4].

To answer this query extensive logical traversal of ontologies is necessary at the time the query is initiated or when data are indexed. Extending the example further, consider a phenotype annotation involving the GO term "neural retina development". If a researcher then makes a query for all phenotypes affecting the zebrafish eye by using the AO term "eye" in the ZFIN mutant search form, extensive logical reasoning must ensure to link the annotation using GO:"neural retina development" to the user query for phenotypes affecting the AO:"eye". ZFIN is just beginning to explore this area of logical ontology traversal and we expect it will become an increasingly important aspect of data retrieval at ZFIN

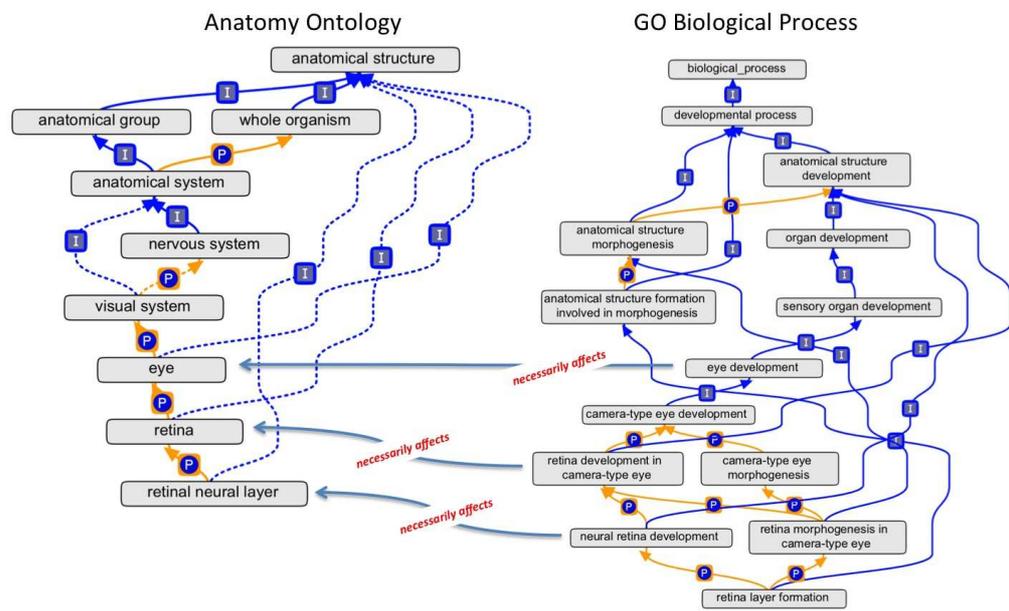


Fig. 1. Anatomy Ontology (left) and Gene Ontology (right) interrelationship.

in the future. We solicit your input on techniques to handle advanced logical ontology traversal, data linking, and data indexing strategies.

3 Future Ontology-Driven Directions

1. Sophisticated logical reasoning to return correct and complete data sets regardless of how annotation was made
2. Faceted data navigation interfaces
3. Data-linked ontology browsing
4. Incorporation of new ontologies (ChEBI, SO[5], others)

References

1. Sprague, J., Bayraktaroglu, L., Bradford, Y., Conlin, T., Dunn, N., Fashena, D., Frazer, K., Haendel, M., Howe, D.G., Knight, J., Mani, P., Moxon, S.A., Pich, C., Ramachandran, S., Schaper, K., Segerdell, E., Shao, X., Singer, A., Song, P., Sprunger, B., Van Slyke, C.E., and Westerfield, M.: The Zebrafish Information Network: The zebrafish model organism database provides expanded support for genotypes and phenotypes. *Nucleic Acids Res.* 36:D768-D772 (2008)
2. Gene Ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* 25: 25-29 (2000)
3. The Open Biological and Biomedical Ontologies. <http://obofoundry.org/>
4. Integrating phenotype ontologies across multiple species <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847714/>
5. Eilbeck K., Lewis S.E., Mungall C.J., Yandell M., Stein L., Durbin R., Ashburner M.: The Sequence Ontology. *Genome Biology* 6:R44 (2005)