Development of an Ontology of Microbial Phenotypes (OMP)

Michelle Giglio[1], Chris Mungall[2], Peter Uetz[3], Lanlan Yin[3], Johannes Goll[3], Deborah Siegele[4], **Marcus Chibucos[1]**, James Hu[4]
[1]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD; [2] Lawrence Berkeley National Laboratory, Berkeley, CA, [3]The J. Craig Venter Institute, Rockville, MD; [4]Texas A&M University, College Station, TX

Abstract

Phenotypic data are routinely used to elucidate gene and protein function in most organisms amenable to experimental manipulation. However, although phenotype ontologies exist for many eukaryotic model organisms, no standardized system exists for the capture of phenotypic information in bacteria. We propose to build an Ontology of Microbial Phenotypes and use it to annotate the prokaryotic model organism *Escherichia coli*.

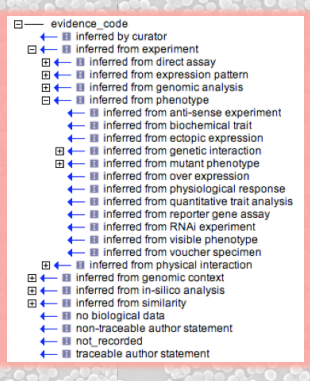
Introduction

Phenotypes are the observable characteristics of an organism that result from the combination of a particular genotype and a particular environment, and thus are a basic and fundamental aspect of the biology of all organisms. The awesome power of genetics is founded on how the phenotypes of mutant genes, alone and in combination, contribute to understanding the biology of affected systems. To fully exploit the power of phenotypes for functional and comparative genomics, the ability to make comparisons across datasets and systems is vital. Making these comparisons either manually or computationally is hindered by the fact that phenotypes are not described consistently for bacteria. Our project aims to develop annotation infrastructure to improve the ability of microbiologists and bioinformaticians to use both existing and new phenotype information and to capture it in a consistent and standardized manner. This will require two key components: 1) an Ontology of Microbial Phenotypes (OMP) that captures phenotype descriptions in a controlled vocabulary, and 2) a set of evidence codes based on extension of the existing Evidence Code Ontology,[1] with links to a database of papers and other resources describing the assays used to "measure" these phenotypes.

Evidence Code Ontology (ECO)

➢ECO is a rich ontology for experimental and other evidence statements currently available from the OBO CVS repository.

➢Some ECO terms.



Conclusion

The manual and cross product efforts were undertaken independently and in parallel by separate members of the group to see what, if any, consistency would be achieved. We found that although the concepts captured were similar, the different researchers chose different PATO quality terms to represent the same concepts. The manual curator chose "abnormal," while the person working on cross products chose "abolished" and "disrupted." The results of this exercise illustrate one reason why the pre-coordinated approach has advantages over the post-coordinated approach. In the post-coordinated approach separate annotators creating phenotype annotations at different points in time may choose different ways of expressing the same concept and thus create inconsistency. In the pre-coordinated approach, one controlled set of PATO terms will be used for term generation, and the fact of storing all the terms in one controlled vocabulary will enforce consistency and uniformity.

Future Directions

If our project is funded, we plan to expand our cross product generation by targeting relevant nodes in the GO and other ontologies. We will extend ECO to include terms that capture the assays used in phenotype analysis. We will apply the OMP and extended ECO to the annotation of *Eschericia coli* and make the data available using EcoliWiki and other resources.

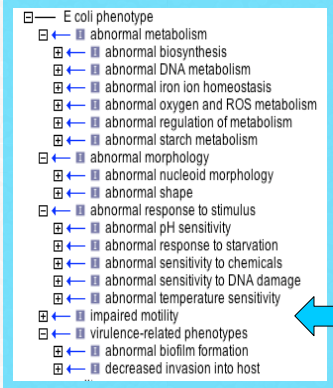
References

1. http://www.obofoundry.org/cgi-bin/detail.cgi?id=evidence_code
2. http://obofoundry.org/wiki/index.php/PATO:Main_Page
3. Day-Richter J, Harris MA, Haendel M, The Gene Ontology OBO-Edit Working Group, and Lewis S. OBO-Edit—an ontology editor for biologists. Bioinformatics. 2007;23(16):2198-2200.

Methods

We have explored two parallel approaches to building the OMP. Both are pre-coordinated approaches that rely on using the terms in the Phenotypic Quality Ontology (PATO) as a basis for building up phenotype terms.[2] In the first approach we read 100 papers and identified 40 phenotypes described in those papers. We organized the 40 phenotypes into a controlled vocabulary using OBO-Edit.[3] While this effort was not comprehensive, we were able to classify the 40 phenotypes into five superclasses and assign PATO entities and qualities. In addition, various assays (biochemical, morphological, and physiological) were collected from the papers that were curated to generate phenotype terms. In the second approach we generated a cross product between a selection of PATO terms and two GO nodes relevant to microbial phenotypes, "GO:0044262 : cellular carbohydrate metabolic process" and "GO:0006520 : cellular amino acid metabolic process." We found the cross product generation method to be quite effective in generating large numbers of relevant terms quickly.

Five superclasses of manually assessed phenotype terms



Manually Derived Cross Products

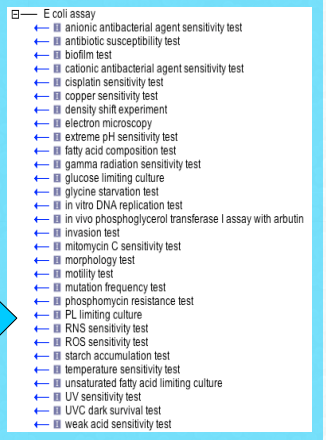
➢Query PubMed for *Escherichia coli* with keywords "mutation AND phenotype" (i.e. "txid562[Organism:exp] AND mutation AND phenotype")

➢Read 100 papers

➢Collect assays

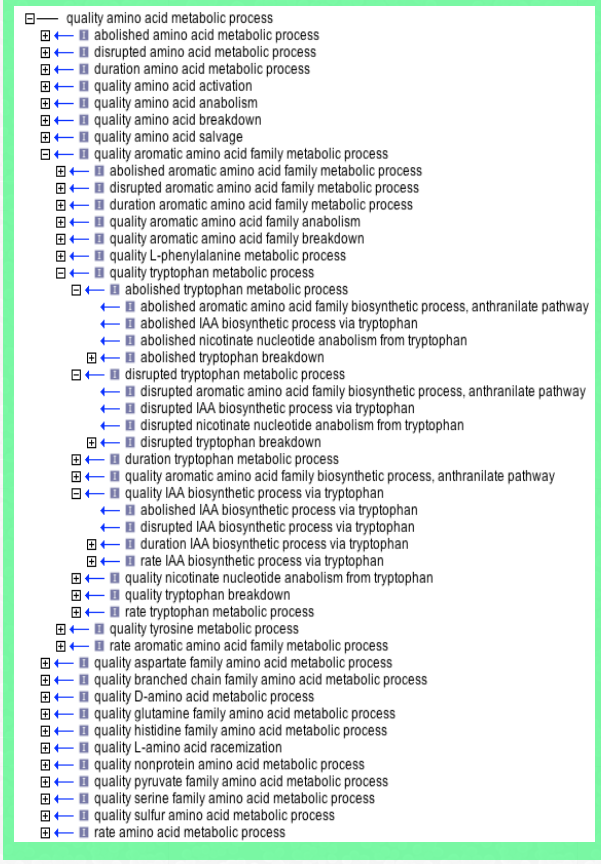
➢Develop terms

E. coli phenotype assays



Automatically Derived Cross Products

Selected PATO-GO cross product terms



➢Combining a quality term from PATO with an entity term from GO creates a new term that describes a phenotype.

➢Example:
"disrupted" x "amino acid metabolic process" = "disrupted amino acid metabolic process"

➢Some PATO terms.