Semantic Index of Phenotypic and Genotypic Data

Charles T. Parker* (chuck.t.parker@namesforlife.com),[1] Nenad Krdzavac,[2] Kevin Petersen,[2] Amber Roberts,[2] Grace Rodriguez,[1] and **George M. Garrity**[1,2]

[1]NamesforLife, LLC; East Lansing, Michigan and [2]Michigan State University; East Lansing, Michigan

http://ontology.namesforlife.com

Project Goals: The goal of this project is to develop a semantic data resource that can serve as a basis for predictive modeling of microbial phenotype. The core technical objectives are twofold: (1) to build a database of normalized phenotypic descriptions (observational data) using the primary taxonomic literature of bacterial and archaeal type strains, and (2) to construct an ontology with reasoning capabilities to make accurate phenotypic and environmental inferences based on that data. This project is tightly coupled with ongoing DOE projects (the Genomic Encyclopedia of Bacteria and Archaea, the Microbial Earth Project, the Community Sequencing Project) and with two key publications, *Standards in Genomic Sciences* (SIGS) and the *International Journal of Systematic and Evolutionary Microbiology* (IJSEM).

The DOE Systems Biology Knowledgebase (KBase) was envisioned to provide a framework for modeling dynamic cellular processes of microorganisms, plants and metacommunities. The KBase will enable rapid iteration of experiments that draw on a wide variety of data and allow researchers to infer how cells and communities respond to natural or induced perturbations, and ultimately to predict outcomes.

Predictive models rely on high quality input data, but not all data are of similar quality nor are they amenable to computational analysis without extensive cleaning, interpretation and normalization. Key among those needed to make the KBase fully operational are phenotypic data, which are more complex than sequence data, occur in a wide variety of forms, often use complex and non-uniform descriptors and are scattered about specialized databases and scientific and technical literature. Incorporating phenotypic information into the KBase requires expertise in harvesting, modeling and interpreting these data.

The Semantic Index of Phenotypic and Genotypic Data will address this problem by providing a resource of reference phenotypic data for all validly published type strains of *Bacteria* and *Archaea*, based on concepts and observational data drawn from the primary taxonomic literature. In the Phase I project we developed software to construct and analyze a corpus of this literature and to extract putative feature domain vocabularies comprising approximately 40,000 candidate phenotypic terms used in 5,750 (now expanded to 11,018 of 17,793 total) new and emended descriptions of the 11,492 distinct type strains of *Bacteria* and *Archaea*. In Phase II, these vocabularies are serving as the basis for developing a phenotypic ontology, a repository of

phenotypic data and normalized phenotypic descriptions for each species. Many of the phenotypes applied to microbes describe a combination of quantitative environmental conditions and qualitative growth and metabolic capabilities. Such terms are challenging to implement in query systems due to their context-based interpretations and conceptual overlap across multiple feature domains. In our past year of research, we have discovered novel design patterns for ontology development [1] that address these problems and remove barriers to machine reasoning over these complex terms.

In our current work, we are applying these novel modeling techniques to encode axioms for automatically resolving ambiguity attributed to the semantic equivalence and imprecision of phenotypic terms arising in literature [2]. These axioms will enable reasoners to make correct inferences over the ontology and phenotypic data. We are also developing a query and retrieval service linked to the ontology that will provide researchers with consistent, accurate interpretations that are usable for predictive modeling and in other research and commercial applications.

Several additional software components were developed to overcome technical barriers that arose during this project [3]. Originally implemented as command-line utilities for vocabulary extraction, annotation and document analysis, we are now developing these into a commercial semantic desktop application for document/corpus analysis and for bootstrapping terminology/ontology development.

Publications

1. Parker, CT, Garrity, GM and Krdzavac, NB. Systems and Methods for Inferring Properties of Objects in the Absence of Direct Observations. U.S. Provisional Patent Application No. 61/880,244. Filed September 20, 2013. Washington, DC: *U.S. Patent and Trademark Office*.
2. Krdzavac, NB, Parker, CT, Garrity, GM: An Observation Ontology. *Bioinformatics*. 2014, *Under review*.
3. Garrity GM: The NamesforLife Semantic Index of Phenotypic and Genotypic Data: Phase I Final Technical Report [Internet]. 2012, doi:10.1601/report.sc0006191p1

Funding for this project was provided through the DOE SBIR/STTR program (DE-SC0006191). Public funding for development of the NamesforLife infrastructure was received from the DOE SBIR/STTR program (DE-FG02-07ER86321), the Michigan Small Business Technology Development Corporation, the Michigan Strategic Fund, and the Michigan Universities Commercialization Initiative.