The NamesforLife Semantic Index of Phenotypic and Genotypic Data

Charles T. Parker[1], Catherine Lyons[1] and **George M. Garrity**[1,2]* (garrity@namesforlife.com)

[1]NamesforLife, LLC, East Lansing, MI and [2]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI

http://www.namesforlife.com

Project goals: Predictive models depend on high quality input data. But not all data are of similar quality nor are all of the data amenable to computational analysis without extensive cleaning, interpretation and normalization. Key among those needed to make the projects such as the DOE Knowledgebase (Kbase) operational are phenotypic data, which are more complex than sequence data, occur in a wide variety of forms, use complex and non-uniform descriptors and are scattered about the literature and specialized databases. Incorporating these data into the Kbase will require expertise in harvesting, modeling and interpreting the data. The *NamesforLife Semantic Index of Phenotypic and Genotypic Data* seeks to address this problem by taking the first steps toward building an ontology of bacterial and archaeal based on the taxonomic literature through the dvelopment of a draft vocabulary of phenotypic features of the taxonomic type strains.

Unlike sequence data, which are essentially universal, uniform and predictable, phenotypic data are inherently complex, noisy and "taxonomically parochial". The same trait may vary significantly under different conditions of growth, at different times during the life of a cell and under different environmental conditions. The language of phenotype is also complex and may be limited in taxonomic scope, requiring expert interpretation, as there is no equivalent to BLAST for searching for phenotypic data, and there is no central repository for such data. In some cases an entire language exists to describe the phenotypic features that apply to a single phylum, class or order (e.g., Cyanobacteria, Actinobacteria) or a particular class of features (e.g., lipids, structural carbohydrates). In addition, phenotypic data must be viewed in a historical perspective (time when the data were collected) to understand what was measured and how it was measured. As was the case with microbial physiology, which had fallen out of fashion as a field of research, so too has "classical" or polyphasic microbial taxonomy resulting in a dwindling community of experts who can readily interpret the existing information in the literature and in various specialized databases. How might that expertise be captured and applied to developing a standardized language and ontology of microbial phenotype?

In 2003, Garrity and Lyons proposed a novel approach to resolving ambiguity of biological nomenclature. Their approach provided a means of resolving the complex relationships that exist among names and the concepts and objects to which names apply. When coupled with Digital Object Identifiers (DOIs) their method provided a means by which names in digital content (e.g. journal articles, technical reports, web pages) and databases could be made actionable and directly linked to expertly curated information about the name, including its history of changes. NamesforLife, LLC has developed a suite of web services and applications based on this method that can be used to semantically enrich or enhance digital in a variety of formats. The Company has already demonstrated that vectors of names (Semiotic Fingerprints) can be used to index and

cross-classify large corpora of scientific and patent literature based on the relationship between named organisms and the underlying subject matter of subsets of documents. The methods and tools are not, however, restricted to biological nomenclature and can be applied to terminologies of all types.

The long-term objective of this STTR project is to develop a semantic index of bacterial and archaeal phenotypes that can be used to augment annotation efforts and to provide a basis for predictive modeling of microbial phenotype. The index will be based on published descriptions of taxonomic type and non-type strains that have been the subject of ongoing genome sequencing efforts as this will provide a mechanism whereby hypotheses can be tested and verified, reproducibly. This project is tightly coupled with ongoing DOE projects (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 first step towards accomplishing this goal, and the primary objective of this Phase I project is the development of a draft vocabulary of phenotypic features.

Our approach towards developing a draft vocabulary of bacterial and archeal phenotype is based on a textual analysis of the richest source of descriptive information; the taxonomic literature. It follows a well-established path used for ontology construction based on derivation of domain-dependent hyponymy (is-a relationships) from a corpus and leverages tools, data resources and expertise that the Company has already developed. For this Phase I project, our target corpus consists of a subset of taxonomic literature of type strains from the IJSEM (2003-2009) and SIGS (2009-2011). These articles have been further subdivided into those pertaining to a single bacterial species or multiple species and higher taxa. The articles have been indexed with Apache Lucene to produce two separate indices; one of the complete documents and one of the descriptions of each new organism alone. We are currently developing a KWIC (KeyWord In Context) interface to permit location and display of a given word in the corpus in its surrounding context to understand usage variations within and across different taxa. Selection of terms for analysis is being done using Apache Luke, which provides facilities for determining usage frequency, coupled with curatorial review for relevance, categorization and synonymy. Then end goal of the Phase I project is to allow end users to view an article with all of the phenotypically relevant terms highlighted based on the KWIC index, to use the KWIC index to auto-populate phenotypic characteristics of a given strain based on the published literature and to allow endusers to adjust/negate/select phenotypic characteristics and their values for a strain, using a normalized set of terms.

Funding 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 Developemt Corporation, the Michican Statgetic Fund, and the Michigan Universities Commercialization Initiative.