**Project Goals**

The goal of this project is to develop a semantic data resource to serve as a basis for predictive modeling of microbial phenotype.

This project is tightly coupled with ongoing DOE projects (the Genomic Encyclopedia of Bacteria and Archaea, the Microbial Genome Project, the Microbial Science Program) and with two high-profile publications, Stendahl in Genomic Sciences (SIS) and the International Journal of Systematic and Evolutionary Microbiology (IJSEM).

The scope of this project covers many technical fields, including text-mining, information extraction, Natural Language Processing, indexing & search, terminology & ontology development, machine reasoning, semantic analysis, sequence analysis and taxonomic classification.

**Background**

The DOE Systems Biology Knowledgebase (KBaSe) was envisioned to provide a framework for modeling dynamic cellular processes of microorganisms, plants and metacommunities. KBaSe will enable rapid iteration of experiments drawing on a variety of data to allow researchers to infer how cells and communities respond to natural and induced perturbations and ultimately 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 variety of forms, often use complex and non-uniform descriptors and are scattered about specialized databases and scientific/technical medical literature. Incorporating phenotypic data into the KBaSe requires expertise in harvesting, modeling, interpreting and validating these data as well as a complete type strain dataset and taxonomy.

**This online resource complements KBaSe by providing a reference set of phenotypic data for nearly all published type strains of Bacteria and Archaea.**

**Our Solution**

The Semantic Index of Phenotypic and Genotypic Data will address this problem by providing a reference of resource 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 will develop software to construct and analyze a corpus of this literature and to extract putative feature domain vocabularies comprising over 40,000 candidate phenotypic terms used in 20,226 new and emended descriptions of the 12,937 distinct type strains of Bacteria and Archaea (NSI, Database, February 20, 2013). In Phase II/III, 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 developing a solution to these problems, we discovered a novel method for establishing concept equivalence that enables precise, consistent, verifiable reasoning over these complex terms.

**Semantic Index of Phenotypic and Genotypic Data**

Charles Parker, 1, Nenad Krzakac, 2, Chuong Vo Phan, 1, Kevin Petersen, 2, Grace Rodriguez 1, and George M. Garrity 1, 2

1 NamesforLife, LLC and 2 Michigan State University (East Lansing, Michigan)