Project Goals

We are developing a standards-compliant semantic desktop data resource to support predictive modeling of microbial phenotype.

Our objectives are:
1. Build a knowledge resource containing genotypic and phenotypic data about prokaryotic strains.
2. Develop a formal ontology capable of making accurate predictions and environmental inferences over this data source.
3. Improve the visibility and accessibility of publicly available genotypic and phenotypic data.

This project is tightly coupled with ongoing ODE projects (Genomic Encyclopedia of Bacteria and Archaea, Microbial Earth Project, Community Science Program) and two key publications (Standards in Genomic Sciences and the International Journal of Systematic and Evolutionary Microbiology).

Background

The Problem

Despite significant improvements in genome annotation, many assertions are hypothetical and may lack experimental support. The taxonomic literature for prokaryotes contains a wealth of experimental phenotypic data, but that knowledge is currently in a form that does not lend itself to integration with databases or ontologies. Predicting phenotypes from high-quality genotypic 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 the types of data needed to support current research are phenotypic data (Table 1), which are more complex than sequence data, occur in a variety of forms, often use complex and non-uniform descriptors, may be taxon-specific and are scattered throughout specialized databases and scientific, technical and medical literature. Integrating phenotypic data from such resources requires expertise in harvesting, model construction, and validation of these data, as well as a complete and actively maintained resource for all of the type strains.

Table 1. Feature classes included in the Prokaryote Knowledge Base, grouped by major feature domain. The features will be available via the Taxonomic Abstracts (http://doi.org/10.1601/ex.4346) and several new services.

<table>
<thead>
<tr>
<th>Feature Class</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>N4L Exemplar DO/ID</td>
<td>Unique identifier for a strain source.</td>
</tr>
<tr>
<td>Host</td>
<td>The species name of the strain.</td>
</tr>
<tr>
<td>Strain Designation</td>
<td>The strain designation used by the original source.</td>
</tr>
<tr>
<td>Collection ID</td>
<td>The collection ID used by the original source.</td>
</tr>
<tr>
<td>Taxonomic status</td>
<td>The taxonomic status assigned to the strain.</td>
</tr>
<tr>
<td>Isolate source</td>
<td>The source of the isolate.</td>
</tr>
<tr>
<td>Physiological</td>
<td>The physiological characteristics of the strain.</td>
</tr>
<tr>
<td>Genotypic</td>
<td>The genotypic characteristics of the strain.</td>
</tr>
</tbody>
</table>

Challenges of Information Extraction (IE)

Extracting information from text is not an easy task. Prior to this stage of the project, we had already produced a curated taxon and strain database covering all of the type strains of prokaryotes, and assembled a complete complex of taxonomic literature, as well as a candidate thesaurus of phenotypic terms. Using these resources, some novel software methods and an extensive curation effort, we are normalizing raw text into phenotypic assertions based on our ontology and controlled vocabulary. These assertions are interpreted by a reasoner to infer phenotype based on all available information that has been reported about a strain. Our method is able to use knowledge at appropriate levels of abstraction to correctly answer queries and produce new knowledge.

Our Solution

Our knowledge base is designed to address these problems by providing reference phenotypic data for nearly all type strains of Bacteria and Archaea based on concepts and observations drawn from the primary taxonomic literature (the corpus of literature that supports our up-to-date taxonomy and strain database). We developed software (Semantic Desktop) to extract putative feature domain vocabularies from this corpus, resulting in the discovery of over 40,000 candidate phenotypic terms used in 20,224 new and emended descriptions of the 12,937 distinct type strains of Bacteria and Archaea (N4L Database, February 20, 2015). We have since developed this vocabulary into a precise thesaurus of phenotypic terms, which will ultimately conform to W3C SKOS-XL semantics, providing a link between the language of microbial phenotype, the semantic web and existing ontologies (N4L-Guide and N4L-Scribe). Our use of existing standards and services, coupled with the broad coverage of prokaryotic taxa, will complement the MGS and RIMMS (MxS) standards by providing a precise and robust vocabulary to use when publishing descriptions of new taxa.

Our thesaurus complements existing taxonomic resources with broad taxonomic coverage.

Our ontology relates reported observations to an organism’s environment and phenotype.

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, imprecision and conceptual overlap across multiple feature domains.

To address this problem, the thesaurus was developed in parallel with a formal ontology that supports inference from observations of an organism under a set of environmental constraints, using a unique meta-model to support queries using these complex terms. By developing a solution to this problem, we discovered a novel method for establishing semantic equivalence among concepts that enables precise, consistent, verifiable reasoning over imprecise terms at multiple levels of abstraction.

Current and Planned Products

We recently deployed a strain finder service (http://doi.org/10.1601/strainfinder) that provides a search interface, persistent landing page and forwarding service for strain identifiers (e.g., http://doi.org/10.1601/strainfinder?rid=93fdd-ATCC+23350). It integrates with the Taxonomic Abstracts and serves as a permanent, electronically traversable link from publications or 3rd party resources directly to specimens held in collections.

We are developing a Strain Registration portal in collaboration with the Joint Genome Institute. NamesforLife will register Digital Object Identifiers and CrossRef metadata for each strain sequenced at JGI. The StrainFinder and Taxonomic Abstracts will provide additional visibility for individual sequencing projects at JGI.

A faceted search engine over the phenotypic characters of prokaryotic strains is under development, which will be the main point of entry to the phenotypic knowledge base. Ontology specialists will be able to query the knowledge base using a GraphQl endpoint, and developer access will be available to named queries via a REST API (in development).

Document Annotation: The N4L::Scribe

The Scribe document annotation service (http://scribe.namesforlife.com) has been significantly updated to recognize (in addition to bacterial and archaean names), esukaryotic names, viral names, GenBank accessions and strain identifiers. This web service embeds links directly into documents (i.e., Microsoft Word [DOCX and .DOCX], Open/LibreOffice [ODF], or any well-formed XML [including XHTML, NLM, JATS, etc.] to the authoritative resources for any recognized names, identifiers or accession numbers. Additionally, summaries of nucleotide or protein sequences are generated so that authors, reviewers or editors may verify the accuracy of the identifiers used in the document. We are also testing delimited text and spreadsheet formats, which can provide nomenclature, taxonomy and strain resolution services for 3rd party databases.

Acknowledgments

Funding for this project was provided through the DOE SBIR/STTR Program (DE-SC0017195). Funding for the NamesforLife infrastructure was received from the DOE SBIR/STTR program (DE-FG52-07ER86302), the Michigan Small Business Technology Development Corporation, the Michigan Strategic Fund, the Michigan Economic Development Corporation, and the Michigan Universities Commercialization Initiative.