Prokaryote.INFO: A semantic knowledge resource for microbial phenotype.

Charles Parker[1], Nenad Krdzavac[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)

NamesforLife
Bringing meaning to life...

Project Goals

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

Our objectives are to:

(1) build a knowledge resource containing standardized phenotypic descriptions of prokaryotic type strains,

(2) develop a formal ontology capable of making accurate phenotypic and environmental inferences over this resource, and

(3) improve the visibility and accessibility of publicly funded research projects that provide these data.

This project is tightly coupled with ongoing DOE 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. 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 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, modeling, interpreting, and validating 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 made available via the Taxonomic Abstracts (http://doi.org/10.1601/about) and several new services.

Strain Metadata	**Morphology**	**Chemotaxonomy**
N4L Exemplar DOI	**Micromorphology**	Fatty Acids
Host	Cell size	Polar Lipids
Strain Designation	Cell shape	Mycolic Acids
Collection ID(s)	Motility	Respiratory quinones
Taxon status (type/non-type)	Sporulation	Peptidoglycan composition
Isolation substrate	Staining characteristics	Polyamines
Isolation source	Intracellular inclusions	**Physiological**
Isolation method	Extracellular features	optimal growth conditions
Geographic location	Life cycle	Cell Images
Environmental information	Other characteristics	sensitivity/tolerance to chemical and physical agents
Genotypic	**Macromorphology**	substrate utilization
16S rRNA sequence	Growth on solid surfaces	terminal electron acceptor
% DNA-DNA similarity	Colony morphology	metabolic end-products
% G+C composition	Growth in liquid	Growth Curves
Whole genome	Pigment production	
Other marker genes	Other features	

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 observational data 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 NamesforLife services (*N4L::Guide* and *N4L::Scribe*). Our use of existing standards and services, coupled with the broad coverage of prokaryotic taxa, will complement the MIGS and MIMS (MIxS) standards by providing a precise and robust vocabulary to use when publishing descriptions of new taxa.

Our thesaurus complements MIxS by providing precise phenotypic language 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. In 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 [1].

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 taxonomy and strain database covering all of the type strains of prokaryotes, and assembled a complete corpus 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.

strain	source	oxygen sensitivity (raw text)	pH sensitivity (raw text)	temperature sensitivity (raw text)
10.1601/ex.3007	rid.516	facultatively anaerobic	Mesophilic and neutrophilic chemoorganotroph: grows between 15 and 30 °C.	Mesophilic and neutrophilic chemoorganotroph: grows between 15 and 30 °C.
10.1601/ex.3857	rid.507	Requiring less than 15%O2 (i.e. 75% air saturation) in the headspace gas (optimum 5–8 %).	pH 4.5–9.0 (optimum pH 6.0–7.5). optimum pH 6.5	The isolate grew at 10–40 °C (optimum 25 °C)
10.1601/ex.4346	rid.500	Strict anaerobe.	pH range for growth 6.3-8.5, pH optimum at 7.0.	T_{min}, 20°C ; T_{opt}, 38°C ; T_{max}, 43°C ;
10.1601/ex.166	rid.490	Obligately anaerobic.	Growth occurs between pH 5.5 and 6.7, with the optimum at around pH 6.5	The temperature range for growth at pH 6.5 was 50–86 °C, with optimum growth at 85 °C.
10.1601/ex.7799	rid.301	Anaerobic, aerotolerant.	Optimal growth at pH 8.8 to 9.75. No growth at pH 8.0 or 10.8.	Optimum temperature for growth, 30 to 37°C; range, 15 to 47°C

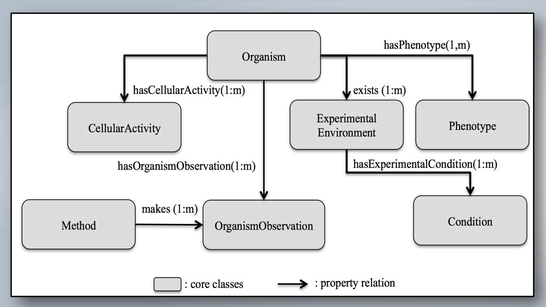
strain	source	oxygen sensitivity (normalized text)	pH sensitivity (normalized text)	temperature sensitivity (normalized text)
10.1601/ex.3007	rid.516	facultative anaerobe	neutrophile	mesophile growth at 15 °C, growth at 30 °C
10.1601/ex.3857	rid.507	growth at 15%, optimal growth at 5%, optimal growth at 8%	optimal growth at pH 6.5	optimal growth at 25 °C
10.1601/ex.4346	rid.500	obligate anaerobe	optimal growth at pH 7.0	optimal growth at 38 °C
10.1601/ex.166	rid.490	obligate anaerobe	optimal growth at pH 6.5	optimal growth at 85 °C
10.1601/ex.7799	rid.301	aerotolerant anaerobe	optimal growth at pH 9.275	optimal growth at 33.5°C

strain	source	oxygen sensitivity (interpreted)	pH sensitivity (interpreted)	temperature sensitivity (interpreted)
10.1601/ex.3007	rid.516	facultative anaerobe	neutrophile	mesophile
10.1601/ex.3857	rid.507	microaerophilic	neutrophile	mesophile
10.1601/ex.4346	rid.500	obligate anaerobe	neutrophile	mesophile
10.1601/ex.166	rid.490	obligate anaerobe	neutrophile	hyperthermophile
10.1601/ex.7799	rid.301	aerotolerant anaerobe	alkalophile	mesophile

(Right) An Orthogonal Semantic Equivalence Map (OSEM) for sensitivity and tolerance to Oxygen. This provides the structure for implementing first order logic (rules and axioms) over three distinct concept taxonomies (SKOS-XL). Assertions may be supplied to an OSEM to infer semantically equivalent representations of phenotype over bi-directional (environment, observation) relations.

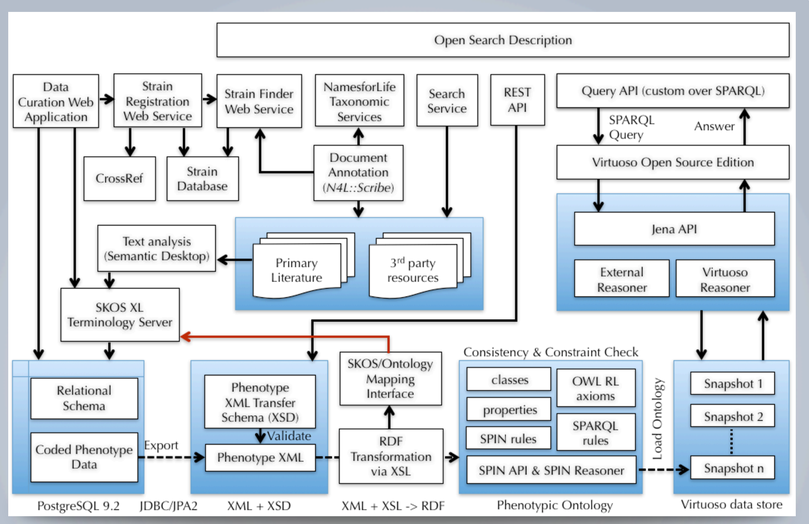
Phenotype		A: anoxic [0,0]	B: aerobic (0,)	
			B1: microaerobic (0,1)	B2: air [1,)
anaerobe	strict anaerobe	+	x	x
	aerotolerant anaerobe	+	G	
aerobe	strict aerobe	x	x	+
	microaerophilic	x	+	x

(G): growth; (x): no growth; (~): growth (suboptimal); (+): growth (optimal); ( ): don't care



(Left) The Upper (Core) Ontology for phenotype. The key to developing a method that can correctly interpret the mixed semantics of phenotypic language was in the separation of the environmental conditions of an organism from the observations of that organism. This separation is enforced at the highest level of our ontology, and individual feature classes extend this ontology.

As additional feature domains are modeled and new relations are discovered, rules and axioms are developed and encoded in the SPIN and OWL RL languages. A reasoner (SPIN reasoner, OWL reasoner and Description Logics reasoners) may infer new knowledge about strains that was not directly reported in the literature.



We adopted a hybrid relational database (PostgreSQL) / thesaurus / formal ontology architecture (above) in order to support curation, reasoning, search and query. The relational schema is mapped to the ontology via an intermediate XML Transfer schema, which also serves as the basis for archiving complete strain records. Records that fail validation or Consistency and Constraint Checking are flagged for curatorial attention.

Snapshots of the relational database are loaded into the ontology data store and individual records are checked for consistency over the ontology. We have migrated away from the Fuseki SPARQL query server (part of the Apache Jena ontology framework) and are now using Virtuoso Open Source Edition as the data store and ontology framework, maintaining our goals of a complete Open Source framework that is free to both commercial and non-commercial use.

The user interface is comprised of a collection of micro services that conform to the Open Search Description standard (where applicable), and allow query of the ontology, data and related resources in a variety of ways, including directly from a browser's native search bar.

Validation (consistency-checking over the ontology) of exported data from DB into phenotype ontology is handled by the SPIN API (TopQuadrant, W3C Draft) and Jena API (HP, Apache License). We use the SPIN reasoner and optionally other OWL reasoners to detect inconsistencies and constraint violations. After that process, data are ready to be stored in a triple store and queried.

Current and Planned Products

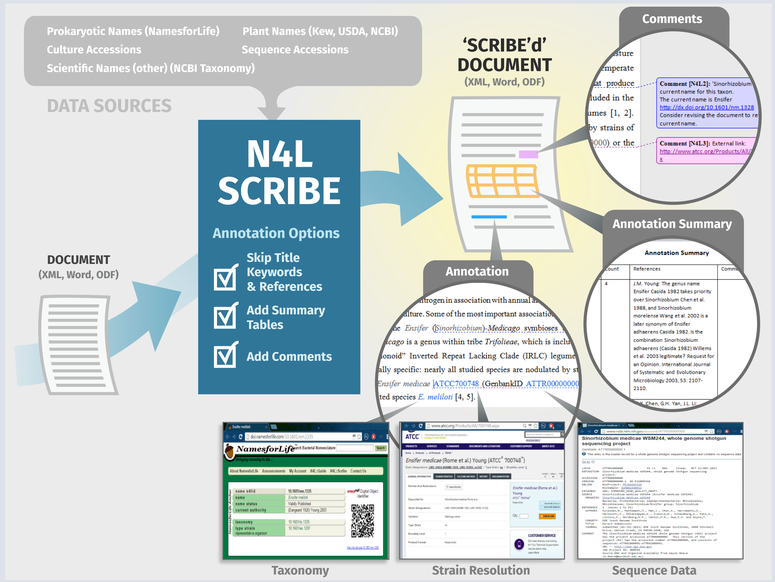
We recently deployed a strain finder service (http://doi.org/10.1601/strainfinder) that provides a search interface, persistent landing page and forwarding service for strain identifiers (e.g., http://doi.org/10.1601/strainfinder?urlappend=%3fid=ATCC+23350). It integrates with the Taxonomic Abstracts 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 Strain Finder 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 query the knowledge directly using a SPARQL 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 archaeal names), eukaryotic names, viral names, GenBank accessions and strain identifiers. This web service embeds links directly into documents (i.e., Microsoft Word [.DOC and .DOCX], Open/Libre Office [ODF], or any well-formed XML [including XHTML, NLM, JATS, etc.]) to the authoritative resources for any recognized names, identifiers or accessions. 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.



(Right) An early working version of our faceted search for strains by phenotype.

The facet query is converted into a SPARQL query and sent to our SPARQL endpoint, a custom Java Web Service built upon the Jena/ARQ API, which interprets the query and invokes a reasoner to infer an answer (a matching set of strains, shown here by their Exemplar DOI).

Using our novel method [1] of Semantic Equivalence, we may construct representations of an environment as a set of environmental constraints and answer questions such as:

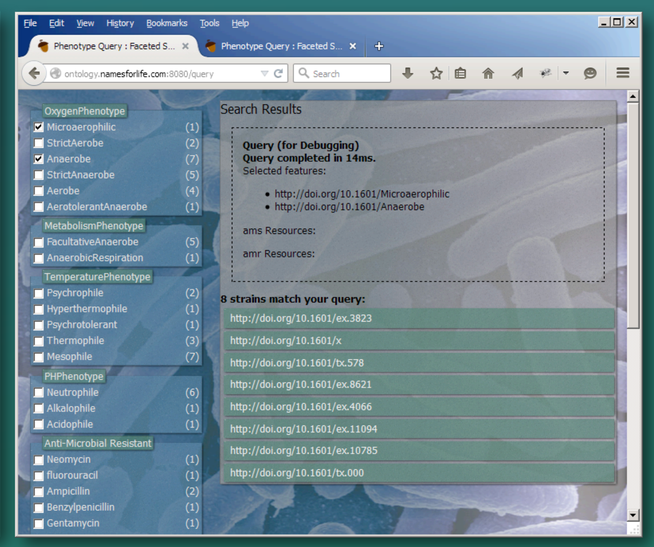
"What strains will grow in this environment?"

"How much phenotypic variation is there within the Streptomycetes?"

"What characteristics does Escherichia share with Salmonella?"

"Which type strains could be considered non-type strains of a different genus?"

"What strains are under-described for their taxa?"



Publications

1. Parker, CT, Garrity, GM and Krdzavac, NB. *Systems and Methods for Establishing Semantic Equivalence Between Concepts.* International Application No. PCT/US2014/056808. Filed September 20, 2014. Washington, DC: U.S. Patent and Trademark Office and Geneva, Switzerland: World Intellectual Property Organization.

2. Sayood, K, Way, S, Ozkan, UN and Garrity, GM. *Classification of Nucleotide Sequences by Latent Semantic Analysis.* International Application No. PCT/US2013/052797. Published June 2, 2014. Washington, DC: U.S. Patent and Trademark Office and Geneva, Switzerland: World Intellectual Property Organization.

3. Parker, CT and Garrity, GM. *Semiotic Indexing of Digital Resources.* U.S. Patent 8,903,825. Issued December 2, 2014. Washington, DC: U.S. Patent and Trademark Office.

Acknowledgments

Funding for this project was provided through the DOE SBIR/STTR program (DE-SC0006191). Funding for 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, the Michigan Economic Development Corporation, and the Michigan Universities Commercialization Initiative.