Developing New Tools for the Old Tree of Life

Kenneth Parnow  
*University of Southern Maine*

William Lambeth  
*University of Southern Maine*

Kelsi Jackson  
*University of Southern Maine*

Jesse Florendo  
*University of Southern Maine*

Jin Mao  
*University of Arizona*

*See next page for additional authors*

Follow this and additional works at: [http://digitalcommons.usm.maine.edu/thinking_matters](http://digitalcommons.usm.maine.edu/thinking_matters)

*Part of the [Biology Commons](http://digitalcommons.usm.maine.edu/biology_commons)*

**Recommended Citation**

Parnow, Kenneth; Lambeth, William; Jackson, Kelsi; Florendo, Jesse; Mao, Jin; Cui, Hong; and Blank, Carrine, "Developing New Tools for the Old Tree of Life" (2017). *Thinking Matters*. 123.

[http://digitalcommons.usm.maine.edu/thinking_matters/123](http://digitalcommons.usm.maine.edu/thinking_matters/123)

This Poster Session is brought to you for free and open access by the Student Scholarship at USM Digital Commons. It has been accepted for inclusion in Thinking Matters by an authorized administrator of USM Digital Commons. For more information, please contact jessica.c.hovey@maine.edu.
Author
Kenneth Parnow, William Lambeth, Kelsi Jackson, Jesse Florendo, Jin Mao, Hong Cui, and Carrine Blank
Developing New Tools for the Old Tree of Life

Kenneth Parnow1, William Lambeth1, Kelsi Jackson1, Jesse Florendo1, Lisa R. Moore1, Jin Mao2, Carrine Blank3, Marcia Ackerman1, Hong Cui4,

1Univ. of Southern Maine, Portland, ME; 2Univ. of Arizona, Tucson, AZ; 3Univ. of Montana, Missoula, MT, United States

Abstract

Millions of species reside in the Tree of Life, making the task of resolving the evolutionary origin of many organisms difficult. Biologists draw on genetic and phenotypic information to sort the Tree of Life, but the study can be slow and complex. Phenomic data (such as cell shape, metabolism and ecology), particularly for microorganisms, is often found in scientific publications and has little digital presence outside of being scanned into an online database. This has been aided by a new text mining computer program, MicroPIE (Microbial Phenomics Information Extractor), that sifts through relevant phenomic data and creates a matrix of key phenomic characters taken from the published descriptions. MicroPIE utilizes multiple natural language processing tools to extract data, along with the knowledge of microbiologists to help developing and verifying the tools. One major challenge to building such a tool is the time it takes to collect and edit phenomic data for tens of thousands of sentences needed to develop a functioning program. We have helped to further the development of MicroPIE to identify new characteristics by providing sentences from published microbial descriptions. We also are creating a “Gold Standard” matrix (GSM) of phenomic information for 100 different bacteria that can then be compared to the MicroPIE output in order to test that MicroPIE has correctly identified and extracted phenomic information. So far MicroPIE has shown potential to aid in resolution of the microbial Tree of Life.

Development and System Architecture of MicroPIE Algorithm

Input Text Descriptions → Input Descriptions in XML

Preprocessor

Sentence Splitting → Sentence Cleaning

Training Sentences

SVM models → Character Predictor

Character Extractor

Linguistic Rule → Value Post-processing

Matrix Generator

Term Lists

Linguistic Rule → Term Matching

Matrix Generator

Taxon-character matrix

Three types of linguistic rules are applied: regular expressions, Part Of Speech (POS) tag patterns, and syntactic patterns. One character extractor can be shared by multiple extractors.

Example of Results From MicroPIE in Phenology Reclassification

Example of How Training Sentences were Corrected for Improving MicroPIE Algorithm

Future Work

- Explore new methods for detecting extraction boundaries and for constructing linguistic rules automatically.
- Replace simple term lists with a new microbial ontology, Micro (Blank et al. 2016).
- Expanding the variety of target characters for extraction.

References


MicroPIE source code in GitHub (https://github.com/biosemantics/micropie2/tree@0.1.0)

This work is funded by the NSF grants DEB-1208256, DEB-1208534, DEB-1208567, DEB-1208685, and DBI-1147266.