Spring 2017

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

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 Symposium Archive*. 123.
https://digitalcommons.usm.maine.edu/thinking_matters/123

Author

Kenneth Parnow, William Lambeth, Kelsi Jackson, Jesse Florendo, Jin Mao, Hong Cui, and Carrine Blank

UNIVERSITY OF SOUTHERN MAINE
Portland • Gorham • Lewiston • Online
usm.maine.edu

MicroPIE

Developing New Tools for the Old Tree of Life

Kenneth Parnow[1], William Lambeth[1], Kelsi Jackson[1], Jesse Florendo[1], Lisa R. Moore[1], Jin Mao[2], Carrine Blank[3], Marcia Ackerman[1], Hong Cui[2],

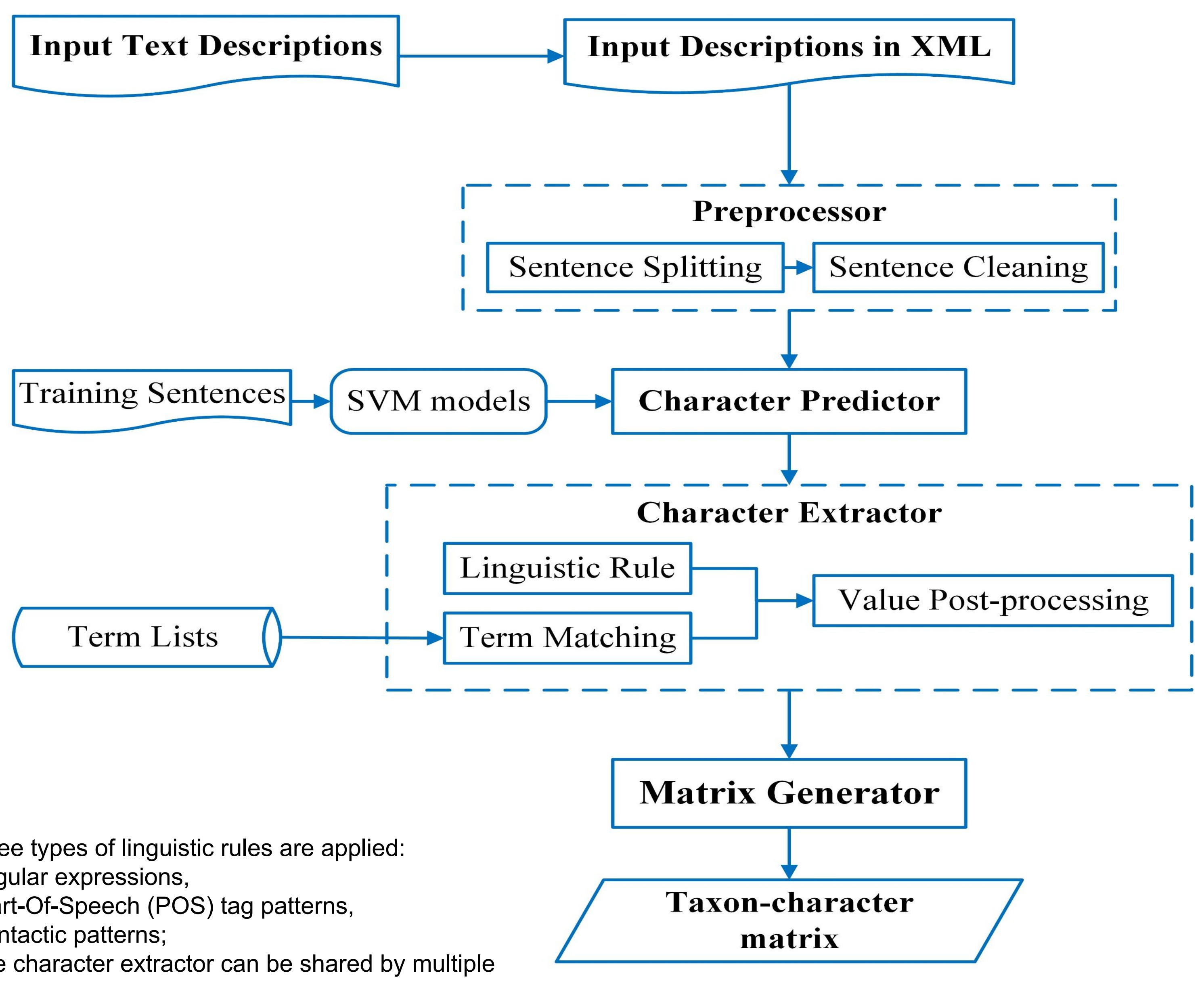
[1]Univ. of Southern Maine, Portland, ME; [2]Univ. of Arizona, Tucson, AZ; [3]Univ. of Montana, Missoula, MT, United States

NSF AVAToL:
Next Generation Phenomics for the Tree of Life
http://avatol.org/ngp/

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 with 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

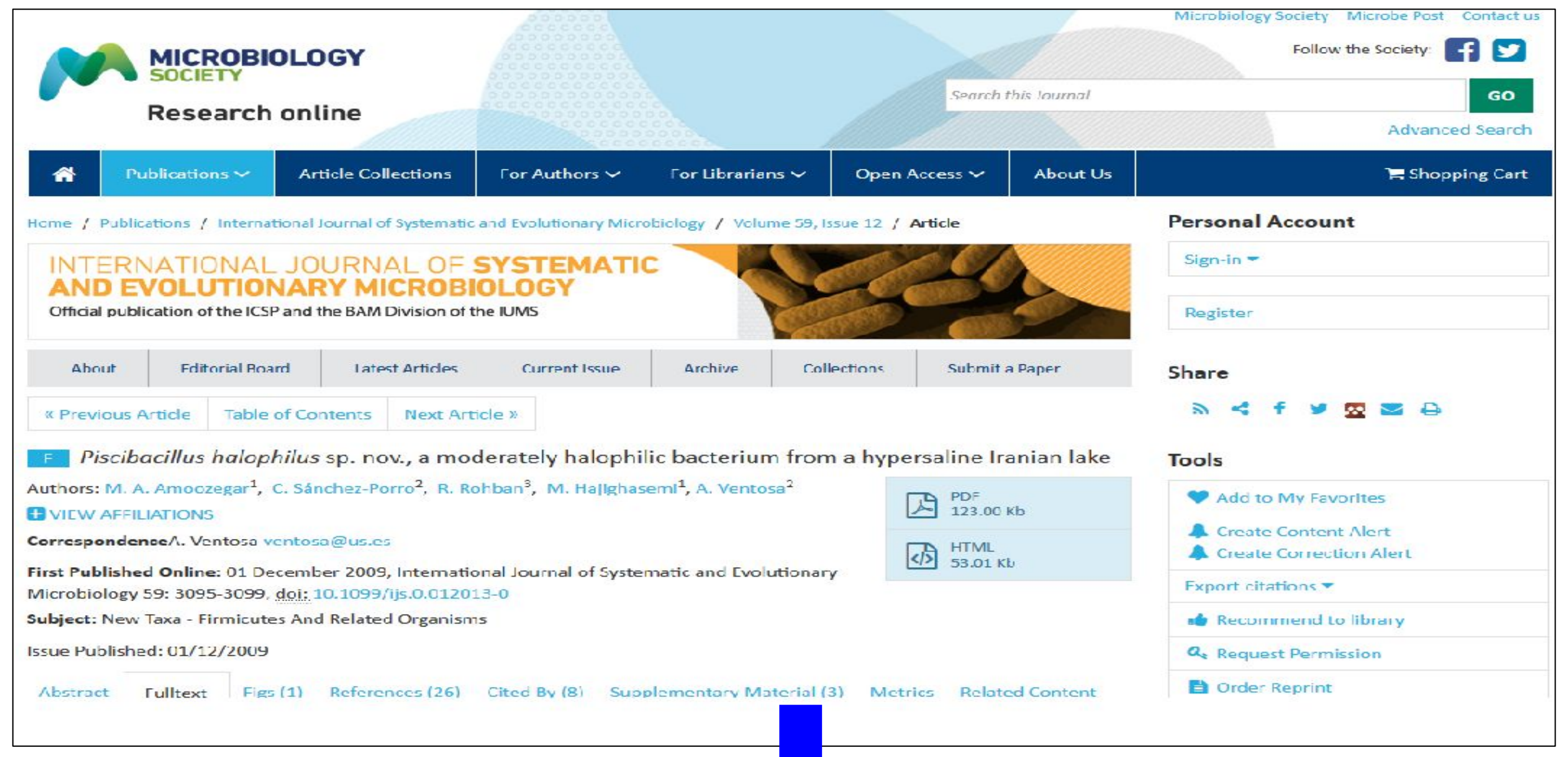

Input Text Descriptions
Input Descriptions in XML
Preprocessor
Sentence Splitting
Sentence Cleaning
Training Sentences
SVM models
Character Predictor
Character Extractor
Linguistic Rule
Term Matching
Term Lists
Value Post-processing
Matrix Generator
Taxon-character matrix

Three types of linguistic rules are applied:
- regular expressions,
- Part-Of-Speech (POS) tag patterns,
- syntactic patterns;

One character extractor can be shared by multiple extractors.

Visual Map of Workflow Used to Improve GSM v2 for MicroPIE

Example of Collecting and Formatting Publications for Use in MicroPIE

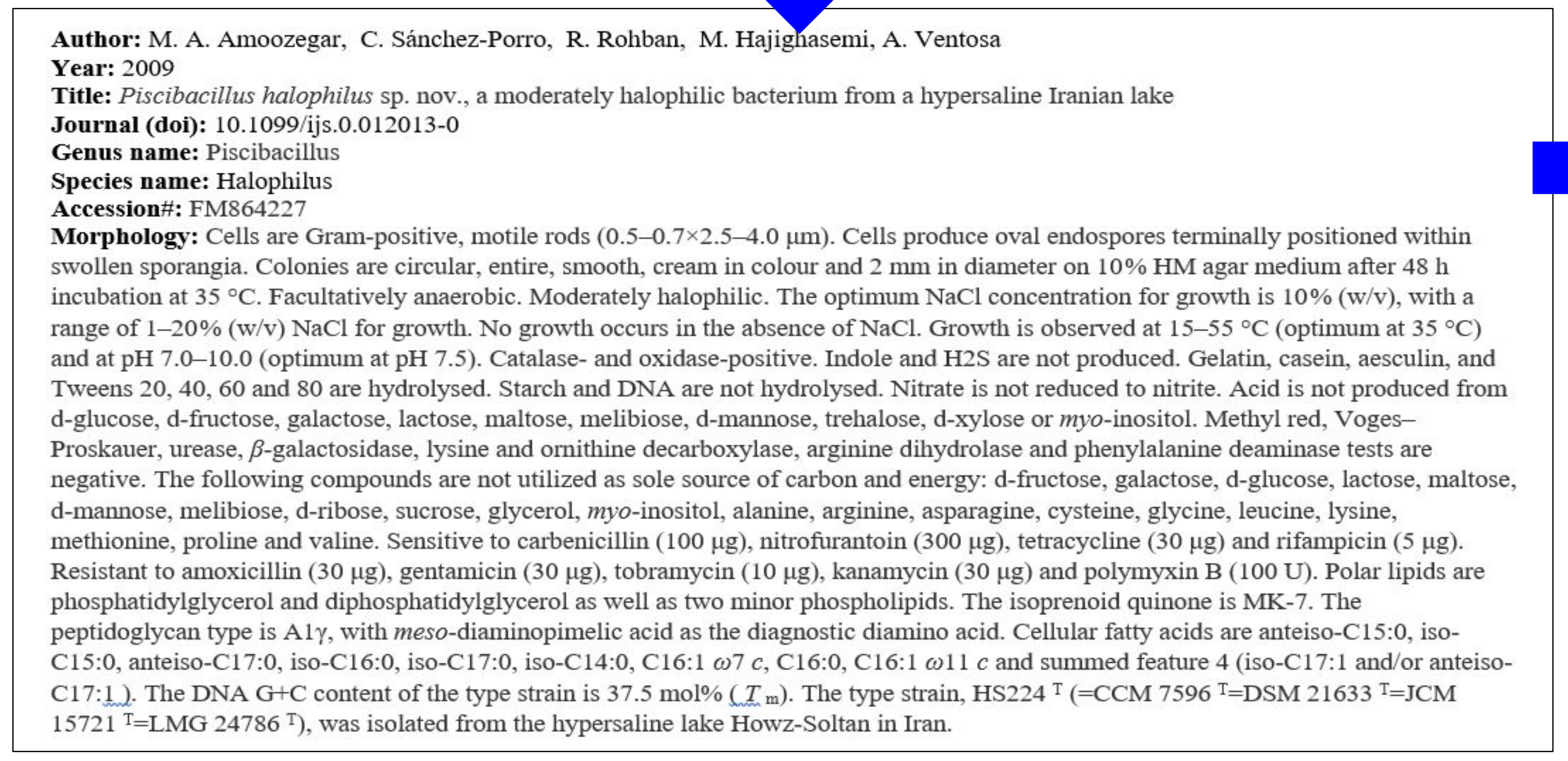


Home / Publications / International Journal of Systematic and Evolutionary Microbiology / Volume 59, Issue 12 / Article

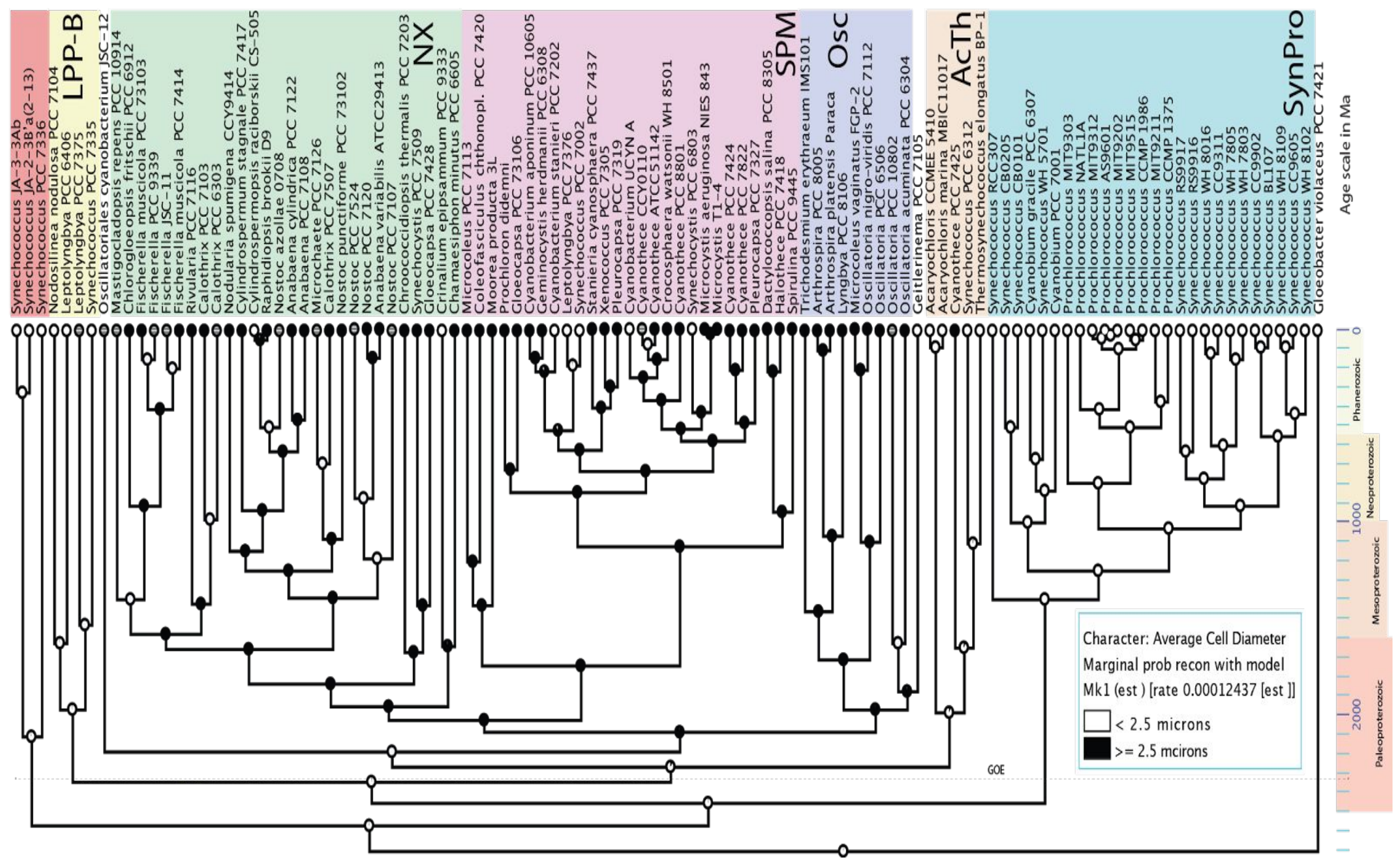
INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY
Official publication of the ICSP and the BAM Division of the IUMS

Piscibacillus halophilus sp. nov., a moderately halophilic bacterium from a hypersaline Iranian lake

Authors: M. A. Amoozegar[1], C. Sánchez-Porro[2], R. Rohban[3], M. Hajighasemi[1], A. Ventosa[2]
VIEW AFFILIATIONS
Correspondence: A. Ventosa ventosa@us.es
First Published Online: 01 December 2009, International Journal of Systematic and Evolutionary Microbiology 59: 3095-3099, doi: 10.1099/ijs.0.012013-0
Subject: New Taxa - Firmicutes And Related Organisms
Issue Published: 01/12/2009


Author: M. A. Amoozegar, C. Sánchez-Porro, R. Rohban, M. Hajighasemi, A. Ventosa
Year: 2009
Title: *Piscibacillus halophilus* sp. nov., a moderately halophilic bacterium from a hypersaline Iranian lake
Journal (doi): 10.1099/ijs.0.012013-0
Genus name: Piscibacillus
Species name: Halophilus
Accession#: FM864227
Morphology: Cells are Gram-positive, motile rods (0.5–0.7×2.5–4.0 µm). Cells produce oval endospores terminally positioned within swollen sporangia. Colonies are circular, entire, smooth, cream in colour and 2 mm in diameter on 10% HM agar medium after 48 h incubation at 35 °C. Facultatively anaerobic. Moderately halophilic. The optimum NaCl concentration for growth is 10% (w/v), with a range of 1–20% (w/v) NaCl for growth. No growth occurs in the absence of NaCl. Growth is observed at 15–55 °C (optimum at 35 °C) and at pH 7.0–10.0 (optimum at pH 7.5). Catalase- and oxidase-positive. Indole and H2S are not produced. Gelatin, casein, aesculin, and Tweens 20, 40, 60 and 80 are hydrolysed. Starch and DNA are not hydrolysed. Nitrate is not reduced to nitrite. Acid is not produced from d-glucose, d-fructose, galactose, lactose, maltose, melibiose, d-mannose, trehalose, d-xylose or *myo*-inositol. Methyl red, Voges–Proskauer, urease, *β*-galactosidase, lysine and ornithine decarboxylase, arginine dihydrolase and phenylalanine deaminase tests are negative. The following compounds are not utilized as sole source of carbon and energy: d-fructose, galactose, d-glucose, lactose, maltose, d-mannose, melibiose, d-ribose, sucrose, glycerol, *myo*-inositol, alanine, arginine, asparagine, cysteine, glycine, leucine, lysine, methionine, proline and valine. Sensitive to carbenicillin (100 µg), nitrofurantoin (300 µg), tetracycline (30 µg) and rifampicin (5 µg). Resistant to amoxicillin (30 µg), gentamicin (30 µg), tobramycin (10 µg), kanamycin (30 µg) and polymyxin B (100 U). Polar lipids are phosphatidylglycerol and diphosphatidylglycerol as well as two minor phospholipids. The isoprenoid quinone is MK-7. The peptidoglycan type is A1γ, with *meso*-diaminopimelic acid as the diagnostic diamino acid. Cellular fatty acids are anteiso-C15:0, iso-C15:0, anteiso-C17:0, iso-C16:0, iso-C17:0, iso-C14:0, C16:1 *ω*7 *c*, C16:0, C16:1 *ω*11 *c* and summed feature 4 (iso-C17:1 and/or anteiso-C17:1). The DNA G+C content of the type strain is 37.5 mol% (T_m). The type strain, HS224 ^{T} (=CCM 7596 ^{T}=DSM 21633 ^{T}=JCM 15721 ^{T}=LMG 24786 ^{T}), was isolated from the hypersaline lake Howz-Soltan in Iran.

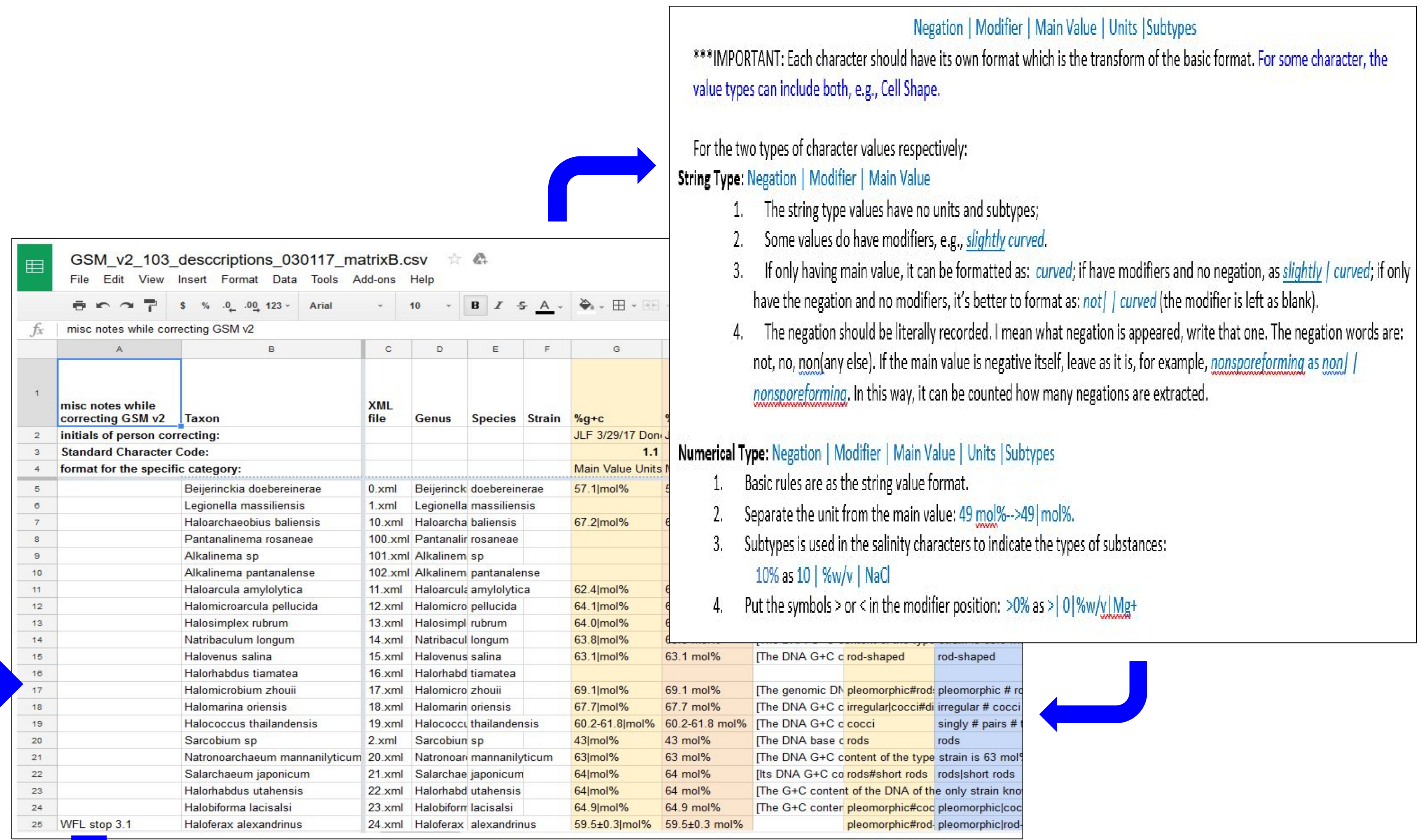
Example of Results From MircoPIE in Phenology Reclassification



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.

Example of How Training Sentences were Corrected for Improving MicroPIE Algorithm


GSM_v2_103_desccriptions_030117_matrixB.csv

misc notes while correcting GSM v2

	A	B	C	D	E	F	G				
1	misc notes while correcting GSM v2	Taxon	XML file	Genus	Species	Strain	%g+c				
2	initials of person correcting:						JLF 3/29/17 Don				
3	Standard Character Code:						1.1				
4	format for the specific category:						Main Value Units				
5		Beijerinckia doebereinerae	0.xml	Beijerinck	doebereinerae		57.1|mol%				
6		Legionella massiliensis	1.xml	Legionella	massiliensis						
7		Haloarchaeobius baliensis	10.xml	Haloarcha	baliensis		67.2|mol%				
8		Pantanalinema rosaneae	100.xml	Pantanalir	rosaneae						
9		Alkalinema sp	101.xml	Alkalinem	sp						
10		Alkalinema pantanalense	102.xml	Alkalinem	pantanalense						
11		Haloarcula amylolytica	11.xml	Haloarcula	amylolytica		62.4|mol%				
12		Halomicroarcula pellucida	12.xml	Halomicro	pellucida		64.1|mol%				
13		Halosimplex rubrum	13.xml	Halosimpl	rubrum		64.0|mol%				
14		Natribaculum longum	14.xml	Natribacul	longum		63.8|mol%				
15		Halovenus salina	15.xml	Halovenus	salina		63.1|mol%	63.1 mol%	[The DNA G+C c	rod-shaped	rod-shaped
16		Halorhabdus tiamatea	16.xml	Halorhabd	tiamatea						
17		Halomicrobium zhouii	17.xml	Halomicro	zhouii		69.1|mol%	69.1 mol%	[The genomic DN	pleomorphic#rod:	pleomorphic # ro
18		Halomarina oriensis	18.xml	Halomarin	oriensis		67.7|mol%	67.7 mol%	[The DNA G+C c	irregular|cocci#di	irregular # cocci
19		Halococcus thailandensis	19.xml	Halococcu	thailandensis		60.2-61.8|mol%	60.2-61.8 mol%	[The DNA G+C c	cocci	singly # pairs # t
20		Sarcobium sp	2.xml	Sarcobiun	sp		43|mol%	43 mol%	[The DNA base c	rods	rods
21		Natronoarchaeum mannanilyticum	20.xml	Natronoar	mannanilyticum		63|mol%	63 mol%	[The DNA G+C content of the type		strain is 63 mol
22		Salarchaeum japonicum	21.xml	Salarchae	japonicum		64|mol%	64 mol%	[Its DNA G+C co	rods#short rods	rods|short rods
23		Halorhabdus utahensis	22.xml	Halorhabd	utahensis		64|mol%	64 mol%	[The G+C content of the DNA of the		only strain kno
24		Halobiforma lacisalsi	23.xml	Halobiforn	lacisalsi		64.9|mol%	64.9 mol%	[The G+C conter	pleomorphic#coc	pleomorphic|coc
25	WFL stop 3.1	Haloferax alexandrinus	24.xml	Haloferax	alexandrinus		59.5±0.3|mol%	59.5±0.3 mol%		pleomorphic#rod-	pleomorphic|rod-

Negation | Modifier | Main Value | Units |Subtypes

***IMPORTANT: Each character should have its own format which is the transform of the basic format. For some character, the value types can include both, e.g., Cell Shape.

For the two types of character values respectively:

String Type: Negation | Modifier | Main Value

1. The string type values have no units and subtypes;
2. Some values do have modifiers, e.g., *slightly curved*.
3. If only having main value, it can be formatted as: *curved*; if have modifiers and no negation, as *slightly | curved*; if only have the negation and no modifiers, it's better to format as: *not| | curved* (the modifier is left as blank).
4. The negation should be literally recorded. I mean what negation is appeared, write that one. The negation words are: not, no, non(any else). If the main value is negative itself, leave as it is, for example, *nonsporeforming* as *non| | nonsporeforming*. In this way, it can be counted how many negations are extracted.

Numerical Type: Negation | Modifier | Main Value | Units |Subtypes

1. Basic rules are as the string value format.
2. Separate the unit from the main value: 49 mol%-->49|mol%.
3. Subtypes is used in the salinity characters to indicate the types of substances:
   10% as 10 | %w/v | NaCl
4. Put the symbols > or < in the modifier position: >0% as >| 0|%w/v|Mg+

Progress Made on GSM for MicroPIE v1 and v2

Category Code	Category Label	Micropie V.1 Sentence Count	Minimal Number of Sentences Needed	Desired Number of Sentences	Micropie V.2 Sentence Count	Total Sentence Count
2.5	Cell relationships & aggregations	114	200	300	**375**	**489**
2.8	External features	136	200	300	**190**	**326**
2.12	Biofilm formation	4	100	200	**4**	**4**
2.13	Filterability	8	100	200	**5**	**5**
2.14	Lysis susceptibility	14	100	200	**41**	**55**
2.15	Cell division pattern & reproduction	12	100	200	**24**	**36**
3.11	Pressure preference	3	100	200	**0**	**0**
3.13	Magnesium requirement for growth	35	100	200	**117**	**152**
3.14	Vitamins/cofactors required for growth	164	300	500	**97**	**261**
3.17	Salinity preference	-	100	200	**135**	**135**
5.5	Film test result	24	100	200	**3**	**27**
5.6	Spot test result	17	100	200	**0**	**0**
6.1	Fermentation products	99	200	300	**1464**	**1563**
6.3	Methanogenesis products	53	100	200	**0**	**53**
6.4	Other metabolic product	169	300	500	**762**	**931**
8.1	Symbiotic relationship	1	100	200	**8**	**9**
8.3	Pathogenic	38	200	300	**112**	**150**
8.4	Disease caused	20	200	300	**78**	**98**
8.5	Pathogen target organ	21	200	300	**67**	**88**
8.6	Haemolytic&haemadsorption properties	76	200	300	**147**	**223**

References

Blank CE, Cui H, Moore LR, Walls RL. MicrO: An Ontology of Phenotypic and Metabolic Characters, Assays, and Culture Media Found in Prokaryotic Taxonomic Descriptions. J Biomed Semantics. 2016; 7: 18.

Cui H, 2012, CharaParser for fine-grained semantic annotation of organism morphological descriptions. Journal of American Society for Information Science and Technology 63(4): 738-754, doi:10.1002/asi.22618.

Vandamme P, Debruyne L, De Brandt E, Falsen E. Reclassification of Bacteroides ureolyticus as Campylobacter ureolyticus comb. nov., and emended description of the genus Campylobacter. Int J Syst Evol Microbiol. 2010; 60(9): 2016-22.

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-1208567, DEB-1208534, DEB-1208685, and DBI-1147266.