Mining associations for organism characteristics in prokaryotes – an integrative approach

G. Pavlovic-Lazetic[1], V. Pajic[2], N. Mitic[1], J. Kovacevic[1], M. Beljanski[3]

[1]*Faculty of Mathematics,University of Belgrade, Belgrade, Serbia*
[2]*Faculty of Agriculture,University of Belgrade, Belgrade, Serbia*
[3]*Institute of General and Physical Chemistry, Serbia*

OVERVIEW

- On the problem – association of phenotype to genotype characteristics
- Related work
- Databases
- Database mining: Example
- Text mining; Method
- Example
- Conclusion

On the problem – association of phenotype to genotype characteristics

- Genotype
- Phenotype
- One of the main problems in the post-genomic era is to associate phenotypic characteristics of an organism to proteins and metabolites encoded by its genome.
- Provide for deeper comprehension of evolutionary processes

- Some prediction possibilities, e.g., trends prediction of some pandemia

- Importance for bio-defense

Related work

- Microbial phenotypes are typically due to the joined action of multiple gene functions
  - Inferring gene function from cross-organismal distribution of phenotypic traits; reliable when the phenotype does not arise from many alternate mechanisms – (Jim et al, 2004)
  - Co-occurrence between sets of genes and the phenotype; association rule mining algorithms; NETCAR, PCAR (Tamura, D'haeseleer, 2008)
  - Thermal adaptation vs. structural disorder and functional complexity (Burra et al, 2010)
  - Genotype-phenotype associations by combining information from a biomedical database with the molecular information from COGs database
  - Association of genes to phenotypes by literature mining and comparative genome analysis (Korbel et al, 2005)

Databases

- Plenty of genotype data and gene sequences for different organisms
- Usually well structured and placed into databases
- Data on phenotypic characteristics of organisms often scattered across different text documents, e.g., scientific papers or encyclopedias
- Public databases (both kinds of information), e.g.,
  - NCBI Entrez Genome database (the most extensive)
  - Comprehensive Microbial Resource
  - Genome Atlas Database
  - DOE-Joint Genome Institute databases
  - Databases and tools for specific types of genotype-phenotype research, etc.

OVERVIEW

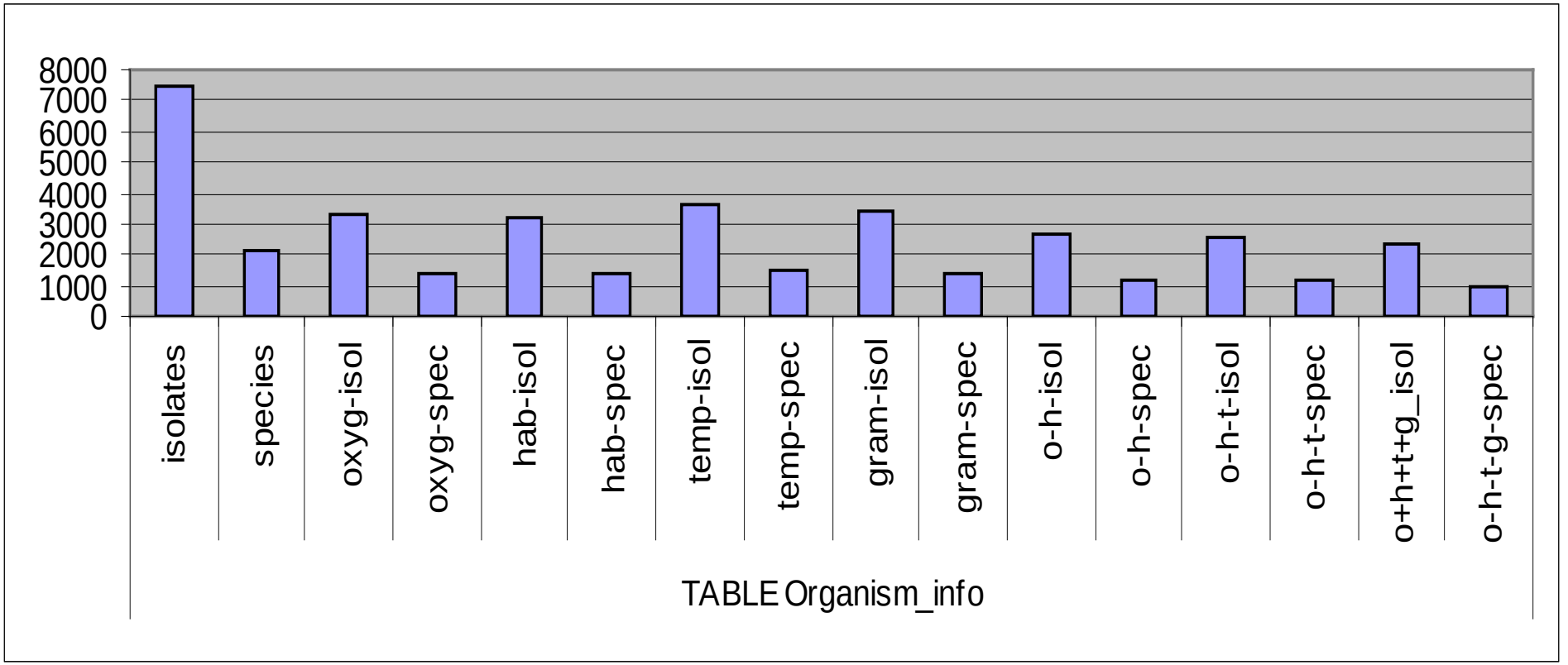
- On the problem – association of phenotype to genotype characteristics
- Related work
- Databases
- Database mining: Example
- Text mining; Method
- Example
- Conclusion

Database mining: Example

- NCBI Entrez Genome database
- An instance (2011)- collection (table) *organism_info*:
- Characteristics:
  - genome size, GC content
  - shape, oxygen, habitat, salinity, temperature, gram, motility, pathogenicity
  - number of isolates / different species : 7467/2163
  - some columns sparse
  - under half-populated

Database mining: Example

- Statistics on organism_info:



Database mining: Example

- Modalities

oxygen	count(*)
	4182
Aerobic	1064
Anaerobic	792
Facultative	1300
Microaerophilic	129

habitat	count(*)
	4273
Aquatic	502
Host-associated	1429
Multiple	856
Specialized	204
Terrestrial	203

Database mining: Example

- Modalities

temp	count(*)
	3835
Cryophilic	1
Hyperthermophilic	78
Mesophilic	3377
Psychrophilic	35
Thermophilic	141

gram	count(*)
	4043
+	1371
-	2047
_	6

Database mining: Example

- Association rule mining
  - *Given a set of transactions consisting of one or more elements (*items*), find rules that predict occurrence of an item based on occurrence of other items in the transaction*
  - Support / Confidence / Support*Confidence / Lift

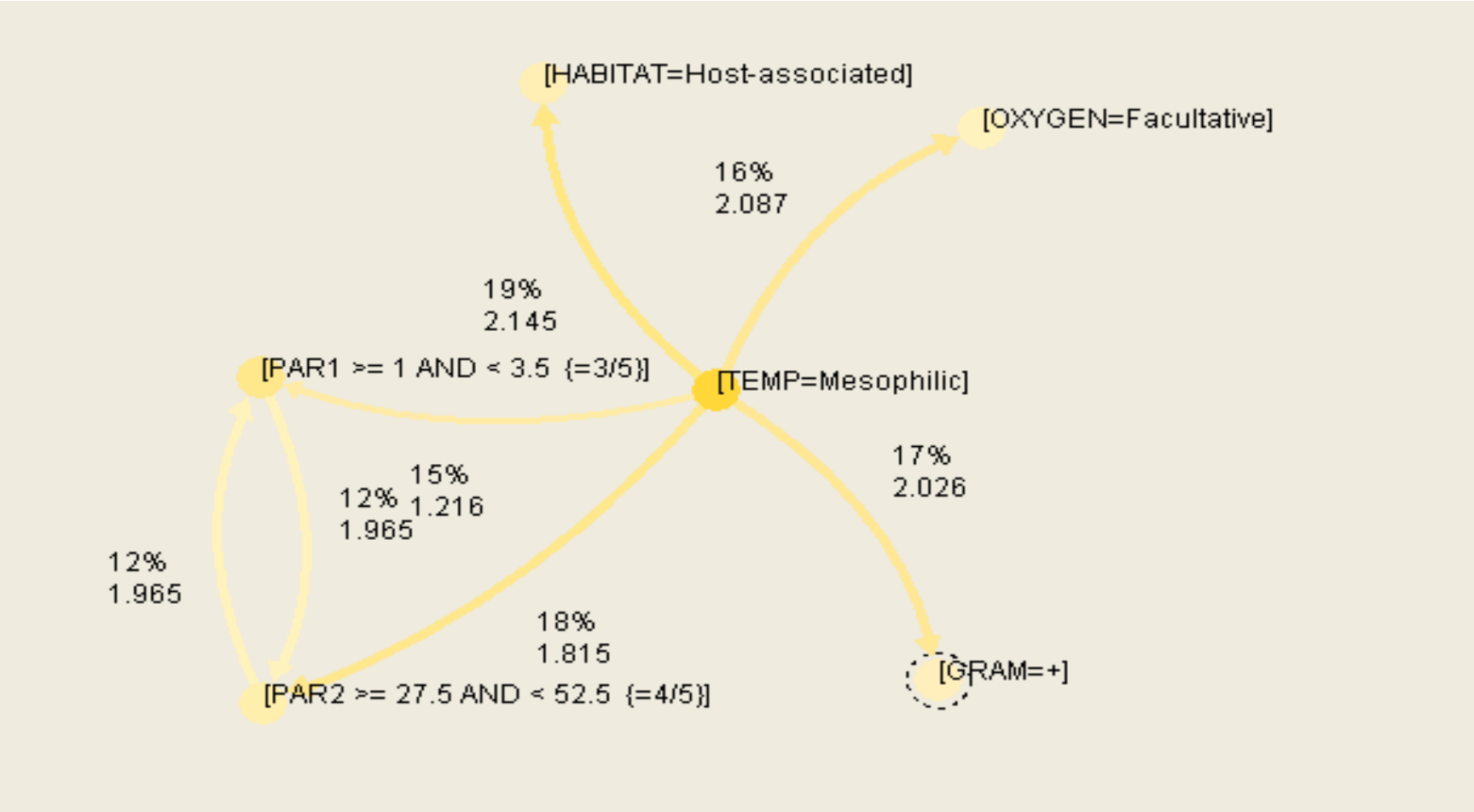
$$s(A \Rightarrow B) = \frac{\sigma(A \cup B)}{N}$$

$$c(A \Rightarrow B) = \frac{\sigma(A \cup B)}{\sigma(A)}$$

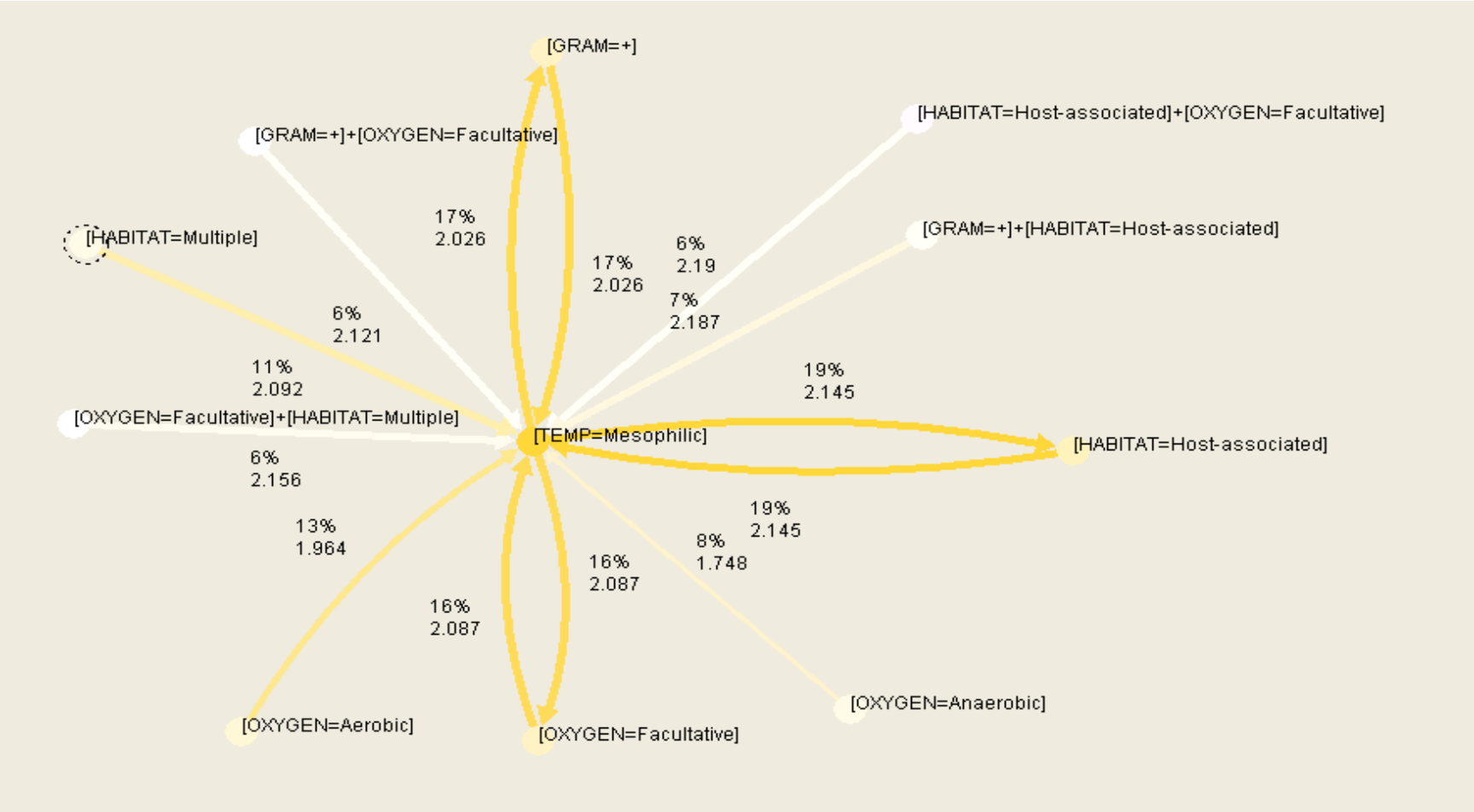
$$\mathrm{lift}\,(X \Rightarrow Y) = \frac{\mathrm{supp}(X \cup Y)}{\mathrm{supp}(X) \times \mathrm{supp}(Y)}$$

- IBM Intelligent Miner - two examples:
  - geno/pheno
  - pheno

Database mining: Example 1



Database mining: Example 2

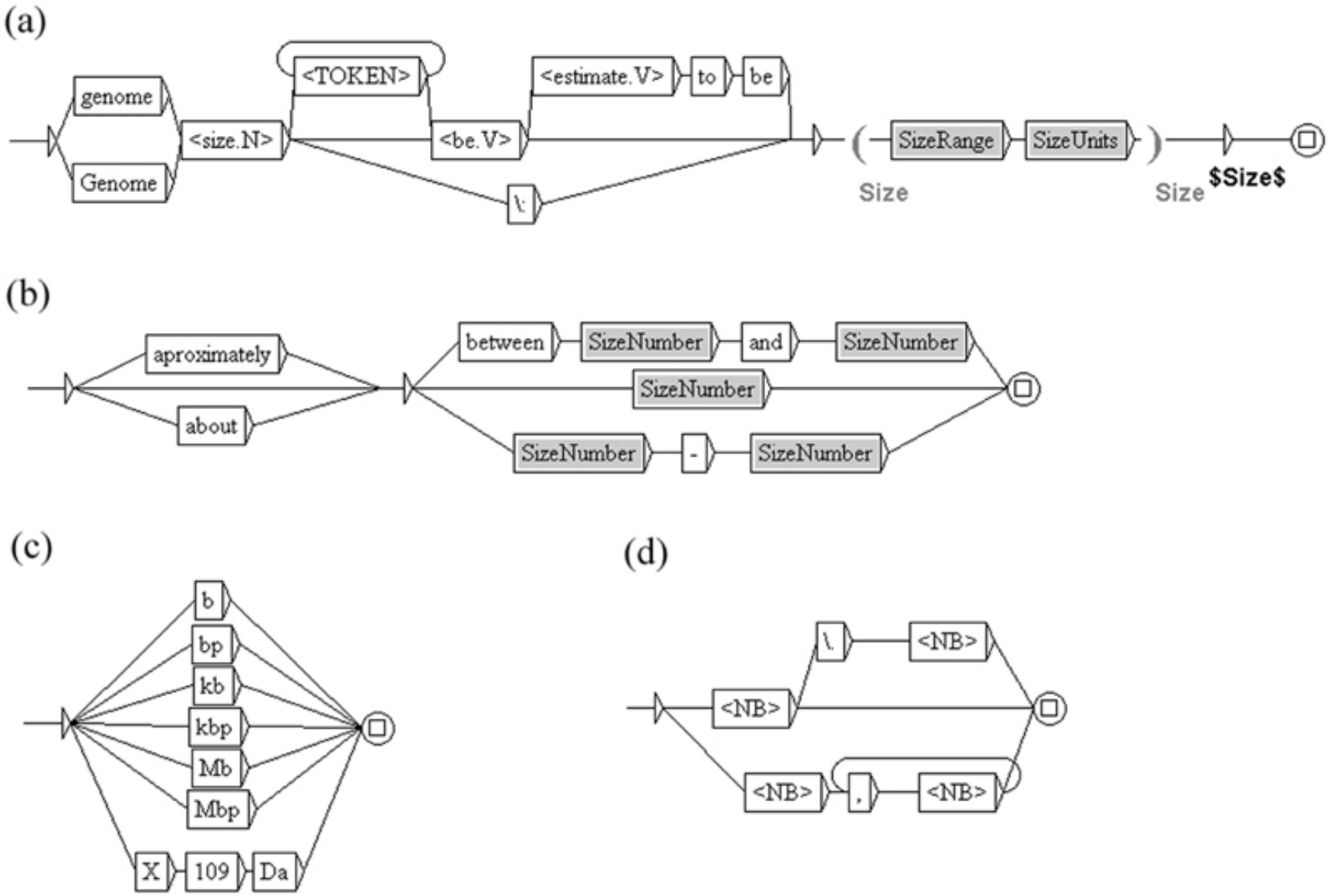


OVERVIEW

- On the problem – association of phenotype to genotype characteristics
- Related work
- Databases
- Database mining: Example
- Text mining; method
- Example
- Conclusion

Text mining; method

- Much more information about microbes and other organisms in unstructured and semi-structured documents
- Text mining: Goal - mining genotype / phenotype organism characteristics from text
- Method:
  - two phase method based on finite state transducers (FST) for information extraction from text
- Tool: system UNITEX; regular expressions; graphs; example



FSTs for genome size

OVERVIEW

- On the problem – association of phenotype to genotype characteristics
- Related work
- Databases
- Database mining: Example
- Text mining; method
- Example
- Conclusion

Example

- Text mining in bioinformatics
  - Materials: encyclopedia of microorganisms
  - *Bergey's Manual of Systematic Bacteriology, Volume 2 : The Proteobacteria (2005)*
  - *Bergey's Manual of Systematic Bacteriology, Volume 3: The Firmicutes (2009)*
  - *Bergey's Manual of Systematic Bacteriology, Volume 4: The Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobacteres, Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae, and Planctomycetes*, (2010)

robium zavarzinii Hirsch 1989b, 495^{VP} (Effective
n: Hirsch 1989, 1903.)
.*i.* M.L. gen. n. *zavarzinii* of Zavarzin, named for
zin, the Russian microbiologist who isolated these

· cells drop- or pear-shaped, somewhat slender,
ae that rarely branch. Mother cells 0.63 × 1.8 μm
5–0.9 × 0.7–2.5 μm). Swarmer cells with 1–3 sub-
ella. In liquid media under most growth condi-
ettes are formed, since mother cells produce a
lfast. Growth in liquids initially as turbidity and
pellicle, with precipitation on the bottom. Colo-
lid media are colorless to light brownish or beige,
id shiny, with entire edges.
organotrophic, aerobic, oligocarbophilic. Good
h the following carbon sources: methanol, meth-
Cl, formate, *n*-butyrate, isovalerate, crotonate, β-
ityrate, ethanol, *n*-propanol, isobutanol, and glyc-
th is stimulated significantly by acetate, *n*-valerate,
rate, galacturonate, formaldehyde, D-glucose, D-
D-melibiose, amygdalin, esculin, chitin, Bacto
DL-lysine, DL-aspartate, and dilute human urine.
sources utilized are: NH_4^+, NO_2^-, NO_3^-, and
Bacto peptone. There is slow growth in the ab-
added nitrogen sources (oligonitrophily). Poor
growth on sheep blood agar with α-hemolysis. T
antibiotics inhibit growth at 30 μg (per disc):
neomycin, and tetracycline. Streptomycin at 1
inhibitory. There is growth in the presence of
Temperature range: 15–37°C. Optimal pH: 6.5
light inhibits growth slightly.

Grow anaerobically with nitrate and gas forn
methanol as the carbon source). With methy
and thioglycolate, there is little growth. Catala
chrome oxidase are positive; gelatin liquefactio
Poly-β-hydroxybutyrate is a storage product.

Not pathogenic for mice or guinea pigs.

Genome size: 2.73 × 10^9 Da (strain ZV-
Boelke et al., 1985).

Habitat: peaty and moist soil near Moscow,

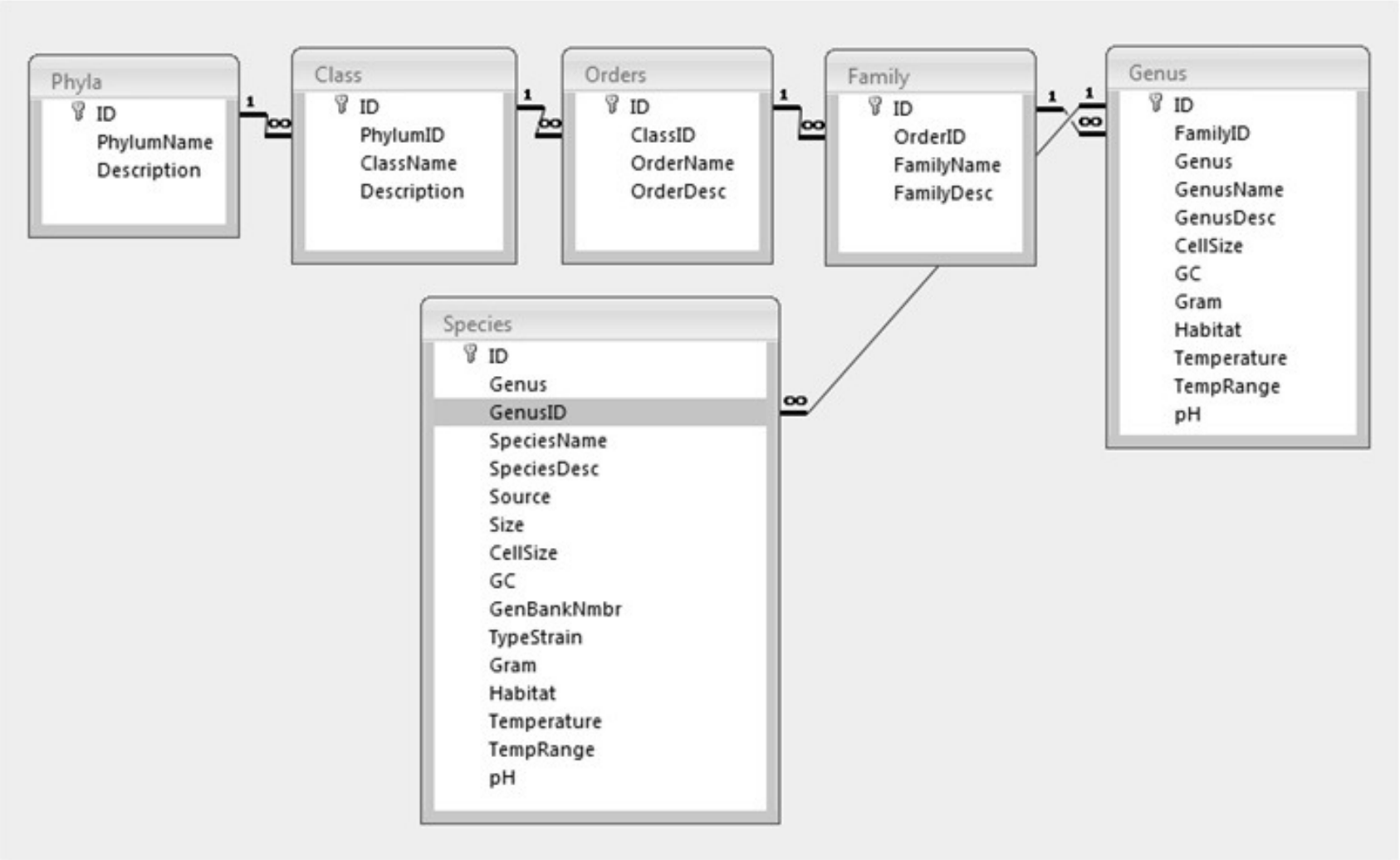
The mol% G + C of the DNA is: 61.8–64.8 (Bd
(Mandel et al., 1972; Gebers et al., 1986; U
Komagata, 1987b; Urakami et al., 1995b).

Type strain: ATCC 27496, IFAM ZV-622.

GenBank accession number (16S rRNA): Y1430

Additional Remarks: Additional strains incluc
580, ZV-620, MY-619, MC-625, MC-629, MC-6
627.

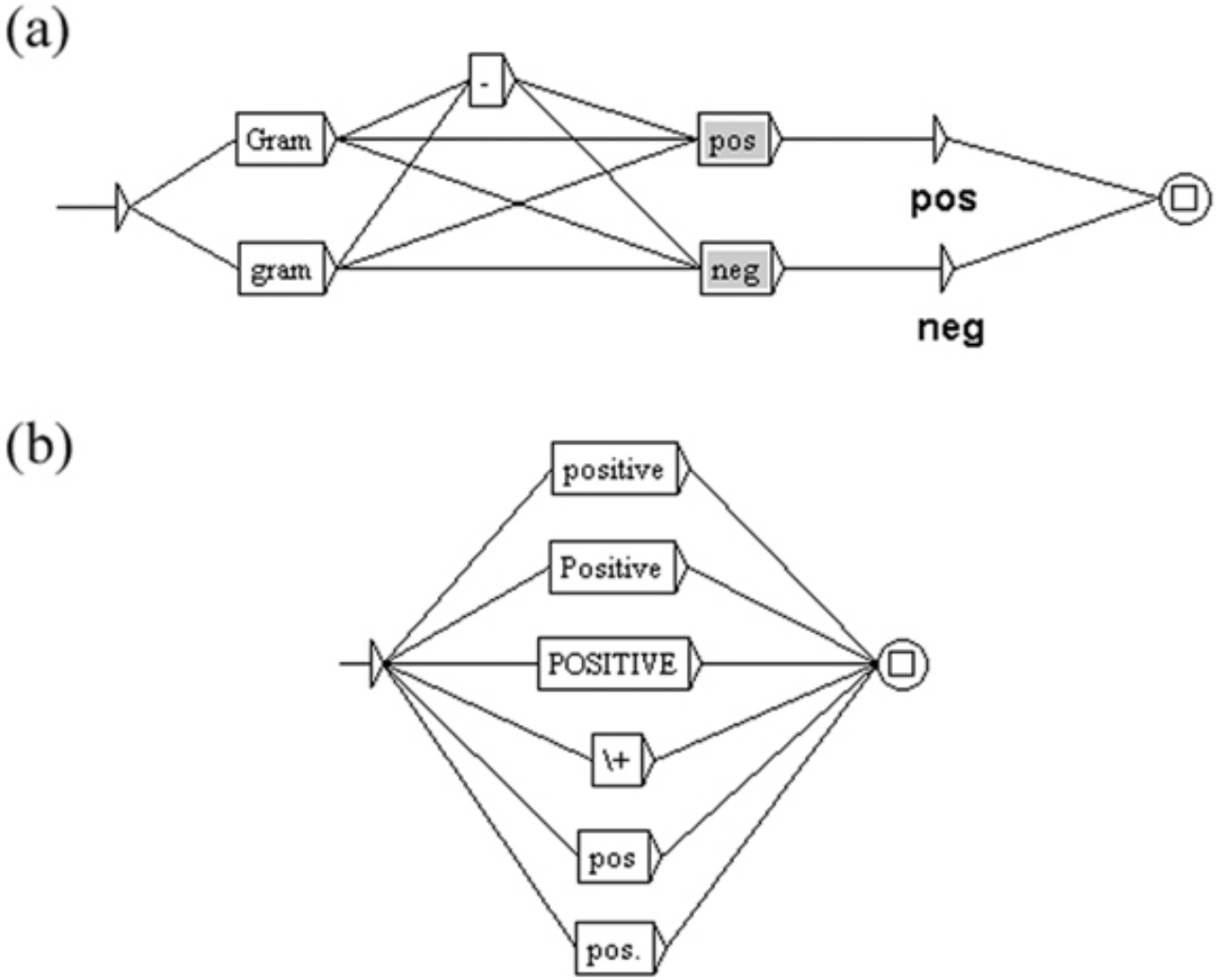
Example: database



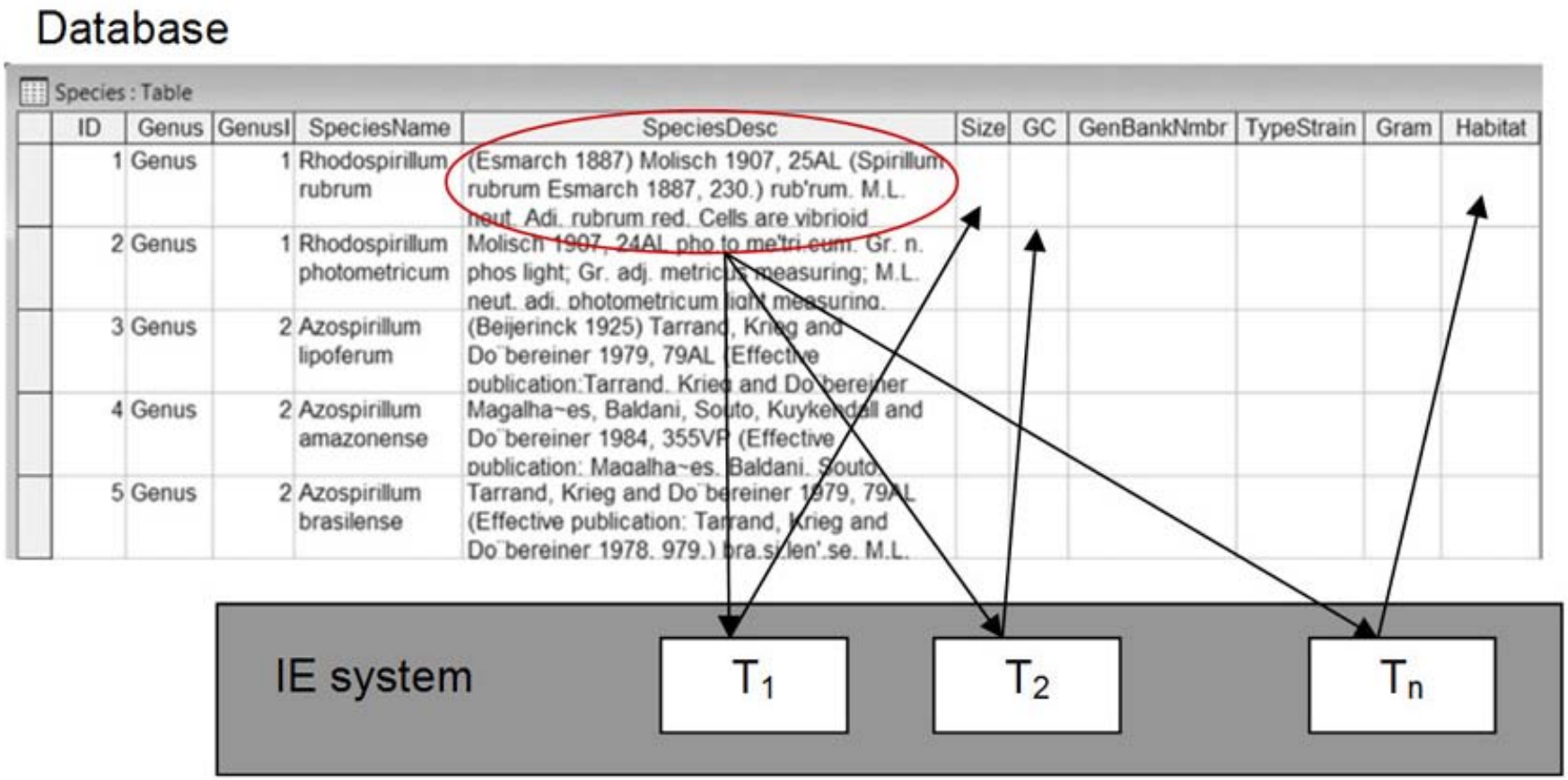
Example: graph "*genome size*"

- *"genome sizes of four G. oxydans strains were estimated to be **between 2240 and 3787 kb**"*
- *"genome size of R. prowazekii is **1,111,523 bp**"*
- *"genome size of R. africae is **1.248 kb**"*
- *"genome size of R. australis is **1256–1276 kbp**"*
- *"genome size is **2.62 X 109 Da**"*
- *"Genome size: **2.73 X 109 Da**"*
- *"genome size is **1.713 Mbp**"*
- *"genome size was estimated to be **approximately 4061 kb**"*
- *"genome size of all the classical strains examined was **about 3000 kb**"*

Example: graph "*gram*"



Example: phase 1



Example: phase 2

Species : Table

ID	GenusID	SpeciesName	SpeciesDesc	Source	Size	CellSize	GC	GenBankNmbr	TypeStrain	Gram	Habitat	Temperature	TempRange	pH
1320	430	Bacillus thermocloacae	Demharter and Hensel 1989a, 495(Effective publication: Demharter and Hensel 1989b, 274.) ther.mo.cloca.cae. Gr. n. therme heat; L. n. cloaca sewer; N.L. gen. n. thermocloacae of a heated sewer. Aerobic, moderately alkaliphilic and thermophilic, Gram-positive, nonmotile rods, 0.5-0.8 mm by 3.0-8.0mm. Description is based upon three isolates. Spore formation only	3		0.5-0.8	42.8-43.7	Z26939 (DSM 5250)	S 6025, DSM 5250	pos	heat-treated sewage sludge	thermophilic	55-60	8-9

Insert Format Records Tools Window Help

Type a qu

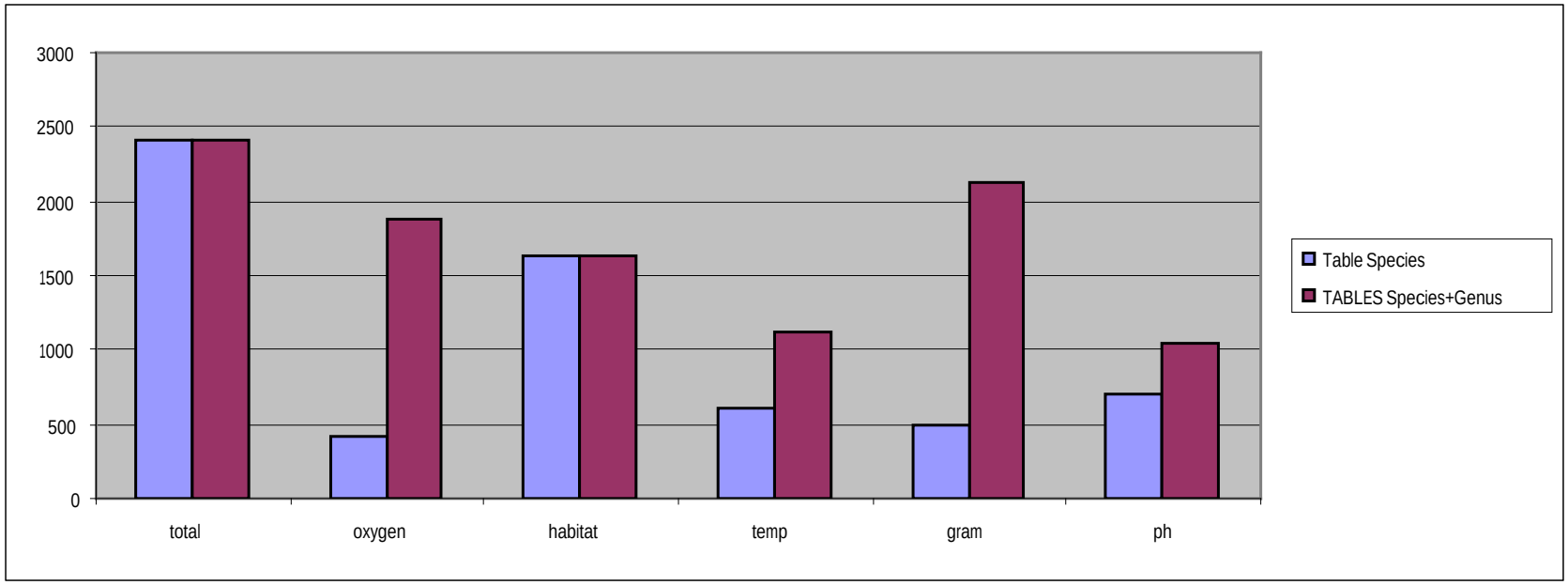
SpeciesName	SpeciesDesc	Source	Size	CellSize	GC	GenBankNmbr	TypeStrain	Gram	Habitat	Temperature	TempRange	pH	Oxygen
Bacillus hermocloacae	Demharter and Hensel 1989a, 495(Effective publication: Demharter and Hensel 1989b, 274.) ther.mo.cloca.cae. Gr. n. therme heat; L. n. cloaca sewer; N.L. gen. n. thermocloacae of a heated sewer. Aerobic, moderately alkaliphilic and thermophilic, Gram-positive, nonmotile rods, 0.5-0.8 mm by 3.0-8.0mm. Description is based upon three isolates. Spore formation only	3		0.5-0.8	42.8-43.7	Z26939 (DSM 5250)	S 6025, DSM 5250	pos	heat-treated sewage sludge	thermophilic	55-60	8-9	aerobic
Bacillus huringiensis	Berliner 1915, 29AL thur.in.gi.encsis. N.L. masc. adj. thuringiensis of Thuringia, the German province from where the organism was first isolated. Facultatively anaerobic, Gram-positive, usually motile rods 1.0-1.2 by 3.0-5.0 mm, occurring singly and in pairs and chains, and forming ellipsoidal, sometimes cylindrical, subterminal, sometimes paracentral, spores	3		1.0-1.2 by 3.0-5.0	33.5-40.1	D16281 (IAM 12077)	IAM 12077, ATCC 10792, NRRL NRS-996, DSM 2046, LMG 7138, NCIMB 9134	neg	all continents, including Antarctica				facultative
Bacillus usciae	Bonjour and Aragno 1985, 223VP (Effectivepublication: Bonjour and Aragno 1984, 400.) tusccia.e. L. gen. n. tusciae from Tuscia, the Roman name for the region of central Italy where the organism was found. Facultatively chemolithoautotrophic, moderately thermophilic, strictly aerobic, motile (by one lateral flagellum). Gram-positive rods 0.8 by 4-5	3		0.8 by 4-5	57-58	AB042062 (IFO 15312)	Aragno T2, DSM 2912, LMG 17940, IFO EMBL/	neg	an acidic pond in a solfatara in Italy	thermophilic		4.2-4.8	aerobic
Bacillus vallismortis	Roberts, Nakamura and Cohan 1996, 474VP val.lis.morctis. L. n. vallis valley; L. fem. n. mors death; N.L. gen. fem. n. vallismortis of Death Valley. Aerobic, Gram-positive, motile rods, forming ellipsoidal spores which lie centrally or paracentrally in unswollen sporangia. Cells 0.8-1.0 by 2.0-4.0 mm, occurring singly and in short chains. Colonies	3		0.8-1.0 by 2.0-4.0	43.0	AB021198 (DSM 11031)	DV1-F-3, NRRL B-14890, DSM 11031, LMG 18725, KCTC 3707	neg	desert soil	28	28-30		aerobic
Bacillus vedderi	Agnew, Koval and Jarrell 1996, 362 (Effective publication: Agnew, Koval and Jarrell 1995, 229.) vedcderi. M.L. gen. n. vedderi of Vedder, named after A. Vedder, the Dutch microbiologist who described Bacillus alcalophilus in 1934. Alkaliphilic, facultatively anaerobic, Gram-positive, motile, narrow rods forming ellipsoidal to spherical spores which lie	3		1.5	38.3	Z48306 (JaH)	JaH, DSM 9768, ATCC 7000130, LMG 17954, NCIM B 13458	neg	red mud bauxite-processing waste, using alkaline oxalate enrichment	40	40	10.0	facultative
Bacillus vietnamensis	Noguchi, Uchino, Shida, Takano, Nakamura and Komagata 2004, 2119VP vi.et.nam.encsis. N.L. adj. vietnamensis referring to Vietnam, the country where the type strain was isolated. Cells are rod-shaped, measuring 0.5-1.0 by 2.0-3.0mm, Gram-positive and aerobic. They are motile with peritrichous flagella. Ellipsoidal spores develop centrally in the cells and	3		0.5-1.0 by 2.0-3.0	43	AB099708		neg	Vietnamese fish sauce and from the Gulf of Mexico		50	6.5-10.0	
Bacillus vireti	Heyrman, Vanparys, Logan, Balcaen, Rodr guez-D az, Felske and De Vos 2004, 54VP vi.reti. L. gen. n. vireti of a field Facultatively	3		0.6-0.9	39.8-40.3	AJ542509 (LMG 21834)	LMG 21834, DSM 15602	neg	soil of Drentse A agricultural research area	30		7 after 48	facultative

368 of 2412

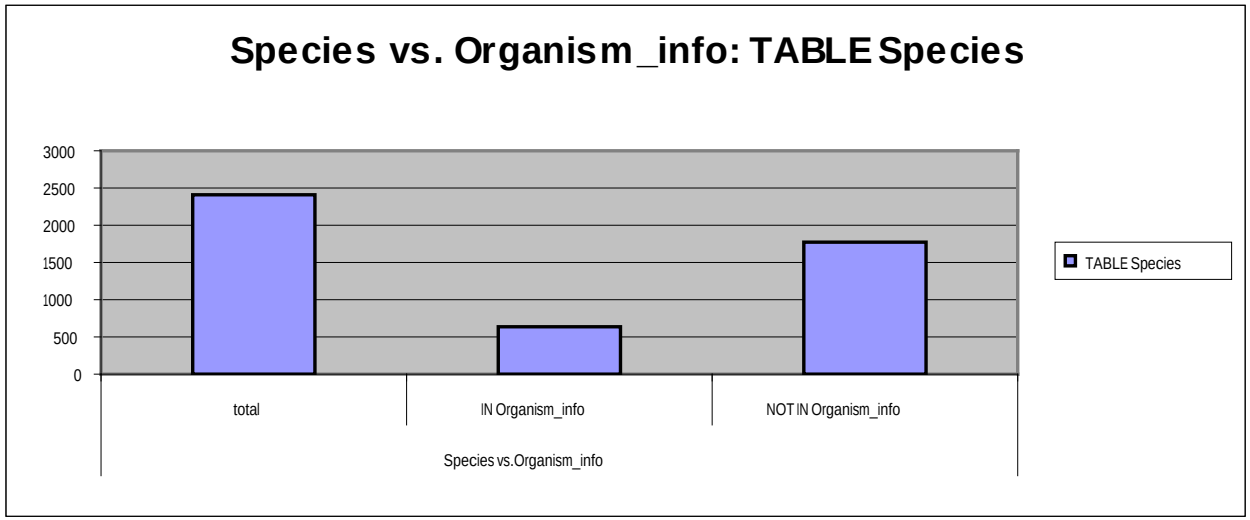
SSH S... | genotype-... | IEEE-IJDMB | preuredjeno | 2 Micros... | bakterije : ... | Species : T... | script - No... | rezultati - ... | komentar-... | Skype™ - ... | EN | 100%

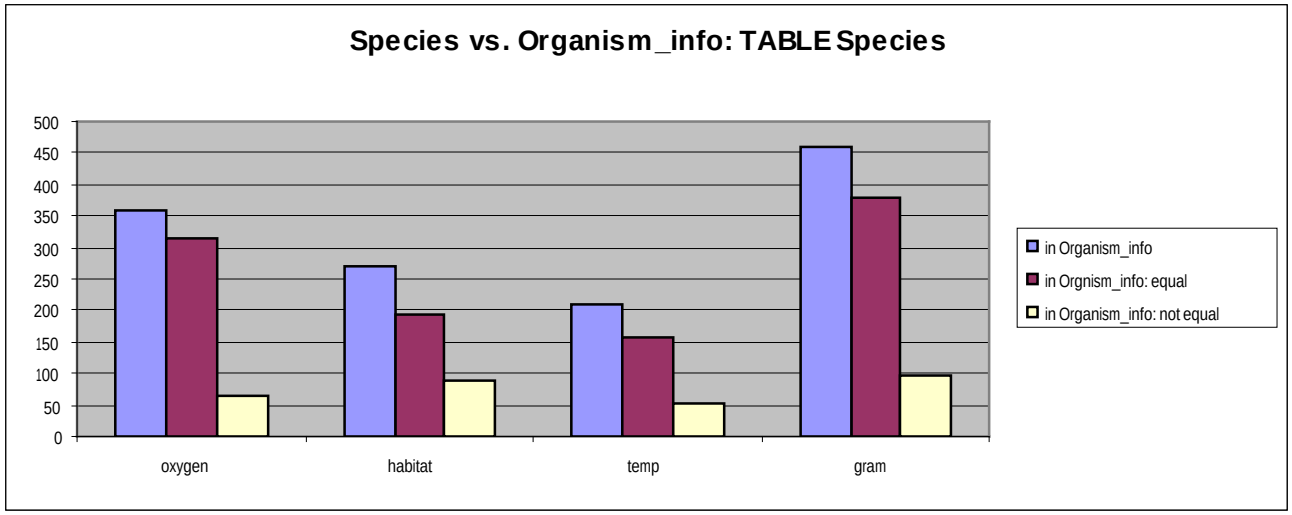
Example: postprocessing

- Manually / biocuration
- E.g., habitat -> aquatic, terrestrial, host-associated,…
- Tables Species, Genus

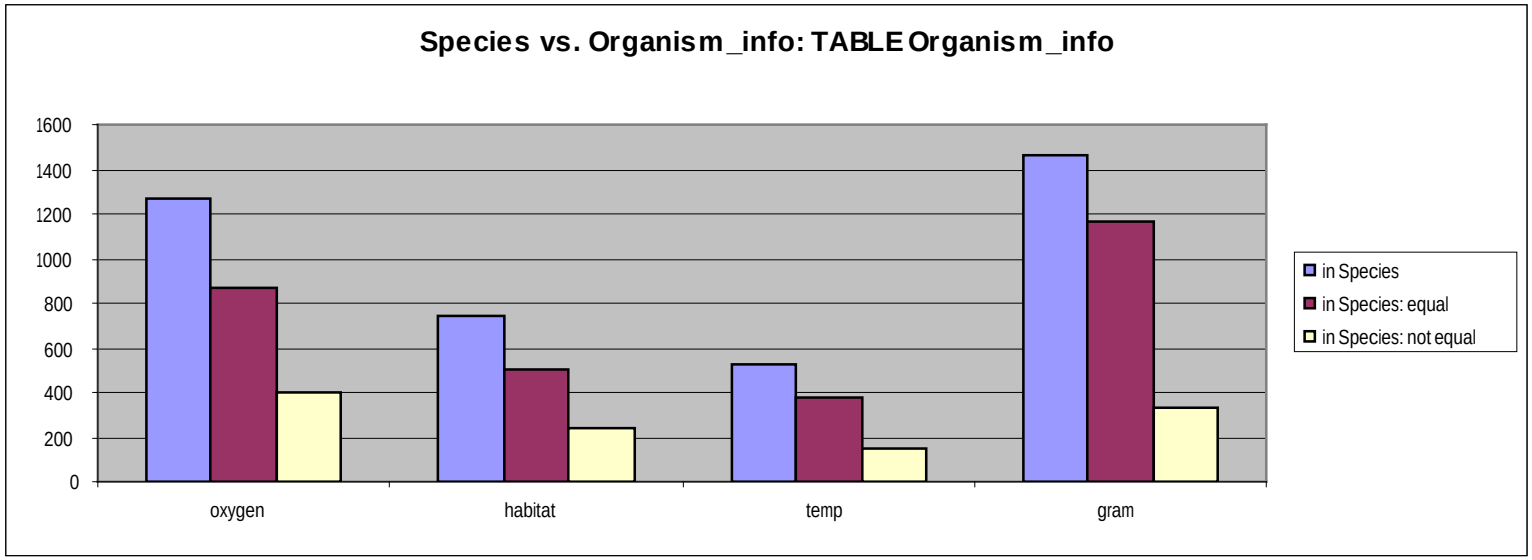


Example: Species vs.Organism_info



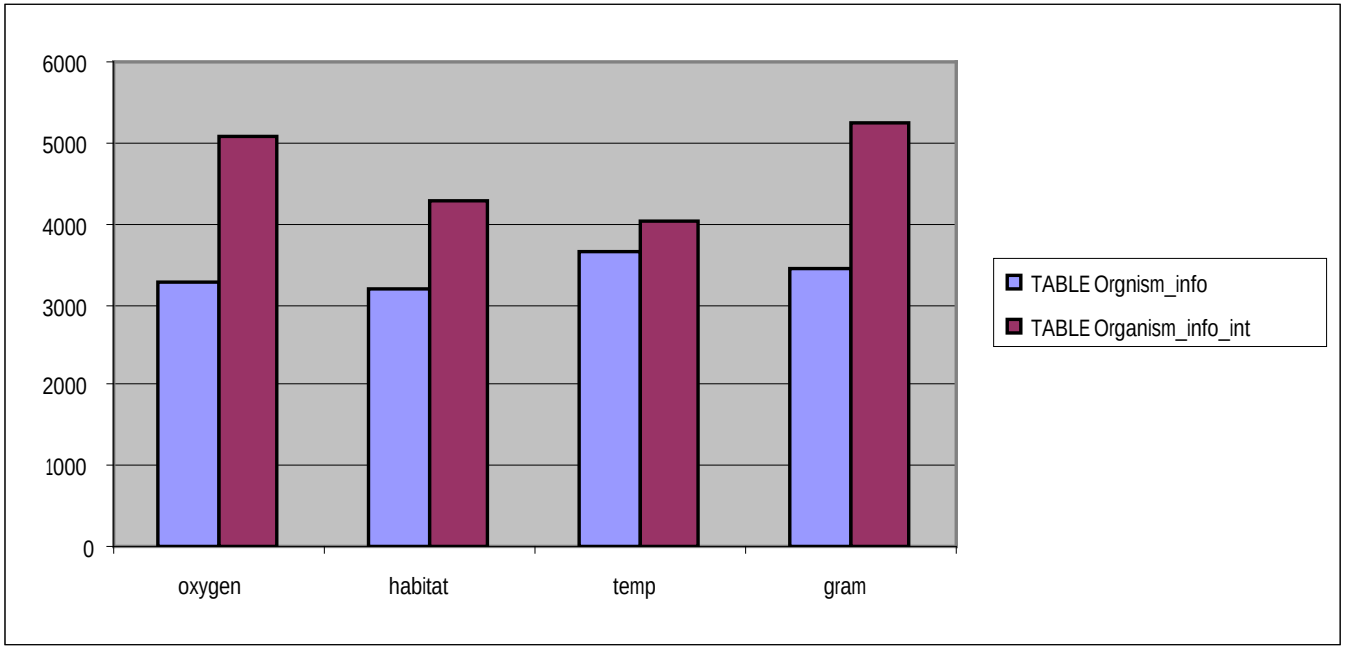


Example: Species vs.Organism_info



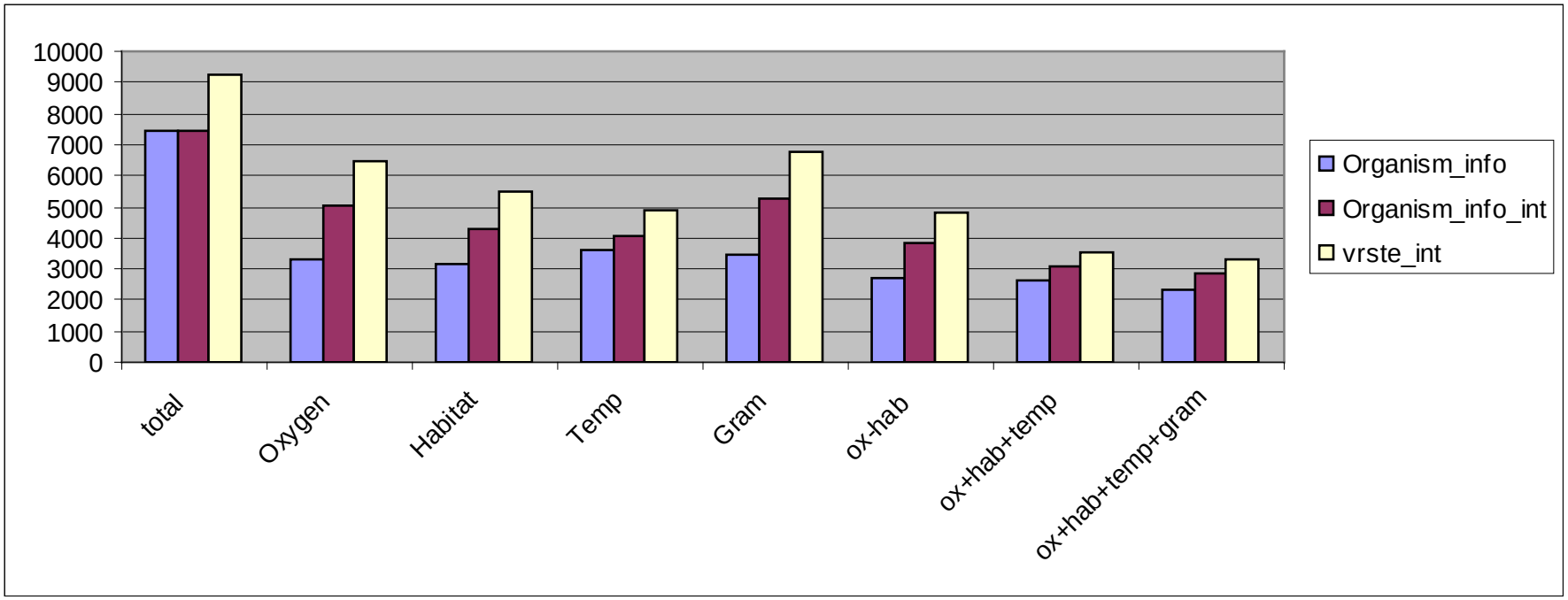
Example: integration of data sources

- Tables: *organism_info, species*
- Populating the table *organism_info* by using data from the table *species*
- Table *organism_info_int*



Example: integration of data sources

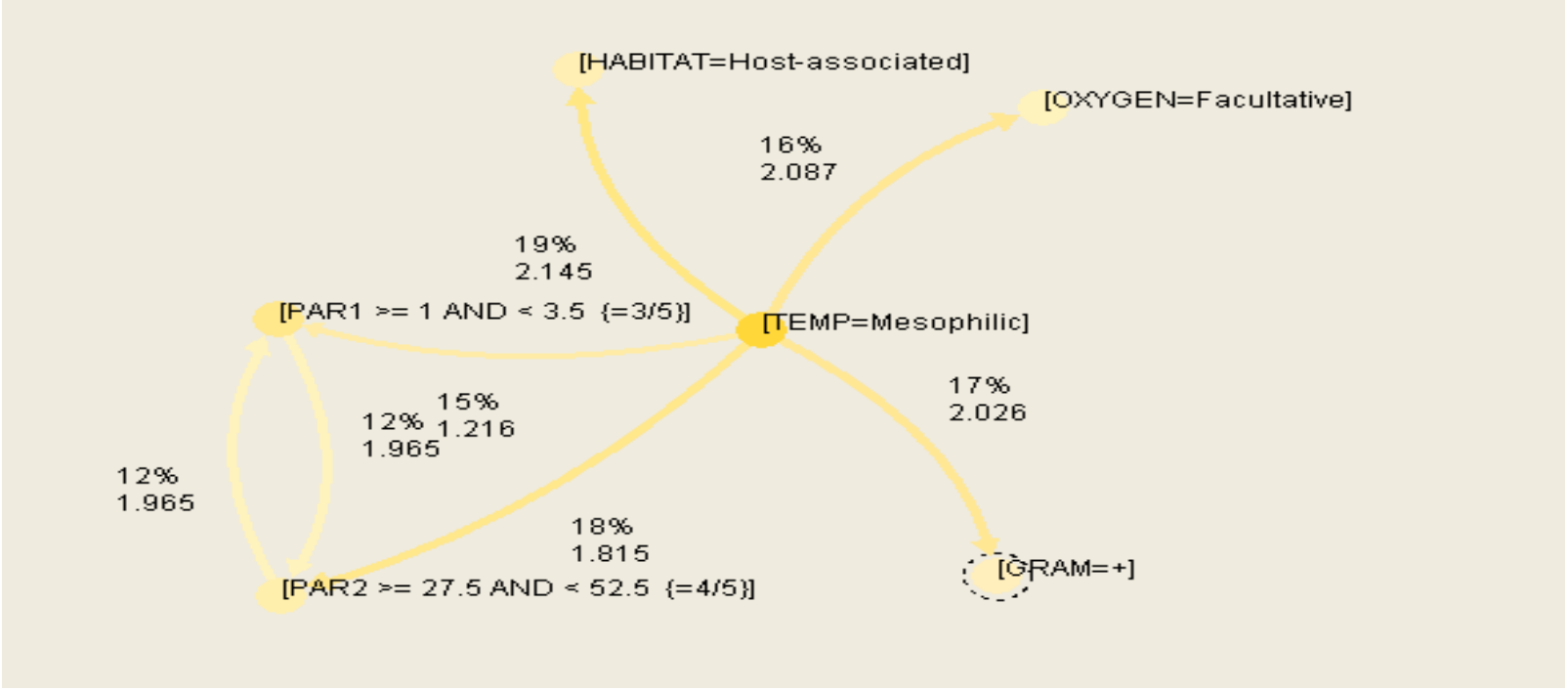
- Tables: *organism_info_int, species*
  - union of the tables *organism_info_int* and *species;*
  - projection to common attribute
- Table *species_int*

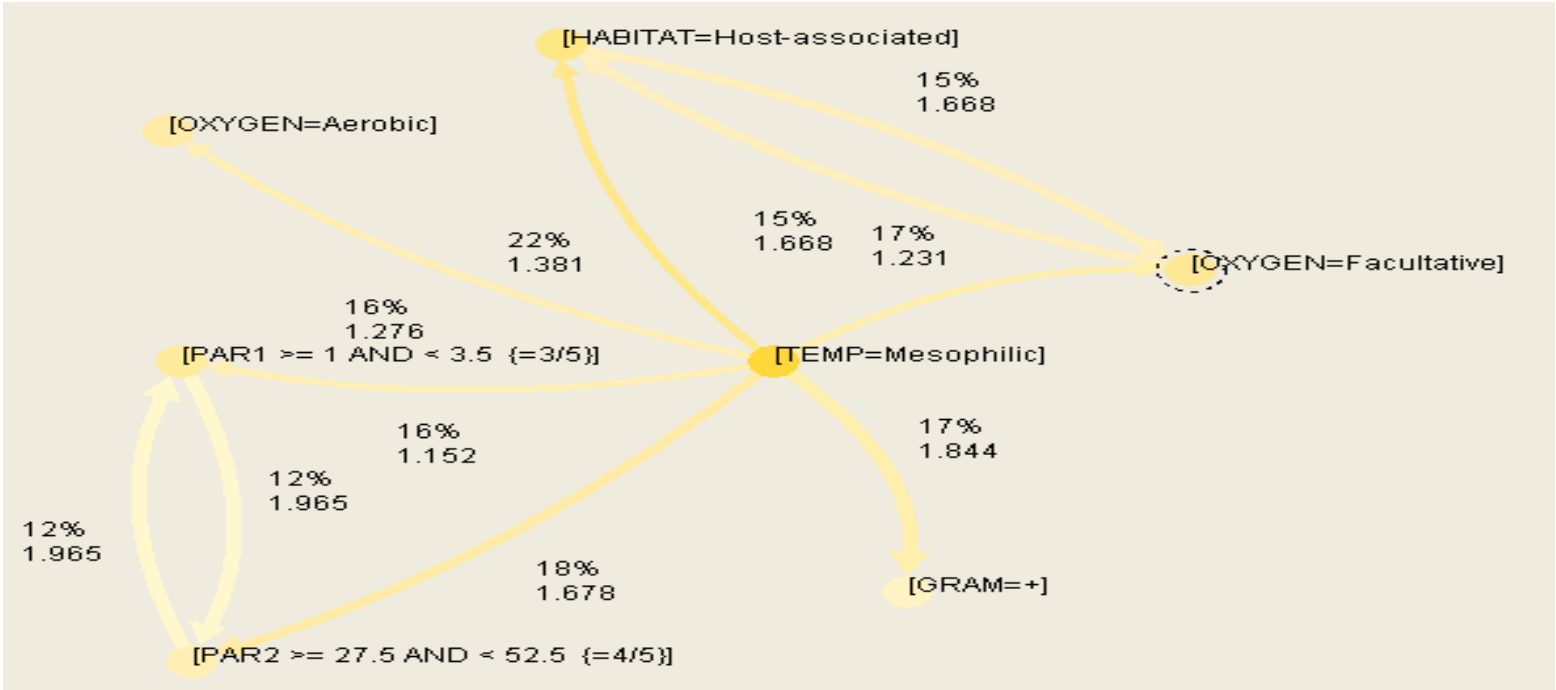


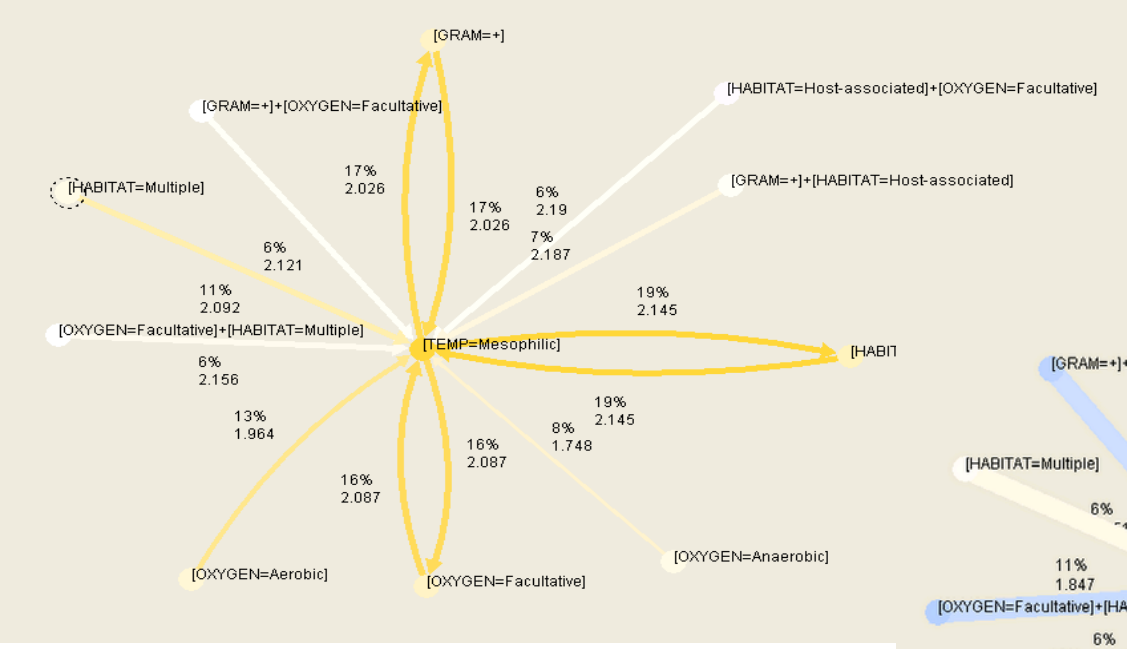
Example: association rule mining

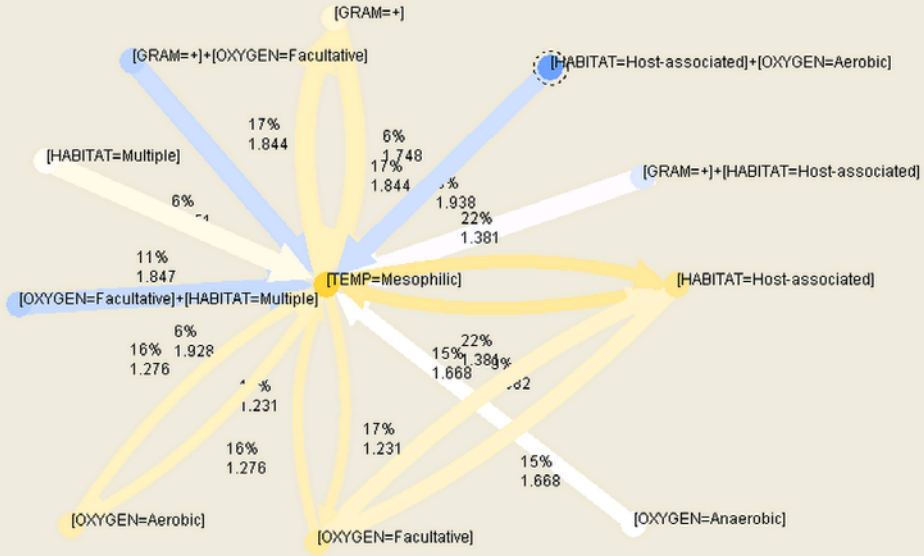
- 1. tables organism_info, organism_info_int
  - (some) genotype / phenotype chracteristics (genome size, GC%, shape, habitat, oxygen, temperature, gram, ph)

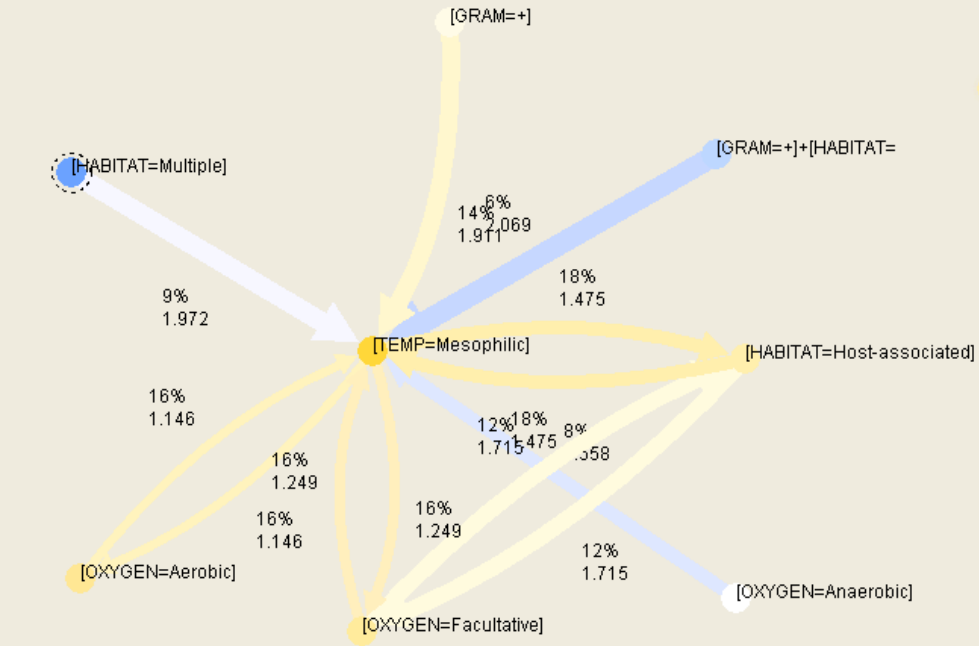
- 2. tables organism_info, organism_info_int, species_int
  - Smaller attribute set; phenotype (eco) characteristics (habitat, oxygen, temperature, gram)


[HABITAT=Host-associated]
[OXYGEN=Facultative]
16%
2.087
19%
2.145
[PAR1 >= 1 AND < 3.5 {=3/5}]
[TEMP=Mesophilic]
17%
2.026
15%
12%
1.216
1.965
12%
1.965
18%
1.815
[PAR2 >= 27.5 AND < 52.5 {=4/5}]
[GRAM=+]


[HABITAT=Host-associated]
15%
1.668
[OXYGEN=Aerobic]
15%
1.668
17%
1.231
22%
1.381
[OXYGEN=Facultative]
16%
1.276
[PAR1 >= 1 AND < 3.5 {=3/5}]
[TEMP=Mesophilic]
16%
1.152
17%
1.844
12%
1.965
12%
1.965
18%
1.678
[GRAM=+]
[PAR2 >= 27.5 AND < 52.5 {=4/5}]


[GRAM=+]
[GRAM=+]+[OXYGEN=Facultative]
[HABITAT=Host-associated]+[OXYGEN=Facultative]
[HABITAT=Multiple]
[GRAM=+]+[HABITAT=Host-associated]
17%
2.026
17%
2.026
6%
2.19
7%
2.187
6%
2.121
11%
2.092
19%
2.145
[OXYGEN=Facultative]+[HABITAT=Multiple]
[TEMP=Mesophilic]
[HABIT
6%
2.156
19%
2.145
13%
1.964
8%
1.748
16%
2.087
16%
2.087
[OXYGEN=Anaerobic]
[OXYGEN=Aerobic]
[OXYGEN=Facultative]


[GRAM=+]
[GRAM=+]+[OXYGEN=Facultative]
[HABITAT=Host-associated]+[OXYGEN=Aerobic]
17%
1.844
6%
1.748
[HABITAT=Multiple]
17%
1.844
[GRAM=+]+[HABITAT=Host-associated]
6%
1.938
22%
1.381
11%
1.847
[TEMP=Mesophilic]
[HABITAT=Host-associated]
[OXYGEN=Facultative]+[HABITAT=Multiple]
6%
1.928
16%
1.276
22%
15%
1.668
1.231
17%
1.231
16%
1.276
15%
1.668
[OXYGEN=Aerobic]
[OXYGEN=Anaerobic]
[OXYGEN=Facultative]


[GRAM=+]
[HABITAT=Multiple]
[GRAM=+]+[HABITAT=
6%
14%
2.069
1.911
18%
1.475
9%
1.972
[TEMP=Mesophilic]
[HABITAT=Host-associated]
16%
1.146
12%
1.715
18%
1.475
16%
1.249
16%
1.146
16%
1.249
12%
1.715
[OXYGEN=Aerobic]
[OXYGEN=Anaerobic]
[OXYGEN=Facultative]

OVERVIEW

- On the problem – association of phenotype to genotype characteristics
- Related work
- Databases
- Database mining: Example
- Text mining; method
- Example
- Conclusion

Conclusion (and beyond…)

- Significant enlargement of databases
- Applicable to many other specific areas and tasks
- Association rules – modest
- Sparse structured data
- Integration of several microbial databases
- Rich set of characteristics
  - Genotypic
  - Phenotypic
  - Structural
  - Protein disorder
  - COGs
  - RNA secondary structures
  - …
- Multivariate analysis

References

- A Cross-Genomic Approch for Systematic Mapping of Phenotypic Traits to Genes, Km Jim, Kush Parmr, Mona Singh, Saeed Tavazoie, Genome Research, 14, 2004, 109-115
- Microbial genotype-phenotype mapping by class association rule mining, Makio Tamura, patrik D'haeseleer, Bioinformatics, 24 (13), 2008, 1523-1529
- **Ontology-guided data preparation for discovering genotype-phenotype relationships, Adrien Coulet12*, Malika Smaïl-Tabbone2, Pascale Benlian3, Amedeo Napoli2 and Marie-Dominique Devignes2** , *BMC Bioinformatics* 2008, **9**(Suppl 4):S3
- **. J. Korbel, T. Doerks, L. J. Jensen, C.Perez-Iratxeta, S. Kaczanowski, S. D. Hooper, M. A. Andrade, P. Bork: Systematic association of genes to phenotypes by genome and literature mining, PLoS Biol, 3, 2005, pp. 134-134.**
- **11. N. J. MacDonald, R. G. Beiko, Efficient learning of microbial genotype-phenotype association rules, Bioinformatics, 26, 2010, pp. 1834-1840.**
- **Reduction in Structurl Disorder and Functional Complexity in the Therml Adpttion of Prokryotes, Prasad V. Burr, Lajos Kalmr, Peter Tompa, Plos One, 2010. 5(8), e12069**
- Goh, C.S., Gianoulis, T.A., Liu, Y., Li, J., Paccanaro, A., Lussier, Y.A. and Gerstein, M. (2006)
- 'Integration of curated databases to identify genotype-phenotype associations', *BMC Genomics,* 7, pp.257–257.

- .

Journal of Integrative Bioinformatics, 8(2):164, 2011 http://journal.imbio.de

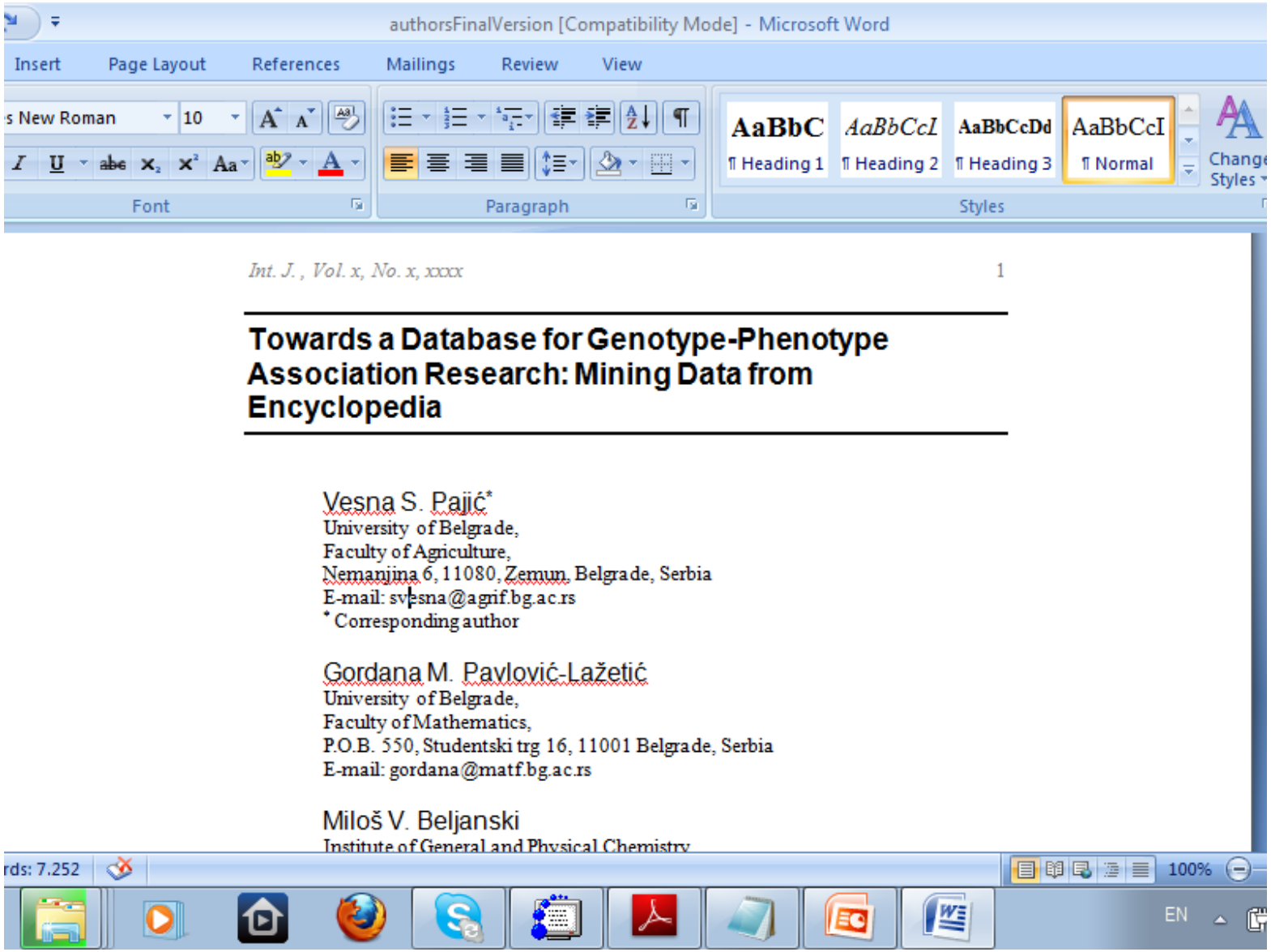
Putting Encyclopaedia Knowledge into Structural Form: Finite State Transducers Approach

Vesna Pajić

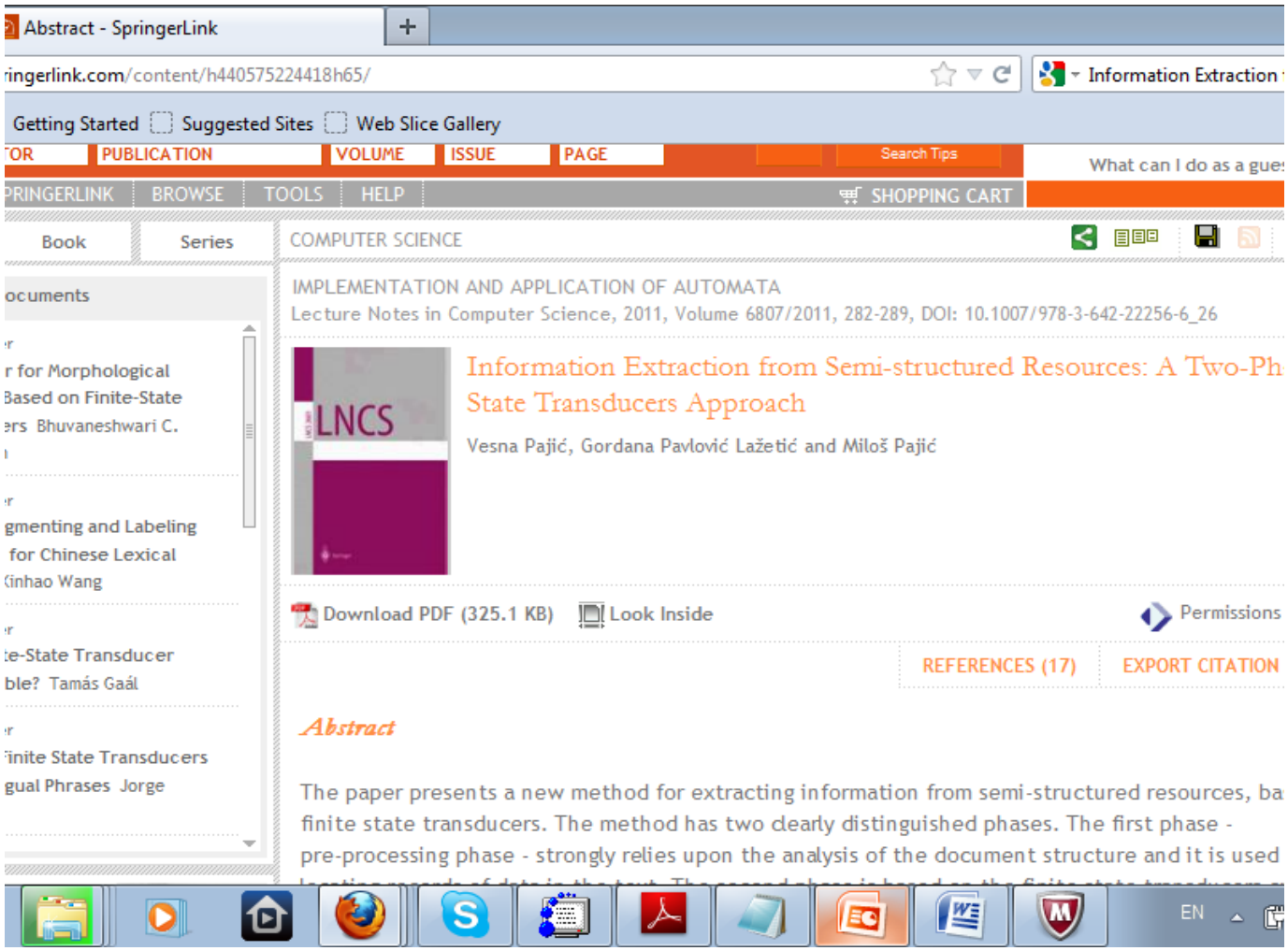
Faculty of Agriculture, University of Belgrade, Nemanjina 6, 11080 Zemun, Belgrade, Republic of Serbia, svesna@agrif.bg.ac.rs

Summary

In biology and functional genomics in particular, understanding the dependence and interplay between different genome and ecological characteristics of organisms is a very challenging problem. There are some public databases which combine this kind of information, but there is still much more information about microbes and other organisms that reside in unstructured and semi-structured documents, such as encyclopaedias. In this



Accepted in International Journal of Data Mining in Bioinformatics, Inderscience Enterprise Ltd. .



[Information Extraction from Semi-structured Resources: A Two-Phase Finite State Transducers Approach]