Mining Associations for Organism Characteristics in Prokaryotes – an Integrative Approach

Gordana Pavlovic-Lazetic[1], Vesna Pajic[2], Nenad Mitic[1], Jovana Kovacevic[1] and Milos Beljanski[3]

[1]University of Belgrade, Faculty of Mathematics, Studentski trg 16, 11000 Belgrade, Serbia
{gordana, nenad, jovana}@matf.bg.ac.rs
[2]University of Belgrade, Faculty of Agriculture, Nemanjina 6, 11080 Belgrade, Serbia
svesna@agrif.bg.ac.rs
[3]Institute for General and Physical Chemistry, Studentski trg 16, 11000 Belgrade, Serbia
mbel@matf.bg.ac.rs

Abstract. Correlations and associations between specific organism characteristics (such as genome size, genome GC content, optimal growth temperature, habitat, oxygen requirements) may provide for deeper comprehension of evolutionary processes as well as for some prediction possibilities, e.g., trends prediction of some pandemia. There is a plenty of genotype data and gene sequences for different organisms, which is usually well structured and deposited into databases. On the other side, data on phenotypic characteristics of organisms are often scattered across different text documents, e.g., scientific papers or encyclopedias. We reconsider correlations between organism characteristics for superkingdoms Archaea and Bacteria and extend the study in a number of ways. We use a larger dataset of prokaryotes as well as a larger set of characteristics by integrating several existing databases with data obtained by literature mining. We recalculate some high-expectation correlations between genomic characteristics (genome size, GC content, distribution among functional groups of proteins) and apply algorithms for association rule mining in order to identify the most confident associations between specific modalities of both genotype and phenotype characteristics.

Keywords. association rule mining; prokaryotes; genotype characteristics; phenotype characteristics

1 Introduction

Rapid developments in molecular biology research technologies produced, as a result, a huge amount of biological data that may be best analyzed using mathematical and computational methods and techniques. Publicly available databases (such as those maintained by the National Center for Biotechnology Information, NCBI [1], DOE's Integrated Microbial Genomes, IMG [2], or PathoSystems Resource Integration Center, PATRIC[3]) provide information for a variety of prokaryotic genomes (both superkingdoms Archaea and Bacteria). Although these databases contain different kind of data (both genotypic and phenotypic, such as: total genome length, number and length of chromosomes and plasmids, GC content, coding / non-coding sequence

ratio, number of protein coding / non-coding genes, codon usage, HGT - number of genomic islands, gram staining, morphological characteristics, habitat characteristic, taxonomic and others), still much larger portion of such information, especially phenotypic ones, resides in semi-structured and unstructured forms such as encyclopedia, science articles, books, web pages. These kinds of resources are unsuitable for computer analysis and need to be transformed into structured forms (e.g. databases).

The overall goal is not only to analyze genomic features and their relations, but also relations between genotypic features and their phenotypic characteristics, as well as taxonomic characteristics, and to do it on a data collection as comprehensive as possible. These relationships provide for deeper comprehension of evolutionary processes and for some prediction possibilities, e.g., trends prediction of some pandemia, by interrelating the features in non-obvious ways.

The main goal of this work is to present a convincing example of how information extraction from a semi-structured source can be integrated with knowledge from structured databases, thus enriching the data collection for mining correlations and associations among genomic and phenotypic characteristics of prokaryotic organisms.

2 Related Work

In bioinformatics, association rule mining has been used primarily in microarray and gene expression data analysis [4-9]. There are several studies that address the challenge of associating genotype to phenotype characteristics [10-17]. In [12] gene function is inferred from cross-organismal distribution of phenotypic traits, which is reliable when the phenotype does not arise from many alternate mechanisms. In [16] co-occurrence between sets of genes and the phenotype has been investigated and association rule mining algorithm netCAR developed and applied in order to extract sets of COGs (clusters of orthologous groups of proteins) associated with a phenotype. MacDonald & Beiko [15] have developed a new genotype–phenotype association approach that uses Classification based on Predictive Association Rules (CPAR), and successfully compare it with the netCAR. In [17], thermal adaptation vs. structural disorder and functional complexity has been investigated, suggesting that adaptation to extreme conditions is achieved by a significant functional simplification, at both the level of the genome and individual genes. In [11], genotype-phenotype associations have been systematically discovered by combining information from a biomedical database GIDEON with the molecular information from NCBI COGs database. Korbel et al [14] reported on systematic association of genes to phenotypes by literature mining and comparative genome analysis. Coulet et al, 2008 [18] employed a bio-ontology for guiding data preparation for discovering genotype-phenotype relationships.

Since biological data are often scattered across different text documents, such as scientific papers or encyclopedias, some of these studies include the use of literature mining and information extraction techniques to uncover such associations and report results along this line [13-14].

3 Materials

Basic source of data for mining associations for organism characteristics are well structured databases such as NCBI database [1] (the most extensive one) IMG [2], PATRIC databases [3], Comprehensive Microbial Resource [19], Genome Atlas Database [20], databases and tools for specific types of genotype-phenotype research, etc. Additional data sources, especially for phenotypic data, are different text documents, e.g., scientific papers, encyclopedias, scientific journals and books (e.g., [21-24]), many of them not being digitized yet.

For the research reported in this paper, we have used two main data sources: (i) NCBI Entrez Genome database [1] - an instance from 2011 (table *organism_info)*, and (ii) *Bergey's Manual of Systematic Bacteriology* [21-23]

3.1 NCBI Entrez Genome Database: Table *organism_info*

The table Organism_info contains 7467 isolates of 2163 different prokaryote species, with data (characteristics) on genome size, GC content, shape, oxygen, habitat, salinity, temperature, gram stain, motility, pathogenicity. Some columns in the table are rather sparse, and the table is overall under half-populated.

Genome size is the total amount of DNA contained within one copy of a genome. It is measured as the total number of nucleotide bases pairs (in megabases -Mb or Mbp). In known prokaryotic organisms genome size vary, for example, between 10,148,695 bp for *Streptomyces scabiei* 87.22 (an important bacterial plant pathogen) to *Candidatus Carsonella ruddii* (an obligate endosymbiotic Gammaproteobacteria) with a genome of 160 000 bp [25]. Distribution of genome size in prokaryotes, calculated by Koonin and Wolf [26], clearly separates two broad genome classes with 4Mb border. We recalculated this distribution on superkingdoms Archaea and Bacteria and confirmed such classifiation in "low" size (length < 4Mb) and "high" size (length > 4Mb) genomes. It has been demonstrated that larger genomes (more than 3 Mb) in free-living organisms, as a result of more complex and varied environments, show trend toward higher GC content than smaller ones, while nutrient limiting and nutrient poor environments dictate smaller genomes of low GC [27].

Guanine-Cytosine (GC) content (or ratio) of a genome refers to the percentage (or ratio) of nitrogenous bases of genome nucleic acids. It may vary between the genomes, as well as in the genome. Average GC content of bacterial genomes varies in range from 25% to 75% [28].

Habitat. Bacteria grow in a wide variety of habitats and conditions. They may be found on the highest mountains, the bottom of the deepest oceans, in the animals guts, and even in the rocks and ice [29]. Modalities for habitat, found in the database (Entrez Genome project, Organism info - Complete genomes [1]), are aquatic, multiple, specialized (i.e., hot springs, salty lakes), host-associated (i.e., symbiotic) and terrestrial.

Oxygen requirement. Bacteria have a wide range of environmental and nutritive requirements. Most bacteria may be placed into one of four groups based on their response to gaseous oxygen. Modalities for oxygen requirement found in (Entrez

Genome project, Organism info – Complete genomes) database are aerobic, facultative anaerobic (facultative for short), anaerobic and microaerophilic. Aerobic bacteria grow in the presence of oxygen and use it as a terminal acceptor of electrons in respiratory chain. Microaerophilic bacteria require lower level of oxygen than present in atmosphere. Anaerobic bacteria instead of oxygen use some other inorganic electron acceptor (sulfur, for example). Facultative anaerobe use oxygen when present, but may grow without oxygen. As compared to anaerobic, aerobic prokaryotes have shown increased GC content [29].

Temperature range. Bacteria grow in many environments from arctic oceans to hot springs. They can be classified into the following modalities: mesophile and extremophile, i.e., thermophile, hyperthermophile and cryophile (or psychrophile). A mesophile grows best in moderate temperature, between 15ºC and 40ºC. The habitats of these organisms include soil, human or animal body, etc. Thermophiles are extremophilic organisms that prefer relatively higher temperatures, between 45ºC and 80ºC. Many of them belong to Archaea. Hyperthermophiles are extreme thermophiles which prefer temperatures above 60ºC. Psychrophiles or Cryophiles are extremophilic organisms that are capable of growth in cold temperatures below 15ºC.

Figure 1. presents statistics on the table *Organism_info* (number of isolates / species with data defined for oxygen requirement, habitat, temperature range, gram stain and their combinations), while Figure 2 represents distribution of modalities for oxygen requirement, habitat, temperature range and gram stain.

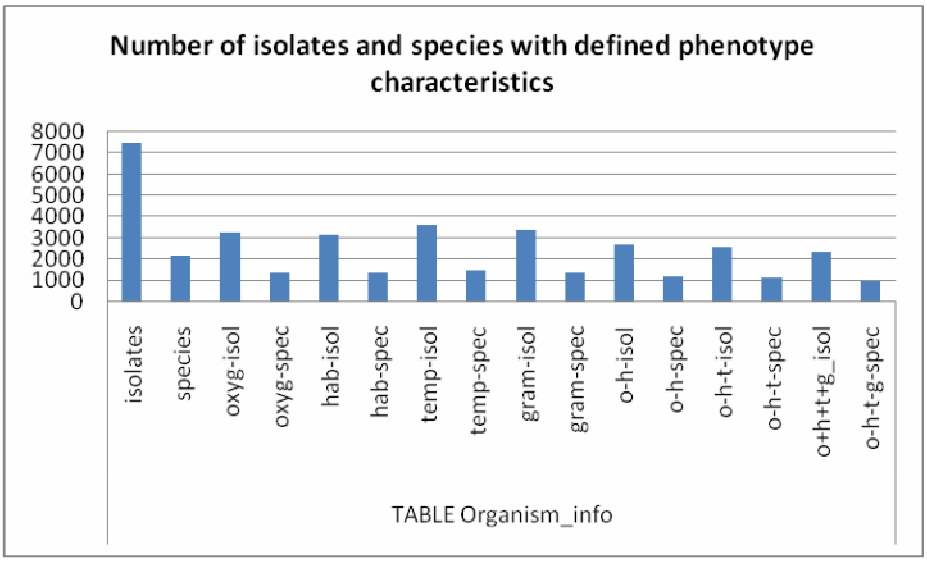


Fig. 1. Statistics on the table *Organism_info*

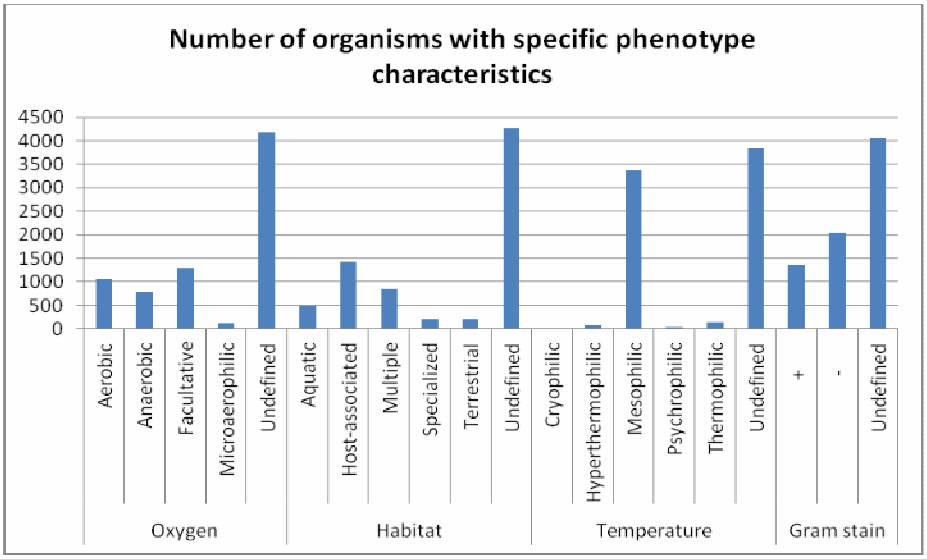


Fig.2 Distribution of modalities: oxygen, habitat, temperature and gram stain

3.2 Encyclopedia of Microorganisms

The data we wanted to transform into structural form reside in the four volumes of the 'Systematic Bacteriology' [21-23], in a form of descriptive, unstructured text in English. An example of a species description is represented in Figure 3; the underlined parts of this text represent the data (or attributes) we wanted to extract.

9. **Hyphomicrobium zavarzinii** Hirsch 1989b, 495[VP] (Effective publication: Hirsch 1989, 1903.)
za.varzin' i.i. M.L. gen. n. *zavarzinii* of Zavarzin, named for G.A. Zavarzin, the Russian microbiologist who isolated these bacteria.

Mother cells drop- or pear-shaped, somewhat slender, with hyphae that rarely branch. Mother cells 0.63 × 1.8 µm (range: 0.5–0.9 × 0.7–2.5 µm). Swarmer cells with 1–3 subpolar flagella. In liquid media under most growth conditions, rosettes are formed, since mother cells produce a polar holdfast. Growth in liquids initially as turbidity and later as a pellicle, with precipitation on the bottom. Colonies on solid media are colorless to light brownish or beige, smooth and shiny, with entire edges.

Chemoorganotrophic, aerobic, oligocarbophilic. Good growth with the following carbon sources: methanol, methylamine·HCl, formate, *n*-butyrate, isovalerate, crotonate, β-hydroxybutyrate, ethanol, *n*-propanol, isobutanol, and glycerol. Growth is stimulated significantly by acetate, *n*-valerate, α-oxoglutarate, galacturonate, formaldehyde, D-glucose, D-mannose, D-melibiose, amygdalin, esculin, chitin, Bacto peptone, DL-lysine, DL-aspartate, and dilute human urine. Nitrogen sources utilized are: NH_4^+, NO_2^-, NO_3^-, and (poorly) Bacto peptone. There is slow growth in the absence of added nitrogen sources (oligonitrophily). Poor growth on sheep blood agar with α-hemolysis. The following antibiotics inhibit growth at 30 µg (per disc): kanamycin, neomycin, and tetracycline. Streptomycin at 10 µg is also inhibitory. There is growth in the presence of 3.5% NaCl. Temperature range: 15–37°C. Optimal pH: 6.5–7.5. Visible light inhibits growth slightly.

Grow anaerobically with nitrate and gas formation (with methanol as the carbon source). With methylamine·HCl and thioglycolate, there is little growth. Catalase and cytochrome oxidase are positive; gelatin liquefaction is negative. Poly-β-hydroxybutyrate is a storage product.

Not pathogenic for mice or guinea pigs.

Genome size: 2.73 × 10^9 Da (strain ZV-580; Kölbel-Boelke et al., 1985).

Habitat: peaty and moist soil near Moscow, Russia.

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

Type strain: ATCC 27496, IFAM ZV-622.

GenBank accession number (16S rRNA): Y14305.

Additional Remarks: Additional strains include IFAM ZV-580, ZV-620, MY-619, MC-625, MC-629, MC-630, and MC-627.

Fig. 3. An excerpt form the encyclopedia 'Systematic Bacteriology'; the underlined parts of the text represent the data to be extracted

4 Methods

For interrelating genotypic and phenotypic characteristics we used data mining techniques, specifically association rule mining [30]. For information extraction from semi-structured source (the Encyclopedia 'Systematic Bacteriology') we applied the two-phased method for information extraction based on finite state transducers [31].

4.1 Association Rule Mining

Data mining [30] is usually defined as:

- discovering hidden information in a database;
- non-trivial extraction of implicit, previously unknown and potentially useful information from database;
- exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns.

Finding association rules is one of the principal methods in data mining. The problem is described in the following way: *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.*

Association rules establishe relationships (associations) among data in large databases, and they are of the form A→B where A and B are sets of elements represented in the data set. A is called *body* of the rule, and B - *head* of the rule. Implication refers to co-occurrence, not to causality.

There are several measures for quality estimation of the rules discovered. The most often used are *support* and *confidence*.

Support for a rule A→B, denoted by $s(A \rightarrow B)$, is defined as

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

where $\sigma(X)$ denotes number of occurrences of an item X in a transaction, and N is total number of items.

Confidence measures how often item B occurs in transactions containing item A, and for the rule A→B, it is defined as

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

Support reflects frequency of set of items occurring in transactions, while confidence measures how often item B occurs in transactions containing item A. We may now restate the problem of finding association rules:

Given a transaction set T, the goal of mining association rules is finding all the rules with support ≥ minsup and confidence ≥minconf.

The higher confidence and support guarantee the more reliable rule. But in certain cases an anomaly arises that both support and confidence are very high but the rule itself does not give a correct result. Because of that, additional measures are used to estimate rule's quality. One of them is *Lift* defined as the ratio between the confidence of the rule ($A \rightarrow B$) and the support of the rule's head (B), i.e. *Lift*($A \rightarrow B$) =

$c(A \rightarrow B)/s(B)$. If A and B are statistically independent, then *Lift=1*. In case *Lift>1*, A and B are said to be positively correlated, while in case *Lift<1*, A and B are said to be negatively correlated. In this context, positive correlation means that the element B (in the head of the rule) is more frequent in transactions containing A (body of the rule) than in transactions not containing A. Analogous holds for negative correlation.

We applied algorithms for association rule mining from the data mining system, IBM Intelligent Miner. It is a part of the programming package IBM InfoSphere Warehouse V9.5 (and later versions). It consists of three components: *Modeling*, used for model creation, *Scoring*, used for testing rules applied to new data in order to estimate benefits, and *Visualization*, used for presentation of results obtained (http://www-01.ibm.com/software/data/infosphere/warehouse/mining.html).

Modeling uses Apriori algorithm to "mine" association rules. Visualization enables to get fast insight into the discovered business rules.

4.2 Information Extraction

Information extraction is a process of identifying some specific data in unstructured texts and assigning a semantic class or a category to them, so it can be transformed into a structural form. There are different methods for automatic extraction of information. Here, we used a method based on finite state transducers (FSTs) – finite state machines that define relations between two sets of strings by transforming one string into another, which has been introduced in [32]. It extracts the relevant data from text segments by applying a collection of FSTs in the form of graphs, describing (most of or all the) possible ways a piece of information we are interested in is expressed in the text (corresponding to the data to be extracted). As a tool for dealing with FSTs we used the system UNITEX [33]. We developed the FSTs for each of the characteristics considered. For example, the FSTs for genome size will recognize the following text sequences and extract the data marked with bold characters:

- *"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. australis is **1256–1276 kbp**"*
- *"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**"*

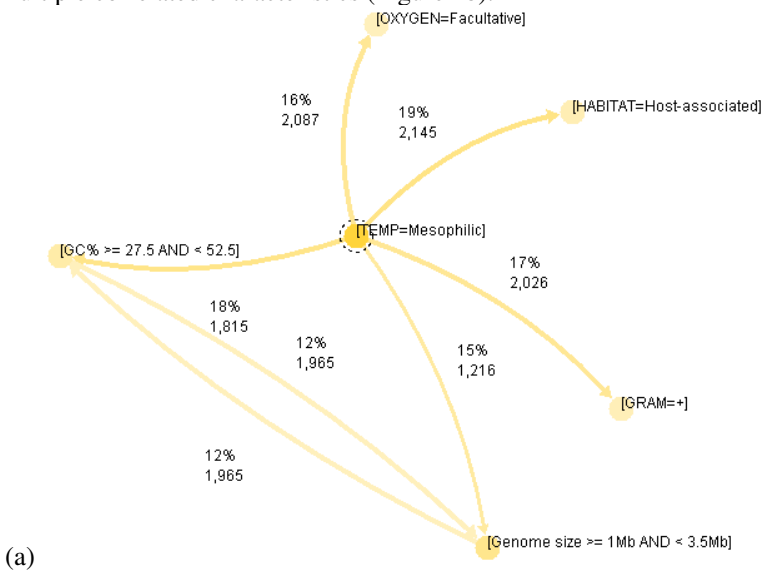
The extracted data is then put into the database, which was used for further analyzes.

5 Results and Discussion

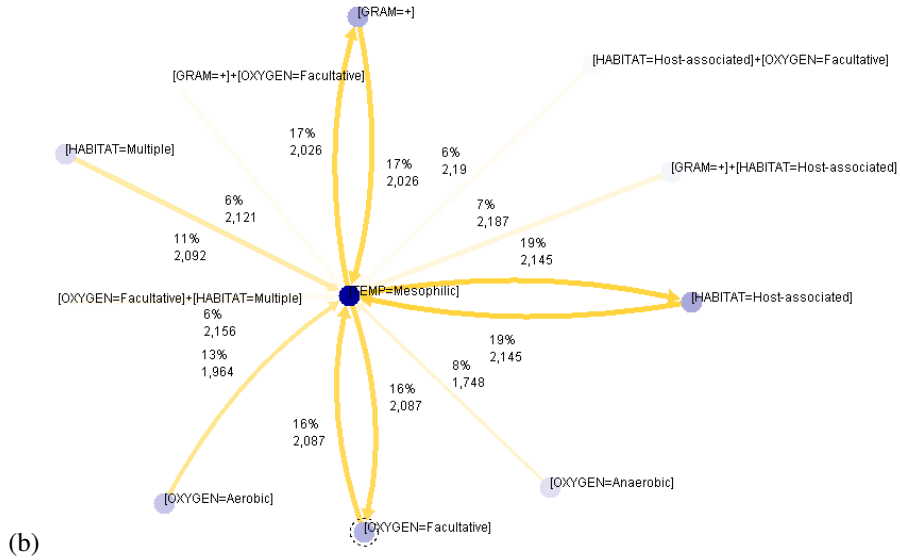
5.1 Mining association from the original NCBI data

Some of the most reliable association rules mined, involving both genotypic and phenotypic characteristics, from the original NCBI data relate mesophilic to host-

associated, facultative oxygen, low GC content organisms and small size genomes (Figure 4a). The rules mined involving phenotypic data only, cross-relate mesophilic with host-associated, facultative oxygen, gram positive organisms, as well as some multiple correlated characteristics (Figure 4b).



(a)



(b)

Fig. 4. Association rules mined from the NCBI table *Organism_info*, involving both genotype and phenotype (4a) and only phenotype characteristics (4b)

5.2 Database – the Transformed Encyclopedia text

As a result, after the application of the information extraction method described in Section 4.2., we created a database from the data extracted from encyclopedia text. The structure of the database is shown on Figure 5.

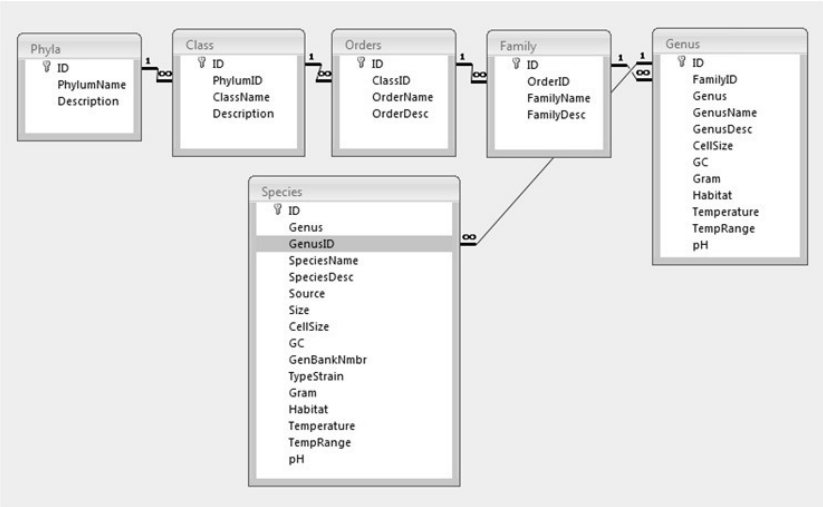


Fig. 5. The database design for the encyclopedia text

The database contains 2412 records in the table Species and 873 records in the table Genus. Numbers of extracted characteristics are as follows (respectively tables Species - Genus): 410 - 554 for Oxygen, 485 - 738 for Gram, 711 - 190 for pH, 1616 - 284 for Habitat, 455 - 257 for Temperature, 638 - 170 for TempRange.

5.3 Integration of data collections

In order to be integrated with the NCBI database, the database produced by structuring the encyclopedia knowledge (specifically tables Species and Genus) had to be manually post-processed - biocurated, supplemented and uniformed. For example, species habitat, recognized from the encyclopedia text as "soil", "fresh water" or "slightly and moderately acid sulfide springs having a high content of elemental sulfur", had to be manually transformed into a habitat modality from the NCBI, "terrestrial" and "aquatic", respectively.

Two types of data integration have been performed: *vertical* and *horizontal* integration.

Vertical data integration. Vertical data integration consists in defining undefined values in columns of the NCBI database (table *Organism_info*). This type of integration involved replacing missing (undefined) data in the NCBI collection by values (if) extracted from the encyclopedia. Since the NCBI collection often contained different isolates of the same species, while the encyclopedia data were at the species level and higher (e.g., genus), extracted values replaced the corresponding missing ones in all

the isolates pertaining to the same species (or even genus). Integration is performed in two steps, thus producing a two-level approximation of organism characteristics:

1. Missing data in the table *Species*, produced by extracting information from the encyclopedia at the species level, replaced by values extracted from the encyclopedia for the corresponding genera (using the table *Genus*)
2. Missing data in the NCBI collection replaced by values extracted from the encyclopedia for the corresponding specie or genus (using the table Species modified in the first step) – resulting with the table *Organism_info_int*

Figure 6. represents the statistics of the output of the first integration step: some characteristics such as oxygen requirement, temperature growth or gram stain became highly enriched.

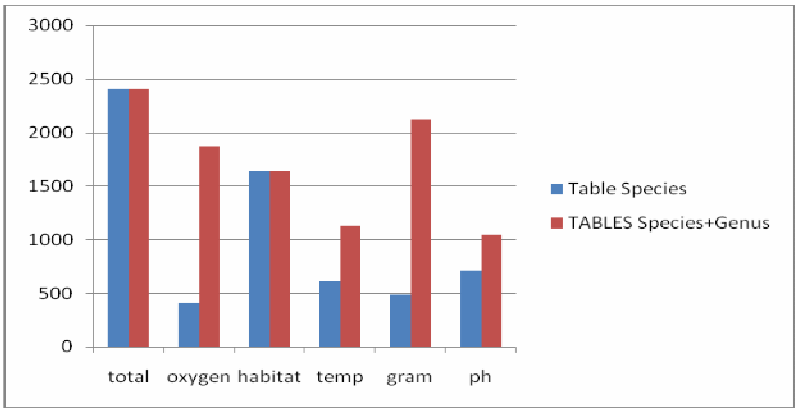


Fig. 6. Vertical data integration – first step: table *Species* supplemented by genera data

Figure 7. represents statistics of the output of the second integration step: all the characteristics for the isolates present in the NCBI database became enriched, some of them up to the one third.

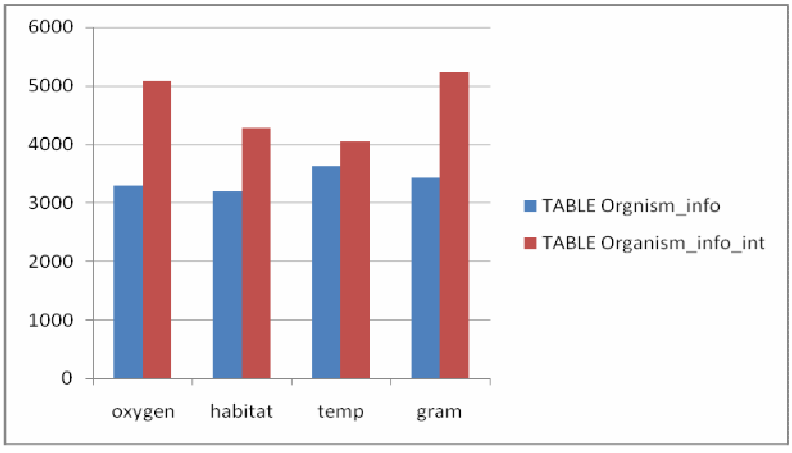


Fig. 7. Vertical data integration – second step: table *Organism_info* supplemented by species / genera data

Horizontal data integration.Horizontal data integration consists in extending the vertically integrated data (explained in the previous section, table *Organism_info_int*) by adding new species extracted from the encyclopedia (table *Species*). In other words, the union operation on the tables *Organism_info_int* and *Species* is performed. Since not all the characteristics from the NCBI database (and thus the table *Organism_info_int*) were present in the encyclopedia (and thus in the table *Species*), projection onto common attributes is then performed and the table *Species_int* obtained. As a result, an enriched set of organisms as well as the enriched set of organisms with defined values for oxygen requirement, habitat, temperature growth, gram stain and their combinations, are obtained (Figure 8).

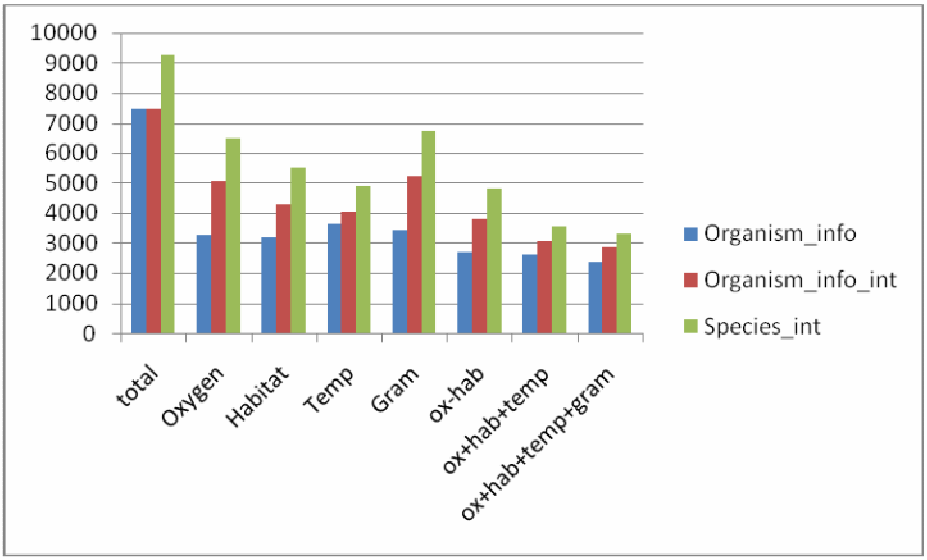


Fig. 8. Horizontal data integration: number of organisms with defined phenotypic characteristics – comparison between basic data collection, vertically integrated and horizontally integrated data

5.4 Mining associations from the integrated data collections

Some of the most reliable association rules mined from the vertically integrated data (table *Organism_info_int*), involving both genotypic and phenotypic characteristics (in addition to those identified for the original NCBI data, Figure 4a), cross-relate host-associated organisms with facultative oxygen requirement, and relate mesophilic to aerobic organisms (Figure 9). Mining association rules involving phenotypic data only (habitat, oxygen, temperature, gram stain), from vertically integrated data (table *Organism_info_int*), confirm these cross-relations (Figure 10), but when applied to horizontally integrated data (table *Species_int*), with enriched set of organisms and attribute values, discard some of the less reliable and multiply related characteristics (e.g., between facultative oxygen requirement and multiple habitat to mesophilic organisms, Figure 11).

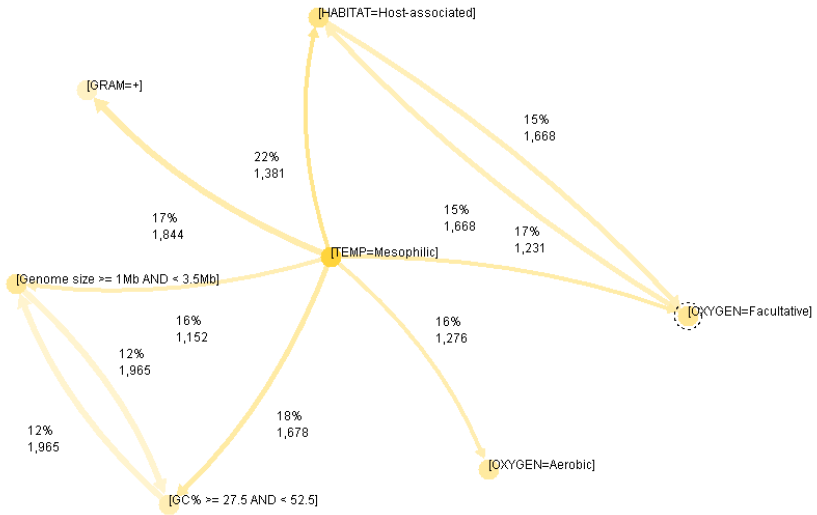


Fig. 9. Association rules mined from the vertically integrated data (table *Organism_info_int*), involving both genotypic and phenotypic characteristics

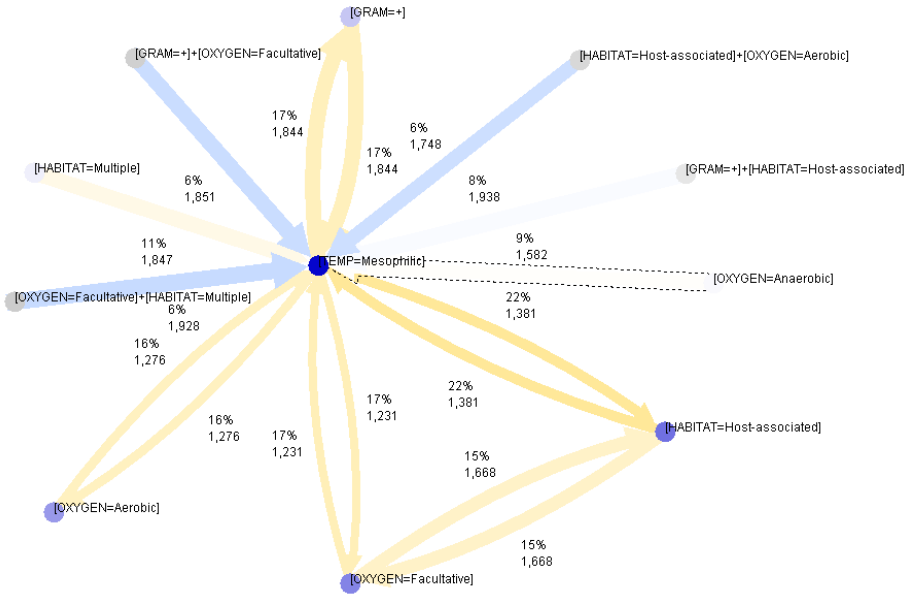


Fig. 10. Association rules mined involving phenotypic data only, from vertically integrated data (table *Organism_info_int*)

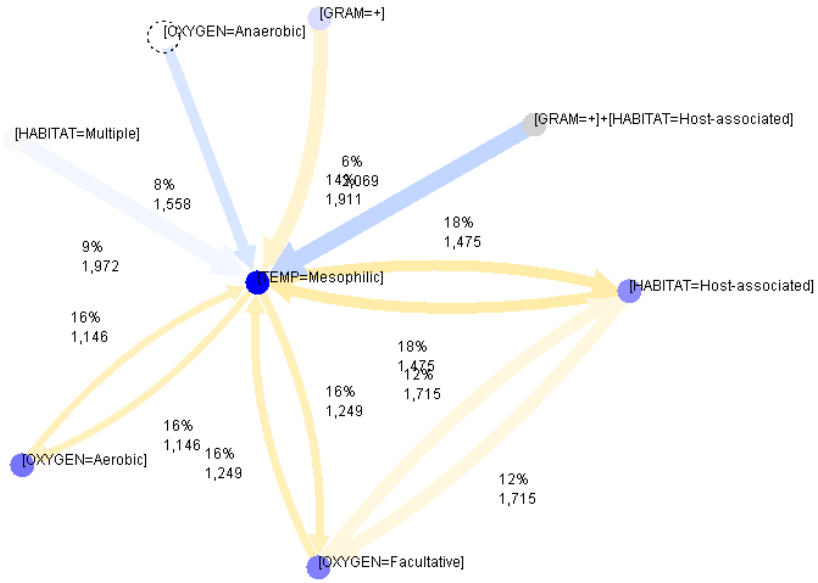


Fig. 11. Association rules mined involving phenotypic data only, from horizontally integrated data (table *Species_int*)

6 Conclusion (and beyond)

By integrating several data sources – structured databases and data extracted from different semi-structured or unstructured repositories such as scientific papers, encyclopedias and other books, web pages, a significant enlargement of the databases can be achieved. We illustrate this fact by integrating one of the most extensive prokaryote databases – the database of the National Center for Biotechnology Information, NCBI, with data extracted from the prokaryote encyclopedia *Bergey's Manual of Systematic Bacteriology*. We show that, although the most extensive, the NCBI database can be significantly enlarged and enriched this way. This sort of integration is applicable to many other specific areas and tasks.

The goal of this work – mining useful and novel association rules between different prokaryote organism characteristics - was only partly fulfilled. We proved that data integration did contribute to the reliability of the association rules mined, which is rather convincing argument for the process of integration itself. On the other side, the association rules mined from the original NCBI database were quite modest and not especially novel; the same holds for mining from the integrated data source, although significantly enlarged and enriched. The main reason for this is the fact that data in the original structured database, as well as in the integrated database is still rather sparse. Another reason may be the fact that association rules were not the most adequate analysis of relationships among the integrated data. So, the very next step in our work will be integration of several structured microbial databases, with richer set of genotypic, phenotypic and taxonomic characteristics, then integration with data

extracted from several different prokaryote encyclopedias, and finally – a multivariate analysis of the integrated data, in addition to association rule mining.

Acknoledgments. This research is a part of the projects IO 174021 and TR 31063, funded by Ministry of Education, Science and Technological Development, Republic of Serbia.

References

1. http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html
2. http://img.jgi.doe.gov/cgi-bin/w/main.cgi
3. http://www.patricbrc.org
4. Creighton, C. and Hanash, S. (2003) `Mining gene expression databases for association rules', *Bioinformatics*, Vol. 19, pp.79{86.
5. Georgii, E., Richter, L., Rckert, U. and Kramer, S. (2005) `Analyzing microarray data using quantitative association rules', *Bioinformatics*, Vol. 21(Suppl 2):II123-II129.
6. Carmona-Saez, P., Chagoyen, M., Rodriguez, A., Trelles, O., Carazo, J.M. and Pascual-Montano, A. (2006) `Integrated analysis of gene expression by association rules discovery', *BMC Bioinformatics*, Vol. 7, doi:10.1186/1471-2105-7-54.
7. Martinez, R., Pasquier, C. and Pasquier, N. (2007) `GenMiner: Mining Informative Association Rules from Genomic Data', *IEEE International Conference 2007 on Bioinformatics and Biomedicine (IEEE BIBM 2007)*, pp 15-22.
8. Martinez, R., Pasquier, N. and Pasquier, C. (2009) `Mining Association Rule Bases from Integrated Genomic Data and Annotations', *In Masulli, F. et al. (Eds.) CIBB 2008. LNBI 5488 Springer-Verlag Berlin Heidelberg*, pp 78-90.
9. Gyenesei, A., Wagner, U., Barkow-Oesterreicher, S., Stolte, E. and Schlapbach, R. (2007) `Mining co-regulated gene pro¯les for the detection of functional associations in gene expression data', *Bioinformatics*, vol. 23, pp.1927-1935.
10. Chang, R., Shoemaker, R. and Wang, W. (2011) 'A novel knowledge-driven systems biology approach for phenotype prediction upon genetic Intervention', *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* Vol. 8, No. 5, pp.1170–1182, oi:10.1109/TCBB.2011.18.
11. 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.
12. A Cross-Genomic Approch for Systematic Mapping of Phenotypic Traits to Genes, K. Jim, K. Parmr, M. Singh, S. Tavazoie, Genome Research, 14, 2004, 109-115
13. Jimeno-Yepes, A., Berlanga-Llavori, R., Rebholz-Schuhmann, D. (2009) 'Exploitation of ontological resources for scientific literature analysis: Searching genes and related diseases', *Engineering in Medicine and Biology Society, EMBC 2009, Annual International Conference of the IEEE*,pp. 7073–7078.
14. 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.
15. N. J. MacDonald, R. G. Beiko, Efficient learning of microbial genotype-phenotype association rules, Bioinformatics, 26, 2010, pp. 1834-1840.
16. Microbial genotype-phenotype mapping by class association rule mining, Makio Tamura, patrik D'haeseleer, Bioinformatics, 24 (13), 2008, 1523-1529
17. Reduction in Structurl Disorder and Functional Complexity in the Therml Adptti on of Prokryotes, Prasad V. Burra, Lajos Kalmr, Peter Tompa, Plos One, 2010. 5(8), e12069

18. Ontology-guided data preparation for discovering genotype-phenotype relationships, A. Coulet, M. Smaïl-Tabbone, P. Benlian, A. Napoli and M.-D. Devignes, *BMC Bioinformatics* 2008, 9(Suppl 4):S3
19. http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi
20. http://www.cbs.dtu.dk/services/GenomeAtlas/
21. Garrity, G. ed. (2005) *Bergey's Manual of Systematic Bacteriology, Volume 2: The Proteobacteria,* ISBN 978-0-387-95040-2.
22. Whitman, W.B, ed (2009) *Bergey's Manual of Systematic Bacteriology, Volume 3: The Firmicutes,* ISBN: 978-0-387-95041-9
23. Whitman, W.B, ed (2010) *Bergey's Manual of Systematic Bacteriology, Volume 4: The Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobacteres, Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae, and Planctomycetes*, ISBN: 978-0-387-95042-6
24. Dworkin, M. (ed.), The Prokaryotes: A Handbook on the Biology of Bacteria, Third Edition, Volumes 1–7, 2006, ISBN-10:0-387-25499-4
25. Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H.E., Moran, N.A. and Hattori, M. (2006) `The 160-kilobase genome of the bacterial endosymbiont Carsonella', *Science*, Vol. 314, p.267.
26. Koonin EV, Wolf YI. (2008) `Genomics of Bacteria and Achaea: the emerging dynamic view of the prokaryotic world', *Nucleic Acids Research*, Vol. 36, pp.6688-6719.
27. Mann, S. and Chen,Y.P.P. (2010) `Bacterial genomic G+C composition-eliciting environmental adaptation', *Genomics*, Vol. 95, pp.7-15.
28. Mann, S. and Chen,Y.P.P. (2010) `Bacterial genomic G+C composition-eliciting environmental adaptation', *Genomics*, Vol. 95, pp.7-15.
29. Schlegel, H.G. and Jannasch, H.W. (2006) `Prokaryotes and Their Habitats', *in The Prokaryotes, Martin Dworkin (Ed.),Springer Science{Business Media, LLC, New York*, Vol. 1, pp.137-184.
30. Tan, P.N., Steinbach, M. and Kumar, V. (2005) `Introduction to Data Mining', *Addison Wesley*
31. Pajić, V.S, Pavlović-Lažetić, G.M, Brandt, B.W, Pajić, M.B, Towards a database for genotype-phenotype association research: mining data from encyclopaedia, *Int. J. Data Mining and Bioinformatics, Vol. 7, No. 2, 2013, pp. 196-213*
32. Pajić, V., Pavlović-Lažetić, G. and Pajić, M. (2011) 'Information Extraction from Semi-structured resources; a Two Phase Finite State Transducers Approach', in Bouchou-Markhoff, B., Caron, P., Champarnaud, J.M. and Maurel, D. (Eds.): *Implementation and Application of Automata: 16th International Conference*, CIAA 2011, Blois, France, 13–16 July, *Lecture Notes in Computer Science*, Springer, Verlag, pp.282–289.
33. Paumier, S. (2006) *Unitex 1.2 User Manual,* Université de Marne-la-Vallée. http://www-igm.univ-mlv.fr/'unitex/UnitexManual.pdf