Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Towards reconstructing a metabolic tree of life

Marina Marcet-Houben[1], Pere Puigbò[1], Antoni Romeu[1] and Santiago Garcia-Vallve[1, *]

[1]Evolutionary Genomics Group, Biochemistry and Biotechnology Department, Rovira i Virgili University, Campus Sescelades, c/ Marcel li Domingo s/n, 43007 TARRAGONA, Spain; Santiago Garcia-Vallve* – E-mail: santi.garcia-vallve@urv.cat; * Corresponding author

received November 17, 2007; accepted November 23, 2007; published online December 11, 2007

Abstract:
Using information from several metabolic databases, we have built our own metabolic database containing 434 pathways and 1157 different enzymes. We have used this information to construct a dendrogram that demonstrates the metabolic similarities between 282 species. The resulting species distribution and the clusters defined in the tree show a certain taxonomic congruence, especially in recent relationships between species. This dendrogram is another representation of the tree of life, based on metabolism that may complement the trees constructed by other methods. For example, the metabolic dissimilarity we demonstrate between *Symbiobacterium thermophilum* (previously defined as Actinobacteria) and the other Actinobacteria species, and the metabolic similarity between *S. thermophilum* and Clostridia, combined with other evidence, suggest that *S. thermophilum* may be re-classified as Firmicutes, Clostridia.

Keywords: metablic pathways; enzymes; dendogram; taxonomy; species

Background:

For many years phylogenetic trees have been used to study the evolution of organisms. Since Charles Darwin first described the evolution of species as a tree, scientist have attempted to create a tree that could represent a hierarchical classification of all known species based on their evolution and at the same time provide information about extinct species and the common ancestry shared by known species. When sequencing technologies were developed, the use of taxonomic marker molecules such as the small subunit ribosomal RNA seemed sufficient to draw consistent phylogenetic trees. Studies using genes or protein sequences led to a classification of microorganisms and recognised the Archaea as the third domain of life. **[1]**

When whole genome sequences of prokaryote organisms became available, everyone hoped that this extended information would help them to build more accurate phylogenies but it was then discovered that different genes produced different trees. It was at this point that doubts were raised as to whether a tree structure was the best representation of evolution. **[2]** Simultaneously, the discovery that horizontal gene transfer events (HGT) between species was more common than previously suspected **[3, 4]** put a strain on the search for the "true tree". **[5]** After all, the gene used in a phylogenetic study may very well have been acquired from an organism that was in no way a direct ancestor. **[6]** In view of the above, some scientists have started to consider that evolution is perhaps better represented by a network than by a tree. **[7]** Studies have also begun into new ways of creating a universal tree of life. If taking a single gene had become insufficient for consistent tree representation, now that hundreds of whole genomic sequences are available, new phylogenomic methods are being developed. **[8]** As it is difficult to align the sequences of two genomes, several methods that use traditional sequence alignment tools have been developed to construct genome trees. **[8, 9, 10]** These methods involve concatenating the homologous sequences from different gene families to construct a single tree **[9, 10, 11]** or comparing different trees to create a supertree. **[12]** Another way to describe the relationships between genomes is to use their gene repertoire. **[13]** New methods based on gene order or gene content have therefore been developed. **[10]** The main problem with these methods is the imbalance in the number of genes between small and large genomes. Two large genomes that are not phylogenetically closely related can have more common genes than a large and a small genome that are closely related. Measures to prevent this must be taken so that the phylogenetic tree does not become biased. **[10]**

Genome trees seem to reveal a phylogenetic signal that supports the three-domain evolutionary scenario and the relationships between some clades of Bacteria. However, deep-level prokaryotic relationships are difficult to infer. **[12]** We have developed a new method for constructing a genome tree based on the metabolic pathways present in each species. The main structure of the metabolic pathways seems to be largely unaffected by HGT. **[14]** This enables us to use them as templates for comparing genomes. Using the orthologous groupings of enzymes found in the KEGG database, we have related genomes and metabolic pathways and created a tree-like representation of a fairly large group of organisms based on their metabolism.

Methodology:

Our aim was to create a dendrogram of different eukaryotic and prokaryotic species based on metabolic data. Here we detail the characteristics of the process used:

Database creation

Starting from the metabolic maps available in the KEGG: Kyoto Encyclopedia of Genes and Genomes **[15]** (http://www.genome.ad.jp/kegg/) and the MetaCyc **[16]** (http://www.metacyc.org) databases, we defined a representative group of pathways and introduced into our database the enzymes that catalyse each of the reactions that form every pathway by their KO number as defined in KEGG. Since a same pathway can follow slightly different routes in different organisms, we added different variants to some of the pathways. For example, we introduced five variants of the glycolysis pathway. At the end, our database contained 434 pathways and 1157 enzymes with different KO numbers.

Percentage matrix

The next step was to relate the data found in our database to a group of organisms. We used the complete genomes found in the KEGG database. For each organism, we created a list of enzymes codified in the genome, listed by their K number. Since the KEGG database is still growing and new genomes are being introduced, some of them still did not have all their KEGG numbers assigned. So, we compared the number of proteins with an assigned KEGG number to the total number of proteins coded in each genome. Those organisms in which the assigned number of proteins in the KEGG database was less than 20 percent were excluded from the list of organisms used to build the dendrogram. Finally we took 282 organisms which are listed in Table 1 (supplementary material) with their abbreviation. Using information from the metabolic database we had previously created, we searched in each genome for the enzymes that completed each pathway. To do so, we made a PERL script that calculated the percentages of enzymes that appeared in a pathway for each organism. The results were presented in a matrix whose rows were the pathways, whose columns were the organisms analysed and in which each element represented the percentage of enzymes of a pathway that one organism contains.

Dendrogram construction

By calculating the Pearson Correlation with the enzyme percentages of all pathways for each pair of organisms, we transformed the percentage matrix into a distance matrix containing the metabolic distance between each pair of organisms. From this distance matrix, and using the PAUP* program version 4.0, we built a dendrogram using the neighbour-joining (NJ) algorithm. This dendrogram graphically represents the relationships between organisms based on their metabolism. We also built the dendrogram with the UPGMA algorithm, but this dendrogram was fairly similar to the one obtained by NJ.

Bootstrap calculation

To verify the dendrogram obtained, we developed a new method based on bootstrap calculations to check how robust each cluster was. From the primary percentage matrix, this method creates a certain number of distance matrices (a thousand in our case) by randomly selecting the metabolic pathways and allowing repetition. Using this group of matrices, we followed the same process as before and obtained a thousand trees. Using the consense program of the Phylip package, we calculated a consensus tree using the majority rule extended option with default parameters. The number of times each node is repeated indicates how reliable that cluster is.

Discussion:

Dendrogram based on metabolism

To ensure that the method developed was suitable for creating a dendrogram that would take into account at least the most basic taxonomic classification, we used it on 282 organisms (9 Eukaryota, 23 Archaea and 250 Bacteria) from the KEGG database. The evolution based on metabolic pathways is represented in the dendrogram in Figure 1. To make comparison easier, we have coloured the branches according to the taxonomic classification of their organism and classified the organisms into fourteen groups. These groups, which differ in size, were defined by taking into account the clusters observed in Figure 1 and their bootstrap values. The result of the groupings and the taxonomic group to which each organism belongs are shown in Table 1 (supplementary material). In general, although this dendrogram does not follow the taxonomic classification perfectly, some large clusters encompass taxonomically related organisms while others appear as mixed clusters. Here we comment two causes that may lead to the grouping of mixed taxonomic clusters.

Reduced genomes

All Archaea are clustered together separately from the bacterial cluster, the only exception is *Nanoarchaeum equitans Kin4-M* (neq). Unlike the other Archaea we used to construct the dendrogram, this organism is an obligate symbiont. **[17]** It appears clustered with most of the intracellular or obligate parasites with a small genome found in our dendrogram (groups 4, 5 and 6). Parasitic organisms have reduced genomes, which means that their metabolic capacity has been lowered to a certain degree. This could explain the clustering of several parasite species even though they are phylogenetically distant. In a tree based on metabolic information, therefore, it should not be surprising to find that the only symbiont Archaea clusters with other parasites due to their particular metabolic characteristics.

Metabolic similarity

The firmicutes are grouped in two main groups, Lactobacillales (Group 9) and Bacillales (Group 10). Between these two groups there are smaller groups of other Firmicutes, one of which contains the Clostridia *Thermoanaerobacter tengcongensis* (tte) and *Clostridium tetani* (ctc) with two other organisms that do not belong to the Firmicutes phylum: *Symbiobacterium thermophilum* (sth) and *Fusobacterium nucleatum* (fnu). The location of *F. nucleatum* among Firmicutes can be explained by their shared metabolic pathways. **[18]** Despite being gram negative, *F. nucleatum* has been found to be more similar to gram positive bacteria than to gram negative ones. This is also true of *S. thermophilum*. The 16S ribosomal DNA-based phylogeny suggested that this bacterium belongs to an unknown taxon in the gram-positive Actinobacteria **[19]**, even though the traditional Gram-stain result indicates that it is gram negative. **[20]** Also, the proteins of *S. thermophilum* show a greater similarity to the proteins found in Firmicutes organisms, in particular to *T. tengcongensis*, than to those found in Actinobacteria. **[20]** The metabolic similarity between *S. thermophilum* and *T. tengcongensis* shown in figure 1 and the metabolic dissimilarity between *S. thermophilum* and the other Actinobacteria, combined with previous evidences **[20, 21]**, suggest that *S. thermophilum* may be re-classified as Firmicutes, Clostridia. **[21]**

Metabolic influence

Not all kinds of metabolism influence our dendrogram in the same way. In Table 2 (supplementary material) we can see a distribution of the enzymes found in the defined groups in the different metabolic groups. For example, Carbohydrate Metabolism has much more influence on the dendrogram than Energy Metabolism, simply because it has many more enzymes and pathways. Also, some of these enzymes are not very useful for classifying organisms into the different clusters. A clear example is the enzymes that catalyse the reactions that produce the different Aminoacyl-tRNAs as they are present in nearly every group, even those with a reduced genome.

Table 2 (under supplementary material) also shows that for several groups some kinds of metabolisms stand out because of the high number of enzymes they possess compared to the main number of enzymes that the metabolic group has in all organisms. For example, Lipid metabolism in Metazoa (Group 13). This is explained by the presence of pathways such as the synthesis of Lecitin or Cholesterol. The contrary is also true. Some groups have fewer enzymes than most. Examples of this are the three parasitic groups (Group 4, 5 and 6). In their low enzyme values, we can clearly see the effects of genome evolutive reduction due to their parasitic nature.

Limitations of metabolic-based methods

By their nature, metabolic pathways databases are human-defined and may be quite inexact, especially when a metabolic pathway found in one species is generalized to another. Several alternative pathways that have not yet been discovered surely exist in different organisms. Therefore, when only one or a few enzymes from a metabolic pathway are missing in one species, an orthologous gene displacement needs to be considered before we can conclude that the pathway is incomplete. Moreover, when a new sequenced genome is annotated, a high percentage of its proteins are not mapped to any pathway. It may therefore be argued that metabolic databases, while extremely useful for reconstructing metabolic properties of organisms, cannot be used to reconstruct the tree of life. However, we have shown that, assuming that any metabolic prediction of a large group of organisms is still incomplete, the phylogenetic signal that it contains partially agrees with the taxonomic information of the species. A metabolic dendrogram of different species can therefore be used as an additional criterion that may help to correctly re-classify some species, as in the case of the *Symbiobacterium thermophilum* we described earlier.

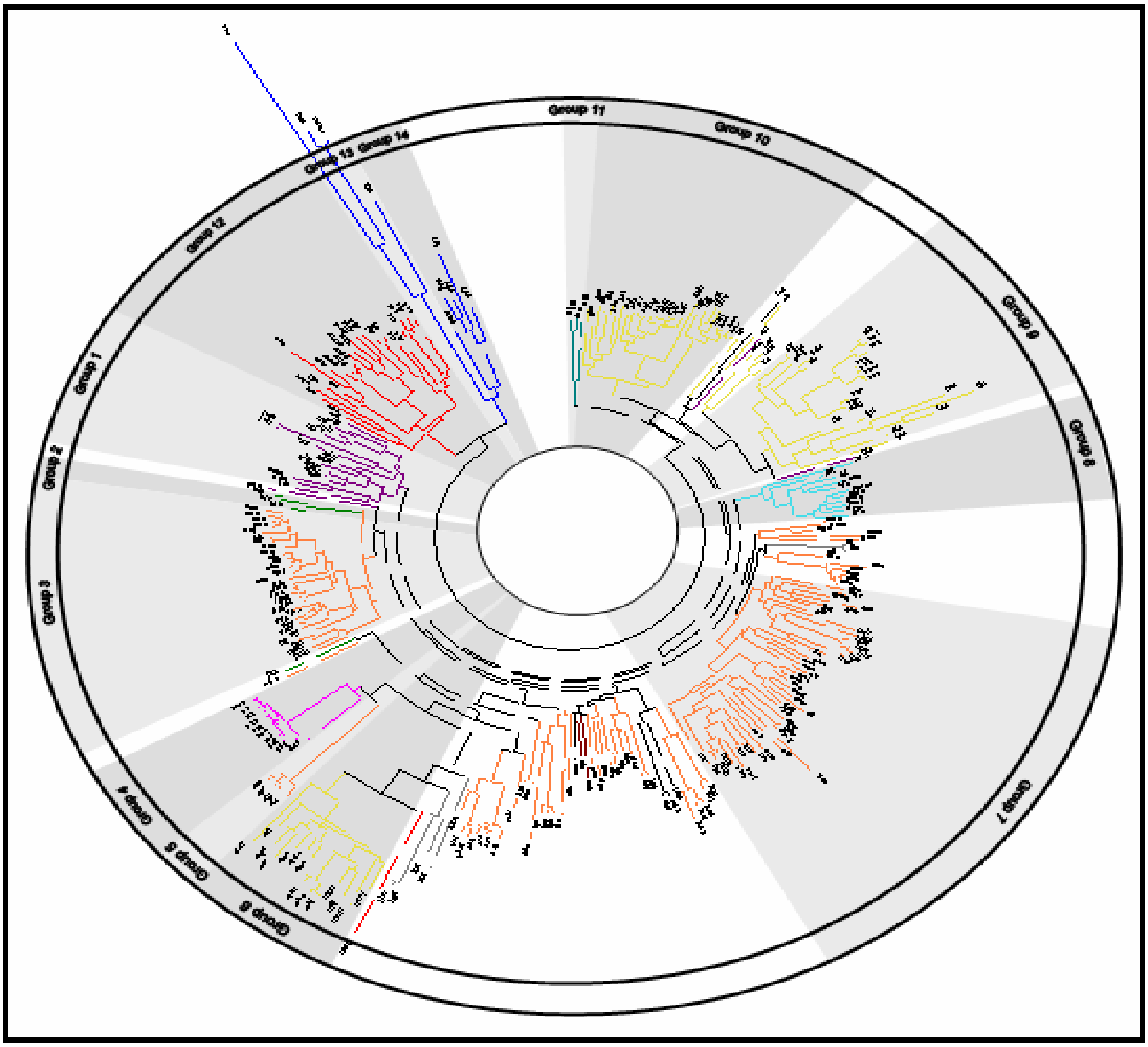

Figure 1: Dendrogram created from metabolic pathways by neighbour joining. The small squares represent nodes with more than 750 repetitions in the bootstrap analysis. The triangles are nodes with more than 900 repetitions. Taxonomic groups are marked by the same colouring: Actinobacteria in purple, Archaea in red, Bacteroidetes in green, Chlamydiae in pink, Cyanobacteria in pale blue, Deinococcus-Thermus in cyan, Eukaryota in dark blue, Firmicutes in yellow, Proteobacteria in orange, Spirochaeta in grey, and others in black.

Conclusion:

We have developed a new method for constructing a dendrogram based on metabolic comparisons between species whose genome has been fully sequenced. Although the evolutionary signal that can be derived from metabolic data is not very strong, it is enough to obtain a rough sketch of the known taxonomic classification. We expect that the reconstruction of metabolic dendrograms may improve as more pathways are discovered and their enzymes are properly situated within those pathways. Until such a time metabolic-based dendrograms may be a useful addition when they are combined with other phylogenetic methods, allowing us to fine-tune dubious classifications that can not be accurately described by other methods.

Acknowledgement:

This work has been financed by the following project: "Ministerio de Ciencia y Tecnología" (Ref. BIO2003-07672). We thank Kevin Costello of the Language Service of the Rovira i Virgili University for his help with writing the manuscript.

References:

[01] C. R. Woese and G. E. Fox, *Proc Natl Acad Sci.,* 74: 5088 (1997) [PMID: 270744]

[02] E. Pennisi, *Science*, 294: 634 (2001) [PMID: 11721026]

[03] S. Garcia-Vallve, *et al.*, *Nucleic Acids Res.*, 31: 187 (2003) [PMID: 12519978]

[04] S. Garcia-Vallve, *et al.*, *Genome Res.*, 10: 1719 (2000) [PMID: 12519978]

[05] W. F. Doolittle, *Science*, 284: 2124 (1999) [PMID: 10381871]

[06] J. P. Gogarten and J. P. Townsend, *Nat Rev Microbiol.*, 3: 679 (2005) [PMID: 16138096]

[07] E. Bapteste, *et al.*, *Trends Microbiol.*, 12: 406 (2004) [PMID: 15337161]

[08] B. E. Duthil, *et al.*, *Bioinformatics*, 23: 815 (2007) [PMID: 17237036]

[09] F. Delsuc, *et al.*, *Nat Rev Genet.*, 6: 361 (2005) [PMID: 15861208]

[10] B. Snel, *et al.*, *Annu Rev Microbiol.*, 59: 191 (2005) [PMID: 16153168]

[11] F. D. Ciccarelli, *et al.*, *Science*, 311: 1283 (2006) [PMID: 16513982]

[12] C. J. Creevey, *et al.*, *Proc R Soc Lond B Biol Sci.*, 271: 2551 (2004) [PMID: 15615680]

[13] M. A. Huynen and P. Bork, *Proc Natl Acad Sci.*, 95: 5849 (1998) [PMID: 9600883]

[14] S. Y. Shi, *et al.*, *Acta Biochim Biophys Sin.*, 37: 561 (2005) [PMID: 16077904]

[15] M. Kanehisa, *et al.*, *Nucleic Acids Res.*, 34: 354 (2006) [PMID: 16381885]

[16] R. Caspi, *et al.*, *Nucleic Acids Res.*, 34: 511 (2006) [PMID: 16381923]

[17] E. Waters, *et al.*, *Proc Natl Acad Sci.*, 100: 12984 (2003) [PMID: 14566062]

[18] V. Kapatral, *et al.*, *J Bacteriol.*, 184: 2005 (2002) [PMID: 11889109]

[19] M. Ohno, *et al.*, *Int. J. Syst. Evol. Microbio.*, 50: 1829 (2000) [PMID: 11034494]

[20] K. Ueda, *et al.*, *Nucleic Acids Res.*, 32: 4937 (2004) [PMID: 15383646]

[21] M. Wu, *et al.*, *PLoS Genet.*, 1: 65 (2005) [PMID: 16311624]

Edited by P. Kangueane

Citation: Marcet-Houben, *et al.*, Bioinformation 2(4): 135-144 (2007)

Supplementary material

Eukaryota

Abbr	Organism Name	Group
cal	*Candida albicans SC5314*	14
cme	*Cyanidioschyzon merolae*	14
ago	*Eremothecium gossypii*	14
hsa	*Homo sapiens*	13
mmu	*Mus musculus*	13
sce	*Saccharomyces cerevisiae*	14
spo	*Schizosaccharomyces pombe*	14
ssc	*Sus scrofa*	-
xla	*Xenopus laevis*	13

Archaea:

Abbr	Organism Name	Group
ape	*Aeropyrum pernix*	12
afu	*Archaeoglobus fulgidus DSM 4304*	12
hma	*Haloarcula marismortui ATCC 43049*	12
hal	*Halobacterium sp. NRC-1*	12
mja	*Methanocaldococcus jannaschii DSM 2661*	12
mmp	*Methanococcus maripaludis*	12
mka	*Methanopyrus kandleri AV19*	12
mac	*Methanosarcina acetivorans C2A*	12
mba	*Methanosarcina barkeri str. fusaro*	12
mma	*Methanosarcina mazei Go1*	12
mth	*Methanothermobacter thermautotrophicus str. Delta H*	12
neq	*Nanoarchaeum equitans Kin4-M*	-
nph	*Natronomonas pharaonis DSM 2160*	12
pto	*Picrophilus torridus DSM 9790*	12
pai	*Pyrobaculum aerophilum*	12
pab	*Pyrococcus abyssi GE5*	12
pfu	*Pyrococcus furiosus DSM 3638*	12
pho	*Pyrococcus horikoshii OT3*	12
sai	*Sulfolobus acidocaldarius DSM 639*	12
sso	*Sulfolobus solfataricus P2*	12
sto	*Sulfolobus tokodaii str. 7*	12
tko	*Thermococcus kodakarensis KOD1*	12
tac	*Thermoplasma acidophilum DSM 1728*	12
tvo	*Thermoplasma volcanium*	12

Actinobacteria:

Abbr	Organism Name	Group
blo	*Bifidobacterium longum NCC2705*	-
cdi	*Corynebacterium diphtheriae NCTC 13129*	1
cef	*Corynebacterium efficiens YS-314*	1
cgb	*Corynebacterium glutamicum ATCC 13032 (Bielefeld)*	1
cgl	*Corynebacterium glutamicum ATCC 13032 (Kyowa Hakko)*	1
cjk	*Corynebacterium jeikeium K411*	1
lxx	*Leifsonia xyli subsp. xyli str. CTCB07*	1

Firmicutes (cont)

Abbr	Organism Name	Group
spn	*Streptococcus pneumoniae TIGR4*	9
spy	*Streptococcus pyogenes M1 GAS*	9
spa	*Streptococcus pyogenes MGAS10394*	9
spg	*Streptococcus pyogenes MGAS315*	9
spz	*Streptococcus pyogenes MGAS5005*	9
spb	*Streptococcus pyogenes MGAS6180*	9
spm	*Streptococcus pyogenes MGAS8232*	9
sps	*Streptococcus pyogenes SSI-1*	9
stc	*Streptococcus thermophilus CNRZ1066*	9
stl	*Streptococcus thermophilus LMG 18311*	9
tte	*Thermoanaerobacter tengcongensis MB4*	-
uur	*Ureaplasma parvum serovar 3 str. ATCC 700970*	6

Proteobacteria:

Abbr	Organism Name	Group
aci	*Acinetobacter sp. ADP1*	7
atc	*Agrobacterium tumefaciens str. C58 (Cereon)*	7
atu	*Agrobacterium tumefaciens str. C58 (U.Washington/Dupont)*	7
ama	*Anaplasma marginale str. St. Maries*	-
eba	*Azoarcus sp. EbN1*	7
bhe	*Bartonella henselae str. Houston-1*	-
bqu	*Bartonella quintana str. Toulouse*	-
bba	*Bdellovibrio bacteriovorus HD100*	-
bbr	*Bordetella bronchiseptica RB50*	7
bpa	*Bordetella parapertussis 12822*	7
bpe	*Bordetella pertussis Tohama I*	7
bja	*Bradyrhizobium japonicum USDA 110*	7
bmb	*Brucella abortus biovar 1 str. 9-941*	7
bme	*Brucella melitensis 16M*	7
bmf	*Brucella melitensis biovar Abortus 2308*	7
bms	*Brucella suis 1330*	7
bab	*Buchnera aphidicola (Baizongia pistaciae)*	-
buc	*Buchnera aphidicola str. APS (Acyrthosiphon pisum)*	-
bas	*Buchnera aphidicola str. Sg (Schizaphis graminum)*	-
bma	*Burkholderia mallei ATCC 23344*	7
bps	*Burkholderia pseudomallei K96243*	7
bur	*Burkholderia sp. 383*	7
cjr	*Campylobacter jejuni RM1221*	-
cje	*Campylobacter jejuni subsp. jejuni NCTC 11168*	-
bfl	*Candidatus Blochmannia floridanus*	-
bpn	*Candidatus Blochmannia pennsylvanicus str. BPEN*	-
pub	*Candidatus Pelagibacter ubique HTCC1062*	7
ccr	*Caulobacter crescentus CB15*	7
cvi	*Chromobacterium violaceum ATCC 12472*	7
cps	*Colwellia psychrerythraea 34H*	7

mpa	*Mycobacterium avium subsp. paratuberculosis K-10*	1
mbo	*Mycobacterium bovis AF2122/97*	1
mle	*Mycobacterium leprae TN*	1
mtc	*Mycobacterium tuberculosis CDC1551*	1
mtu	*Mycobacterium tuberculosis H37Rv*	1
nfa	*Nocardia farcinica*	1
pac	*Propionibacterium acnes KPA171202*	-
sma	*Streptomyces avermitilis MA-4680*	1
sco	*Streptomyces coelicolor A3(2)*	1
sth	*Symbiobacterium thermophilum IAM 14863*	-
tfu	*Thermobifida fusca YX*	1
twh	*Tropheryma whipplei str. Twist*	1
tws	*Tropheryma whipplei TW08/27*	1

Bacteroidetes:

Abbr	Organism Name	Group
bfs	*Bacteroides fragilis NCTC 9343*	2
bfr	*Bacteroides fragilis YCH46*	2
bth	*Bacteroides thetaiotaomicron VPI-5482*	2
pgi	*Porphyromonas gingivalis W83*	-

Chlamydiae:

Abbr	Organism Name	Group
cmu	*Chlamydia muridarum Nigg*	4
cta	*Chlamydia trachomatis A/HAR-13*	4
ctr	*Chlamydia trachomatis D/UW-3/CX*	4
cab	*Chlamydophila abortus S26/3*	4
cca	*Chlamydophila caviae GPIC*	4
cpa	*Chlamydophila pneumoniae AR39*	4
cpn	*Chlamydophila pneumoniae CWL029*	4
cpj	*Chlamydophila pneumoniae J138*	4
cpt	*Chlamydophila pneumoniae TW-183*	4
pcu	*Parachlamydia sp. UWE25*	4

Cyannobacteria:

Abbr	Organism Name	Group
ava	*Anabaena variabilis ATCC 29413*	8
gvi	*Gloeobacter violaceus*	8
ana	*Nostoc sp. PCC 7120*	8
pmt	*Prochlorococcus marinus str. MIT 9313*	8
pmn	*Prochlorococcus marinus str. NATL2A*	8
pma	*Prochlorococcus marinus subsp. marinus str. CCMP1375*	8
pmm	*Prochlorococcus marinus subsp. pastoris str. CCMP1986*	8
syc	*Synechococcus elongatus PCC 6301*	8
syw	*Synechococcus sp. WH 8102*	8
syn	*Synechocystis sp. PCC 6803*	8
tel	*Thermosynechococcus elongatus BP-1*	8

Deinococcus-Thermus:

Abbr	Organism Name	Group
dra	*Deinococcus radiodurans R1*	11
tth	*Thermus thermophilus HB27*	11

cbu	*Coxiella burnetii RSA 493*	-
dar	*Dechloromonas aromatica RCB*	7
dps	*Desulfotalea psychrophila LSv54*	-
dvu	*Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough*	-
ecn	*Ehrlichia canis str. Jake*	-
erg	*Ehrlichia ruminantium str. Gardel*	-
eru	*Ehrlichia ruminantium str. Welgevonden (South Africa)*	-
erw	*Ehrlichia ruminantium str. Welgevonden (France)*	-
eca	*Erwinia carotovora subsp. atroseptica SCRI1043*	3
ecc	*Escherichia coli CFT073*	3
ecj	*Escherichia coli K12 W3110*	3
eco	*Escherichia coli K12 MG1655*	3
ecs	*Escherichia coli O157:H7*	3
ece	*Escherichia coli O157:H7 EDL933*	3
ftu	*Francisella tularensis subsp. tularensis*	-
gsu	*Geobacter sulfurreducens PCA*	-
gox	*Gluconobacter oxydans 621H*	-
hdu	*Haemophilus ducreyi 35000HP*	-
hit	*Haemophilus influenzae 86-028NP*	3
hin	*Haemophilus influenzae Rd KW20*	3
hhe	*Helicobacter hepaticus ATCC 51449*	-
hpy	*Helicobacter pylori 26695*	-
hpj	*Helicobacter pylori J99*	-
ilo	*Idiomarina loihiensis L2TR*	7
lpf	*Legionella pneumophila str. Lens*	-
lpp	*Legionella pneumophila str. Paris*	-
lpn	*Legionella pneumophila subsp. pneumophila str. Philadelphia 1*	-
msu	*Mannheimia succiniciproducens MBEL55E*	3
mlo	*Mesorhizobium loti MAFF303099*	7
mca	*Methylococcus capsulatus str. Bath*	7
ngo	*Neisseria gonorrhoeae FA 1090*	-
nme	*Neisseria meningitidis MC58*	-
nma	*Neisseria meningitidis Z2491*	-
nwi	*Nitrobacter winogradskyi Nb-255*	7
noc	*Nitrosococcus oceani ATCC 19707*	7
neu	*Nitrosomonas europaea ATCC 19718*	7
pmu	*Pasteurella multocida subsp. multocida str. Pm70*	3
pca	*Pelobacter carbinolicus DSM 2380*	-
ppr	*Photobacterium profundum*	3
plu	*Photorhabdus luminescens subsp. laumondii TTO1*	3
pha	*Pseudoalteromonas haloplanktis TAC125*	7
pae	*Pseudomonas aeruginosa PAO1*	7
pfl	*Pseudomonas fluorescens Pf-5*	7
pfo	*Pseudomonas fluorescens PfO-1*	7
ppu	*Pseudomonas putida KT2440*	7
psp	*Pseudomonas syringae pv. phaseolicola 1448A*	7
psb	*Pseudomonas syringae pv. syringae B728a*	7
pst	*Pseudomonas syringae pv. tomato str. DC3000*	7

ttj	*Thermus thermophilus HB8*	11

Firmicutes:

Abbr	Organism Name	Group
baa	*Bacillus anthracis str. A2012*	10
ban	*Bacillus anthracis str. Ames*	10
bar	*Bacillus anthracis str. 'Ames Ancestor'*	10
bat	*Bacillus anthracis str. Sterne*	10
bca	*Bacillus cereus ATCC 10987*	10
bce	*Bacillus cereus ATCC 14579*	10
bcz	*Bacillus cereus E33L*	10
bcl	*Bacillus clausii KSM-K16*	10
bha	*Bacillus halodurans*	10
bld	*Bacillus licheniformis DSM13*	10
bli	*Bacillus licheniformis ATCC 14580*	10
bsu	*Bacillus subtilis subsp. subtilis str. 168*	10
btk	*Bacillus thuringiensis serovar konkukian str. 97-27*	10
cac	*Clostridium acetobutylicum ATCC 824*	-
cpe	*Clostridium perfringens str. 13*	9
ctc	*Clostridium tetani E88*	-
efa	*Enterococcus faecalis V583*	9
gka	*Geobacillus kaustophilus HTA426*	10
lac	*Lactobacillus acidophilus NCFM*	9
ljo	*Lactobacillus johnsonii NCC 533*	9
lpl	*Lactobacillus plantarum WCFS1*	9
lsa	*Lactobacillus sakei subsp. sakei 23K*	9
lla	*Lactococcus lactis subsp. lactis Il1403*	9
lin	*Listeria innocua Clip11262*	-
lmo	*Listeria monocytogenes EGD-e*	-
lmf	*Listeria monocytogenes str. 4b F2365*	-
mfl	*Mesoplasma florum L1*	6
mga	*Mycoplasma gallisepticum R*	6
mge	*Mycoplasma genitalium G37*	6
mhy	*Mycoplasma hyopneumoniae 232*	6
mhp	*Mycoplasma hyopneumoniae 7448*	6
mhj	*Mycoplasma hyopneumoniae J*	6
mmo	*Mycoplasma mobile 163K*	6
mmy	*Mycoplasma mycoides subsp. mycoides SC str. PG1*	6
mpe	*Mycoplasma penetrans*	6
mpn	*Mycoplasma pneumoniae M129*	6
mpu	*Mycoplasma pulmonis UAB CTIP*	6
msy	*Mycoplasma synoviae 53*	6
oih	*Oceanobacillus iheyensis HTE831*	10
poy	*Onion yellows phytoplasma*	6
sab	*Staphylococcus aureus RF122*	10
sac	*Staphylococcus aureus subsp. aureus COL*	10

par	*Psychrobacter arcticus 273-4*	7
reu	*Ralstonia eutropha JMP134*	7
rso	*Ralstonia solanacearum GMI1000*	7
rsp	*Rhodobacter sphaeroides 2.4.1*	7
rpa	*Rhodopseudomonas palustris CGA009*	7
rco	*Rickettsia conorii str. Malish 7*	5
rfe	*Rickettsia felis URRWXCal2*	5
rpr	*Rickettsia prowazekii str. Madrid E*	5
rty	*Rickettsia typhi str. Wilmington*	5
sec	*Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67*	3
spt	*Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC 9150*	3
sty	*Salmonella enterica subsp. enterica serovar Typhi str. CT18*	3
stt	*Salmonella enterica subsp. enterica serovar Typhi Ty2*	3
stm	*Salmonella typhimurium LT2*	3
son	*Shewanella oneidensis MR-1*	7
sfx	*Shigella flexneri 2a str. 2457T*	3
sfl	*Shigella flexneri 2a str. 301*	3
ssn	*Shigella sonnei Ss046*	3
sil	*Silicibacter pomeroyi DSS-3*	7
sme	*Sinorhizobium meliloti 1021*	7
tbd	*Thiobacillus denitrificans ATCC 25259*	7
vch	*Vibrio cholerae O1 biovar eltor str. N16961*	3
vfi	*Vibrio fischeri ES114*	3
vpa	*Vibrio parahaemolyticus RIMD 2210633*	3
vvu	*Vibrio vulnificus CMCP6*	3
vvy	*Vibrio vulnificus YJ016*	3
wbr	*Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis*	-
wol	*Wolbachia endosymbiont of Drosophila melanogaster*	-
wbm	*Wolbachia endosymbiont strain TRS of Brugia malayi*	-
wsu	*Wolinella succinogenes DSM 1740*	-
xac	*Xanthomonas axonopodis pv. citri str. 306*	-
xcb	*Xanthomonas campestris pv. campestris str. 8004*	-
xcc	*Xanthomonas campestris pv. campestris str. ATCC 33913*	-
xcv	*Xanthomonas campestris pv. vesicatoria str. 85-10*	-
xoo	*Xanthomonas oryzae pv. oryzae KACC10331*	-
xfa	*Xylella fastidiosa 9a5c*	-
xft	*Xylella fastidiosa Temecula1*	-
ypm	*Yersinia pestis biovar Medievalis str. 91001*	3
ype	*Yersinia pestis CO92*	3
ypk	*Yersinia pestis KIM*	3
yps	*Yersinia pseudotuberculosis IP 32953*	3
zmo	*Zymomonas mobilis subsp. mobilis ZM4*	-

Spirochaetas:

Abbr	Organism Name	Group
lic	*Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130*	-

sar	*Staphylococcus aureus subsp. aureus MRSA252*	10	lil	*Leptospira interrogans serovar Lai str. 56601*	-
sas	*Staphylococcus aureus subsp. aureus MSSA476*	10	bbu	*Borrelia burgdorferi B31*	-
sav	*Staphylococcus aureus subsp. aureus Mu50*	10	bga	*Borrelia garinii PBi*	-
sam	*Staphylococcus aureus subsp. aureus MW2*	10	tde	*Treponema denticola ATCC 35405*	-
sau	*Staphylococcus aureus subsp. aureus N315*	10	tpa	*Treponema pallidum subsp. pallidum str. Nichols*	-
sep	*Staphylococcus epidermidis ATCC 12228*	10		**Others:**	
ser	*Staphylococcus epidermidis RP62A*	10	**Abbr**	**Organism Name**	**Group**
sha	*Staphylococcus haemolyticus JCSC1435*	10	aae	*Aquifex aeolicus VF5*	-
ssp	*Staphylococcus saprophyticus subsp. saprophyticus*	10	cch	*Chlorobium chlorochromatii CaD3*	-
sag	*Streptococcus agalactiae 2603V/R*	9	cte	*Chlorobium tepidum TLS*	-
sak	*Streptococcus agalactiae A909*	9	det	*Dehalococcoides ethenogenes 195*	-
san	*Streptococcus agalactiae NEM316*	9	deh	*Dehalococcoides sp. CBDB1*	-
smu	*Streptococcus mutans UA159*	9	fnu	*Fusobacterium nucleatum subsp. nucleatum ATCC 25586*	-
spr	*Streptococcus pneumoniae R6*	9	tma	*Thermotoga maritima MSB8*	-

Table 1: Abbreviation and taxonomic classification of the 282 organisms included in the analysis

	Taxonomic classification (organisms belonging to classification/ total organisms group)	**Carbohydrate Metabolism**		**Energy Metabolism**		**Lipid Metabolism**		**Nucleotide Metabolism**		**Amino Acid Metabolism**		**Metabolism of Other Amino Acids**		**Glycan Biosíntesis and Metabolism**		**Polyketides and Non ribosomal Peptides**		**Metabolism of Cofactors and Vitamins**		**Biosíntesis of Secondary Metabolites**		**Biodegradation of Xenobiotics**		**tRNAs**		**Total Enzymes**
Group1	Actinobacteria (17/17)	176	(25.9)	28	(4.1)	22	(3.2)	51	(7.5)	207	(30.5)	20	(2.9)	8	(1.2)	4	(0.5)	116	(17.0)	15	(2.3)	14	(2.0)	20	(2.9)	679
Group2	Bacteroidetes (3/3)	201	(29.2)	34	(4.9)	12	(1.8)	48	(7.0)	199	(28.9)	18	(2.7)	23	(3.4)	4	(0.6)	113	(16.4)	13	(1.9)	0	(0.0)	22	(3.2)	688
Group3	Enterobacteria, Vibrionales, Pasteurellales (29/29)	216	(27.8)	38	(4.9)	21	(2.8)	52	(6.6)	224	(28.8)	27	(3.4)	25	(3.2)	2	(0.3)	129	(16.6)	14	(1.8)	7	(0.9)	22	(2.8)	777
Group4	Chlamidiae (10/10)	117	(33.2)	11	(3.1)	9	(2.6)	20	(5.6)	82	(23.3)	5	(1.6)	19	(5.4)	0	(0.1)	63	(17.8)	6	(1.7)	0	(0.0)	20	(5.7)	352
Group5	Rickettsia (4/4)	71	(28.1)	8	(3.4)	6	(2.6)	20	(8.0)	53	(21.0)	7	(2.8)	20	(8.0)	1	(0.4)	38	(15.3)	5	(1.9)	1	(0.6)	20	(8.0)	251
Group6	Mollicutes (14/14)	96	(50.9)	10	(5.3)	4	(1.9)	14	(7.2)	19	(10.0)	4	(2.1)	0	(0.0)	0	(0.0)	21	(10.9)	1	(0.6)	1	(0.3)	21	(10.9)	189
Group7	Proteobacteria (43/43)	188	(24.3)	36	(4.7)	27	(3.5)	52	(6.8)	240	(31.1)	25	(3.3)	21	(2.8)	3	(0.4)	120	(15.6)	16	(2.0)	21	(2.7)	21	(2.7)	771
Group8	Cyanobacteria (11/11)	160	(24.7)	29	(4.4)	17	(2.6)	48	(7.4)	190	(29.4)	18	(2.8)	15	(2.3)	4	(0.5)	123	(19.0)	19	(3.0)	4	(0.6)	20	(3.2)	647
Group9	Lactobacillales (21/22)	152	(31.6)	25	(5.2)	15	(3.1)	47	(9.8)	126	(26.3)	14	(2.9)	7	(1.5)	3	(0.7)	53	(11.1)	13	(2.8)	2	(0.5)	21	(4.4)	479
Group10	Bacillales (26/26)	193	(26.9)	32	(4.4)	27	(3.7)	50	(7.0)	229	(31.9)	27	(3.7)	7	(1.0)	2	(0.3)	108	(15.0)	16	(2.2)	7	(1.0)	21	(2.9)	718
Group11	Deinococcus-Thermus (3/3)	175	(25.7)	34	(5.0)	23	(3.3)	50	(7.3)	227	(33.2)	18	(2.6)	7	(1.0)	2	(0.3)	102	(15.0)	13	(2.0)	9	(1.4)	22	(3.2)	682

Group12	Archaea (23/23)	122	(25.2)	19	(4.0)	11	(2.4)	44	(9.0)	168	(34.9)	11	(2.2)	2	(0.5)	3	(0.6)	64	(13.2)	13	(2.7)	6	(1.3)	19	(4.0)	483
Group13	Metazoa (3/3)*	202	(28.8)	22	(3.1)	71	(10.1)	52	(7.4)	194	(27.7)	20	(2.9)	14	(2.0)	1	(0.1)	77	(11.1)	17	(2.5)	8	(1.1)	21	(3.0)	699
Group14	Fungi (4/5)	183	(25.9)	29	(4.1)	31	(4.4)	49	(7.0)	237	(33.6)	21	(3.0)	14	(2.0)	1	(0.1)	93	(13.2)	16	(2.3)	9	(1.2)	22	(3.1)	706

Table 2: Metabolic influence over the different clusters. For each group of organisms the mean number of enzymes involved in each kind of metabolism and the percentage of enzymes that belong to a determined metabolism in comparison to the total number of enzymes used to create the dendrogram are shown. Green and red numbers or percentages indicate a group of organisms that has more or lower enzymes of a kind of metabolism than most of the other groups. * Sus scrofa (ssc) was excluded from these data because of the lack of KEGG numbers on important metabolic protein