Comparative genomic analysis by microbial COGs self-attraction rate

Daniele Santoni, Vincenzo Romano-Spica

Comparative genomic analysis by microbial COGs self-attraction rate

Daniele Santoni, Vincenzo Romano-Spica

PII: S0022-5193(09)00047-2
DOI: doi:10.1016/j.jtbi.2009.01.035
Reference: YJTBI5452

To appear in: *Journal of Theoretical Biology*

Received date: 4 July 2008
Revised date: 22 December 2008
Accepted date: 27 January 2009

Cite this article as: Daniele Santoni and Vincenzo Romano-Spica, Comparative genomic analysis by microbial COGs self-attraction rate, *Journal of Theoretical Biology* (2009), doi:10.1016/j.jtbi.2009.01.035

Comparative genomic analysis by microbial COGs self-attraction rate.

Daniele Santoni[1,2], Vincenzo Romano-Spica[2]

1 Institute for Computing Applications "M. Picone", National Research Council (CNR) Rome - Italy.

2Department of Health Sciences - Laboratory of Microbiology and Bioinformatics, University of Rome – "Foro Italico", Rome, Italy.

Corresponding author: Vincenzo Romano Spica.

Email: vincenzo.romanospica@iusm.it

tel/fax: +39 06 36733223.

Piazzale L. de Bosis 6, 00194, Rome, Italy.

Abstract

Whole genome analysis provides new perspectives to determine phylogenetic relationships among microorganisms. The availability of whole nucleotide sequences allows different levels of comparison among genomes by several approaches. In this work, self-attraction rates were considered for each cluster of orthologous groups of proteins (COGs) class in order to analyse gene aggregation levels in physical maps. Phylogenetic relationships among microorganisms were obtained by comparing self-attraction coefficients. 18-dimensional vectors were computed for a set of 168 completely sequenced microbial genomes (19 archea, 149 bacteria). The components of the vector represent the aggregation rate of the genes belonging to each of 18 COGs classes. Genes involved in non-essential functions or related to environmental conditions showed the highest aggregation rates. On the contrary genes involved in basic cellular tasks showed a more uniform distribution along the genome, except for translation genes. Self-

attraction clustering approach allowed classification of Proteobacteria, Bacilli and other species belonging to Firmicutes. Rearrangement and Lateral Gene Transfer events may influence divergences from classical taxonomy. Each set of COG classes aggregation values represents an intrinsic property of the microbial genome. This novel approach provides a new point of view for whole genome analysis and bacterial characterization.

Keywords: Evolution; Phylogeny; Whole genome analysis; Gene order.

Introduction

To date a large amount of microbial organisms have been completely sequenced. Advances in genomic research and bioinformatics allow the acquisition of new points of view to compare different microorganisms and perform phylogenetic analysis. Traditionally bacterial phylogenetic trees have been built on the basis of sequence similarity of small ribosomal subunit, especially 16S rRNA (Woese, 1987; Woese, 1990). Whole genome analysis is now able to compare microbial organisms not only focusing on individual gene families, but also considering all the genes and their relative order. Gene order conservation is generally well preserved at close phylogenetic distance and it could be partially lost during evolution (Tamames *et al.*,1997; Huynen *et al.*,1998).

Lots of conserved clusters of genes can be found within microbial genomes and often correspond to operons. This is the basic element in microbial gene order analysis and consists of co-oriented and functionally related genes, that form a transcription unit with a unique promoter, enabling simultaneous and equimolar expression. Operons appear to be conserved during evolution because coordinated regulation may provide a selective benefit. It has also been observed that highly conserved clusters of genes are composed of Open Reading Frames (ORF) belonging to the same functional class. Moreover, genes responsible for related functions are frequently located close together on genetic maps (Tamames, 2001).

The notion of functional class was described by Tatusov and colleagues introducing the system of Clusters of Orthologous Groups of proteins (COGs) (Tatusov *et al.,*1997; Tatusov *et al.,*2000, Tatusov *et al.,* 2001).. Each set of clusters responsible for a common given cellular task, such as translation, transcription, cell mobility and secretion, energy production and conversion represents a functional category or a class. The analysis of COGs distribution may provide taxonomic information and contribute new insights in gene regulation and function. The presence of clusters of genes, conserved among species and belonging to the same functional class, suggests new strategies also for phylogenetic tree analysis. In 2001 Tamames showed how measuring gene order by comparative analysis of conserved “runs” can represent a valid

instrument to evaluate phylogenetic distances in prokaryotes (Tamames, 2001). Several other strategies were proposed based on different approaches (Wolf *et al.*, 2001a; Kunin *et al.*, 2005). In this paper, self-attraction methods have been applied using aggregation coefficients of COG classes to compare gene order conservation rate among prokaryotes. The general strategy was previously applied for the identification of keywords in literary texts and is based on the principle that words with a relevant meaning tend to attract themselves (Ortuno *et al.*, 2002). The computed self-attraction coefficients provide valuable information on the aggregation rate of genes belonging to functional classes and can support phylogenetic genome analysis.

Methods

A set of 168 (19 archea, 149 bacteria) completely sequenced organisms (additional data - Organisms list), with a genome size greater than 1.5 Mb and with a unique chromosome was retrieved from GeneBank database (http://www.ncbi.nlm.nih.gov). The maps of Clusters of Orthologous Groups of proteins (COGs) classes were extracted for each of these genomes, to the aim of studying the distribution of gene categories. The choice of single chromosome organisms, even if representing a limitation, allowed the comparison of genomes by a unique source data. The circular structure of bacterial chromosomes simplified the application of the algorithm. Only for a few exceptions with linear chromosomes, an artificial closure at the 3' and 5' ends was performed to circularize the sequence, such as in the literarature text analysis (*Borrelia burgdoferi* B31, *Streptomyces avermitilis* MA-4680, *Streptomyces coelicolor* A3). A set of Perl scripts were developed in order to parse and analyse the obtained data as reported below. All genes of each genomes were numbered, assigning a label 1 to the first gene, 2 to the second and so on untill the last one. For each COG class the distribution of the distances between every two consecutive occurrences of genes belonging to the same category was considered. Every class distribution was normalized with respect to its mean. The standard deviation of the normalized distribution was computed for 18 COGs categories for each bacterial genome. This value was defined as the coefficient of self-attraction $\alpha_i(X)$ where X is

one of the considered COG class and *i* identifies the *i*-th organism in our list. Formally, the algorithm can be defined as follows. All the genes of a given genome were considered; a number was assigned progressively to each gene with respect to its position along the sequence. Let *n* be the total number of genes of the *i*-th organism. Let $\{x_k\}_j$ be the succession of gene positions, where *j* identifies a COG class belonging to the set Ω = { C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, T, U, V } and $1 \leq k \leq n(j)$ where $n(j)$ is the total number of genes of class *j* in the sequence. The distribution $\{s_k\}_j$ of the distances (that is to say how many genes between x_k e x_{k+1}) of all the couples of consecutive occurrences of genes of the given class *j* was computed: $s_k = x_{k+1} - x_k$ for $1 \leq k < n(j)$ and $s_{n(j)} = (n - x_{n(j)}) + x_1$. In order to eliminate both the dependence on frequency for different classes and the dependence on the length of the genome, $\{s_k\}_j$ was normalized by dividing each element for the mean of the distribution, that is to say the ratio of the total number of genes and the number of gene occurrences of the class. The self- attraction coefficient $\alpha(X)$, with $X \in \Omega$, was defined as the standard deviation of the above distribution. We used the notation $\alpha_i(X)$ to refer to a specific microorganism in our set.

A 18-dimension vector $\alpha_i = (\alpha_i(C), \alpha_i(D), \ldots, \alpha_i(V))$ was obtained for each organism, each scalar value representing the aggregation rate of the genes belonging to a COG class in that genome. The mean and standard deviation inter- and intra-organism were considered. The mean value and the corresponding standard deviation were computed for each function class in order to evaluate the rate of aggregation of each cluster of genes in the considered pool of genomes (inter-organism). On the other hand, the mean value and the corresponding standard deviation were computed in order to evaluate the aggregation rate of the genes for each genome (intra-organism). The 18-dimension space was considered with respect to Euclidean metric; the 168 points in this space, corresponding to the 168 bacterial organisms, were clustered with a Perl implementation of a k-means algorithm (Jagota, 2000). Several tests were performed with different number of clusters (k=5,6,7,..,40) randomly picking initial means. For each fixed number of clusters 10000 simulations were performed selecting those having the best fit, that is

to say the smallest measure referring to the sum of distances of each point from the mean of its cluster. The measure φ of a simulation is defined as follows:

$$\varphi = (1/|D|) \sum d(x, \mu_{C(x)})$$

where $\mu_{C(x)}$ denotes the mean of the cluster to which datum x is assigned in the simulation, *D* is the set of all considered organisms and *d* is the Euclidean distance in the 18-dimension space. The best fit clustering simulations were analysed to retrieve similarity and identify the sets of organisms falling in the same cluster. A phylogenetic tree was built considering the distances of each couple of clusters as the distance between their means. Fitch-Margoliash and least-squares distance methods from Phylip 3.65 package (Felsenstein, 1993; Felsenstein, 1989) were used to obtain the tree data and Phylodendron to generate the tree drawing (http://iubio.bio.indiana.edu/treeapp/).

Results

COGs maps belonging to a set of 168 microorganisms (19 archea, 149 bacteria) were extracted from GenBank database (http://www.ncbi.nlm.nih.gov). The distribution profiles of the COGs classes in the genomes were examined in order to assess the presence of cluster of genes belonging to the same functional class. To this aim the aggregation rate of every class in the microbial genomes was investigated by considering the self-attraction coefficients as parameters of how genes of a class "attract" themselves. A system built on algorithms used for keywords detection in literary text was implemented (Ortuno *et al.*, 2002). The general principle is based on the analysis of frequency fluctuations of word occurrences. It was adapted to provide self-attraction rate for COGs classes in bacterial genomes. The standard deviation of the distances

between successive occurrences of genes belonging to the same functional class, resulted an effective parameter to quantify self-attraction rates, as it works for keywords identification within a text. The greater the coefficient of the standard deviation the stronger the class self-attraction rate will be.

A definitive phylogenetic tree was obtained based on this criteria (Figure 1).

Self-attraction rate for organisms and classes.

Eighteen self-attraction COGs class coefficients were computed for each bacterial genome (additional data - Complete table of self-attraction values). The mean and the standard deviation between all the 168 organisms for each COGs class were analysed (table 1). The mean values ranged between 1.01 and 1.55. The computation provided the greatest mean values for the following gene clusters: translation (mean = 1.55), cell motility (mean = 1.52), energy production and conversion (mean = 1.43). The lowest values were found for posttranslational modification, protein turnover and chaperones (mean = 1.20), cell cycle control, mitosis (mean = 1.13) and transcription (mean = 1.01). The analysis of standard deviation showed a dispersion around the mean that ranged from 0.08 for transcription and 0.47 for cell mobility, supporting a specific aggregation rate for each functional category. The observed values showed that genes involved in a certain task are distributed more or less homogenously, along the genome, depending on the class they belong to. The genes of the translation class appear to be the most grouped in all organisms with a standard deviation of 0.27. Genes involved in cell mobility have a high rate of aggregation in the motile organisms, while the fewer genes present in not-motile ones show a uniform distribution. This is also revealed by the standard deviation for cell mobility (0.47), the highest between all the function classes. For example, *Zymomonas mobilis* subs. Mobilis ZM4 has a very high value (3.04); while *Methanothermobacter thermautotrophicus* str. Delta H and *Picrophilus torridus* DSM 9790 have the lowest values (0.66 and 0.40, respectively). The genes of the transcription class appear to be homogenously

distributed with respect to the others (1.01) and constantly in all the organisms because of a very low standard deviation (0.08).

The mean value and the standard deviation among all classes were also computed for each organism (additional data - Organisms means values). The mean values of all organisms ranged from a minimum around 1 in *Ehrlichia ruminantium* str. Welgevonden (mean = 1.00) and *Gloeobacter violaceus* PCC 7421 (mean = 1.03), up to more than 1.40 in *Yersinia pseudotuberculosis* IP 32953 (mean = 1.40), *Escherichia coli* CFT073 (mean = 1.42) and *Geobacillus kaustophilus* HTA426 (mean = 1.45). In general those microorganisms, close in classical taxonomy, showed similar aggregation values. This can be due to conservation in gene order or operon regions, supporting the feasibility of the approach in inferring phylogenetic classification.

Description of the clusters.

In order to group microbes with respect to self-attraction parameters, an iterative k-means algorithm was implemented considering the distance between each couple of genomes as Euclidean distances between their components. Every organism was attributed to a group considering the distances with respect to the means of the clusters. A number of ten thousand simulations seemed to be appropriate in providing a suitable performance for each number of clusters (figure 2). When considering different number of clusters (from 5 up to 40), a best performance at 26 (s = 0.5051) was identified with a reasonable range between 22 and 30 clusters (figure 3 and additional data). Simulations with a lower or higher number of clusters did not change the general distribution of the microbial genomes, but reduced the accuracy due to excess in aggregation or overspreading, respectively. Comparison of different simulations revealed that cluster distributions do not change for most homogeneous groups such as strains of *Streptococcus pyogenes*, *Staphylococcus spp.*, *Bacillus spp*., but also members of the Euryarchaeota, or Gammaproteobacteria including *Salmonella spp., E. coli* and *Shigella spp*.

In order to estimate the evolutionary relationship between clusters, a phylogenetic tree was built considering the distance between clusters as the distance between their means (figure 1).

The microbial genomes belonging to each cluster are shown in table 2. Cluster 1 was very conserved in the simulations and comprises ten organisms all belonging to Firmicutes phylum: *Staphylococci* (*Staphylococcus aureus*, *Staphylococcus haemoliticus*, *Staphylococcus epidermidis*) and *Clostridium perfringens*. Cluster 2 comprises two different Crenarchaeota, belonging to the genus *Sulfolobus*, with a considerable distance from the mean (D=0.60). Cluster 3 comprises four Firmicutes *Streptococci* and two *Bacteroides fragilis* at higher distance (D>0.60). Cluster 4 contains the V583 strain of *Enterococcus faecalis,* a Firmicutes. Cluster 5 comprises Actinobacteria (n=7, including four Corynebacteria), Firmicutes (n=3, Lactobacilli), Bacteroides (n=1), Deinococcus-Thermus (n=1) and Fusobacterium (n=1). Cluster 6 comprises Actinobacteria (n=4, including *Mycobacterium tuberculosis CDC1551*) and Firmicutes (n=2, *Streptococcus thermophilus*). Cluster 7 comprises Archeobacteria (n=3, including two Crenarchaeota and one Euryarchaeota), Alphaproteobacteria (n=1, *Thermotoga maritima* strain MSB8); the closest element to the mean is *Haemophilus influenzae* 86-028NP belonging to Gammaproteobacteria; it is remarkable that strain Rd KW20 of *Haemophilus influenzae* is located in cluster 22, supporting the hypothesis of a different genomic structure with respect to strain 86-028NP (Harrison et al. 2005). Cluster 8 comprises two Bacteroidetes, one Epsilonproteobacteria and one Euryarchaeota. Cluster 9 comprises seven strains of *Streptococcus pyogenes* and *Sulfolobus tokodaii* that belongs to Crenarchaeota but is distant from *Sulfolobus acidocaldarius* and *solfataricus* located in cluster 2. Interestingly, *Streptococcus pyogenes* SSI is far from the other Streptococci (D=0.67 vs 0.20≤D≤0.47) and in several simulations it fell into different clusters. Cluster 10 is the most heterogeneous comprising Planctomycetes (n=1), Cyanobacteria (n=3), Chlamydiae (n=1), [1]Epsilonproteobacteria (n=3), Deinococcus-Thermus (n=1) and Betaproteobacteria (n=1). Cluster 10 interrupts the predominance of Firmicutes and begins that of Proteobacteria, with the exception of cluster 18 and 19 containing Bacilli and Listeria. Cluster 11, 12 and 13 comprise Alpha, Beta, Gamma, Delta and Epsilon Proteobacteria. Cluster 12 is a singleton: *Idiomarina loihiensis* L2TR ,a Gammaproteobacteria. Cluster 14 is heterogeneous with the prevalence of

1

Proteobacteria and Euryarchaeota, it also includes one Actinobacteria and one Spirochaetes (*Borrelia bugdoferi*). Cluster 15 is composed by three *Legionellae* very close to each other (0.19≤D≤0.25). Cluster 16 and 17 comprise Gammaproteobacteria (n=14, including enterobacteriaceae *Salmonella spp.*, *Shigella spp.*, *Escherichia coli* strain K12 and CFT073, *Yersinia spp.*), Betaproteobacteria (n=3, *Bordetella spp*.), Alphaproteobacteria (n=1, *Zymomonas mobilis*). Cluster 16 is one of the most conserved with all the organisms close to the mean (0.29≤D≤0.54). Cluster 18 and 19 harbor the majority of *Bacillus spp.* and *Listeria spp.,* belonging to the Firmicutes phylum, and several Proteobacteria including *Escherichia coli* O157. Cluster 20 comprises five *Pseudomonas,* other Alpha, Beta and Gamma Proteobacteria and Firmicutes *Bacillus licheniformis* that often clustered with Proteobacteria and not with other Bacillaceae. Cluster 21 is composed by one Euryarchaeota and one Spirochaetes. Cluster 22 comprises Alpha (n=2), Beta (n=2, *Neisseria meningitidis*), Gamma (n=2, *Haemophilus influenzae* Rd KW20) and Delta (n=1) Proteobacteria. Cluster 23 is composed by seven Euryarchaeota. Cluster 24 comprises Acquificae, Gamma and Epsilon Proteobacteria. Cluster 25 comprises one Gammaproteobacteria (*Haemophilus ducreyi*) and one Euryarchaeota. Cluster 26 is a singleton: *Ehrlichia ruminantium* str. Welgevonden Alphaproteobacteria, the organism having the smallest mean aggregation value (mean=1.00).

A general representation of the distribution of the different clusters is reported in figure 1, showing also the distribution of Gram-positive, Gram-negative, bacilli and extremophiles species. It evidences the distance among different clusters and the closure of taxonomically related microbial genomes, with some exceptions.

Discussion

Insight in bacterial genome organization today is one of the major challenges in computational sequence analysis. Recombination events and Lateral Gene Transfer (LGT) may deeply influence the evolution of bacterial genomes, so that relationships between species could be depicted as a network rather than a tree. Comparison of whole genome sequences shows a loss

of gene order conservation beyond the level of operons, even among relatively close species (Tamames *et al.,* 1997; Huynen *et al.*, 1998). Analysis of conserved clusters of genes, often involved in operons, may represent a strategy to investigate evolution and predict possible functional associations (Wolf *et al.*, 2001b). Maps of Clusters of Orthologous Groups of proteins (COGs) provide a tool for estimating genomic regions considering a functional point of view (Ling *et al.*, 2002).

When computing the self-attraction COGs class values in all 168 microbial genomes, lowest aggregation rates were observed for those genes responsible for essential functions such as transcription, cell cycle control, mitosis, post-translational modifications, and highest values for non-essential functions related to specific environmental conditions, such as motility, transport and catabolism (Table 1). Even if belonging to essential functional class, translation represents an exception, since it is highly clustering in all microorganisms, with the highest aggregation value (Mean=1.55). In according to the selfish model for operon formation described by Lawrence and Roth, weakly-selected genes responsible for not-essential functions are clustered; on the contrary, genes for essential processes are not clustered, with several exceptions including the ribosomal genes involved in translation (Lawrence and Roth 1996). However, recent reports using genome wide analysis found essential genes in operons, suggesting a more complex scenario (Pal, 2004).

Based on the self-attraction rates computed for each bacterial genome, a k-means algorithm was implemented to obtain clusters of phylogenetically related species. Analysis of the aggregation rate allowed us to cluster different microbial genomes including Proteobacteria, Bacilli, and other species belonging to Firmicutes (Figure 1).

Self-attraction rate, LGT and genomic rearrangement.

Interestingly, LGT or major rearrangements were found in strains that diverge in clusterization, but are close in traditional taxonomic classification. For example, three Euryarchaeota belonging to hyperthermophilic Pyrococcus genus are located in different clusters: *Pyrococcus horikoshii* in cluster 14, *Pyrococcus abyssi* in cluster 21 and *Pyrococcus furiosus* in cluster 23.

There are evidences of insertion of mobile genetic elements in *P. horikoshii* and *P. abyssi*; GC content and codon usage analysis of diversity of several homologous genes support a possible recent acquisition by horizontal transfer (Chinen *et al.*, 2000; Zivanovic *et al.*, 2002). Also other three Crenarchaeota belonging to the genus Sulfolobus do not group together: *Sulfolobus tokodaii* str. 7 in cluster 9, *Sulfolobus acidocaldarius* DSM 639 and *Sulfolobus solfataricus* P2 in cluster 2, even if cluster 2 contains only two elements in opposite location at high distance from the mean (D=0.60). Indeed, *Sulfolobus* genomes contain large numbers of putatively mobile elements, both IS elements (Insertion Sequence elements) and MITEs (Miniature Inverted-repeat Transposable Elements), suggesting major structural rearrangements occurred since the three organisms diverged (Brugger *et al.*, 2004).

Microbial genomes belonging to the species of *Haemophilus influenzae* have been included in our set: RD KW20 and 86-028NP, a non pathogenic and a nontypeable pathogenic strain respectively. The former belongs to cluster 22, close to other Proteobacteria, the latter is in cluster 7, far from other proteobacteria and closer to Firmicutes and Actinobacteria. Despite large regions of similarity, strain 86-028NP's contains major rearrangements in the genome architecture and expresses 208 genes absent in strain RD (Harrison *et al.,* 2005).

Streptococcus pyogenes strain SSI-1 belongs to cluster 9 together with other members of the same species. However, the SSI-1 strain shows a higher distance from the mean and in several simulations fall in separated clusters. The SSI-1 genome is highly conserved with respect to other strains, but a large genomic rearrangement has been described to occur in specific regions involving genes encoding superantigens and mitogenic factors (Nakagawa *et al.*, 2003).

Four strain belonging to the species *Escherichia coli* have been included in our set. The non pathogenic ones, K12 and CFT073 fall in cluster 16, the pathogenic strains, O157:H7 and O157:H7 EDL933, fall in cluster 19. Comparative analysis between *E. coli* K12 and *E. coli* O157:H7 revealed a surprising level of diversity

between the two genomes. Most differences in overall gene content are attributable to horizontal transfer, that may indicate candidate genes involved in pathogenesis (Perna *et al.*, 2001). Indeed, 1,387 new genes encoded in strain-specific regions were found in O157:H7, including

alternative metabolic capacities. This is in agreement with the observed self-attraction values, mainly divergent within the COG classes E, H, I, P involved in metabolic cellular activities.

Microorganisms belonging to the phylum Actinobacteria are located in cluster 5 (n=7) and 6 (n=4), with the exception of *Leifsonia xyli* subsp. xyli str. CTCB07 in cluster 14 and *Symbiobacterium thermophilum* IAM 14863 in cluster 18. Despite *Symbiobacterium thermophilum* is attributed in the classical taxonomy to the phylum of Actinobacteria, the analysis of its proteome demonstrated a greater similarity with Firmicutes, including Bacilli and Clostridia (Ueda *et al.*, 2004). Most of *S. thermophilum* proteins (47%) showed top match similarity with proteins from Firmicutes, in particular with *Thermoanaerobacter tengcongensis*. In addition, *S. thermophilum* genome contains several transposons and insertion sequences, and a variety of respiratory systems including Nap nitrate reductase, which were found only in Gram-negative bacteria. (Ueda *et al.*, 2004). The ability of *S.thermophilum* leading to tryptophanase and tyrosinase production, typical of enterobacteria, further suggests that this bacterium may be considered a Gram negative (Hirahara *et al.*, 1992; Hirahara *et al.*, 1993). However, *S.thermophilum* lacks the major Gram-negative membrane biosynthesis proteins, in agreement with 16S rDNA phylogeny data, that assign *S.thermophilum* to the Gram-positive bacterial group. Otherwise, *S.thermophilum* produces endospores, a property found only in two classes of firmicutes: Bacilli and Clostridia, where Gram-variability has been described (Beveridge, 1990; Ueda *et al.*, 2004). All these similarities between *S.thermophilum*, Bacilli, Clostridia and the close genomic relationships with *Thermoanaerobacter tengcongensis* are consistent with the observed self-attraction clusterization reported in table 2, supporting the effectiveness of self-attraction analysis in discriminating microbial organisms with respect to their genome architecture and functional features.

Divergence from classical taxonomy.

However, some microorganisms, closely related in classical taxonomy and belonging to the same species or genus, appear to be distant in the phylogenetic tree without any apparent clear reason. This can infer the presence of divergent patterns of gene clusters. For example

Clostridium perfringens st. 13 is located in cluster 1 with several species of Staphylococci while other species of Clostridium, *C. tetani* E88 and *C. acetobutylicum* ATCC 824 belongs to cluster 18 with several Bacilli and gammaproteobacteria. Even if *Coxiella burnetii* genome was reported to be very closely related to *Legionella pneumophila*, it is included in cluster 11, while all the three fully sequenced strains of *L. pneumophila* are located in cluster 15 (Chien et al. 2004). Other microbial genomes locates distantly from all the others, such as *Picrophilus torridus* DSM 9790 or *Bartonella henselae* str. Houston-1 in cluster 7, *Bradyrhizobium japonicum* USDA 110 in cluster 14, showing a very high distance from the mean (0.95, 0.98, 1.13, respectively), *Enterococcus faecalis* V583, *Idiomarina loihiensis* L2TR and *Ehrlichia ruminantium* str. Welgevonden, even segregated in single-element clusters (n. 4, 12, 26, respectively) (Figure 1).

This could be a starting point to investigate this exception in order to better understand their evolution and their rearrangement in their gene order.

These and other observations, suggest that each set of COGs aggregation values represents an intrinsic property of the bacterium. Microbial self-attraction rates seem to represent a relevant feature suitable in enlightening evolutionary properties and can support genome comparison of microbial species.

Acknowledgments

This study was supported by grants IUSM 03/05337 and MIUR 2005068489_004.

References

Beveridge, T.J., 1990. Mechanism of gram variability in select bacteria. J Bacteriol 172, 1609-1620.

Brugger, K., Torarinsson, E., Redder, P., Chen, L. & Garrett, R.A., 2004. Shuffling of Sulfolobus genomes by autonomous and non-autonomous mobile elements. Biochem Soc Trans. 32, 179-183.

Chien, M., Morozova, I., Shi, S. & other 34 authors, 2004. The genomic sequence of the accidental pathogen Legionella pneumophila. Science. 305, 1966-1968.

Chinen, A., Uchiyama, I. & Kobayashi, I., 2000. Comparison between *Pyrococcus horikoshii* and *Pyrococcus abyssi* genome sequences reveals linkage of restriction-modification genes with large genome polymorphisms. Gene. 259, 109-121.

Felsenstein, J., 1989. PHYLIP -- Phylogeny Inference Package (Version 3.2). Cladistics. 5, 164-166.

Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

Harrison, A., Dyer, D.W., Gillaspy, A. & other 10 authors (2005). Genomic sequence of an otitis media isolate of nontypeable *Haemophilus influenzae*: comparative study with *H. influenzae* serotype d, strain KW20. J Bacteriol. 187, 4627-4636.

Hirahara, T., Horinouchi, S. & Beppu, T., 1993. Cloning, nucleotide sequence, and overexpression in *Escherichia coli* of the beta-tyrosinase gene from an obligately symbiotic thermophile, *Symbiobacterium thermophilum*. Appl Microbiol Biotechnol. 39, 341-346.

Hirahara, T., Suzuki, S., Horinouchi, S. & Beppu, T., 1992. Cloning, nucleotide sequences, and overexpression in *Escherichia coli* of tandem copies of a tryptophanase gene in an obligately symbiotic thermophile, *Symbiobacterium thermophilum*. Appl Environ Microbiol. 58, 2633-2642.

Huynen, M.A. & Bork, P., 1998. Measuring genome evolution. Proc Natl Acad Sci U S A. 95, 5849-5856.

Jagota, A., 2000. *Data Analysis and Classification for Bioinformatics*. Bioinformatics By The Bay Press.

Koonin, E.V., Tatusov, R.L. & Galperin, M.Y., 1998. Beyond complete genomes: From sequence to structure and function. Curr Opin Struct Biol. 8, 355-363.

Kunin, V., Ahren, D., Goldovsky, L., Janssen, P. & Ouzounis, C.A., 2005. Measuring genome conservation across taxa: divided strains and united kingdoms. Nucleic Acids Res. 33, 616-621.

Lawrence, J.G. & Roth, J.R., 1996. Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics. 143, 1843-1860.

Lin, J. & Gerstein, M., 2000. Whole-genome Trees Based on the Occurences of Folds and Orthologs: Implications for Comparing Genomes on Different Levels. Genome Res. 10, 808-818.

Ling, L., Wang, J., Cui, Y., Li, W. & Chen R., 2002. Proteome-wide analysis of protein function composition reveals the clustering and phylogenetic properties of organisms. Mol Phylogenet Evol. 25, 101-111.

Nakagawa, I., Kurokawa, K., Yamashita, A. & other 10 authors, 2003. Genome Sequence of an M3 Strain of *Streptococcus pyogenes* Reveals a Large-Scale Genomic Rearrangement in Invasive Strains and New Insights into Phage Evolution. Genome Res. 13, 1042-1055.

Ortuño, M., Carpena, P., Bernaola-Galvan, P., Muñoz, E. & Somoza, A.M., 2002. Keyword detection in natural languages and DNA. Europhys Lett. 57, 759–764.

Perna, N.T., Plunkett, G. 3rd, Burland, V. & 25 other authors, 2001. Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. *Nature.* 409, 529-533.

Tamames, J. (2001). Evolution of gene order conservation in prokaryotes. *Genome Biology* 2, research0020.1-0020.11.

Tamames, J., Casari, G., Ouzounis, C. & Valencia, A., 1997. Conserved clusters of functionally related genes in two bacterial genomes. J Mol Evol. 44, 66-73.

Tatusov, R.L., Koonin, E.V. & Lipman, D.J., 1997. A genomic perspective on protein families. Science. 24, 631-637.

Tatusov, R.L., Galperin, M.Y., Natale, D.A. & Koonin, E.V., 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 28, 33-36.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D. & Koonin, E.V.,. 2001. The COG database: new

developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 29, 22-28.

Ueda, K., Yamashita, A., Ishikawa, J., Shimada, M., Watsuji, T.O., Morimura, K., Ikeda, H., Hattori, M. & Beppu, T., 2004. Genome sequence of *Symbiobacterium thermophilum*, an uncultivable bacterium that depends on microbial commensalism. Nucleic Acids Res. 32, 4937-4944.

Woese, C.R., 1987. Bacterial evolution. Microbiol Rev. 51, 221-271

Wolf, Y.I., Rogozin, I.B., Grishin, N.V., Tatusov, R.L. & Koonin, E.V., 2001a. Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evol Biol. 20, 1-8.

Wolf, Y.I., Rogozin, I.B., Kondrashov, A.S., & Koonin, E.V., 2001b. Genome Alignment, Evolution of Prokaryotic Genome Organization, and Prediction of Gene Function Using Genomic Context. *Genome Res.* 11, 356-372.

Zivanovic, Y., Lopez, P., Philippe, H. & Forterre, P., 2002. Pyrococcus genome comparison evidences chromosome shuffling-driven evolution. Nucleic Acids Res. 30, 1902-1910.

Tables

Table 1: COGs classes mean self-attraction value and standard deviation.

COGs class	Function class description	Mean value	Standard deviation
J	Translation	1.55	0.27
N	Cell motility	1.52	0.47
C	Energy production and conversion	1.43	0.17
H	Coenzyme transport and metabolism	1.36	0.19
P	Inorganic ion transport and metabolism	1.34	0.14
G	Carbohydrate transport and metabolism	1.34	0.21
E	Amino acid transport and metabolism	1.33	0.12
U	Intracellular trafficking and secretion	1.32	0.24
M	Cell wall/membrane biogenesis	1.28	0.12
F	Nucleotide transport and metabolism	1.28	0.22
I	Lipid transport and metabolism	1.27	0.18
V	Defense mechanisms	1.27	0.23
L	Replication, recombination and repair	1.24	0.16
T	Signal transduction mechanisms	1.22	0.16
Q	Secondary metabolites biosynthesis, transport and catabolism	1.21	0.19
O	Posttranslational modification, protein turnover, chaperones	1.20	0.11
D	Cell cycle control, mitosis	1.13	0.20
K	Transcription	1.01	0.08

Table 2: 26 cluster composition with associated the Phylum/Class and the distance from the mean for each organism.

C	Phylum/Class	Organism	D	C	Phylum/Class	Organism	D
1	Firmicutes	Staphylococcus aureus subsp. aureus COL	0,16	14	Alphaproteobacteria	Bradyrhizobium japonicum USDA 110	1,13
1	Firmicutes	Staphylococcus aureus subsp. aureus MSSA476	0,25	15	Gammaproteobacteria	Legionella pneumophila str. Lens	0,19
1	Firmicutes	Staphylococcus aureus subsp. aureus MW2	0,27	15	Gammaproteobacteria	Legionella pneumophila subsp. pneumophila str. Philadelphia 1	0,20
1	Firmicutes	Staphylococcus aureus subsp. aureus MRSA252	0,30	15	Gammaproteobacteria	Legionella pneumophila str. Paris	0,25
1	Firmicutes	Staphylococcus aureus subsp. aureus Mu50	0,34	16	Gammaproteobacteria	Salmonella enterica subsp. enterica serovar Typhi str. CT18	0,29
1	Firmicutes	Staphylococcus aureus subsp. aureus N315	0,36	16	Gammaproteobacteria	Salmonella enterica subsp. enterica serovar Typhi Ty2	0,29
1	Firmicutes	Staphylococcus haemolyticus JCSC1435	0,36	16	Gammaproteobacteria	Escherichia coli K12	0,48
1	Firmicutes	Staphylococcus epidermidis RP62A	0,49	16	Gammaproteobacteria	Shigella flexneri 2a str. 2457T	0,49
1	Firmicutes	Staphylococcus epidermidis ATCC 12228	0,50	16	Gammaproteobacteria	Salmonella typhimurium LT2	0,49
1	Firmicutes	Clostridium perfringens str. 13	0,56	16	Gammaproteobacteria	Shigella flexneri 2a str. 301	0,51
2	Crenarchaeota	Sulfolobus acidocaldarius DSM 639	0,60	16	Gammaproteobacteria	Escherichia coli CFT073	0,52
2	Crenarchaeota	Sulfolobus solfataricus P2	0,60	16	Gammaproteobacteria	Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67	0,54
3	Firmicutes	Streptococcus pneumoniae TIGR4	0,43	17	Gammaproteobacteria	Yersinia pestis CO92	0,40
3	Firmicutes	Streptococcus agalactiae NEM316	0,52	17	Gammaproteobacteria	Yersinia pestis biovar Medievalis str. 91001	0,46
3	Firmicutes	Streptococcus pneumoniae R6	0,52	17	Gammaproteobacteria	Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC 9150	0,49
3	Firmicutes	Streptococcus agalactiae 2603V-R	0,61	17	Gammaproteobacteria	Yersinia pseudotuberculosis IP 32953	0,51
3	Bacteroidetes-Chlorobi	Bacteroides fragilis NCTC 9343	0,64	17	Betaproteobacteria	Bordetella parapertussis 12822	0,52
3	Bacteroidetes-Chlorobi	Bacteroides fragilis YCH46	0,71	17	Betaproteobacteria	Bordetella bronchiseptica RB50	0,56
4	Firmicutes	Enterococcus faecalis V583	0,00	17	Gammaproteobacteria	Xanthomonas oryzae pv. oryzae KACC10331	0,59
5	Actinobacteria	Streptomyces coelicolor A3(2)	0,39	17	Gammaproteobacteria	Photorhabdus luminescens subsp. laumondii TTO1	0,60
5	Actinobacteria	Corynebacterium diphtheriae NCTC 13129	0,44	17	Betaproteobacteria	Bordetella pertussis Tohama I	0,73
5	Actinobacteria	Corynebacterium glutamicum ATCC 13032	0,49	17	Alphaproteobacteria	Zymomonas mobilis subsp. Mobilis ZM4	0,89
5	Deinococcus-Thermus	Thermus thermophilus HB8	0,52	18	Firmicutes	Bacillus anthracis str. Sterne	0,36
5	Actinobacteria	Corynebacterium jeikeium K411	0,57	18	Firmicutes	Bacillus anthracis str. Ames	0,36
5	Firmicutes	Lactobacillus johnsonii NCC 533	0,57	18	Gammaproteobacteria	Pseudomonas putida KT2440	0,37
5	Bacteroidetes-Chlorobi	Bacteroides thetaiotaomicron VPI-5482	0,58	18	Firmicutes	Bacillus anthracis str. 'Ames Ancestor'	0,37
5	Actinobacteria	Corynebacterium efficiens YS-314	0,61	18	Firmicutes	Bacillus cereus ATCC 10987	0,38
5	Fusobacteria	Fusobacterium nucleatum subsp. nucleatum ATCC 25586	0,61	18	Firmicutes	Bacillus thuringiensis serovar konkukian str. 97-27	0,39
5	Actinobacteria	Thermobifida fusca YX	0,64	18	Firmicutes	Clostridium tetani E88	0,42
5	Firmicutes	Lactobacillus acidophilus NCFM	0,67	18	Firmicutes	Bacillus cereus E33L	0,45
5	Firmicutes	Lactobacillus plantarum WCFS1	0,73	18	Deltaproteobacteria	Geobacter sulfurreducens PCA	0,46
5	Actinobacteria	Bifidobacterium longum NCC2705	0,82	18	Gammaproteobacteria	Xanthomonas axonopodis pv. citri str. 306	0,49
6	Actinobacteria	Mycobacterium tuberculosis CDC1551	0,41	18	Betaproteobacteria	Nitrosomonas europaea ATCC 19718	0,51
6	Actinobacteria	Streptomyces avermitilis MA-4680	0,42	18	Betaproteobacteria	Chromobacterium violaceum ATCC 12472	0,52
6	Firmicutes	Streptococcus thermophilus LMG 18311	0,45	18	Firmicutes	Thermoanaerobacter tengcongensis MB4	0,52
6	Firmicutes	Streptococcus thermophilus CNRZ1066	0,49	18	Gammaproteobacteria	Xylella fastidiosa 9a5c	0,57
6	Actinobacteria	Nocardia farcinica IFM 10152	0,60	18	Deltaproteobacteria	Desulfotalea psychrophila LSv54	0,58
6	Actinobacteria	Propionibacterium acnes KPA171202	0,69	18	Gammaproteobacteria	Mannheimia succiniciproducens MBEL55E	0,77
7	Gammaproteobacteria	Haemophilus influenzae 86-028NP	0,43	18	Firmicutes	Clostridium acetobutylicum ATCC 824	0,79
7	Crenarchaeota	Pyrobaculum aerophilum str. IM2	0,52	18	Actinobacteria	Symbiobacterium thermophilum IAM 14863	0,82
7	Crenarchaeota	Aeropyrum pernix K1	0,66	19	Firmicutes	Bacillus subtilis subsp. subtilis str. 168	0,42
7	Thermotogae	Thermotoga maritima MSB8	0,69	19	Firmicutes	Bacillus cereus ATCC 14579	0,48
7	Euryarchaeota	Picrophilus torridus DSM 9790	0,95	19	Firmicutes	Oceanobacillus iheyensis HTE831	0,50
7	Alphaproteobacteria	Bartonella henselae str. Houston-1	0,98	19	Gammaproteobacteria	Escherichia coli O157:H7	0,52
8	Bacteroidetes/Chlorobi	Chlorobium tepidum TLS	0,51	19	Gammaproteobacteria	Escherichia coli O157:H7 EDL933	0,53
8	Bacteroidetes-Chlorobi	Porphyromonas gingivalis W83	0,53	19	Firmicutes	Bacillus halodurans C-125	0,54
8	Epsilonproteobacteria	Wolinella succinogenes DSM 1740	0,55	19	Firmicutes	Listeria monocytogenes str. 4b F2365	0,54
8	Euryarchaeota	Methanothermobacter thermautotrophicus str. Delta H	0,73	19	Firmicutes	Listeria monocytogenes EGD-e	0,58
9	Firmicutes	Streptococcus pyogenes MGAS5005	0,20	19	Firmicutes	Bacillus clausii KSM-K16	0,61
9	Firmicutes	Streptococcus pyogenes MGAS8232	0,40	19	Firmicutes	Listeria innocua Clip11262	0,65
9	Firmicutes	Streptococcus pyogenes MGAS315	0,41	19	Firmicutes	Geobacillus kaustophilus HTA426	0,73
9	Firmicutes	Streptococcus pyogenes M1 GAS	0,42	20	Gammaproteobacteria	Pseudomonas syringae pv. phaseolicola 1448A	0,24
9	Firmicutes	Streptococcus pyogenes MGAS10394	0,45	20	Gammaproteobacteria	Pseudomonas fluorescens Pf-5	0,37
9	Firmicutes	Streptococcus pyogenes MGAS6180	0,47	20	Firmicutes	Bacillus licheniformis ATCC 14580	0,38
9	Firmicutes	Streptococcus pyogenes SSI-1	0,67	20	Gammaproteobacteria	Pseudomonas syringae pv. tomato str. DC3000	0,38
9	Crenarchaeota	Sulfolobus tokodaii str. 7	0,76	20	Gammaproteobacteria	Pseudomonas syringae pv. syringae B728a	0,42
10	Planctomycetes	Rhodopirellula baltica SH 1	0,37	20	Gammaproteobacteria	Pseudomonas aeruginosa PAO1	0,43
10	Cyanobacteria	Prochlorococcus marinus str. MIT 9313	0,40	20	Gammaproteobacteria	Erwinia carotovora subsp. atroseptica SCR11043	0,44
10	Deinococcus-Thermus	Thermus thermophilus HB27	0,44	20	Betaproteobacteria	Dechloromonas aromatica RCB	0,46
10	Epsilonproteobacteria	Helicobacter pylori J99	0,48	20	Alphaproteobacteria	Silicibacter pomeroyi DSS-3	0,47
10	Epsilonproteobacteria	Campylobacter jejuni RM1221	0,48	20	Betaproteobacteria	Ralstonia solanacearum GMI1000	0,54
10	Chlamydiae-Verrucomicrobia	Candidatus Protochlamydia amoebophila UWE25	0,50	20	Alphaproteobacteria	Sinorhizobium meliloti 1021	0,54
10	Epsilonproteobacteria	Helicobacter pylori 26695	0,57	21	Euryarchaeota	Pyrococcus abyssi GE5	0,39
10	Betaproteobacteria	Neisseria gonorrhoeae FA 1090	0,59	21	Spirochaetes	Treponema denticola ATCC 35405	0,39
10	Cyanobacteria	Synechococcus sp. WH 8102	0,61	22	Betaproteobacteria	Neisseria meningitidis MC58	0,36
10	Cyanobacteria	Glocobacter violaceus PCC 7421	0,88	22	Betaproteobacteria	Neisseria meningitidis Z2491	0,44
11	Deltaproteobacteria	Bdellovibrio bacteriovorus HD100	0,36	22	Gammaproteobacteria	Haemophilus influenzae Rd KW20	0,52
11	Gammaproteobacteria	Coxiella burnetii RSA 493	0,41	22	Deltaproteobacteria	Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough	0,54
11	Gammaproteobacteria	Psychrobacter arcticus 273-4	0,46	22	Gammaproteobacteria	Methylococcus capsulatus str. Bath	0,57
11	Epsilonproteobacteria	Helicobacter hepaticus ATCC 51449	0,51	22	Alphaproteobacteria	Caulobacter crescentus CB15	0,59
12	Gammaproteobacteria	Idiomarina loihiensis L2TR	0,00	22	Alphaproteobacteria	Bartonella quintana str. Toulouse	0,74
13	Gammaproteobacteria	Acinetobacter sp. ADP1	0,29	23	Euryarchaeota	Methanosarcina barkeri str. fusaro	0,40
13	Gammaproteobacteria	Colwellia psychrerythraea 34H	0,35	23	Euryarchaeota	Methanosarcina acetivorans C2A	0,45
13	Betaproteobacteria	Azoarcus sp. EbN1	0,40	23	Euryarchaeota	Thermococcus kodakarensis KOD1	0,46
13	Alphaproteobacteria	Rhodopseudomonas palustris CGA009	0,44	23	Euryarchaeota	Pyrococcus furiosus DSM 3638	0,53
13	Gammaproteobacteria	Shewanella oneidensis MR-1	0,46	23	Euryarchaeota	Methanosarcina mazei Go1	0,60
14	Gammaproteobacteria	Xanthomonas campestris pv. campestris str. ATCC 33913	0,44	23	Euryarchaeota	Methanosphaera stadtmanae DSM 3091	0,60
14	Actinobacteria	Leifsonia xyli subsp. xyli str. CTCB07	0,46	23	Euryarchaeota	Thermoplasma volcanium GSS1	0,62
14	Gammaproteobacteria	Xanthomonas campestris pv. campestris str. 8004	0,48	24	Epsilonproteobacteria	Campylobacter jejuni subsp. jejuni NCTC 11168	0,44
14	Euryarchaeota	Methanococcus maripaludis S2	0,59	24	Gammaproteobacteria	Pasteurella multocida subsp. multocida str. Pm70	0,57
14	Euryarchaeota	Pyrococcus horikoshii OT3	0,60	24	Aquificae	Aquifex aeolicus VF5	0,63
14	Spirochaetes	Borrelia burgdorferi B31	0,68	25	Gammaproteobacteria	Haemophilus ducreyi 35000HP	0,39
14	Alphaproteobacteria	Gluconobacter oxydans 621H	0,82	25	Euryarchaeota	Methanopyrus kandleri AV19	0,39
14	Euryarchaeota	Archaeoglobus fulgidus DSM 4304	0,83	26	Alphaproteobacteria	Ehrlichia ruminantium str. Welgevonden	0,00

Figure Legend

Figure 1: Phylogenetic tree of the 26 clusters reported in table 2. Each cluster is identified with its mean and the distances are computed wit respect to Euclidean metric. Clusters are renamed clockwise. The diagram shows the distribution of Gram positive, Gram negative, extremophiles (thermophiles and psycrophiles). In Gram positive group, species belonging to Bacilli, are divided from the other. For each cluster the mean distance of the organisms from the corresponding seed is shown.

Figure 2: Number of simulations with respect to the best score. For each considered number of clusters (k=15,26,30) the best score (Y-axis) is picked out of a set with a given number of simulations (X-axis). The plots indicate how the trends of the best score decrease until reaching a plateau around 10.000 simulations.

Figure 3: Number of clusters against best score. For each number of clusters (k=5-40 X-axis) the best score (Y-axis) is plotted out of ten thousand simulations.

Figure 1

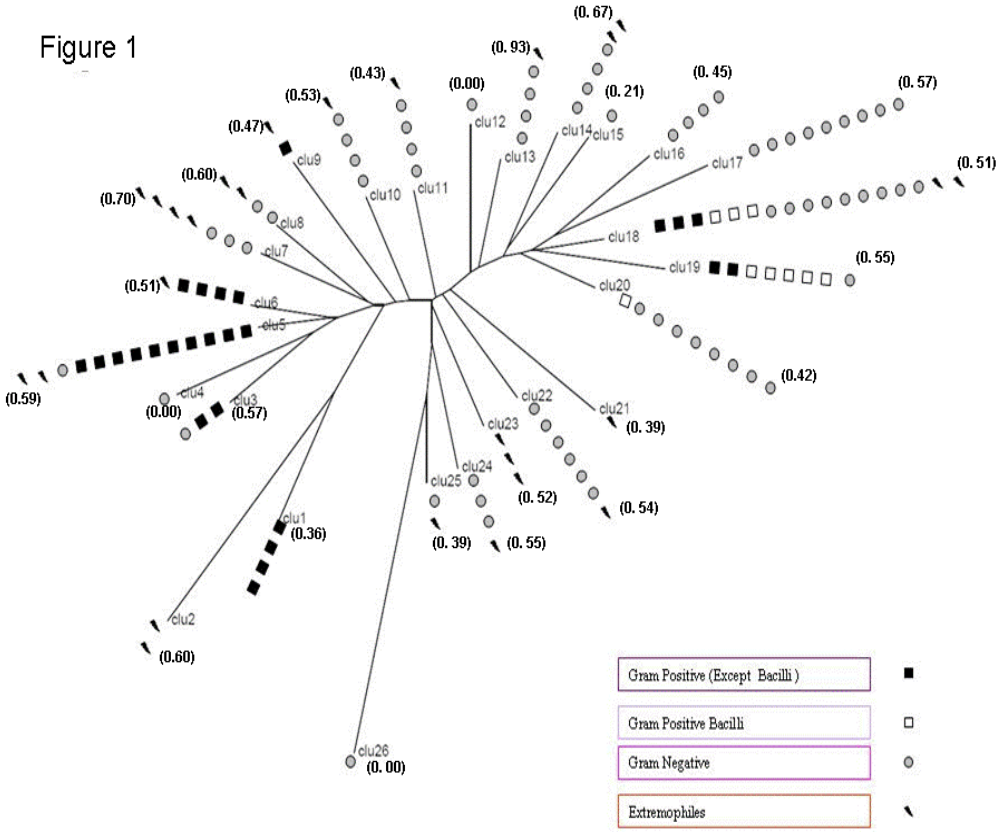

Figure 2

Figure 2

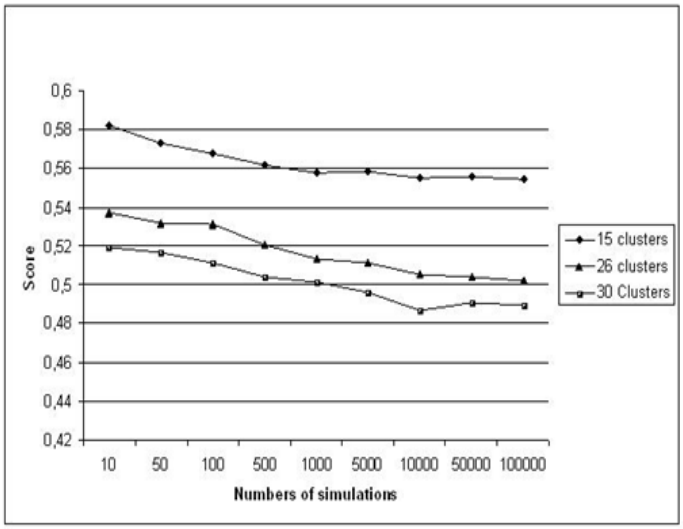

Figure 3