Quentin SCULO, Olivier LESPINET and Bernard LABEDAN

Evolution Moléculaire et Bioinformatique des Génomes, Institut de Génétique et Microbiologie, CNRS UMR 8621, Université Paris-Sud, Bâtiment 400, 91405 Orsay Cedex, France

Abstract: *The full proteomes of 109 microbial species have been exhaustively compared to find out common orthologues. The closest orthologues were grouped in families after filtering at the level of both the evolutionary distance separating these orthologues and the minimal number of genomes present in these families. Two methods were used to reconstruct a genomic tree. The distance separating a pair of genomes was computed as the mean of the evolutionary distances separating each pair of orthologues common to these two genomes and belonging to the same families. Alternatively, trees were built using a triplet method. Each triplet will vote for a pair of genomes. The vote was assessed using two approaches, a yes/no one and a proportion one. Trees obtained using both methods appear to be rather congruent and they display some remarkable features such as the position of the hyperthermophilic bacteria. Remarkably, several bacterial deep nodes appear reliable and better defined than in other trees including the one based on 16s RNA sequences.*

Keywords: families of closest orthologues, triplet of genomes, genomic trees

1 Introduction

As it has been repeatedly underlined by Carl Woese for the last 25 years, Microbiology has long suffered from an absence of sound taxonomic relationships between prokaryotic species [1]. The comparison of 16s ribosomal RNA sequences has been an immense progress in the (re)definition of the main groups of bacteria and the creation of the concept of archaea [1,2]. However, the phylogenetic tree built using these RNA sequences for the whole set of available species is dramatically unresolved at the level of its deeper nodes, whichever the method used to reconstruct this tree. Moreover, it has been suggested by several groups that such tree reconstruction is meaningless due to a very high rate of horizontal transfer between prokaryotic species [3].

The publication of the complete sequences of many prokaryotic genomes has been a unique opportunity for checking the actual topology of the Tree of Life, and especially that of microorganisms. Different approaches have been proposed in order to cope with the large datasets obtained through the systematic sequencing of complete genomes. These approaches have been based mainly on gene content, gene order, supertrees, large sets of concatenated genes, protein structural domains [4-9]. See [10] for an authoritative review.

In this paper, we present new approaches that help to increase the resolution of the internal nodes of the prokaryotic tree while giving a satisfying topology. These approaches are based on the analysis of the whole proteomes encoded by a set of 109 completely sequenced microorganisms. After an exhaustive search for orthologues which are common to these different proteomes, we group them into families. As detailed below, our families approach seems to give a better topology than similar approaches previously published [10], although it is still sensitive to several drawbacks proper to this approach (i.e. lateral transfer, genome size, etc). We also propose another new approach based on decision votes using a triplet method with two variants. We further show that this triplet approach allows to get rid of some of the previously mentioned drawbacks.

2 Finding out all sound orthologues

As it has been underlined by Fitch, reconstructing history of orthologues allows to infer species history [11]. To ascertain that the search for orthologues is both exhaustive and exact we used the DARWIN AllAll program [12] to compare each protein of each proteome of entirely sequenced genomes of microorganisms

with all the other ones. This AllAll search was made by comparing progressively all proteomes using a pair approach. As already showed [13], this program, based on a maximum likelihood approach, is very efficient to detect – in one step (using successively the dynamic programming algorithms of Needleman and Wunsch and of Smith and Waterman) - all segments of homology (SOH) even when they are very distant. This is crucial when comparing numerous genomes located at very disparate distances in the taxonomy space. A pair of SOH was retained only if they were separated by less than 250 PAM units (according to [14]) and if each segment was extending on at least 80% of the length of the shorter matching protein. To further ascertain that we get at any rate the *best* orthologue, we used a reciprocal approach as initially proposed by Overbeek *et al*. [15]. The homology detected between protein *a* encoded by genome G_A and protein *b* encoded by genome G_B, was kept only if the PAM distance separating *a* from *b* was smaller than that separating either *a* from any other protein encoded by G_B or *b* from any other protein encoded by G_A. Accordingly, we selected, for each pair of genomes, the orthologues having the shortest PAM distance. This step was crucial in eliminating the huge majority of the unwanted paralogues separated by a speciation event. Using this approach we obtained a list of 217,881 best reciprocal orthologues amounting to 71 % of the 306,938 proteins encoded by the 109 genomes (92 bacteria, 14 archaea and 3 Eucaryotes) studied in this paper.

3 Reconstructing a genomic tree of microorganisms from the distances separating their families of orthologues

3.1 Experimental design to define sound families of orthologues.

The pairs of orthologues were first sorted by their ascending PAM distance and then examined from the closest to the farthest. If none of the proteins belonged to a family, they were put in a new one. If only one protein was a member of a family, the other was added to this family. If both proteins were included in different families, both families were fused except if they shared a common species as illustrated on Fig. 1.

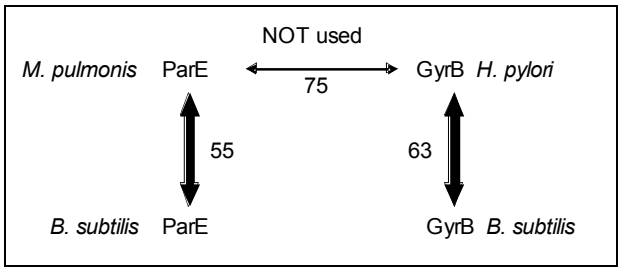

Figure 1. Filtering the pairs of homologous proteins which are not belonging to families of orthologues. The respective evolutionary distances separating the different homologues are given in PAM units.

The two paralogues GyrB and ParE encoded by *Bacillus subtilis* are matching respectively with GyrB from *Helicobacter pylori* and ParE encoded by *Mycoplasma pulmonis*. When the pair matching those two last proteins (corresponding to a higher PAM distance) was further treated it was discarded, preventing the fusion of the family of GyrB orthologues with that of ParE orthologues. Using such a filter, 813,643 pairs (corresponding to 28.7 % of the total of so-called orthologous pairs) appeared to be non orthologous and were ignored. The remaining 206,417 orthologues were grouped in 25,114 families that display a very asymmetric distribution in size. A large majority was made of very small families (e.g. 11,749 families of two members, 2,734 families of three members, etc…). We further kept only families containing a number of genomes greater than a threshold of n. For the present set of data, we tried various values of n as discussed

below (see Figs. 2 and 3).

3.2 Reconstructing trees

The evolutionary distance between a pair of genomes was calculated as the mean of the PAM distances separating each pair of orthologues common to these two genomes and belonging to the same previously defined families. These evolutionary distances were further used to build a matrix and to derive a distance tree using the Neighbor-Joining algorithm [16].

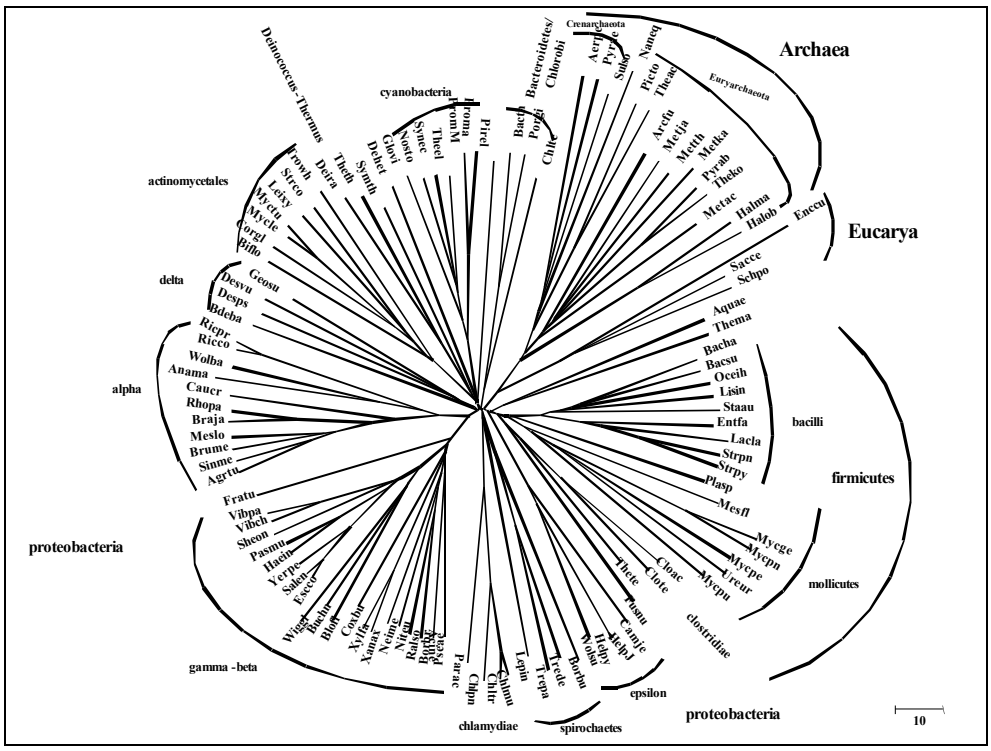

Figure 2. A genomic tree made with all families containing members belonging to at least 20 genomes. In this figure, as well as in the next ones, the species names have been abbreviated as the concatenation of the three first letters of the genus name and the two first letters of the species name. e.g. ***Bac**illus **su**btilis* = Bacsu

In order to get genomic trees reflecting the diversity of the set of studied genomes we had to use a sampling of families in terms of functions and life styles as large and assorted as possible. This was found to be the case, for instance, of the tree shown on Fig. 2 where we asked for the presence in each family of members belonging to at least 20 genomes. In these conditions, 100,373 orthologues (48.6 % of the total) that grouped in 2,295 families (9.1 % of the total) were used to compute the distance matrix and the derived tree.

Alternatively, one can use a subset of this dataset in order to increase the phylogenetic signal by enlarging the number n of genomes per families. In this case, the subset that correspond to the most ubiquitous families would be less noisy but poorer in protein sampling. Fig.3 shows an example of such a tree where the number n of genomes was increased to being at least 80 per family. This subset was made of only 20,530 orthologues (9.95 % of the total) grouped in 208 families (0.8 % of the total).

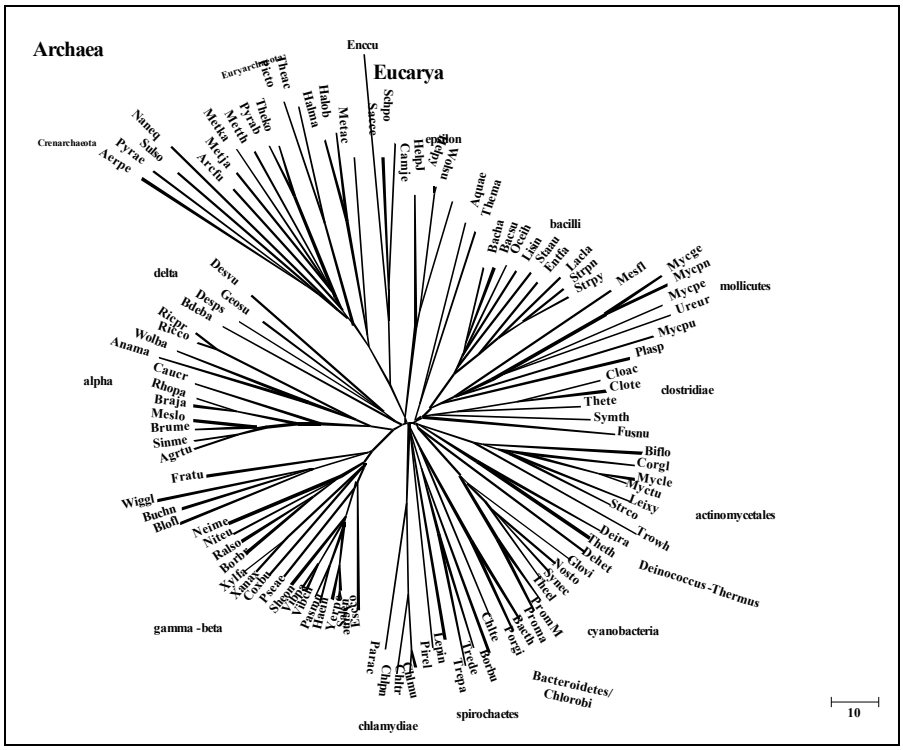

Figure 3. A genomic tree made with all families containing members belonging to at least 80 genomes.

Interestingly, and remarkably, the general topology obtained with the subset of at least 80 genomes (Fig.3) is very similar to the subset of at least 20 ones (Fig. 2). However, it can be seen that the increase of the number of genomes per family enlarged the distance separating the node common to the grouping of archaea and eukaryotes from bacteria and the distances between the main bacterial branches.

These family trees (Figs 2 and 3) display a topology where the positions of many groups of organisms, as well as that of their most external nodes is very similar to their accepted taxonomic distribution. We note however two differences: (i) the epsilon are not grouping with the other proteobacteria, (ii) the methanogen *Methanosarcina acetivorans* and the two halophiles *Halobacterium species NRC-1* and *Haloarcula marismortui*, are emerging at the basis of the archaeal subtree and are not grouped with the other euryarchaeota. This last grouping may reflect an adaptation to a non thermophilic life style and could be interpreted as cases of massive lateral transfers. Therefore, we tried to design another approach that might be less susceptible to such drawbacks.

4 Reconstructing a genomic tree using triplet approaches

Let's define a triplet of genomes G_A, G_B and G_C. In order to determine the relative distances separating each genome from the two others, the PAM distances separating each orthologue common to G_A, G_B and G_C were compared for each triplet. Table 1 shows, as an example, a set of five bacterial genomes : three proteobacteria, *Agrobacterium tumefaciens*, *Escherichia coli*, *Vibrio cholerae* and two firmicutes *Bacillus subtilis*, *Lactobacillus lactis*. For each triplet G_A, G_B and G_C, we determined which pair of orthologues was the closest for every triplet of orthologues. We further required that the difference between the two best pairs

of closest orthologues must be larger than 1%. Accordingly, the relative distances separating each genome from the two others will be estimated as the number of closest orthologues n_{AB}, n_{AC} and n_{BC}. We further used these numbers to compute a score for each pair of genomes. Two alternative approaches were tried, the "yes/no" one and the "proportion" one to determine the vote of the triplet of genomes, as detailed below.

G_A	G_B	G_C	Total[a]	$n_{1\%}$[b]	n_{AB}	$\%_{AB}$	n_{AC}	$\%_{AC}$	n_{BC}	$\%_{BC}$	yes/no vote
Agrtu	Lacla	Vibch	320	15	33	10.31	**198**	61.87	74	23.12	(Agrtu,Vibch)
Agrtu	Escco	Vibch	669	2	36	5.38	32	4.78	**599**	89.54	(Escco,Vibch)
Agrtu	Escco	Lacla	340	17	**211**	62.06	37	10.882	75	22.06	(Agrtu,Escco)
Agrtu	Bacsu	Vibch	433	19	76	17.55	**181**	41.80	157	36.26	(Agrtu,Vibch)
Agrtu	Bacsu	Lacla	374	7	48	12.83	13	3.48	**306**	81.82	(Bacsu,Lacla)
Agrtu	Bacsu	Escco	466	18	75	16.09	**216**	46.35	157	33.69	(Agrtu,Escco)
Escco	Lacla	Vibch	442	1	17	3.84	**412**	93.21	12	2.71	(Escco,Vibch)
Bacsu	Lacla	Vibch	405	4	**329**	81.23	59	14.57	13	3.21	(Bacsu,Lacla)
Bacsu	Escco	Vibch	587	2	24	4.09	26	4.43	**535**	91.14	(Escco,Vibch)
Bacsu	Escco	Lacla	449	6	77	17.15	**353**	78.62	13	2.89	(Bacsu,Lacla)

Table 1. Voting for the closest pair of genomes among the different ones for each triplet. [a] Total number of triplets of orthologues common to the triplet of genomes [b] Total of triplets of orthologues for which the difference between their respective pairwise distances are less than 1%.

4.1 The "yes/no" approach

Each triplet G_A, G_B and G_C voted for the pair of genomes whose number of closest orthologues was the largest. Moreover, votes were validated only if this number was at least 10% larger than the two other ones. Such non-voting triplets were found to be rather scarce, corresponding to less than 5.4% of the 209,934 triplets obtained in the case of the 109 studied genomes. First line of Table 1 shows that in the case of the triplet of genomes Agrtu, Lacla and Vibch, the highest number of triplets of common orthologues (198, in bold) was found for the pair (Agrtu,Vibch). The score for a pair of genomes was further computed as the total of votes this pair received. Since the vote for the pair (Agrtu,Vibch) was found two times, its score amounted to 2. Accordingly, a scoring matrix (Fig. 4A) was derived from Table 1 (last column on the right). Then, a distance matrix (Fig. 3B) was derived from this scoring matrix by subtracting each score from an arbitrary value defined as the largest score + 1.

	Agrtu	*Bacsu*	*Escco*	*Lacla*
Bacsu	0	-	-	-
Escco	1	0		-
Lacla	1	3	1	-
Vibch	2	0	2	0

A Scoring matrix

	Agrtu	*Bacsu*	*Escco*	*Lacla*
Bacsu	**4**			
Escco	**3**	**4**		
Lacla	**3**	**1**	**3**	
Vibch	**2**	**4**	**2**	**4**

Distance matrix **B**

Figure 4. Computing a scoring matrix and derivating a distance matrix.

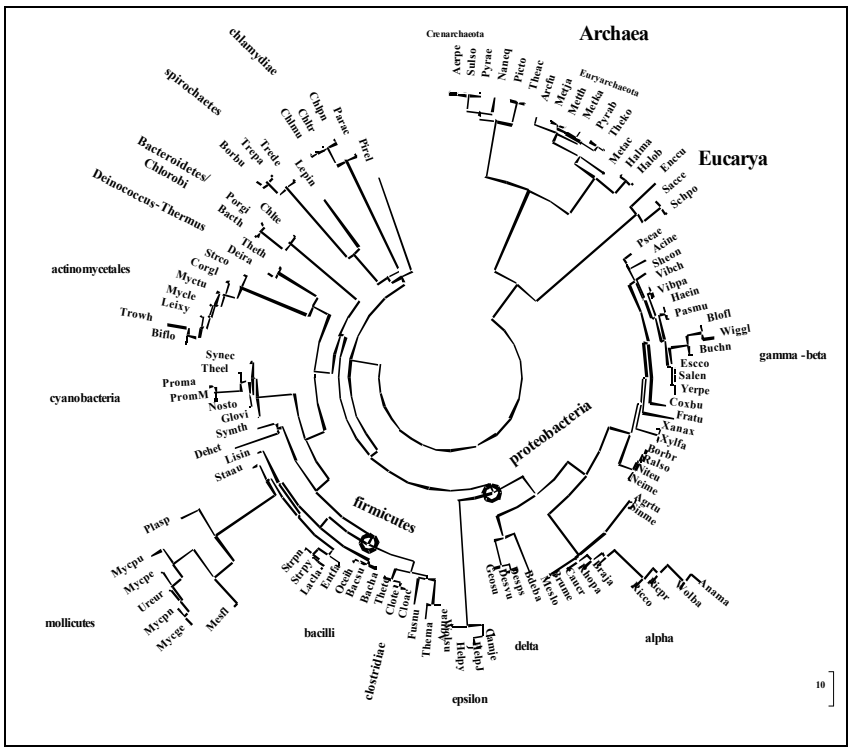

Figure 5. A genomic tree made with the method of triplets using the yes/no approach.

Finally, this distance matrix was used to reconstruct an evolutionary tree applying the neighbor-joining algorithm [16]. This yes/no approach applied to the whole set of triplets found in the case of the 109 studied genomes leads to the tree shown on Fig. 5.

4.2 The " proportion " approach

Here, the score for a pair G_AG_B was rated directly as an evolutionary distance as follows. First, the proportion of triplets where the PAM distance between the common orthologues was shorter for G_AG_B than for G_AG_C and G_BG_C was computed for every genome G_C. As it appears in the first line of Table 1, 61.87 % (shaded cell) of the triplets are in favour of the pair (Agrtu,Vibch). Then, the complement to 1 of the relative proportions (grouping genomes G_A and G_B) were added up for all triplets of genomes where G_A and G_B were involved (G_C being variable). For example, the score of the pair (Agrtu,Vibch) in Table 1 (shaded cells) was computed as shown below and directly used as the distance separating both genomes. Table 2 shows the distance matrix obtained for this set of five genomes.

Computing the score of the pair (Agrtu,Vibch) :

(1 – 0.6187) + (1 – 0.0478) + (1 – 0.418) = 1.915

	Agrtu	*Bacsu*	*Escco*	*Lacla*
Bacsu	**2.535**			
Escco	**1.862**	**2.451**		
Lacla	**2.753**	**0.583**	**2.712**	
Vibch	***1.915***	**2.447**	**0.261**	**2.710**

Table 2. Distance matrix obtained using the proportion approach.

From this distance matrix, a neighbor-joining tree [16] was derived. When applied to the 109 studied genomes this proportion approach gave the tree shown on Fig. 6.

5 Discussion

Determining the nature of deep prokaryotic relationships is one of the most fundamental issues (micro)biologists have to answer. If we assume (1) that a core of genes is shared between microorganisms as a result of vertical descent and (2) that we can still detect the old events of gene duplication and gene differentiation, it should be possible to reconstruct an evolutionary tree of this core. Several approaches have been already tried over the last ten years [4-10] in order to address this crucial point, with mixed success. Here we have described two new approaches that seem to give rather congruent trees and that display some deep nodes that appear to be reliable.

5.1 Congruent phylogenetic trees

The different genomic trees (Figs. 2, 3, 5 and 6) we obtained share many common features, independently of the method used, and their topology appears to agree with a large majority of the present taxonomy data. The three Domains defined by Woese et al [2] (i) form three clades that are well separated, (ii) the Archaea and Eucarya share a common branch after separating from Bacteria. The further examination of both prokaryotic Domains shows the following features.

In the Domain Archaea, the three phyla (Crenarchaeota, Euryarchaeota, Nanoarchaeota) are themselves well recognized in all trees. However, several discrepancies appear. (1) Except for the tree made with at least 80 genomes per family (Fig. 3), the two Thermoplasmata (*Picrophilus torridus*, *Thermoplasma acidophilum*) are found to branch far from the other Euryarchaeota. They are found to form a paraphyletic group with Crenarchaeota and Nanoarchaeota. This result seems to confirm previous observations that *T. acidophilum* and *P. torridus* share numerous genes with the crenarchaeons such as *Sulfolobus solfataricus* inhabiting the same environment [17,18]. (2) The halobacteria and *M. acetivorans* are emerging at the basis of the archaeal tree. This apparently aberrant position could be interpreted as reflecting well-known cases of massive lateral transfer between these species and bacteria [3,19]. The fact that the yes/no tree (Fig. 5) did not display this aberrant position would confirm that this methodological approach is less sensitive to lateral transfer than the other ones, despite the position of the Thermoplasmata.

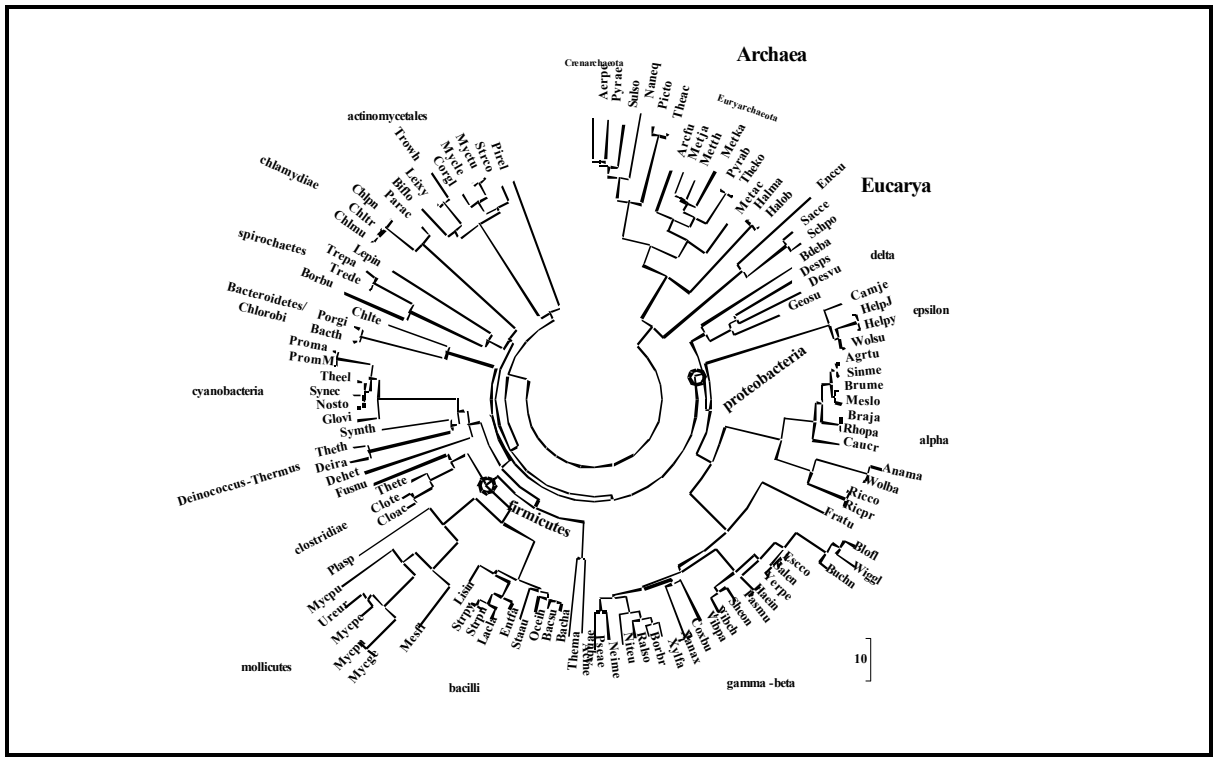

Figure 6. A genomic tree made with the method of triplets using the proportion approach.

In the case of the Domain Bacteria several important points are emerging when comparing the trees obtained using the different methods.

1. The great majority of the bacterial branches that are defined on a taxonomic basis are well recognized. Thus, in contrast with many approaches based on gene content (see [10]), these methods of tree reconstruction appear to be poorly sensitive to adverse effects brought in by genome sizes, lateral transfer or long branch attraction. For example, small genomes of symbionts such as *Buchnera* are clustering with their recent relatives such as *Escherichia coli*. Moreover, inside each phylum the relationships seems well conserved at the different taxonomic levels (including family). However, we note (as already observed by other groups [10]) a tendency to mix up the beta and the gamma proteobacteria.

2. The position of the two hyperthermophiles *Aquifex* and *Thermotoga* displays two interesting features : both hyperthermophiles are always branching (i) on a common node, (ii) but not at the basis of the bacterial tree, contrarily to what is observed in the 16s RNA tree and in several genomic trees [e.g. 10,20].

5.2 Reliable deep nodes of the Domain Bacteria

After collapsing their external branches it was possible to compare the simplified topologies of the four trees shown in Figs 2, 3, 5 and 6, respectively. Fig. 7 shows several groupings which are found constantly in the four trees as well as a few differences.

The main differences correspond to variable positions of *Pirellula* and of the Bacteroidetes/Chlorobi group. Note also that in the two families trees, the epsilon species (black arrow) are not grouping with the other proteobacteria (alpha, beta, gamma and delta). This aberrant position has been already observed in other trees

based on a gene content approach [10].

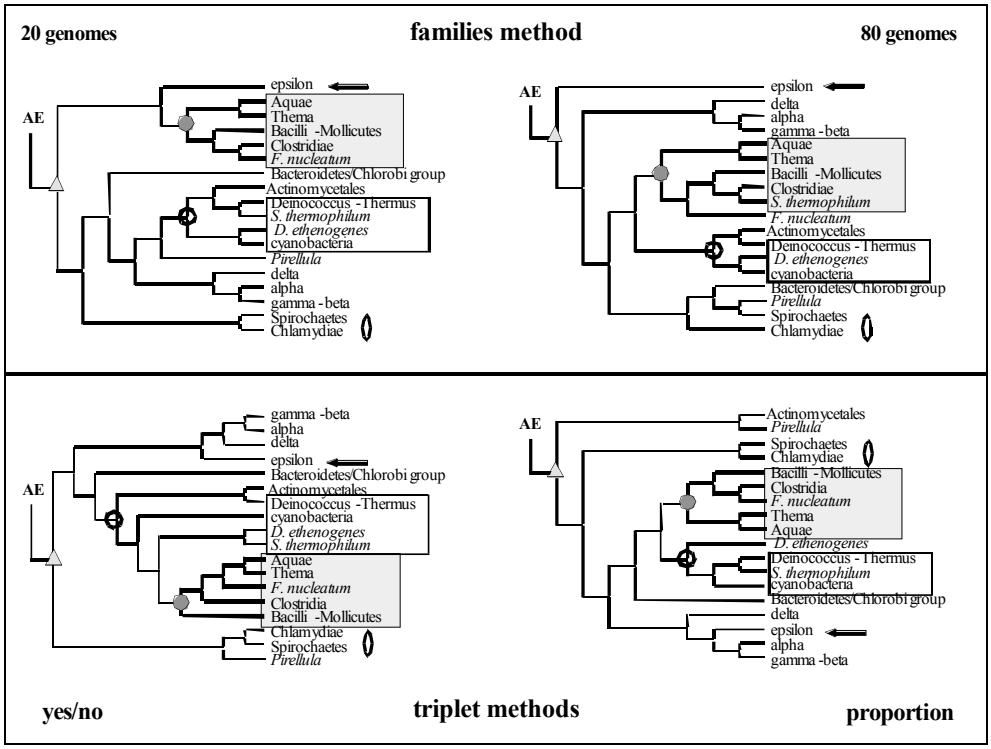

Figure 7. Comparing the genomic trees obtained with our new methods at the level of their simplified topologies. All external nodes have been collapsed. Bacterial trees are rooted (triangle) by the outgroup AE made of archaea and eukaryotes.

Moreover, we can note the remarkable following constant points about bacterial deep nodes :

o grouping of Firmicutes, *Fusobacterium* and the hyperthermophiles *Aquifex* and *Thermotoga* (shaded rectangle in Fig. 7). Note that the clustering of *Fusobacterium* (*Fusobacteria*) within the *Firmicutes* has been recently described in the case of a specific protein [21].

o grouping of Spirochaetes and Chlamydiae, generally as a clade (oval in Fig. 7). This grouping has been occasionally observed [10].

o remarkably, the actinobacterium *Symbiobacterium thermophilum* is never grouping with the other Actinobacteria (which are all belonging to the order of Actinomycetales). Moreover, it is often grouping with the cluster described below.

o there is a strong tendency to clustering the following branches : *Deinococcus-Thermus*, cyanobacteria and chloroflexi (*Dehalococcoides ethenogenes*). Moreover, either Actinomycetales and/or *S. thermophilum* are frequently associated to this cluster (open rectangle in Fig. 7) either in a monophyletic or a paraphyletic way.

The soundness of the results obtained with the new methods described herein have to be confirmed by increasing the number and the biodiversity of the studied species. However, the ability of these methods to disclose new relations at ancient nodes of the tree of life appear already promising.

Acknowledgements

This work is supported by the CNRS and the PPF "*Bioinformatique et Génomique*" of the Université Paris-Sud. We thank the *Centre de Ressources Informatiques* of the Université Paris-Sud for allowing the use of its PC Linux cluster system for intensive comparison of microbial proteomes.

References

[1]. Woese CR. Bacterial evolution. 1987 *Microbiol Rev*. 51:221-271
[2]. Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci U S A*; 87:4576-9
[3]. Doolittle WF. 1999. Phylogenetic classification and the universal tree. *Science*; 284:2124-9.
[4]. Snel, B. Bork P, Huynen MA 1999. Genome phylogeny based on gene content. *Nat. Genet*. 21, 108-110
[5]. Tekaia F, Lazcano A, Dujon B. 1999. The genomic tree as revealed from whole proteome comparisons. *Genome Res*. 9, 550-557
[6]. Korbel JO, Snel B, Huynen MA, Bork P. 2002. SHOT: a web server for the construction of genome phylogenies. *Trends in Genetics*. 18, 158-162
[7]. Clarke, G.D. et al. 2002. Inferring Genome Trees by Using a Filter To Eliminate Phylogenetically Discordant Sequences and a Distance Matrix Based on Mean Normalized BLASTP Scores. *J. Bacteriol*. 184: 2072-2080.
[8]. Daubin V, Gouy M, Perriere G 2001. Bacterial molecular phylogeny using supertree approach. Genome Inform Ser Workshop Genome Inform.;12:155-64
[9]. Deeds EJ, Hennessey H, Shakhnovich EI. 2005. Prokaryotic phylogenies inferred from protein structural domains. *Genome Res* 15:393-402
[10]. Wolf Y, et al. 2002. Genome trees and the tree of life. *Trends in Genetics* 18:472-479
[11]. Fitch WM. Distinguishing homologous from analogous proteins. *Syst Zool*. 1970 Jun;19(2):99-113.
[12]. Gonnet, G.H. et al. 1992. Exhaustive matching of the entire protein sequence database. *Science*, 256, 1443-1445
[13]. Le Bouder-Langevin S, Capron-Montaland I, De Rosa R and B. Labedan. (2002). A Strategy to retrieve the whole set of protein modules in microbial proteomes. *Genome Res* 12:1961-1973
[14]. Altschul, S.F. 1991. Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol*.; 219:555-65
[15]. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. 1999. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A*. 96, 2896-901
[16]. Saitou, M. L., and M. Nei. 1987. The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol. Evol*. 4, 406-425
[17]. Ruepp A, Graml W, Santos-Martinez ML, Koretke KK, Volker C, Mewes HW, Frishman D, Stocker S, Lupas AN, Baumeister W. 2000. The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature*.;407:508-13
[18]. Futterer O, Angelov A, Liesegang H, Gottschalk G, Schleper C, Schepers B, Dock C, Antranikian G, Liebl W. 2004. Genome sequence of *Picrophilus torridus* and its implications for life around pH 0. *Proc Natl Acad Sci USA*.101:9091-6
[19]. Deppenmeier U, Johann A, Hartsch T, Merkl R, Schmitz RA, Martinez-Arias R, Henne A, Wiezer A, Baumer S, Jacobi C, Bruggemann H, Lienard T, Christmann A, Bomeke M, Steckel S, Bhattacharyya A, Lykidis A, Overbeek R, Klenk HP, Gunsalus RP, Fritz HJ, Gottschalk G. 2002. The genome of *Methanosarcina mazei*: evidence for lateral gene transfer between bacteria and archaea. *J Mol Microbiol Biotechnol*. 4:453-61
[20]. Gophna U, W. Ford Doolittle, and R. L. Charlebois. 2005. Weighted Genome Trees: Refinements and Applications. *J. Bacteriol*. 187:1305-1316
[21]. Wolf M, T. Müller, T. Dandekar and J. D. Pollack. 2004. Phylogeny of Firmicutes with special reference to Mycoplasma (Mollicutes) as inferred from phosphoglycerate kinase amino acid sequence data. *Int J Syst Evol Microbiol* 54:871-875