

Bacterial Molecular Phylogeny Using Supertree Approach

Vincent Daubin

daubin@biomserv.univ-lyon1.fr

Manolo Gouy

gouy@biomserv.univ-lyon1.fr

Guy Perrière

perriere@biomserv.univ-lyon1.fr

Laboratoire de Biométrie et Biologie Évolutive, UMR CNRS 5558, Université Claude Bernard - Lyon 1, 43 bd. du 11 Novembre 1918, 69622 Villeurbanne Cedex, France

Abstract

It has been claimed that complete genome sequences would clarify phylogenetic relationships between organisms but, up to now, no satisfying approach has been proposed to use efficiently these data. For instance, if the coding of presence or absence of genes in complete genomes gives interesting results, it does not take into account the phylogenetic information contained in sequences and ignores hidden paralogy by using a similarity-based definition of orthology. Also, concatenation of sequences of different genes takes hardly in consideration the specific evolutionary rate of each gene. At last, building a consensus tree is strongly limited by the low number of genes shared among all organisms. Here, we use a new method based on supertree construction, which permits to cumulate in one supertree the information and statistical support of hundreds of trees from orthologous gene families and to build the phylogeny of 33 prokaryotes and four eukaryotes with completely sequenced genomes. This approach gives a robust supertree, which demonstrates that a phylogeny of prokaryotic species is conceivable and challenges the hypothesis of a thermophilic origin of bacteria and present-day life. The results are compatible with the hypothesis of a core of genes for which lateral transfers are rare but they raise doubts on the widely admitted “complexity hypothesis” which predicts that this core is mainly implicated in informational processes.

Keywords: prokaryotes, phylogeny, supertree, complete genomes, horizontal gene transfers

1 Introduction

Though it seems sensible to consider that genes remain associated in genomes for long periods in eukaryotes, recent data suggest that it is not the case in prokaryotes where a large number of horizontal transfers is believed to have occurred [32]. Methods using comparisons of base or codon composition in DNA reveals that bacterial genomes may have up to 17% of their genes of alien origin, with only few of them identifiable as mobile elements. Yet, unexpected sequence patterns may not be proofs of alien origin. Such kind of evidence assumes that bacteria and archaea have an homogeneous base composition along their chromosome and requires to fix a non-objective threshold of “normality”. An objective proof of alien origin should be given by phylogenetic analysis. Nevertheless, this raises other problems such as reconstruction artifacts and though phylogeneticists steadily warn against these problems [4, 34], the difficulty to obtain congruent gene phylogenies is often seen as a result of lateral exchanges. Hence, the prokaryotic world is now often seen as a “genome space” [4] in which horizontal transfer between organisms appear to be the rule. However, transfers probably do not concern every kind of genes in the same way. For instance, it has been suggested that genes having much macromolecular interactions are less likely to be transferred [21]. One could thus imagine that a core of genes remains more or less stable through evolution. If so, a phylogeny of bacterial species

proportionally to the approximate bootstrap probability associated to each internal branch. Because bootstrap values may not be linearly related to the number of sites supporting a node, we tested different coding schemes, with linear and non-linear relation of weighting to bootstrap values. The supertree method appears to be weakly sensitive to the different coding schemes and tends to give exactly the same topology. We finally decided to weight linearly all bootstrap values over 70%. The matrices obtained are concatenated into a supermatrix in which species absent from a gene family are encoded as unknown state (?). The supertree is calculated on the supermatrix using Neighbour-Joining from observed divergences under the pairwise gap removal option of CLUSTALW [19]. We have empirically found that this method which is much faster, gives the same result as parsimony since every character is informative in the matrix. Bootstrap values of the supertree were computed by re-sampling 500 times among orthologous gene families. All the data used to build the trees as well as all supertrees mentioned in this article are available at <ftp://pbil.univ-lyon1.fr/pub/datasets/GIW2001>.

3 Results

3.1 Building a Reliable Supertree

It has been shown that phylogenetic methods are highly sensitive to the number of studied taxa, small sets being subject to strong reconstruction artifacts [25]. To reduce such problems, supertrees were calculated only from our ML trees built with families encompassing more than a minimum number of species (Table 1). The supertree topology was actually affected by this threshold criterion. Supertrees at each threshold were compared with each other and a conserved topology emerged for thresholds between 15 and 19 species (which corresponds to sets of 196 to 130 trees). Only bootstrap values were different.

It is worth noting that the topology of the archaeal part of the tree is less stable and statistically robust than the bacterial part, which remains stable for thresholds between 12 and 24 (*i.e.* 249 to 75 trees). This means that there is a supertree that is relatively robust to addition or withdrawal of orthologous gene families, especially for bacteria, which are the most, represented species. This supertree (Fig. 2) is thus a compromise between the quantity (*i.e.*, the number of orthologous gene families) and the quality (reduced by artifacts acting on small families) of the information.

3.2 The Phylogeny of Bacteria

The supertree presents an interesting topology particularly for bacteria. Some well known groups are monophyletic and well supported (low G+C Gram-positives, high G+C Gram-positives, Proteobacteria, Chlamydiales and Spirochaetes). However, there are some substantial differences with rRNA phylogenies which place hyperthermophilic and radioresistant bacteria very close to the root [45]. The supertree gives no evidence for such early emergence of these groups and tend to give them positions close to mesophilic bacteria. Instead, the basal position in the bacterial tree is occupied by vertebrate parasites (Chlamydiales and Spirochaetes). Though not always supported by bootstrap, this position remains robust to threshold variation, which may be an indication of its evolutionary significance. The archaea have a less supported topology which may result from the low number of species sampled and perhaps also from to the low quality of the predicted sequences of *Aeropyrum pernix* [6].

3.3 Informational and Operational Genes

Supertrees separating informational genes (implicated in replication, transcription, translation and related processes) and operational genes [21] were built. With the same threshold selection, the operational supertree appears to be very similar to the global supertree with a tendency for Chlamydiales and Spirochaetes to cluster together at the basis of bacteria. This peculiar topology of the bacteria

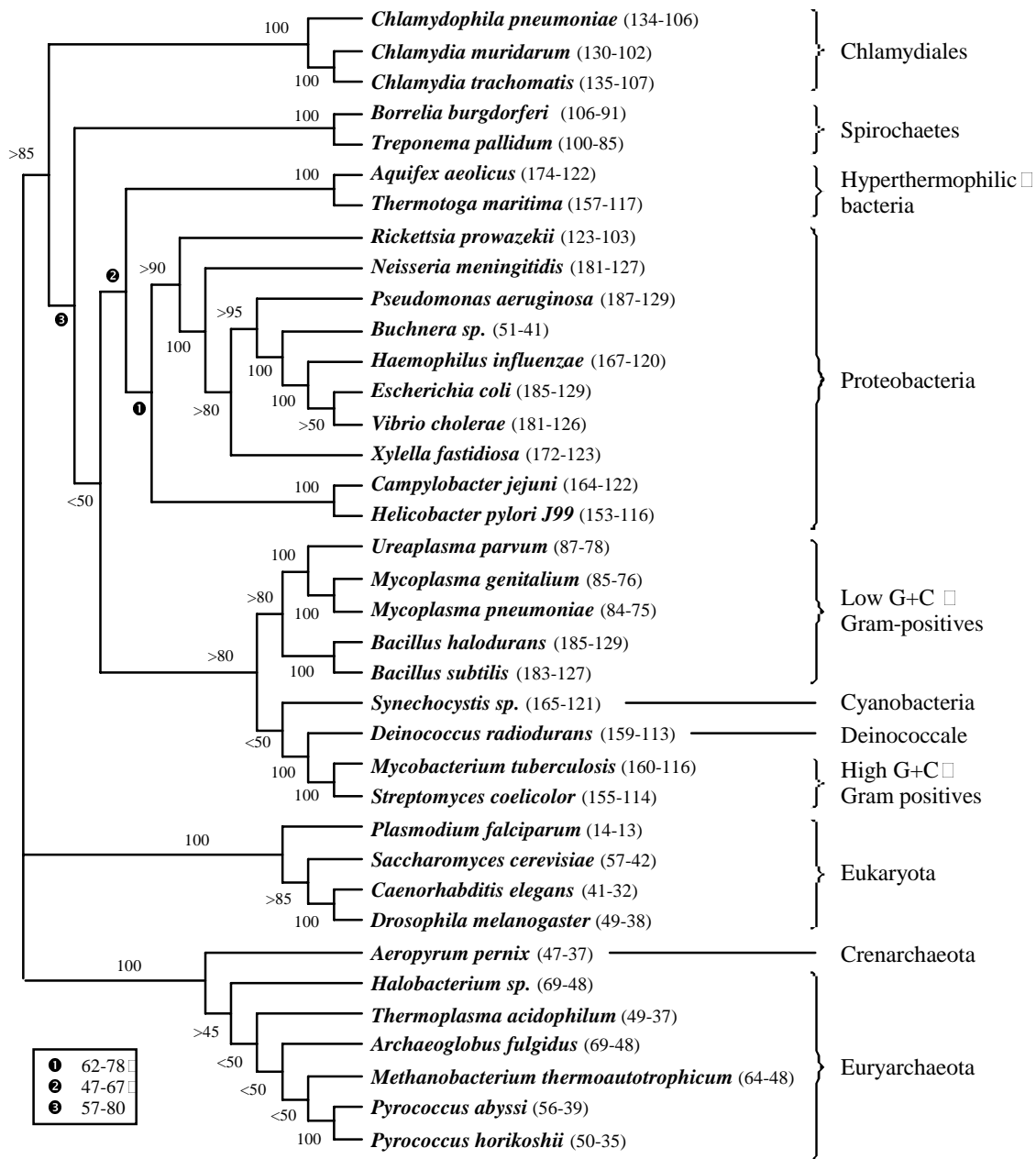


Figure 2: Supertree of 37 species. This supertree represents the stable topology obtained with families encompassing a minimum number of species of 15 to 19. This supertree is a cladogram since the coding scheme used does not produce distances that are directly related to genetic divergence. Bootstrap values (obtained from 500 replicates) are given next to branches. Bootstrap values, which were moderately variable to threshold, are defined by upper or lower bounds. The three branches which fluctuates more drastically are defined by a number for which the minimum and maximum values are given at the bottom of the figure. The number of families in which each species is represented is shown between parentheses for the 15 and 19 thresholds respectively.

tree remains stable for threshold values between 15 and 24. The informational supertree give a less stable topology but remains in agreement with the global supertree as far as statistically supported branches are concerned. These results are not consistent with those obtained by Jain *et al.* [21] which detected more difficulty for operational genes to give congruent phylogenetic information.

3.4 Other Supertrees

If some positions of bacterial species in the supertree are artefactuals, they are likely to be sensitive to the presence or absence of outgroups [37]. To test whether the topology of the bacterial part of the tree was dependent on the presence of the outgroups, we built a supertree with ML trees made from the bacterial sequences of the orthologous gene families. This bacterial supertree gives exactly the same stable topology as the global supertree for a large range of threshold values (9 to 22 species that corresponds to 281 to 73 orthologous gene families). Remarkably, most groups have a higher bootstrap support in the bacterial than in the global supertree. This happens with Proteobacteria (bootstrap values between 66 and 96) and with the *Thermotoga-Aquifex*-Proteobacteria clade (bootstrap values between 68 and 95). The branch separating Spirochaetes and Chlamydiales from all other bacteria has a significant bootstrap support for threshold values 17 to 22. The same experiments, made with Archaea, yielded trees that differed from global supertree, but without significantly higher bootstrap support.

Table 1: Number of orthologous gene families with respect to the number of taxa they contain. The number of species defines a minimum threshold for which a supertree is built. Sp. = number of species considered; Fam. = number of families; Trees = number of trees. The number of informational and operational families or trees are given between parentheses. The portion from 15 to 19 corresponds to the thresholds that give the same supertree topology.

Sp.	Fam.	Trees	Sp.	Fam.	Trees
7	67 (10/57)	459 (74/385)	23	7 (0/7)	82 (36/46)
8	36 (2/34)	392 (64/328)	24	14 (4/10)	75 (36/39)
9	42 (5/37)	356 (62/294)	25	20 (10/10)	61 (32/29)
10	35 (8/27)	314 (57/257)	26	11 (6/5)	41 (22/19)
11	30 (2/28)	279 (49/230)	27	6 (1/5)	30 (16/14)
12	23 (0/23)	249 (47/202)	28	2 (2/0)	24 (15/9)
13	16 (0/16)	226 (47/179)	29	0	22 (13/9)
14	14 (0/14)	210 (47/163)	30	2 (0/2)	22 (13/9)
15	18 (2/16)	196 (47/149)	31	2 (0/2)	20 (13/7)
16	18 (0/18)	178 (45/133)	32	1 (0/1)	8 (13/5)
17	17 (0/17)	160 (45/115)	33	2 (1/1)	17 (13/4)
18	13 (0/13)	143 (45/98)	34	2 (1/1)	15 (12/3)
19	13 (2/11)	130 (45/85)	35	7 (5/2)	13 (11/2)
20	12 (3/9)	117 (43/74)	36	3 (3/0)	6 (6/0)
21	14 (2/12)	105 (40/65)	37	3 (3/0)	3 (3/0)
22	9 (2/7)	91 (38/53)			

4 Discussion

4.1 Supertree Compared to Other Genome Trees

The existence of such a supertree, its stability to threshold variation and the support of several of its key nodes by significant bootstrap values is an important result since it has been claimed that, because of extremely frequent horizontal transfers, no species phylogeny could be found [4, 7, 41]. It seems likely that, even with several lateral transfers and long-branch artefacts affecting each gene tree, subgroups of trees may share fragmentary information on species phylogeny that emerges through supertree computation. Some nodes appear to be only moderately supported by bootstrap, but remain stable under variation of the threshold value, especially for bacteria for which the topology stays the same through addition of more than 150 trees. This stability may be a good indication of which topology is the most likely. Moreover, the congruence of supertrees made with completely distinct sets of genes (informational and operational) is very striking and constitutes a strong argument for

the method.

Another way to compute genome trees has been proposed [41]. It is based on concatenation of sequences from different genes. This approach has been shown to be extremely sensitive to addition or withdrawal of genes for a given number of species and Teichmann et al. concluded for an absence of phylogenetic signal in bacterial genes. The problem of this method may be that it does not consider the diversity of evolutionary rates among genes. Indeed, if two genes are informative at different levels of the phylogeny, their concatenation will attenuate this information rather than bring it out. In contrast to the concatenation method, the supertree method allows to take evolutionary rates in account and to reduce the impact of unresolved phylogeny by bootstrap weighting. The result is a tree that is remarkably stable to orthologous gene family sampling.

Jain *et al.* [21] suggested that informational genes, which interact with many macromolecular partners are less frequently transferred than operational genes. The building of two independent supertrees for these two categories of genes gives no evidence for such a difference. The supertree seems to be more affected by threshold variation than by functional separation. This suggests another explanation: the results of Jain *et al.* rest on rather small trees (in term of number of species species), which we know are more likely subject to phylogeny artefacts. The differences brought to light by their work could be tendencies of operational genes to be less conserved than informational ones, leading to difficulty for inferring phylogeny with them when species sampling is small. Moreover, they used a similarity based definition of orthology that may raise problems in the case of multigene families. This definition of orthology has also been used in several gene-content based trees [10, 42] that give results that may be due to unidentified paralogies.

4.2 Which Artifacts May Affect the Supertree

The sample of completely sequenced bacterial genomes is currently strongly biased toward species of medical interest. Hence, the supertree contains many parasites from which certain are endocellular (such as *Rickettsia* and *Chlamydia*) and display particular evolutionary patterns. Based only on topology and statistical support, our method of supertree is predicted to be sensitive to systematical artifacts of reconstruction. Nevertheless, though systematical bias exist, the artefacts are not likely to gather systematically the same species, depending on the species sampling, which may, by definition, be different between gene families in our approach. In this case, even weak congruent information due to phylogenetic signal would be stronger than conflicting artefactual information's. For instance, *Mycoplasma* species have a very low genomic G+C content (only 25% for *Ureaplasma parvum* and 32% for *Mycoplasma pneumoniae*) and are known to have a very reduced genome and fast evolutionary rate [31]. Such a fast rate is probably related to the absence of mut genes and of the *lexA* repressor of the SOS system in these genomes [36]. This is probably why this species tend to have very basal position in several single gene [17, 23] and multiple gene [18, 26, 41] phylogeny. Therefore, the fact that *Mycoplasma* species are unambiguously grouped with *Bacillus* in the supertree suggests that the supertree method is robust against biases related to G+C content and evolutionary rates. The same remarks can be made for *Helicobacter pylori*, which shows a high level of genetic variation between strains [44]. Remarkably, no such defects in DNA repair systems have been found in Chlamydiales [39] and Spirochaetes [24, 40].

Archaea and Eukaryotes are very distant outgroup organisms to Bacteria. Therefore, the problem arises of whether the topology of the bacterial part of the tree given in Fig. 2 results from long branch attraction artefact. The complete identity between supertrees computed without and with archaeal and eukaryotic outgroups suggests that relations between bacterial phyla in the tree is not determined by the presence of distant outgroup sequences.

4.3 The Supertree of Life: New Insights into Bacterial History

The topology of the supertree strongly supports the monophyly of the three domains of life (Bacteria, Archaea and Eukaryota). The phylogeny of Proteobacteria appears to be well resolved at this level and is in agreement with the rRNA phylogeny. Their monophyly (including *H. pylori* and *Campilobacter jejunii*) is well supported and this last result is particularly valuable because it has rarely been found with genomic tree methods [26, 41, 42]. Equally interesting is the position of the thermophilic bacteria *Aquifex aeolicus* and *Thermotoga maritima* that are strongly grouped together but not at the base of the bacteria, as placed by small subunit (SSU) ribosomal RNA phylogeny [45]. The position of these organisms challenges the hypothesis of a hyperthermophilic origin of life [2, 45]. Although their branching with Proteobacteria stays at the edge of being significant in the global supertree, it is one of the most stable nodes under variation of threshold values. Moreover, it presents significant bootstrap supports for several threshold values in the bacterial supertree. Thus, the genomic supertree brings no evidence for an early divergence of thermophilic lineages and is more consistent with a mesophilic Last Universal Common Ancestor (LUCA) [11, 13]. This view interprets the early emergence of these lineages in SSU rRNA trees as a reconstruction artifact [11, 23] and suggests that *Thermotoga* and *Aquifex* have been secondarily adapted to high temperature [11, 29]. Several studies have already reported a clustering of these bacteria with Proteobacteria [23] or Gram-positives [15, 43].

The monophyly of low G+C Gram-positives (including *Bacillus* and *Mycoplasma*) on one side and of high G+C Gram-positives on the other side appear to be very robust, although the strongly supported position of *Deinococcus radiodurans* claims for the polyphyly of the Gram-positives. This position is very striking since *Deinococcus* is usually considered to have a much more basal position in the bacteria [45]. Huang and Ito [20] have already noted such a position, close to Gram-positives, with a DNA polymerase C phylogeny. This radioresistant bacterium was first identified as a Gram-positive [30], but has been excluded from this group after molecular phylogenetic analysis [45]. *Deinococcus* is now considered as a close parent of *Thermus aquaticus*, which is a Gram-negative thermophilic bacterium. Another point is the close relation of *Synechocystis* sp. with the Gram-positives, which has already been noted [12, 45].

The basal position of Chlamydiales seems to have some level of support, especially in comparison to other nodes at this level of the phylogeny. The bootstrap of the deep nodes of the supertree are indeed rather low and may reveal the attenuation of phylogenetic signal through evolution and the increase of probability of lateral transfers with divergence time. Nevertheless, the robustness of these deep nodes to threshold variation is an indication that they may have an evolutionary significance. This position needs to be confirmed, by including in the supertree species that are reasonably close to Chlamydiales. Particularly, the supertree of operational genes suggests a clade including Chlamydiales and Spirochaetes emerging at the basis of the tree. The fact that these bacteria are vertebrate parasites does not preclude their basal position since they may possess close free living relatives.

Though displaying rather low bootstrap values, the archaeal part of the tree supports the monophyly of Euryarchaeota as widely admitted. It gives also no evidence for a polyphyly of archaea since they appear as a strongly supported clade. This part of the supertree though, appears to be very poorly resolved in comparison to the other two domains. Its stability and bootstrap supports are low. Therefore, this topology must be taken carefully. Our experience of supertrees suggests that this problem will be resolved by increasing the number of archaeal species in the study. Particularly, the fact that Crenarchaeota are represented only by *A. pernix* for which sequence prediction quality has been discussed reduces confidence in this part of the tree [6].

4.4 Horizontal Transfers: “Genome Space” or Core of Genes?

The level of resolution of the supertree is in strong disagreement with the “genome space” [4] vision of the prokaryotic world that predicts a “star phylogeny”. One could argue that grouping of species in the supertree would only reflect the frequency of gene exchanges between these species. This interpretation

can be excluded for two reasons. First, the supertree method would then not be expected to give a tree topology radically different from gene-content based trees [26, 38, 42] which are phenetic trees. Second, the supertree would not be expected to be sensitive to the number of taxa per family. It is worth noting that we have made a particularly stringent selection on trees to build the supertree. In particular, we took a phylogenetic definition of orthology and not a similarity-based definition as is often the case for practical reasons. Hence, we have excluded from analysis all gene trees where a species was represented more than once. This selection allowed us to make absolutely no a priori assumption on the topology of the tree and to reduce the probability of taking into account hidden paralogies. The supertree reveals that the class of genes so defined contains congruent information on the phylogeny of prokaryotic life. This pleads for a vision of prokaryotic evolution where a "core" of genes tends to remain stable through evolution [9, 38]. The variation of the supertree topology when small values of the threshold are used reveals that genes less widely spread among organisms give no congruent phylogenetic information. It may be because this core is made of widely represented genes, but it may also be due to failure of phylogenetic methods to extract information from small orthologous gene families.

Although our results support the core hypothesis, the separation of operational and informational genes is in contradiction with the hypothesis of Jain *et al.* [21], since these supertrees are very similar and equally resolved. Though there may certainly be operational genes having many macromolecular interactions, one could also turn round the complexity hypothesis : genes having many macromolecular interactions tend to be more conserved through evolution and require hence less *de novo* adaptations when horizontally transferred. More and more examples are now found that show that horizontal transfers may concern informational genes [1, 5, 46]. Conversely, the interpretation of incongruent phylogeny of operational genes as horizontal transfer events may have been overemphasised [14]. We argue that the most common problem in telling prokaryotic story may not be horizontal transfers, as widely admitted, but phylogenetic artefacts and perhaps hidden paralogies, which are considered *a priori* as unlikely events.

Very few molecular markers give at the same time arguments for the monophyly of Proteobacteria and the monophyly of low G+C Gram-positives probably because of common long branch artefacts affecting this species. This problem appears to be even stronger when combining data from different genes [16, 18, 26, 41, 42]. The present supertree method appears as a good tool to infer phylogeny since it does take into account molecular phylogenetic information, is insensitive to phenetic characters such as gene content and shows great robustness to gene sampling. Since it is independent to artefacts acting on single genes, this majority supertree may be used as a reference for detecting horizontal transfers with phylogenetic methods, and to quantify the importance of this phenomenon in prokaryotic evolution.

References

- [1] Asai, T., Zaporozjets, D., Squires, C., and Squires, C.L., An *Escherichia coli* strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria, *Proc. Natl. Acad. Sci. USA*, 96:1971–1976, 1999.
- [2] Barns, S.M., Delwiche, C.F., Palmer, J.D., and Pace, N.R., Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences, *Proc. Natl. Acad. Sci. USA*, 93:9188–9193, 1996.
- [3] Baum, B.R., Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees, *Taxon*, 41:3–10, 1992.
- [4] Bellgard, M.I., Itoh, T., Watanabe, H., Imanishi, T., and Gojobori, T., Dynamic evolution of genomes and the concept of genome space, *Ann. NY Acad. Sci.*, 870:293–300, 1999.
- [5] Brochier, C., Philippe, H., and Moreira, D., The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome, *Trends Genet.*, 16:529–533, 2000.

- [6] Cambillau, C. and Claverie, J.M., Structural and genomic correlates of hyperthermostability, *J. Biol. Chem.*, 275:32383–32386, 2000.
- [7] Doolittle, W.F., Lateral genomics, *Trends Cell Biol.*, 9(12):M5–8, 1999.
- [8] Eisen, J.A., Assessing evolutionary relationships among microbes from whole-genome analysis, *Curr. Opin. Microbiol.*, 3:475–480, 2000.
- [9] Eisen, J.A., Horizontal gene transfer among microbial genomes: new insights from complete genome analysis, *Curr. Opin. Genet. Dev.*, 10:606–11, 2000.
- [10] Fitz-Gibbon, S.T. and House, C.H., Whole genome-based phylogenetic analysis of free-living microorganisms, *Nucleic Acids Res.*, 27:4218–4222, 1999.
- [11] Forterre, P., A hot topic: the origin of hyperthermophiles, *Cell*, 85:789–792, 1996.
- [12] Galtier, N. and Gouy, M., Molecular phylogeny of Eubacteria: a new multiple tree analysis method applied to 15 sequence data sets questions the monophyly of Gram-positive bacteria, *Res. Microbiol.*, 145:531–541, 1994.
- [13] Galtier, N., Tourasse, N., and Gouy, M., A nonhyperthermophilic common ancestor to extant life forms, *Science*, 283:220–221, 1999.
- [14] Glansdorff, N., About the last common ancestor, the universal life-tree and lateral gene transfer: a reappraisal, *Mol. Microbiol.*, 38:177–185, 2000.
- [15] Gribaldo, S., Lumia, V., Creti, R., de Macario, E.C., Sanangelantoni, A., and Cammarano, P., Discontinuous occurrence of the hsp70 (*dnaK*) gene among Archaea and sequence features of HSP70 suggest a novel outlook on phylogenies inferred from this protein, *J. Bacteriol.*, 181:434–443, 1999.
- [16] Grishin, N.V., Wolf, Y.I., and Koonin, E.V., From complete genomes to measures of substitution rate variability within and between proteins, *Genome Res.*, 10:991–1000, 2000.
- [17] Gupta, R.S., Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaeobacteria, eubacteria, and eukaryotes, *Microbiol. Mol. Biol. Rev.*, 62:1435–1491, 1998.
- [18] Hansmann, S. and Martin, W., Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis, *Int. J. Syst. Evol. Microbiol.*, 50:1655–1663, 2000.
- [19] Higgins, D.G., Thompson, J.D., and Gibson, T.J., Using CLUSTAL for multiple sequence alignments, *Methods Enzymol.*, 266:383–402, 1996.
- [20] Huang, Y.-P. and Ito, J., DNA polymerase C of the thermophilic bacterium *Thermus aquaticus*: classification and phylogenetic analysis of the family C DNA polymerases, *J. Mol. Evol.*, 48:756–769, 1999.
- [21] Jain, R., Rivera, M.C., and Lake, J.A., Horizontal gene transfer among genomes: the complexity hypothesis, *Proc. Natl. Acad. Sci. USA*, 96:3801–3806, 1999.
- [22] Kishino, H., Miyata, T., and Hasegawa, M., Maximum likelihood inference of protein phylogeny and the origin of chloroplasts, *J. Mol. Evol.*, 30:151–160, 1990.
- [23] Klenk, H.P., Meier, T.D., Durovic, P., Schwass, V., Lottspeich, F., Dennis, P.P., and Zillig, W., RNA polymerase of *Aquifex pyrophilus*: implications for the evolution of the bacterial *rpoBC* operon and extremely thermophilic bacteria, *J. Mol. Evol.*, 48:528–541, 1999.
- [24] Lafay, B., Lloyd, A.T., McLean, M.J., Devine, K.M., Sharp, P.M., and Wolfe, K.H., Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases, *Nucleic Acids Res.*, 27:1642–1649, 1999.
- [25] Lecointre, G., Philippe, H., Van Le, H.L., and Le Guyader, H., Species sampling has a major impact on phylogenetic inference, *Mol. Phyl. Evol.*, 2:205–224, 1993.

- [26] Lin, J. and Gerstein, M., Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels, *Genome Res.*, 10:808–818, 2000.
- [27] Lyons-Weiler, J., Hoelzer, G.A., and Tausch, R.J., Relative apparent synapomorphy analysis (RASA). I: The statistical measurement of phylogenetic signal, *Mol. Biol. Evol.*, 13:749–757, 1996.
- [28] Martin, W. and Muller, M., The hydrogen hypothesis for the first eukaryote, *Nature*, 392:37–41, 1998.
- [29] Miller, S.L., and Lazcano, A., The origin of life - did it occur at high temperatures? *J. Mol. Evol.*, 41:689–692, 1995.
- [30] Murray, R.G.E., Family II. *Deinococcaceae* Brooks and Murray 1981. 356VP, *Bergey's Manual of Systematic Bacteriology*, Williams and Wilkins, 1035–1043, 1986.
- [31] Ochman, H., Elwyn, S., and Moran, N.A., Calibrating bacterial evolution, *Proc. Natl. Acad. Sci. USA*, 96:12638–12643, 1999.
- [32] Ochman, H., Lawrence, J.G., and Groisman, E.A., Lateral gene transfer and the nature of bacterial innovation, *Nature*, 405:299–304, 2000.
- [33] Perrière, G., Duret, L., and Gouy, M., HOBACGEN: database system for comparative genomics in bacteria, *Genome Res.*, 10:379–385, 2000.
- [34] Philippe, H. and Laurent, J., How good are deep phylogenetic trees? *Curr. Opin. Genet. Dev.*, 8:616–623, 1998.
- [35] Ragan, M.A., Phylogenetic inference based on matrix representation of trees, *Mol. Phyl. Evol.*, 1:53–58, 1992.
- [36] Razin, S., Yogev, D., and Naot, Y., Molecular biology and pathogenicity of mycoplasmas, *Microbiol. Mol. Biol. Rev.*, 62:1094–1156, 1998.
- [37] Robinson, M., Gouy, M., Gautier, C., and Mouchiroud, D., Sensitivity of the relative-rate test to taxonomic sampling, *Mol. Biol. Evol.*, 15:1091–1098, 1998.
- [38] Snel, B., Bork, P., and Huynen, M.A., Genome phylogeny based on gene content, *Nature Genet.*, 21:108–110, 1999.
- [39] Shirai, M., Hirakawa, H., Kimoto, M., Tabuchi, M., Kishi, F., Ouchi, K., Shiba, T., Ishii, K., Hattori, M., Kuhara, S., and Nakazawa, T., Comparison of whole genome sequences of *Chlamydia pneumoniae* J138 from Japan and CWL029 from USA, *Nucleic Acids Res.*, 28:2311–2314, 2000.
- [40] Subramanian, G., Koonin, E.V., and Aravind, L., Comparative genome analysis of the pathogenic spirochetes *Borrelia burgdorferi* and *Treponema pallidum*, *Infect. Immun.*, 68:1633–1648, 2000.
- [41] Teichmann, S.A. and Mitchison, G., Is there a phylogenetic signal in prokaryote proteins? *J. Mol. Evol.*, 49:98–107, 1999.
- [42] Tekaia, F., Lazcano, A., and Dujon, B., The genomic tree as revealed from whole proteome comparisons, *Genome Res.*, 9:550–557, 1999.
- [43] Tiboni, O., Cammarano, P., and Sanangelantoni, A.M., Cloning and sequencing of the gene encoding glutamine synthetase I from the archaeum *Pyrococcus woesei*: anomalous phylogenies inferred from analysis of archaeal and bacterial glutamine synthetase I sequences, *J. Bacteriol.*, 175:2961–2969, 1993.
- [44] Wang, G., Humayun, M.Z., Taylor, D.E., Mutation as an origin of genetic variability in *Helicobacter pylori*, *Trends Microbiol.*, 7:488–493, 1999.
- [45] Woese, C., Bacterial evolution, *Microbiol. Rev.*, 51:221–271, 1987.
- [46] Yap, W.H., Zhang, Z., Wang, Y., Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon, *J. Bacteriol.*, 181:5201–5209, 1999.