Genetics and Molecular Biology, 30, 1, 270-276 (2007)

www.sbg.org.br

Research Article

A phylogenomic appraisal of the evolutionary relationship of Mycoplasmas

Karla S.C. Yotoko and Sandro L. Bonatto

Centro de Biologia Genômica e Molecular, Faculdade de Biociências,
Pontifícia Universidade Católica do Rio Grande do Sul, 90619-900 Porto Alegre, RS, Brazil.

Abstract

Several genomes of mycoplasmas have been sequenced and here we tried to retrieve the evolutionary relationships of nine species using a phylogenomic approach. Several methods were used to build phylogenetic trees based on protein sequence information, gene-order, and gene-content. We also utilized datasets composed of individual and concatenated sets of orthologous proteins, as well as with reduced unreliable alignment regions. Most of our results converge to a single topology, except for the trees built with both the maximum parsimony method and with the gene-order dataset. The gene-content dataset presented trees consistent with most nodes of the convergent tree, but in the gene-order dataset most internal branches were clearly saturated and unreliable. The topological difference between the trees obtained by the diverse methods could not be explained by regions with unreliable alignments or attributed to horizontal gene transfer among the genomes. It is possible that the incongruence between the methods could be associated with their differential sensibility in relation to certain evolutionary factors. Further analysis using other empirical genomic datasets would be necessary in order to better understand the basis of such conflicts.

Key words: phylogenomic, Mycoplasma, phylogenetic methods, maximum likelihood, maximum parsimony.

Received: April 12, 2006; Accepted: October 5, 2006.

Introduction

Mycoplasma bacteria comprise a large number of obligatory parasites of a wide spectrum of hosts that includes animals (vertebrates and arthropods) and plants. These species are distinguished phenotypically from other bacteria by their very small size and lack of a cell wall. Together with the extremely reduced size of the genomes of some species, this led several authors to consider them as the smallest self-replicating organisms (reviewed in Razin *et al*. 1998). This genome simplicity, associated with the losses caused to human health and livestock production by some species, stimulated the sequencing of the whole genomes of several species. In fact, the second organism whose genome was entirely sequenced was *Mycoplasma genitalium* (Fraser *et al*. 1995). Currently, there are 12 complete genomes of mycoplasmas available in GenBank (including *Ureaplasma*), three of those (two strains of *M. hyponeumonieae* and one strain of *M. synoviae*, Vasconcelos *et al*. 2005) were sequenced by the Brazilian Genome Programs (Southern Network for Genome Analysis and Brazilian National Genome Project Consortium).

Send correspondence to Sandro L. Bonatto. Centro de Biologia Genômica e Molecular, Faculdade de Biociências, Pontifícia Universidade Católica do Rio Grande do Sul, 90619-900 Porto Alegre, RS, Brazil. E-mail: slbonatto@pucrs.br.

The availability of whole genomes allows the construction of phylogenetic trees based on a large set of genes, which supposedly may reveal, with an elevated probability, the long sought "correct tree", in contrast with the conflicting phylogenies obtained with the use of individual genes separately (Nei and Kumar 2000). Rokas *et al*. (2003) concluded, from a phylogenomic study in yeast, that analyses of more than a hundred concatenated genes yielded a single, fully resolved species tree with maximum support and that such an approach may resolve incongruence in phylogenies. Eisen and Fraser (2003) suggested that combining in a phylogenomic approach the perspectives of genomic and evolutionary studies would greatly help the construction of the true tree of life.

Notwithstanding these optimist scenarios, phylogenetic trees based on genomic-scale analysis are not free of problems, such as the identification of the orthologous sequences, which in practice may be very difficult in some taxonomic groups (Bapteste *et al*. 2005; Hughes *et al*. 2005); and the occurrence of horizontal gene transfers (HGT) among genomes, whose incidence in nature and consequence for tree estimation are still under great debate (*e.g.*, Gogarten and Townsend 2005 *vs.* Ochman *et al*. 2005). Finally, the limitations of the methods of tree building currently used in phylogenomics are largely unknown,

producing another constraint to correctly assemble the tree of life (Delsuc *et al*. 2005).

In this paper, we revisited the original phylogenomic analysis of the mycoplasmas performed by Vasconcelos *et al*. (2005) in order to compare, in a more detailed way, the phylogenetic tree reconstruction under different methods and sets of protein sequences.

Materials and Methods

Taxa studied

We focused our attention on a set of nine complete genomes of mycoplasmas studied by Vasconcelos *et al*. (2005) to compare our results with their findings. All these nine species belong to the groups Hominis and Pneumonieae, and their accession numbers in GenBank are listed in Table 1. We avoided the inclusion of outgroups (*Mesoplasma florum* and *Mycoplasma mycoides,* Entomoplasmatales) in order to extend the number of genes that could be used in the analyses. Therefore, all phylogenetic trees presented here may be considered as unrooted trees. However, Vasconcelos *et al*. (2005) have found very consistently that the genomes studied here are rooted between the clades Hominis and Pneumonieae. We consequently draw our trees rooted in that branch.

Detection of orthologous sequences

The clusters of orthologous protein coding genes were detected using the bidirectional best hit (BBH) method (Overbeek 1999), and only those genes presented as a single copy in all the genomes studied were used in the analyses except for the presence/absence method (see below).

Sequences alignment

Protein sequences were aligned separately for each orthologous set using Clustalw 1.8 (Thompson *et al*. 1994). The phylogenetic analyses described below were performed on these individual genes (partitions) as well as on datasets in which the proteins were concatenated to form "supergenes" (*e.g.*, Gontcharov *et al*. 2004). The Gblocks program (Castresana 2000) was used in some analyses to remove unreliable regions in the alignments, characterized as ambiguous in the alignments. This should reduce the noise in the phylogenetic signal caused by such regions.

Table 1 - The taxonomic classification and the access numbers in GenBank of the nine species of mycoplasmas included in this work.

Species	ID	Access number
Mycoplasma gallisepticum	R	NC_04829
Mycoplasma genitalium	G-37	NC_000908
Mycoplasma hyopneumoniae	J	NC_007295
Mycoplasma hyopneumoniae	7448	NC_007332
Mycoplasma penetrans	HF-2	NC_004432
Mycoplasma pneumoniae	M129	NC_000912
Mycoplasma pulmonis	UAB CTIP	NC_002771
Mycoplasma synoviae	53	NC_007294
Ureaplasma parvum	ATCC700970	NC_002162

JKL subset of sequences

Jain *et al*. (1999) have argued that the genes related to the processing and storing of information in cells should be less prone to HGT and therefore may be the most appropriate set to estimate a reliable bacterial phylogeny. To test this hypothesis in our dataset, we used the System for Automated Bacterial Integrated Annotation (SABIA, Almeida *et al*. 2004) in order to search for information storage and processing genes (represented here by the J, K, and L functional categories of COG, Clusters of Orthologous Groups, Tatusov *et al*. 2000). We performed additional analyses using only these JKL sequences and as a contrast, we performed also the analyses using only the non-JKL sequences. These analyses were compared with those performed with the full set of orthologous sequences.

Phylogenetic methods

Three main methods were applied to build the phylogenetic trees using the concatenated alignments: Maximum Likelihood (ML) (Felsenstein 1981), performed with the ProtML program of the Molphy Package (using the JTT-F substitution model); Neighbor Joining (NJ) (Saitou and Nei 1987), achieved with the programs PROTML and NJDIST (Molphy package), using the ML-distance; and Maximum Parsimony (MP), calculated with the program PAUP* 4.0 b10 (Swofford 2002), under a heuristic search with default parameters.

The phylogenetic trees for each individual gene partition were estimated using the Tree-Puzzle 5.1 ML program (Schmidt *et al*. 2002). We estimated the α parameter of the γ-distribution from the data using the JTT+F distance and 1000 puzzle steps for each gene partition. We also used these individual substitution models in order to construct a combined ML tree with the program COMBINE (Pupko *et al*. 2002).

In addition, we inferred a "bootstrap-gene tree" by randomly re-sampling and concatenating the orthologous proteins throughout 500 replications. This tree was calculated with the NJ method (JTT+F distance) and a majority rule consensus was constructed using the Molphy package.

Gene-content and gene -order phylogenies

"Gene-content" phylogenies were estimated using the presence/absence of the genes on the aligned genomes considering 873 orthologous clusters and performed with the MP and the NJ approaches (using the parameters described above), as suggested by Korbel *et al*. (2002).

Phylogenies based on the rearrangement distances between the genomes (gene-order trees) were estimated by

NJ using distances calculated using the GRIMM server with the default option settings and by the conserved gene-pairs method (Korbel *et al*. 2002).

Estimation of branch robustness

To estimate the robustness of each internal branch of the trees, we used the non-parametric bootstrap test (Efron *et al*. 1996; Felsenstein 1985) with 100 replications for each method used in this work, except for the COMBINE tree.

Topology comparisons

In order to test if some candidate trees (produced with different methods and datasets) could be significantly rejected by the concatenated data set under the ML criterion, we applied the Shimodaira-Hasegawa (1999) (SH) test using Paup* 4.0 b10.

Results

Selected sequences

The numbers of estimated genes in each of the nine genomes included in this work range from 484 in *M. genitalium* to 1037 in *M. penetrans* (see Vasconcelos *et al*. 2005 for details). However, only 227 genes with a total of 92,083 amino acids passed our criteria for inclusion, that is, they are putative orthologs and exist as single copies in all genomes. When we considered the JKL sequences only, 124 proteins were selected, totalizing 45,674 amino acids. Conversely, 103 genes were considered in non-JKL analyses (42,542 aa). In the analyses where the putatively unreliable alignment regions were removed by the Gblocks method, only 51,516 amino acids (~55%) of the 227 gene partitions were taken in consideration. The gene content analysis was performed with a set of 873 orthologous gene sequences.

Phylogenetic analyses

Figure 1A shows the phylogenetic tree based on the concatenated 227 orthologous genes sequences found by the majority of the methods (ML, NJ, MP, and the bootstrap-gene tree), as well as the bootstrap values of the nodes inferred under all these methods. This tree agrees with the topology presented by Vasconcelos *et al*. (2005) and divided the species into the groups Hominis and Pneumoniae. Almost all internal nodes presented maximum bootstrap values for all the methods. This tree was also identical with the tree constructed with separated substitution models for each protein partition revealed by the ML COMBINE method. The only exception to the convergent topology represented in Figure 1A was found under the MP method, as also reported by Vasconcelos *et al*. (2005). In the MP tree, instead of node six (Figure 1A), we found node eight (represented in Figure 1B), with *Ureaplasma parvum* as a sister species of the remaining Pneumoniae species.

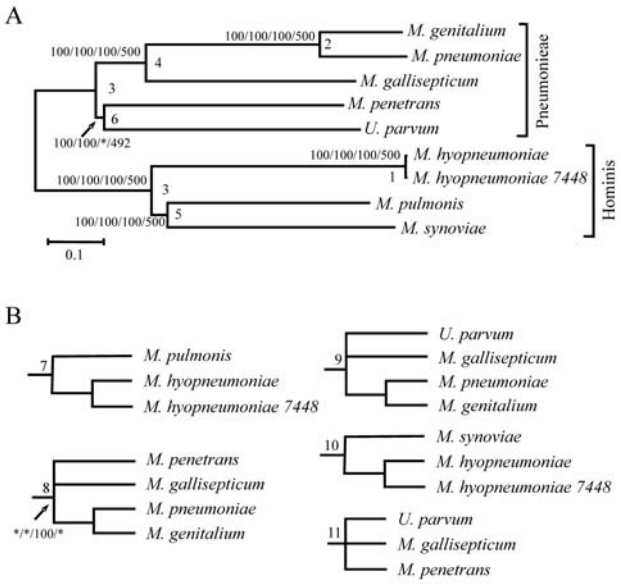

Figure 1 - (A) ML tree of the nine species of *Mycoplasma*. Each node was assigned a number 1-6. Above each node are the bootstrap values of the different sequence methods in the following order: ML topology, NJ topology, MP topology, all with 100 replications, and the bootstrap gene tree, with 500 replications. (B) The alternative nodes found in some gene partition analyses under the ML approach. These nodes were assigned with numbers 7-11. Node 8 was also found in the MP analysis, and its bootstrap support was indicated under it. Asterisks mean that a given node was not supported by a given analysis in the order showed in (A).

Although we have represented in this tree only those analyses performed with the entire 227 gene data set, most of the topologies of the individual gene partitions, constructed with the ML method, also reached the same topology. The main alternative nodes found are represented in Figure 1B. Figure 2 shows the bootstrap confidence values for the 11 nodes represented in Figure 1 for all topologies of the 227 individual gene partitions. There is a wide variation in the support values for the different nodes. For example, nodes 1 and 2 presented support above 50% for all genes and these, plus node 3, presented 100% bootstrap support for most partitions. On the other hand, the other nodes, such as node 6, hardly ever have a gene that alone presented 100% support.

The bootstrap-gene tree supports the ML/NJ topology, with maximum bootstrap values (500) in all nodes, except for node 6 (492). The tree estimated with the ML COMBINE program (Pupko *et al*. 2002), which takes into account different substitution models for each gene partition, also supports the node 6. It is interesting to investigate how much data (gene partitions and amino acids) are necessary for obtaining high confidence values for the eleven main nodes found in the analyses. To assess the relationship between the size of the dataset and the confidence values of each node, we randomly re-sampled and concatenated an increasing number of genes draw from the 227 genes dataset (Figure 3B). The nodes differ widely in the number of genes that are necessary to achieve high support values.

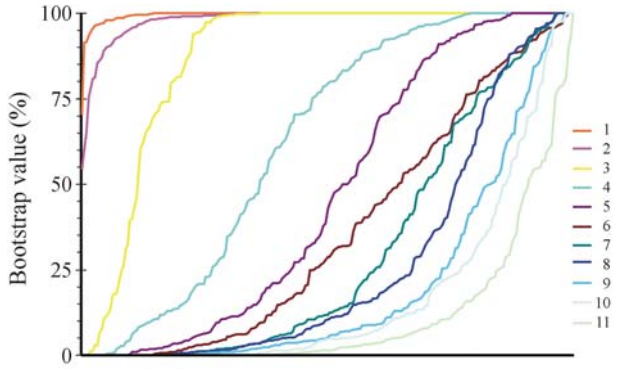

Figure 2 - Ranked distribution of the percent bootstrap values of the eleven nodes presented in Figure 1, recovered from each of the 227 partitions under the ML approach.

For nodes 1 to 3, three genes are usually enough to get 100% confidence values, while for node 4, about 10 genes are sufficient to achieve 95% confidence. In contrast, nodes 5 and 6 reached high support values much more slowly. More than 100 genes are necessary for node 5 to reach consistently > 95% values and for node 6 this value is achieved only with about 200 genes. Conversely, the support values for the alternative nodes (from Figure 1B) decline with the inclusion of more genes, although node 8 maintains bootstrap values around 15% even when about 100 genes are used. The features described above were also found when we re-sampled the amino-acid positions of the concatenated dataset, irrespective of the genes (partitions) they originate (Figure 3A). Very interestingly, in this case, higher bootstrap values were achieved more readily, with only about 35,000 amino acids (≈ 40%) being sufficient for all branches to achieve > 95% support values. These results suggest, as expected, that positions within a gene are not independent and that using unlinked genes or positions may be a better strategy to more efficiently recover a genome tree (Rokas *et al*. 2003).

When we restricted the set of analyzed sequences to the categories J, K, and L from the COG database (JKL dataset) or, conversely, to the non-JKL sequences (non-JKL dataset), or when we restricted the set of amino acids to those with alignments that we were more confident with (ambiguous positions removed with the GBlock algorithm, Gb dataset), most of the resulting trees were identical to those showed in Figure 1A. The only exceptions were the MP trees built with JKL and Gb sequences, which supported node 8 instead of node 6, concordant with the MP tree obtained with the complete set of orthologous sequences (Figure 1B). In all our trees, the bootstrap values for the nodes 1, 2 and 3 were 100%. The values of the remaining nodes are shown in Table 2. Curiously, the use of the non-JKL sequences with the MP method recovered the tree presented in Figure 1A, but with a relatively low bootstrap value in node 6 (62%).

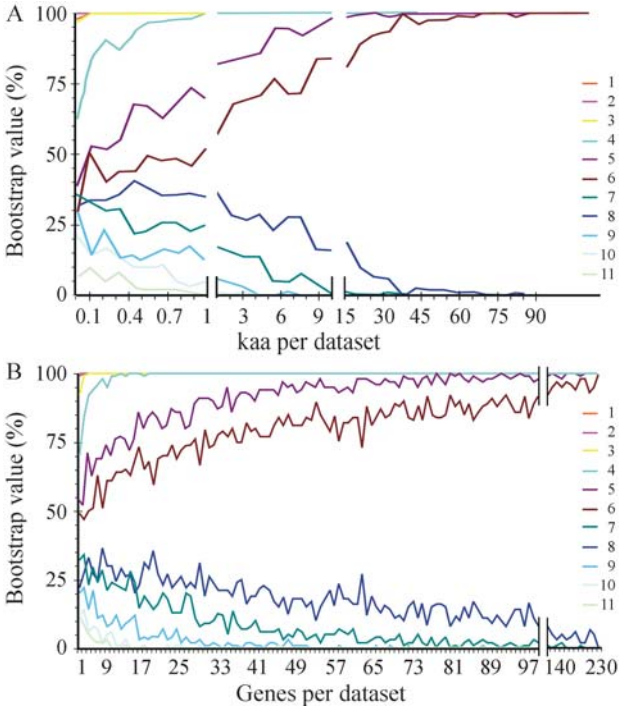

Figure 3 - Bootstrap support values of the eleven nodes indicated in Figure 1 with increasing number of (A) amino acids or (B) genes used to construct the phylogenetic tree. The values were estimated with the NJ method from concatenation of randomly re-sampled subsets of data. kaa: thousands amino-acids. Double vertical bars indicate changes in axis scale.

Table 2 - Bootstrap values of the nodes 4, 5, 6, and 8 (Figure 1) obtained with different methods of tree building and different datasets throughout 100 replications.

Phylogenetic method	Dataset	Bootstrap values (%) of nodes			
		4	5	6	8
ML	JKL[a]	100	100	81	-*
	Non JKL[b]	98	73	73	-
	Gb[c]	97	76	75	-
NJ	JKL	100	100	97	-
	Non JKL	100	100	99	-
	Gb	100	100	97	-
MP	JKL	100	98	-	100
	Non JKL	100	99	62	-
	Gb	100	100	-	65

*Node not present in this tree.
[a]the J, K, and L COG sequences (124 genes).
[b]only the non J, K, and L sequences (103 genes).
[c]only the reliable alignment regions from GBlocks (51,516 amino acids).

In the phylogeny constructed using the presence/absence of genes (the gene-content tree), two different trees (Figure 4) were recovered with the two different methods used. The NJ distance tree places *U. parvum* as sister species to the remaining Pneumoniae species (node 8) while the MP tree places *M. penetrans* in this position (node 9). In

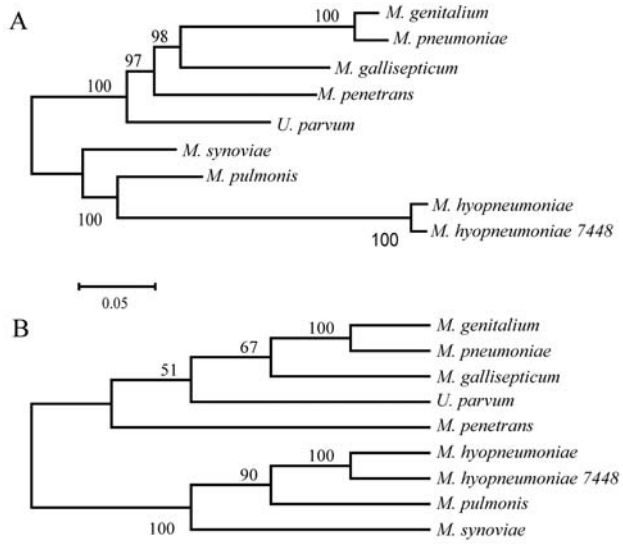

Figure 4 - Gene-content phylogenies estimated using the presence/absence of 873 homologous proteins. (A) NJ tree based on gene-content distance calculated as suggested by Korbel *et al*. (2002); (B) MP tree.

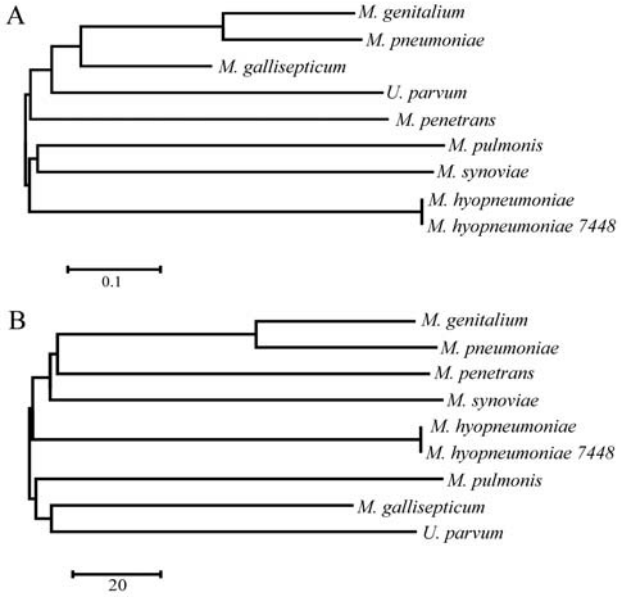

Figure 5 - NJ trees estimated using gene-order distances based on 227 homologous proteins. (A) GRIMM distance; (B) gene-pair distance.

addition, they put *M. pulmonis* as sister species to *M. hyopneumonieae* (node 7). The bootstrap values for these nodes (except for node 7), however, are lower than the confidence values presented in Figure 1A.

The phylogenies constructed using the differences in the order of genes in each genome (gene-order trees) (Figure 5) are the most different from those estimated here, presenting several conflicts with the other trees. Most of the internal branches of these trees (except branches 1 and 2) are so short that the relationship could be considered unresolved, while the terminal branches are very long. Taken together, this suggests that the phylogenetic signal of the genome organization (gene-order) was significantly lost at the earliest branches.

Finally, the Shimodaira-Hasegawa (1999) test (Table 3) showed that these two trees are not statistically different, even though the topology recovered with the MP approach is different from the topology obtained under the ML and NJ approaches. The trees obtained by the other methods (gene-content and gene-order) were significantly different from the ML tree.

Discussion

In this paper we have expanded considerably the phylogenomic analysis of the mycoplasmas performed by Vasconcelos *et al*. (2005), applying new methods and trying to better understand the evolution of this group based on their genomes and the origin of the topological differences in the phylogenetic trees found by different phylogenetic approaches.

Although most of our analyses resulted in topologies identical to the tree presented in Vasconcelos *et al*. (2005), we found some differences that should be addressed here. Specifically, the analyses based on gene-content and gene-order recovered four other alternative topologies (Figures 4 and 5) that are significantly different from the tree represented in Figure 1A (Table 3). In all of them, node 6 is absent, while node 8 is present only in the NJ tree based on gene content (Figure 4A, bootstrap value = 97%). It is interesting to note that even though genome sizes vary widely in the mycoplasmas, which has been considered as a major problem for gene-content trees (see Snel *et al*. 2005 for a review); the gene-content phylogenies are very similar to the tree presented in Figure 1A. Indeed, both methods used to build these trees retrieved with high support four of the six nodes of that figure (see Figure 5). The two remaining nodes (5 and 6) correspond to those which required the largest number of genes and amino acids to reach high bootstrap values, that is, they are the ones with the lowest

Table 3 - Shimodaira-Hasegawa test results of the best alternative trees obtained in this work.

	Log likelihood	Difference	p
ML/NJ/ Bootstrap Gene tree[a]	-686977.09	(Best)	-
MP topology[b]	-687028.34	51.25	0.4950
Gene content (NJ) topology[c]	-687289.25	312.17	0.0000*
Gene content (MP) topology[c]	-687472.68	495.59	0.0000*
Gene order (NJ-Distance) topology[d]	-687199.46	222.37	0.0160*

*$p < 0.05$
[a]presented in Figure 1A.
[b]MP topology, represented in Figure 1 (present the node 8 - Figure 1B- instead of the node 6 - Figure 1A).
[c]represented in Figure 4.
[d]represented in Figure 5.

phylogenetic signal. Thus, as pointed out by Snel *et al*. (2005), the use of shared gene content may be a useful tool to recover phylogenies.

The gene-order trees (Figure 5) seem clearly saturated and therefore unreliable, as most of the internal branches (except for branches 1 and 2) were extremely short, while those leading to the individual genomes were very long, suggesting that the phylogenetic signal of the genome order was almost lost at the earliest branching of the phylogeny. This result suggests high genome structure plasticity in the mycoplasmas which contrasts with the results of Suyama and Bork (2001), which suggested a well conserved gene order among Mollicutes; and agrees with previous analyses indicating rapid chromosomal rearrangements in this group (Rocha and Blanchard 2002).

All trees obtained from the 227 concatenated orthologous sequences by the ML and NJ methods point to a hypothesis for the evolutionary relationships of these species, represented by the phylogeny in Figure 1A, while the MP method recovered a phylogeny containing node 8 of Figure 1B (instead of node 6). The existence of conflicting organismal phylogenies is expected among individual gene trees (Figure 2), but it was somewhat unexpected in the present case in which we used a set of more than two hundred concatenated sequences and almost a hundred thousand characters in some analyses. Although these two trees are not statistically different (Table 3), the topologies recovered by the two sets of methods still tell different histories for the evolution of the group. Therefore, the use of even a huge data set of concatenated genes does not always yield a "single, fully resolved species tree" as suggested by Rokas *et al*. (2003).

Among the causes for the occurrence of conflicting topologies one could cite: unreliable alignments, horizontal gene transfer, and differences in sensibility among methods of phylogenetic inference in relation to certain evolutionary factors. To test the first possibility we used only the reliable alignment regions (GBlocks dataset), but the results did not change (Table 2). To test the second hypothesis, we inferred the trees using only those sequences that are theoretically less prone to horizontal transfer events (the information processing and storage genes, JKL dataset) as suggested by Jain *et al*. (1999). The topologies obtained with this dataset were identical with those obtained with the full dataset for each method: the ML and NJ methods recovered the topology presented in Figure 1A while the MP method recovered node 8 (Figure 1B). It has been suggested that HGTs may be so common as to obscure the concept of a species phylogeny (Doolittle 1999; Gogarten and Townsend 2005). However, no consistent phylogenetic signal for HGT was found among the single copy genes studied here. This corroborates, in an evolutionarily more recent group, the results of the γ -proteobacteria (Lerat *et al*. 2003) and other studies (Daubin *et al*. 2003), which suggest that HGT is not a common result in the set of orthologous genes used for phylogenetic analysis when stringent criteria are used, such as using only single copy genes that are present in all genomes.

In conclusion, nodes 1 to 5 were well supported by all the phylogenetic methods and sets of sequences used here and thus are likely to be part of the true tree. Considering that the SH tests could not discriminate between trees containing node 6 and trees containing node 8, it is, however, not yet possible to assert which grouping is closer to the true one.

While on the one hand our results suggest that the ML and NJ approaches perform better than the MP approach to resolve topologies in a genomic scale, recent simulations have on the other hand shown that heterogeneously evolving genes can bias ML methods but not MP methods (Kolaczkowski and Thornton 2004, but see Gadagkar and Kumar 2005). Therefore, it is not yet clear how sensitive the different inference methods are to a wide range of evolutionary factors, especially in real datasets. Consequently, we think that further analysis must be performed with several real genomic datasets trying to better understand the basis for these differences in the performance of the methods and how to minimize them (*e.g.*, Phillips *et al*. 2004). Currently, our best alternative seems to follow Opperdoes' (2003) suggestion that a tree should be considered robust and thus reliable only when broadly different methods infer similar or identical tree topologies, and when such topologies are supported by good confidence values (*i.e.* more than 95%).

Acknowledgments

We gratefully thank Ana Teresa Ribeiro de Vasconcelos for her support and encouragement and Rangel C. Souza and Roger F.C. Paixão (Labinfo, LNCC/CNPq) for their help with the identification of the orthologous clusters and COG categories. This work was part of the Brazilian National Genome Program (Southern Network for Genome Analysis and Brazilian National Genome Project Consortium) with funding provided by MCT/CNPq and SCT/FAPERGS (RS).

References

Almeida LGP, Paixao R, Souza RC, Costa GC, Barrientos FJA, Santos MT, Almeida DF and Vasconcelos AT (2004) A system for automated bacterial (genome) integrated annotation - SABIA. Bioinformatics 20:2832-2833.

Bapteste E, Susko E, Leigh J, MacLeod D, Charlebois RL and Doolittle WF (2005) Do orthologous gene phylogenies really support tree-thinking? BMC Evol Biol 5:33-43.

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540-552.

Daubin V, Moran NA and Ochman H (2003) Phylogenetics and the cohesion of bacterial genomes. Science 301:829-832.

Delsuc F, Brinkmann H and Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. Nature Rev Genet 6:361-375.

Doolittle WF (1999) Lateral genomics. Trends Biochem Sci 24:M5-M8.

Efron B, Harlloran E and Holmes S (1996) Bootstrap confidence levels for phylogenetic trees. Proc Natl Acad Sci USA 93:13429-13434.

Eisen JA and Fraser CM (2003) Phylogenomics: Intersection of evolution and genomics. Science 300:1706-1707.

Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol 17:368-376.

Felsenstein J (1985) Confidence limits on phylogenies - An approach using the bootstrap. Evolution 39:783-791.

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleichmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, *et al*. (1995) The minimal gene complement of *Mycoplasma genitalium*. Science 270:397-403.

Gadagkar S and Kumar S (2005) Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. Mol Biol Evol 22:2139-2141.

Gogarten JP and Townsend JP (2005) Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol 3:679-687.

Gontcharov AA, Marin B and Melkonian M (2004) Are combined analyses better than single gene phylogenies? A case study using SSU rDNA and *rbc*L sequence comparisons in the Zygnematophyceae (Streptophyta) Mol Biol Evol 21:612-624.

Hughes AL, Ekollu V, Friedman R and Rose JR (2005) Gene family content-based phylogeny of prokaryotes: The effect of criteria for inferring homology. Syst Biol 54:268-276.

Jain R, Rivera MC and Lake JA (1999) Horizontal gene transfer among genomes: The complexity hypothesis. Proc Natl Acad Sci USA 96:3801-3806.

Kolaczkowski B and Thornton JW (2004) Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431:980-984.

Korbel JO, Snel B, Huynen MA and Bork P (2002) SHOT: A web server for the construction of genome phylogenies. Trends Genet 18:158-162.

Lerat E, Daubin V and Moran NA (2003) From gene trees to organismal phylogeny in prokaryotes: The case of the gamma proteobacteria. Plos Biol 1:101-109.

Nei M and Kumar S (2000) Molecular Evolution and Phylogenetics. Oxford University Press, New York, 333 pp.

Ochman H, Lerat E and Daubin V (2005) Examining bacterial species under the specter of gene transfer and exchange. Proc Natl Acad Sci USA 102:6595-6599.

Opperdoes FR (2003) Phylogenetic analysis using protein sequences. In: Salemi M and Vandamme A-M (eds) A Pratical Approach to DNA and Protein Phylogeny. Cambridge University Press, UK, pp 237-235.

Overbeek R, Fonstein M, D'Souza M, Pusch GD and Maltsev N (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci USA 96:2896-2901.

Phillips MJ, Delsuc F and Penny D (2004) Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol 21:1455-1458.

Pupko T, Huchon D, Cao Y, Okada N and Hasegawa M (2002) Combining multiple data sets in a likelihood analysis: Which models are the best? Mol Biol Evol 19:2294-2307.

Razin S, Yogev D and Naot Y (1998) Molecular biology and pathogenicity of mycoplasmas. Microbiol Mol Biol Rev 62:1094-1156.

Rocha EPC and Blanchard A (2002) Genomic repeats, genome pasticity and the dynamics of *Mycoplasma* evolution. Nucleic Acids Res 9:2031-2042.

Rokas A, Williams BL, King N and Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798-804.

Saitou N and Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406-425.

Schmidt HA, Strimmer K, Vingron M and von Haeseler A (2002) Tree-puzzle: Maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502-504.

Shimodaira H and Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol 16:1114-1116.

Snel B, Huynen MA and Dutilh BE (2005) Genome trees and the nature of genome evolution. Annu Rev Microbiol 59:191-209.

Suyama M and Bork P (2001) Evolution of prokaryotic gene order: Genome rearrangements in closely related species. Trends Genet 17:10-13.

Swofford DL (2002) PAUP* phylogenetic analysis using parsimony (*and other methods) Version 4. Sinauer Associates, Sunderland, Massachusetts.

Tatusov RL, Galperin MY, Natale DA and Koonin EV (2000) The COG database: A tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28:33-36.

Thompson JD, Higgins DG and Gibson TJ (1994) Clustal W - Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673-4680.

Vasconcelos ATR, Ferreira HB, Bizarro CV, Bonatto SL, Carvalho MO, Pinto PM, Almeida DF, Almeida LGP, Almeida R, Alves L, *et al*. (2005) Swine and poultry pathogens: The complete genome sequences of two strains of *Mycoplasma hyopneumoniae* and a strain of *Mycoplasma synoviae*. J Bacteriol 187:5568-5577.

Internet Resources

GRIMM: A tool for analyzing rearrangements in pairs of genomes, including unichromosomal and multichromosomal genomes, and signed and unsigned data, http://www.cse.ucsd.edu/groups/bioinformatics/GRIMM/index.html (verified 11/29/2005).

ProtML and Njdist programs to infer phylogenetic trees, http://www.ism.ac.jp/ismlib/softother.e.html (verified 09/11/2006).

Associate Editor: Darcy F. de Almeida