The Genomic Tree as Revealed from Whole Proteome Comparisons

Fredj Tekaia,[1,3] Antonio Lazcano,[2] and Bernard Dujon[1]

[1]*Unité de Génétique Moléculaire des Levures [URA1300 Centre National de la Recherche Scientifique (CNRS) and UFR927 University Pierre and Marie Curie], Institut Pasteur, 75724 Paris Cedex 15, France;* [2]*Facultad de Ciencias, UNAM, Apdo. Cd. Universitaria, 04510 Mexico City, Mexico*

The availability of a number of complete cellular genome sequences allows the development of organisms' classification, taking into account their genome content, the loss or acquisition of genes, and overall gene similarities as signatures of common ancestry. On the basis of correspondence analysis and hierarchical classification methods, a methodological framework is introduced here for the classification of the available 20 completely sequenced genomes and partial information for *Schizosaccharomyces pombe*, *Homo sapiens*, and *Mus musculus*. The outcome of such an analysis leads to a classification of genomes that we call a genomic tree. Although these trees are phenograms, they carry with them strong phylogenetic signatures and are remarkably similar to 16S-like rRNA-based phylogenies. Our results suggest that duplication and deletion events that took place through evolutionary time were globally similar in related organisms. The genomic trees presented here place the Archaea in the proximity of the Bacteria when the whole gene content of each organism is considered, and when ancestral gene duplications are eliminated. Genomic trees represent an additional approach for the understanding of evolution at the genomic level and may contribute to the proper assessment of the evolutionary relationships between extant species.

The determination of complete genome sequences from ≥20 organisms offers an unprecedented opportunity for the study of evolutionary problems in molecular biology and at a highly integrated level. One of the first problems to address in such a context concerns the derivation of the universal tree of life, which should reflect the global evolutionary relationships of whole organisms, and not only single-gene phylogenies. The universal tree of life was based on the 16S-like rRNA genes (Woese 1987; Woese et al. 1990) and led to the proposal of the three primary kingdoms or domains (Eukarya, Bacteria, and Archaea). However, this proposal has been criticized on different grounds (Gupta 1998; Mayr 1998). Although other molecular phylogenies have confirmed this analysis, many genes (particularly those encoding metabolic enzymes) give different topologies or even fail to support the three-domain classification of living organisms (Cavalier-Smith 1989; Forterre et al. 1992; Brown and Doolittle 1997; Doolittle 1998; Gupta 1998). Within the three-domain classification itself, a recurrent question concerns the controversial proximity of Archaea to either Eukarya or Bacteria (Brinkmann and Philippe 1999). Archaeal organisms appear to be close to Eukarya when the protein synthesis machinery (transcription and translation) is considered but close to Bacteria if metabolic genes are compared (Doolittle and Logsdon 1998).

Such unsettling differences not only reflect classical problems in phylogenetic reconstruction due to horizontal transfer (which may have been more intense during early cellular evolution, Woese 1998), unequal rates of nucleotide substitution, and gene displacement but also underline the fact that trees depict the evolutionary distances between genes and not between organisms or entire genomes.

Previous attempts to analyze the macrostructure of genomes for phylogenetic reconstruction have been based on a number of well-known techniques such as DNA hybridization studies and restriction enzyme fragment analyses (Li 1997). As in the case of gene-based phylogenies, such approaches are ultimately dependent on the degree of sequence divergence. On the contrary, analysis of comparative gene order provides quantitative models of genome evolution that become independent from the degree of sequence divergence once orthologs have been defined (Sankoff et al. 1992; Boore and Brown 1998). Likewise, a more integrative view of genome evolution is feasible with the shared gene trees proposed recently by Snel et al. (1999). Here we present a different but complementary approach, not on the basis of evolutionary descent but on a hierarchical classification of genomes involving their gene content and overall similarity. We call the resulting phenograms the genomic tree.

RESULTS

Construction of Genomic Trees by Comparisons of All Predicted ORF Products for Completely Sequenced Genomes

In this work, we aim to derive the genomic tree from all of the available completely sequenced genomes

[3]Corresponding author.
E-MAIL tekaia@pasteur.fr; FAX 33 1 40 61 34 56.

9:550–557 ©1999 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/99 $5.00; www.genome.org

through whole proteome comparisons, taking into account the predicted gene product content of each organism and their similarity. The construction of such a tree requires an appropriate methodological approach. The full set of predicted gene products of a completely sequenced organism is compared with itself and with that of every other organism considered. The possible similarity of a given open reading frame (ORF) product to any other is determined by appropriately defined statistical limits (see Methods). Comparison of organism *j* with organism *i* determines the proportion of ORFs in organism *j* that have at least one similar ORF in organism *i* (T_{ij}). We call this proportion the "weight" of the common ancestry of *j* with respect to *i* (see Methods). The validation of the statistical limits used in such comparisons is discussed in Methods. The overall pairwise comparison of *n* organisms leads to a *nxn* matrix of T_{ijs}. The appropriate method to handle such data matrices as a whole is correspondence analysis (Benzecri 1973; Greenacre 1984). The rationale of this method is to derive an orthogonal system of axes, called factors and denoted F_1, F_2,... F_{n-1} (a maximum of $n - 1$ such axes can be determined), which pass through the barycenter of the observations and correspond to a decreasing order of the amount of information each factor represents. Each organism is represented by its coordinates in this system. Thus, distances between organisms can be calculated, and their subsequent classification according to their neighborhood leads to a hierarchical tree, or the genomic tree. Such a tree is a graphical representation of the relationship between sets of organisms, which includes indirectly genome sizes, levels of internal redundancy due to ancestral duplications, and overall gene loss or acquisition events. This tree is independent of functional identity. Instead, it is based on the sole presence or absence of genes of common ancestry, as defined by comparison with all other genomes.

This method was applied to the data set, obtained from the comparison of the 20 completely sequenced organisms, plus the data available from human, mouse, and *Schizosaccharomyces pombe* (see Table 1). The results of our analysis are presented in Figure 1, in which organisms are represented on the best factorial space (i.e., the first and second factors). The distances between the surveyed organisms were calculated from their factorial coordinates and used to construct the genomic tree shown in Figure 2a. Four well-defined groups of organisms with similar profiles appear on this tree: (1) An archaeal cluster formed by *Methanococcus jannaschii*; *Archaeoglobus fulgidus*, *Methanobacterium thermoautotrophicum*, and *Pyrococcus horikoshii*; (2) a (eu)bacterial group formed by *Escherichia coli*, *Synechocystis* sp., *Bacillus subtilis*, *Aquifex aeolicus*, *Mycobacterium tuberculosis*, *Campylobacter jejuni*, *Haemophilus influenzae*, *Helicobacter pylori*, *Rickettsia prowazekii*, *Chlamydia trachomatis*, *Treponema pallidum*, and *Borrelia burgdorferi*; (3) a mycoplasma cluster (*Mycoplasma pneumoniae* and *Mycoplasma genitalium*) that groups with the Bacteria cluster; and (4) the eukaryotic group (*Caenorhabditis elegans*, *Mus musculus*, *Homo sapiens*, *S. pombe*, and *Saccharomyces cerevisiae*).

As indicated in Figure 2, the different species are not distributed at random in our trees, but their overall clustering follows the three-domain distribution, whose general topology is remarkably similar to unrooted 16S-like rRNA-based and gene-shared phylogenies (Woese 1987; Woese et al. 1990; Snel et al. 1999). Note, however, that although this tree is the outcome of a hierarchical classification, it carries a strong phylogenetic signature and can thus be considered a genomic tree of considerable assistance in understanding the evolutionary relationships between genomes.

In the approach discussed here, genome size, levels of ancestral gene redundancy due to duplications, and overall loss or acquisition of genes all contribute indirectly to the position of a given organism in the factorial space. An obvious concern is that the inclusion of small genomes, such as those of *M. genitalium* or *M. pneumoniae*, which may have undergone massive gene losses, may drastically alter the genomic tree by the limitation imposed on the proportions of genes of common ancestry in other genomes. To test this possibility, we eliminated the two *Mycoplasma* genomes from the data set and recomputed a novel tree (Fig. 2b). As shown in Figure 2b, whereas the exclusion of the mycoplasma produces no major changes in the overall tree topology, it affects the internal branching of the Bacteria, displacing *A. aeolicus* from a cluster that includes *E. coli*, *B. subtilis*, *Synechocystis* sp., and *M. tuberculosis*, to another branch with *R. prowazekii* and *C. trachomatis*. These changes probably are due to the small branch lengths among the inner nodes of the Bacteria. Removal of the two mycoplasma genomes affects slightly the Archaea, in which *P. horikoshii* is displaced by *M. janaschii*.

Because the positions of *B. burgdorferi* (850 genes), *C. trachomatis* (877 genes), *R. prowazekii* (837 genes), and *T. pallidum* (1031 genes) in the genomic tree are within the major bacterial branch, and not with the *M. genitalium–M. pneumoniae* cluster (468 and 677 genes, respectively), genome size is not overemphasized in the genomic tree. The fact that the two mycoplasma genomes form a deep branch within the Bacteria that is far removed from their close relative *B. subtilis* (as shown, e.g., by their 16S rRNA phylogeny) demonstrates that more realistic assessments of the different contributions of the variables defining the position of a given organism in the genomic tree are still required. We have tested the impact of the insertion of additional sequences on the topology of the tree by simu-

Table 1. Completely Sequenced Organisms and Other Fragmentary Data Considered in this Analysis

Organism	Domain[a]	Code[b]	ORFs[c]	Partitions[d]
H. influenzae	B	HI	1713	1377
M. genitalium	B	MG	468	361
Synechocystis sp.	B	Ssp	3168	2002
M. pneumoniae	B	MP	677	424
H. pylori	B	HP	1577	1226
E. coli	B	EC	4290	2473
B. subtilis	B	BS	4100	2573
B. burgdorferi	B	BB	850	696
A. aeolicus	B	AE	1522	1157
M. tuberculosis	B	MT	3924	2329
T. pallidum	B	TP	1031	852
C. trachomatis	B	CT	877	718
C. jejuni	B	CJ	1731	1323
R. prowazekii	B	RP	837	653
M. jannaschii	A	MJ	1735	1180
M. thermoautotrophicum	A	MTH	1871	1227
A. fulgidus	A	AF	2437	1423
P. horikoshii OT3	A	PH	2061	1373
S. cerevisiae	E	SC	6182	4437
C. elegans	E	CE	19,099	7558
S. pombe[e]	E	SP	3579	2248
H. sapiens	E	Hs		
M. musculus[e]	E	Mm		

Predicted ORF products considered in this study are essentially as described in the original publications: *H. influenzae* (Fleischman et al. 1995), *M. genitalium* (Fraser et al. 1995), *M. jannaschii* (Bult et al. 1996), *Synechocystis* sp. strain PCC6803 (Kaneko et al. 1996), *M. pneumoniae* (Himmelreich et al. 1996), *H. pylori* (Tomb et al. 1997), *E. coli* (Blattner et al. 1997), *M. thermoautotrophicum* (Smith et al. 1997), *B. subtilis* (Kunst et al. 1997), *A. fulgidus* (Klenk et al. 1997), *B. burgdorferi* (Fraser et al. 1997), *A. aeolicus* (Deckert et al. 1998), *M. tuberculosis* (Cole et al. 1998), *P. horikoshii* (Kawarabayasi et al. 1998), *T. pallidum* (Fraser et al. 1998), *C. trachomatis* (Stephens et al. 1998), *R. prowazekii* (Andersson et al. 1998), and *C. elegans* (The *C. elegans* Sequencing Consortium, 1998). Yeast *S. cerevisiae* ORF products (Goffeau et al. 1997) correspond to those indicated in the MIPS server: http://www.mips.biochem.mpg.de/, with a few modifications. Preliminary complete proteome of *C. jejuni* (ftp.sanger.ac.uk in/pub/pathogens/Cj/) was considered.

[a](B) Bacteria; (A) Archaea; and (E) Eukarya.

[b]Organism abbreviations used in Figs. 1 and 3.

[c]The total number of predicted ORF products.

[d]The total number of distinct partitions.

[e]*S. pombe* correspond to those at the Sanger ftp server: ftp.sanger.ac.uk under/pub/yeast/sequences/pombe/pompep/), human (*H. sapiens*) and mouse (*M. musculus*) sequences Hsuniq, Mmuniq (Boguski et al. 1995). An incomplete set of data was used, containing 3579 ORF products representing at least 68% of total proteome (V. Wood pers. comm.); 43,088, and 8,821 sets of clustered ESTs derived from GenBank release 106, respectively. Hsuniq and Mmuniq were used solely as targets for comparisons with the other organisms.

lating the inclusion of artificial genomes with different degrees of ancestral conservation with the actual genomes (0, 20, 100) and found essentially the same results (data not shown).

Construction of Genomic Trees by Comparison of the Minimized Sets of Predicted ORF Products from Completely Sequenced Genomes

In the approach presented here, gene families are represented by their respective weights, and this corresponds to the whole genome picture. This approach can be complemented by highlighting the functional classification of genes. Because the organisms included here have genomes with different sizes and may exhibit important variations in the degree of gene acquisition and loss, it is important to construct a tree derived from genomes reduced to their minimal content by eliminating ancestral gene duplications that are still recognizable. Thus, in a second alternative approach, we have reduced each organism to its minimum genomic content by eliminating ancestral gene duplications and derived a second genomic tree. This was achieved with the following approximation: Each organism is represented by its partitions (i.e., genes with common ancestry; see Methods). Accordingly, instead of considering ancestry weight as the outcome of gene similarity, we considered it as the result of partition similarity. This neutralizes the variable rates of gene acquisition and losses because now each given gene family is represented by only one member (i.e., the corresponding partition). The resulting data set was analyzed following the previous methodology for the construction of a hierarchical tree, which represents the constituent set of genes in the organisms considered. The conservation rate of organism *j* in organism

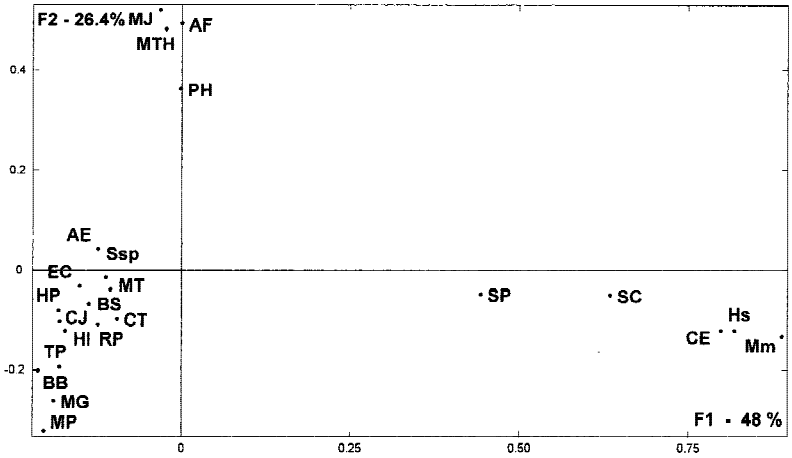

Figure 1 Factorial representation of the weight of ancestral duplication and common ancestry in each genome, obtained by the multidimensional correspondence analysis method. First and second factorial axes (F_1 and F_2) represent, respectively, 48% and 26.4% of total information included in the ancestry weight matrix resulting from predicted gene product comparisons (see Methods). Dots represent the distribution of the surveyed organisms (abbreviations are as in Table 1).

i is defined by P_{ij}, which is obtained by dividing the number of distinct partitions of *j* including members having at least one significant match in *i*, by the total number of distinct partitions in *j*. Thus, for instance, 11.4% of yeast partitions share ancestral conservation with *B. subtilis*, 9.6% with *H. influenzae*, and so on.

Figure 3a shows the organisms' distribution on the first factorial space, and Figure 3b shows the genomic tree obtained by the hierarchical classification of their genomes by their distances as calculated in the whole factorial space. This new genomic tree represents the synthesis of the minimal content relationships in the considered organisms, and can be mapped onto the small subunit (SSU) rRNA phylogeny discussed by Woese (1987) and Woese et al. (1990). This result bears upon the current debates on the major divisions in the living world (Gupta 1998; Mayr 1998).

DISCUSSION

Genomic and Gene Trees

On the basis of correspondence analysis and a hierarchical classification of gene content and overall gene similarities as ancestry weight, we have developed a new approach for genomic analysis that allows the construction of genomic trees that carry a strong phylogenic signature and whose overall topology strongly resembles the SSU rRNA-based evolutionary trees (Woese et al. 1990). Although our approach provides an excellent equivalent to the 16S-like rRNA-based branching orders of Archaea and the Eukarya (Fig. 2a), it leads to a different branching order within the Bacteria domain. This is particularly true of Gram-positive bacteria, from which the two mycoplasma are widely separated, forming a branch distant from both *B. subtilis* and *M. tuberculosis*. The latter species group into a non-natural cluster together with *E. coli*, *Synechocystis* sp., and *A. aeolicus*. It is of interest that *A. aeolicus*, whose exact phylogenetic position has been debated (Deckert et al. 1998), is firmly located in our tree within the Bacteria, as in the case of rRNA phylogenies (Burggrof et al. 1992; Pitulle et al. 1994; Reysenbach et al. 1996). However, it does not branch off early and instead clusters with *M. tuberculosis*.

In contrast to the rooted universal phylogenies that pair Archaea with the eukaryotic branch (Gogarten et al. 1989; Iwabe et al. 1989; Brown and Doolittle 1995), our methodology places the two prokaryotic kingdoms closer to each other than any one of them is to Eukarya (Fig. 2a). This is in accordance with the reconstruction of the universal tree, which eliminates artifacts due to long branch attraction and places Archaea as a sister group of Bacteria (Brinkmann and Philippe 1999).

Because the genomic tree shown in Figure 2a is based solely on the ancestral duplication and conservation proportions, its coherence at a gross level with the small subunit rRNA tree suggests that the average duplication and loss events that have taken place through evolutionary time are statistically similar in related organisms. That is, the weight of ancestry contributes to define the overall properties of a genome and groups it in a way that is strongly reminiscent of rRNA-based phylogenies over extended periods of evolutionary time.

The strong similarity between our genomic trees, which embody sequence divergence, gene losses, and acquisitions, with the 16S-like RNA phylogenies, which are based solely on sequence divergence, raises the issue of their consistency with phylogenetic trees constructed from other genes common to all the surveyed organisms. To analyze this issue, 75 partitions of universal genes were determined [i.e., each partition of structural orthologous genes (see Methods) includes at least one member from each of the organisms considered]. Phylogenetic trees were constructed by the neighbor-joining method with a bootstrap value of 1000, by use of the Clustal W program (Thompson et al. 1994). These 75 trees are scarcely consistent with each other and with the rooted universal tree. The resulting gene trees can be divided roughly as follows: (1) 36% are consistent with the rooted universal tree (i.e., that branch archaeal genes with eukaryal ones) and include, among others, those of genes encoding ribosomal proteins, as well as those involved in DNA me-

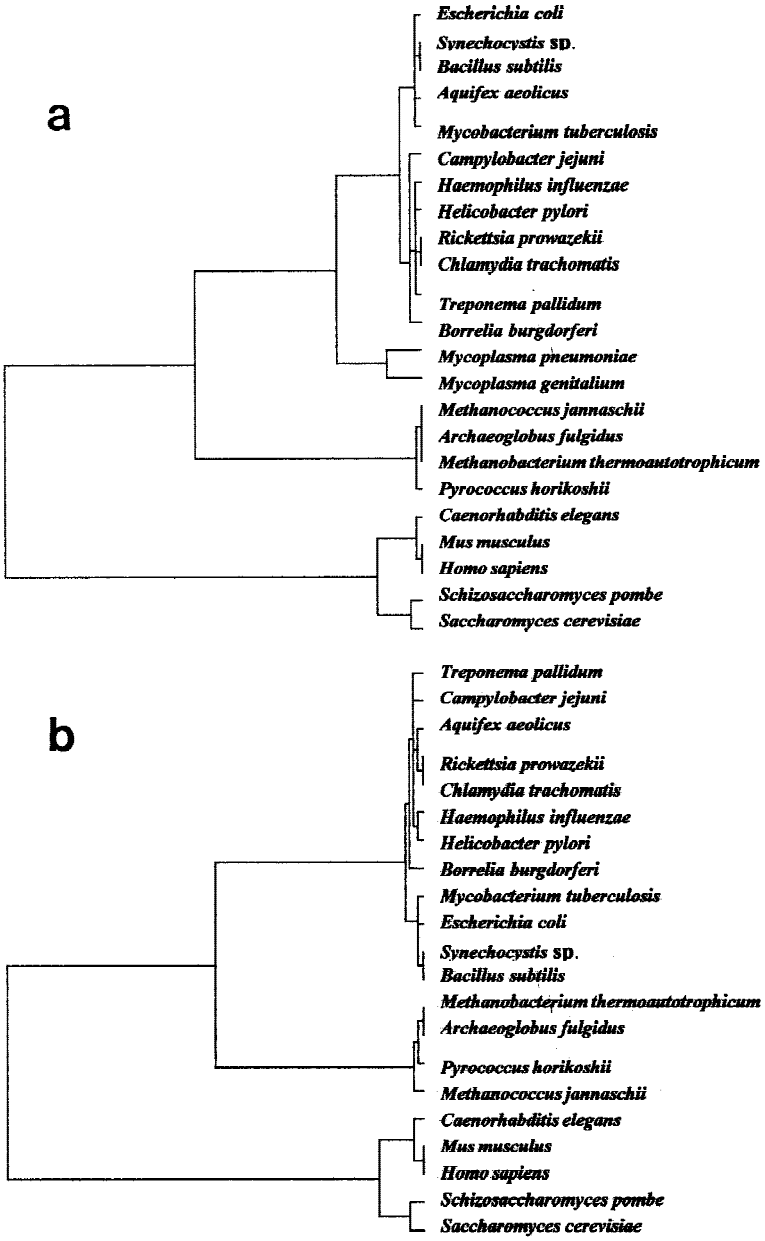

Figure 2 Genomic tree. (*a*) This tree is obtained by a hierarchical classification of the organisms on the basis of their neighborhood distances. Distances between all pairs of organisms are calculated in the factorial space obtained by correspondence analysis. Horizontal lines between nodes are proportional to their similarity. (*b*) Same tree excluding data from *M. genitalium* and *M. pneumoniae*.

tabolism and a few metabolic pathways; (2) in 21%, archaeal genes branch with bacterial ones, and this group includes, among others, genes involved in electron transport, gluconeogenesis/glycolysis, and RNA processing; and (3) 43% are a mixture of the previous topologies, that is, some genes branch with eukaryal genes, whereas others branch with bacterial genes, and eukaryal sometimes branch with bacterial genes (F. Tekaia, A. Lazcano, and B. Dujon, in prep.). These results show the phylogenetic distortion effects on gene trees, and emphasize the conflict between species and gene trees. In light of these distortions, the last genomic tree shown in Figure 3b may be considered as the average tree of all orthologous genes.

Future genome sequences will allow further refinement of the genomic trees presented here, and critical comparison with the sequence-based (Woese 1987) and shared-gene (Snel et al. 1999) trees will lead to a proper assessment of the value of our results. The trees presented here are less likely to suffer from the pitfalls of traditional methods such as variable changes in sequences and reliability of sequence alignments (Gupta 1998), because our approach is insensitive to such problems. However, our methodology is not intended to substitute for evolutionary inference on the basis of sequence comparisons but, rather, to provide a snapshot of the molecular evolution whether the large variations of genome sizes between organisms, the level of internal redundancy in each genome, and the losses or acquisitions of genes during evolution are considered or not. The observed differences between the topology of the genomic trees very likely are due to the different weights of gene families and their ancestry. The proximity between the Archaea and the Bacteria observed in the two genomic trees has to be confirmed once more completely sequenced eukaryal and archaeal genomes are available. Nevertheless, the statistical analysis of the degree of ancestral duplication and evolutionary conservation discussed here may help in the development of novel approaches to the management and understanding of large volumes of genomic data. Thus, our results represent an additional approach for the understanding of evolution at the genomic level and may contribute to the proper assessment of the evolutionary relationships between extant species.

METHODS

The rationale for the construction of genomic trees is based on the systematic comparison of the predicted translation products of all surveyed organisms (data taken from original publications; see Table 1) as a means to determine the presence or absence of genes of common ancestry with an internally calculated threshold of significance. For each organism included in the analysis, every gene product is successively used as a query sequence against all the gene products of the same organism, and against all the gene products of each of

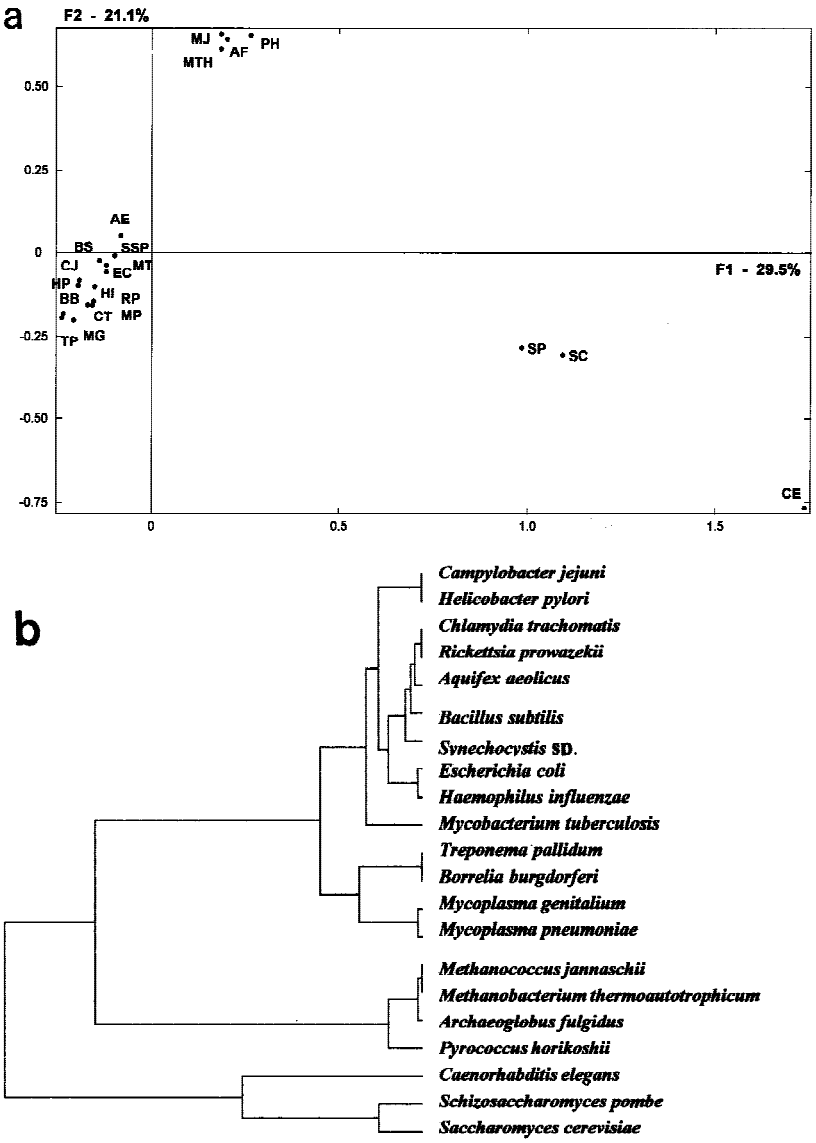

Figure 3 (*a*) Factorial representation of the constituent ancestry in each genome. First and second axes (F_1 and F_2) represent, respectively, 29.5% and 21.1% of the total information included in the ancestry weight matrix resulting from the organism's partitions (see Methods). Organismal distribution on this factorial plane is very similar to that in Fig. 1. Human and mouse, for which no accurate ancestral gene duplications can be presently calculated, were not considered in this analysis (abbreviations are as in Table 1). (*b*) Genomic tree for the considered organisms (minus human and mouse) as obtained from the whole factorial space resulting from the corresponding analysis (see Fig. 2a for methods of analysis).

the other organisms considered; the former is used to define partitions of genes, and the latter to measure the weight of common ancestry (see below).

Definition of a Partition

A set of ORF products in a given organism defines a partition if, and only if, the following three properties are verified: (1) Each member of the set has at least one highly significant match with one other member of the set; (2) no member of the set has highly significant matches with members not included in the set; and (3) the set cannot be partitioned into subsets verifying (1) and (2) (i.e., the set is minimal).

Note that an ORF product that has no significant match in its own organism fulfills these properties and therefore is considered as a single member partition. Thus, a partition includes all ORF products with common ancestry in a given organism (the number of distinct partitions are shown in Table 1). Note that such construction of partitions is sometimes referred to as the single-linkage clustering method.

Definition of the Weight of Common Ancestry

The weight of common ancestry is the proportion of gene products that share a common ancestry with the gene products of other organisms, with all species being examined serially. The presence of homolog(s) defines the degree of ancestral duplication within each genome and of conservation between genomes.

Organism-Specific Comparisons

Comparisons of each query sequence with the complete proteome databases of the same and every other organism were performed with BLASTP (Altschul et al. 1990), version 1.4.8, using the *pam250* substitution matrix, which favors large segment pairs and, hence, detects distantly related ORFs and the *seg* (Wootton and Federhen 1993) filter to eliminate compositionally biased regions in the query sequence. The *M. musculus* (Mmuniq) and human (Hsuniq) databases (see Table 1) serve solely as targets for comparison by queries of other genomes, with the TBLASTN (Altschul et al. 1990) program.

Because the genomes analyzed here exhibit important differences in size and complexity, we first determined a limit of significance of the BLASTP probability scores for each of the genomes considered (Tekaia and Dujon 1999). This was achieved by use of sets of random sequences, equivalent in number to the number of ORFs of each genome, and generated with sizes and amino acid compositions equal to the average size and composition of the actual proteome of each organism. Each of these random sequences was compared against the entire database of the cognate organism, and the best probability scores were recorded as for actual sequences. For each organism, the highest BLASTP probability score leaving <5% of pseudosignificant matches was considered as the limit of significance when that organism is used as target. Probability score limits were set at 10^{-9} for *S. cerevisiae*, 10^{-5} for *B. subtilis*, *B. burgdorferi*, *M. tuberculosis*, *M. jannaschii*, and *C. elegans*, 10^{-3} for *S. pombe*, *T. pallidum*, *P. horikoshii*, and *R. prowazekii*, 10^{-2} for *C. trachomatis*, and 10^{-4} for all other genomes.

Ancestry Weight Matrix using ORF Products

The data table *T* resulting from the pairwise comparisons of the organisms considered here can be found in http://www-alt.pasteur.fr/~tekaia/dupcons.html. In this table, T_{ij} is the proportion of ORF products from organism *j* having a common ancestry with one or several ORF product(s) of organism *i* (note that T_{ij} is normalized because it is divided by the total number of ORFs in *j*). T_{jj} is the proportion of ORF products in organism *j* having a common ancestry with one or several other ORF product(s) of the same organism. In this study, $i = 1, 23$ and $j = 1, 21$, the difference between *i* and *j* correspond to the sequences from man (Hsuniq) and mouse (Mmuniq), which serve solely as targets for comparisons not as queries. As an example, in the *S. cerevisiae* genome, 16.7% of the ORFs share ancestral conservation with the *B. subtilis* genome, 12.7% with the *H. influenzae* genome, and so on. We refer to such proportions as the weight of common ancestry in the *S. cerevisiae* genome when compared with *B. subtilis*, *H. influenzae*, and others. Because of variable genome sizes and internal redundancy, the matrix is not symmetrical, for example, the weight of common ancestry in the *B. subtilis* genome when compared with *S. cerevisiae* is 15.5%.

Ancestry Weight Matrix Using Partitions

The data table *P* resulting from the pairwise comparisons of the organisms can also be found in http://www-alt.pasteur.fr/~tekaia/dupconsparts.html. In this table, P_{ij} is the proportion of distinct partitions in organism *j* having ancestry with at least one predicted gene product in organism *i*. Because each partition is unique in its organism, $P_{jj} = 100.\%$ (i.e., when comparing a given organism with itself, each partition is its unique match).

Structural Orthologous Genes

Two genes belonging to two distinct organisms are called structural orthologs, if and only if each shows the most similarity to the other when comparing it with its counterpart organism.

Partitions are obtained by applying the same definition (as in organisms) and by considering the whole set of orthologous genes obtained from the considered organisms.

ACKNOWLEDGMENTS

We thank Stewart Cole for helpful discussion. We are indebted to Henri Buc, Edouard Yeramian, and the members of the Unité de Génétique Moléculaire des Levures for their encouragement and several useful discussions. Support from the Manlio Cantarini Foundation (Paris) and Universidad Autonoma de Mexico—Direction General de Asunto del Personal Academico (UNAM–DGAPA) project PAPIIT-IN213598 support to A.L. is gratefully acknowledged. B.D. is Professor of Molecular Genetics at University Pierre et Marie Curie and a member of the Institut Universitaire de France.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Andersson, S.G., A. Zomorodipour, J.O. Andersson, T. Sicheritz-Ponten, U.C. Alsmark, R.M. Podowski, A.K. Naslund, A.S. Eriksson, H.H. Winkler, and C.G. Kurland. 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396:** 133–140.

Benzecri, J-P. 1973. *L'analyse des données. Vol 2: L'analyse des correspondances*, Dunod, Paris, France.

Blattner, F.R., G. Plunkett III, C.A. Bloch, N.T. Perna, V. Burland, M. Riley, J. Collado-Vides, J.D. Glasner, C.K. Rode, G.F. Mayhew et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277:** 1453–1462.

Boguski, M.S. and G.D. Schuler. 1995. Establishing a human transcript map. *Nat. Genet.* **10:** 369–371.

Boore, J.L. and W.M. Brown. 1998. Big trees from little genomes: Mitochondrial gene order as a phylogenetic tool. *Curr. Opin. Genet. Dev.* **8:** 668–674.

Brinkmann, H. and H. Philippe. 1999. Archaea sister-group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. *Mol. Biol. Evol.* (In press).

Brown, J.R. and W.F. Doolittle. 1995. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc. Natl. Acad. Sci.* **92:** 2441–2445.

———. 1997. Archaea and the prokaryote-to-eukaryote transition. *Microbiol. Mol. Biol. Rev.* **61:** 456–502.

Bult, C.J., O. White, G.J. Olsen, L. Zhou, R.D. Fleischmann, G.G. Sutton, J.A. Blake, L.M. FitzGerald, R.A. Clayton, J.D. Gocayne et al. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273:** 1058–1073.

Burggraf, S., G.J. Olsen, K.O. Stetter, and C.R. Woese. 1992. A phylogenetic analysis of *Aquifex pyrophilus*. *Syst. Appl. Microbiol.* **15:** 353–356.

Cavalier-Smith, T. 1989. Molecular phylogeny. Archaebacteria and Archezoa. *Nature* **339:** 100–101.

Cole, S.T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S.V. Gordon, K. Eiglmeier, S. Gas, C.E. 3rd Barry, F. Tekaia et al. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393:** 537–544.

Deckert, G., P.V. Warren, T. Gaasterland, W.G. Young, A.L. Lenox, D.E. Graham, R. Overbeek, M.A. Snead, M. Keller, M. Aujay et al. 1998. The complete genome of the hyperthermophilic bacterium Aquifex aeolicus. *Nature* **392:** 353–358.

Doolittle, R.F. 1998. Microbial genomes opened up. *Nature* **392:** 339–342.

Doolittle, W.F. and J.M. Logsdon Jr. 1998. Archaeal genomics: Do archaea have a mixed heritage? *Curr. Biol.* **8:** R209–R211.

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.-F. Tomb, B.A. Dougherty, J.M. Merrick et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269:** 496–512.

Forterre, P., N. Benachenhou-Lahfa, F. Confalonieri, M. Duguet, C. Elie, and B. Labedan. 1992. The nature of the last universal ancestor and the root of the tree of life, still open questions. *Biosystems* **28:** 15–32.

Fraser, C.M., J.D. Gocayne, O. White, M.D. Adams, R.A. Clayton, R.D. Fleischmann, C.J. Bult, A.R. Kerlavage, G. Sutton, J.M. Kelley et al. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* **270:** 397–403.

Fraser, C.M., S. Casjens, W.M. Huang, G.G. Sutton, R. Clayton, R. Lathigra, O. White, K.A. Ketchum, R. Dodson, E.K. Hickey et al. 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390:** 580–586

Fraser, C.M., S.J. Norris, G.M. Weinstock, O. White, G.G. Sutton, R. Dodson, M. Gwinn, E.K. Hickey, R. Clayton, K.A. Ketchum et al. 1998. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281:** 375–388.

Goffeau, A., R. Aert, M.L. Agostini-Carbone, A. Ahmed, M. Aigle, L. Alberghina, K. Albermann, M. Albers, M. Aldea, D. Alexandraki et al. 1997. The Yeast Genome Directory. *Nature*(Suppl) **387:** 5–105.

Gogarten, J.P., H. Kibak, P. Dittrich, L. Taiz, E.J. Bowman, B.J.

Bowman, M.F. Manolson, R.J. Poole, T. Date, T. Oshima et al. 1989. Evolution of the vacuolar H^+-ATPase: Implications for the origin of eukaryotes. *Proc. Natl. Acad. Sci.* **86:** 6661–6665.

Greenacre, M. 1984. *Theory and application of correspondence analysis*, Academic Press, London, UK.

Gupta, R.S. 1998. Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among Archaebacteria, Eubacteria, and Eukaryotes. *Microbiol. Mol. Biol. Rev.* **62:** 1435–1491.

Himmelreich, R., H. Hilbert, H. Plagens, E. Pirkl, B.-C. Li, and R. Herrmann. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acid Res.* **24:** 4420–4449.

Iwabe, N., K. Kuma, M. Hasegawa, S. Osawa, and T. Miyata. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci.* **86:** 9355–9359.

Kaneko, T., S. Sato, H. Kotani, A. Tanaka, E. Asamizu, Y. Nakamura, N. Miyajima, M. Hirosawa, M. Sugiura, S. Sasamoto et al. 1996. Sequence analysis of the genome of the unicellular *cyanobacterium Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3:** 109–136.

Kawarabayasi, Y., M. Sawada, H. Horikawa, Y. Haikawa, Y. Hino, S. Yamamoto, M. Sekine, S. Baba, H. Kosugi, A. Hosoyama et al. 1998. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. *DNA Res.* **5:** 55–76.

Klenk, H.P., R.A. Clayton, J.F. Tomb, O. White, K.E. Nelson, K.A. Ketchum, R.J. Dodson, M. Gwinn, E.K. Hickey, J.D. Peterson et al. 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390:** 364–370.

Kunst, F., N. Ogasawara, I. Moszer, A.M. Albertini, G. Alloni, V. Azevedo, M.G. Bertero, P. Bessieres, A. Bologin, S. Borchert et al. 1997. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* **390:** 390: 249–259.

Li, W-H. 1997. *Molecular evolution*, Sinauer, Sunderland, MA.

Mayr, E. 1998. Two empires or three? *Proc. Natl. Acad. Sci.* **95:** 9720–9723.

Pitulle, C., Y. Yang, M. Marchiani, E.R. Moore, J.L. Siefert, M. Aragno, P. Junrtshuk, and G.E. Fox. 1994. Phylogenetic position of the genus *Hydrogenobacter*. *Int. J. Syst. Bact.* **44:** 620–626.

Reysenbach, A.L., G.S. Wickham, and N.R. Pace. 1994. Phylogenetic analysis of the hyperthermophilic pink filament community in Octopus Spring, Yellowstone National Park. *Appl. Environ. Microbiol.* **60:** 2113–2119.

Sankoff, D., G. Leduc, N. Antoine, B. Paquin, B.F. Lang, and R. Cedergren 1992. Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. *Proc. Natl. Acad. Sci.* **89:** 6575–6579.

Smith, D.R., L.A. Doucette-Stamm, C. Deloughery, H. Lee, J. Dubois, T. Aldredge, R. Bashirzadeh, D. Blakely, R. Cook, K. Gilbert et al. 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: Functional analysis and comparative genomics. *J. Bacteriol.* **179:** 7135–7155.

Snel, B., P. Bork, and M.A. Huynen. 1999. Genome phylogeny based on gene content. *Nat. Genet.* **21:** 108–110.

Stephens, R.S., S. Kalman, C. Lammel, J. Fan, R. Marathe, L. Aravind, W. Mitchell, L. Olinger, R.L. Tatusov, Q. Zhao et al. 1998. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* **282:** 754–759.

Tekaia, F. and B. Dujon. 1999. Pervasiveness of gene conservation and persistence of duplicates in cellular genomes. *J. Mol. Evol.* (In press).

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *Caenorhabditis elegans*: A platform for investigating biology. *Science* **282:** 2012–2018.

Thompson, J.D., D.G. Higgins, and T.J. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.

Tomb, J.F., O. White, A.R. Kerlavage, R.A. Clayton, G.G. Sutton, R.D. Fleischmann, K.A. Ketchum, H.P. Klenk, S. Gill, and B.A. Dougherty. 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388:** 539–547.

Woese, C.R. 1987. Bacterial evolution. *Microbiol. Rev.* **51:** 221–271.

———. 1998. The universal ancestor. *Proc. Natl. Acad. Sci.* **95:** 6854–6859.

Woese, C.R., O. Kandler, and M.L. Wheelis. 1990. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci.* **87:** 4576–4579.

Wootton, J.C. and S. Federhen. 1993. Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* **17:** 149–163.

Received November 30, 1998; accepted in revised form March 30, 1999.