Genome phylogeny based on gene content

Berend Snel[1,2,3], Peer Bork[1,2] & Martijn A. Huynen[1,2]

Species phylogenies derived from comparisons of single genes are rarely consistent with each other, due to horizontal gene transfer[1], unrecognized paralogy and highly variable rates of evolution[2]. The advent of completely sequenced genomes allows the construction of a phylogeny that is less sensitive to such inconsistencies and more representative of whole-genomes than are single-gene trees. Here, we present a distance-based phylogeny[3] constructed on the basis of gene content, rather than on sequence identity, of 13 completely sequenced genomes of unicellular species. The similarity between two species is defined as the number of genes that they have in common divided by their total number of genes. In this type of phylogenetic analysis, evolutionary distance can be interpreted in terms of evolutionary events such as the acquisition and loss of genes, whereas the underlying properties (the gene content) can be interpreted in terms of function. As such, it takes a position intermediate to phylogenies based on single genes and phylogenies based on phenotypic characteristics. Although our comprehensive genome phylogeny is independent of phylogenies based on the level of sequence identity of individual genes, it correlates with the standard reference of prokaryotic phylogeny based on sequence similarity of 16s rRNA (ref. 4). Thus, shared gene content between genomes is quantitatively determined by phylogeny, rather than by phenotype, and horizontal gene transfer has only a limited role in determining the gene content of genomes.

When we compared the protein sequences encoded by 13 completely sequenced genomes with each other and recorded the number of genes shared between the genomes using an operational definition of orthology[5], two patterns emerged (Table 1). Not unexpectedly, the first one is that large genomes have many genes in common; for example, the highest number of shared genes can be observed between *Escherichia coli* and *Bacillus subtilis*, which have the largest genomes among the Bacteria. This effect of size is reflected in the numbers of genes that the four archaeal genomes share with bacteria of various sizes (Fig. 1). The second emerging pattern is a phylogenetic one: the number of genes two genomes have in common depends on their evolutionary distance[2]. *Haemophilus influenzae* for example shares more genes with its close relative *E. coli* than with *B. subtilis*. We created a phylogeny of the genomes using the neighbour-joining algorithm[3], with the fraction of shared genes in the smallest of the two genomes as a similarity criterion using random subsets of the genes per genome for bootstrapping (Fig. 2*a*). The resulting tree reflects the standard phylogeny as based on 16s rRNA (with some minor exceptions; Fig. 2*b*; refs 4,6). The two major lineages of cellular life that are represented here by multiple species, the Archaea and Bacteria, are monophyletic with maximal bootstrap values, with the third lineage (Eukarya) being equidistant between them. In the bacterial branch, *Aquifex aeolicus* appears at the root of the tree, and the purple bacteria, the γ subdivision within the purple bacteria and the 'low G+C' Gram-positive bacteria are all monophyletic. The sequences of both *Mycoplasma genitalium* and *Helicobacter pylori* evolve at relatively high rates[2]. Distance-based phylogenetic methods tend to move highly divergent sequences towards the root of the tree, however, the method used here is relatively insensitive to such variations in rates of evolution of gene sequences. The four archaeal genomes in this analysis are all Euryarchaeota. The location of *Pyrococcus horikoshii* at the root of the Euryarchaeota is confirmed in the 16s rRNA phylogeny. The

Table 1• Common gene content in genomes

	AF	MT	MJ	PH	AQ	SY	BS	MG	BB	EC	HI	HP	SC
AF	**2,407**	*48.1*	*50.1*	*40.2*	*38.2*	*26.3*	*26.8*	*33.3*	*25.2*	*28.1*	*26.4*	*23.6*	*23.1*
MT	900	**1,871**	*55.7*	*37.4*	*35.3*	*31.1*	*30.9*	*30.3*	*24.8*	*32.0*	*24.2*	*22.3*	*27.9*
MJ	870	966	**1,735**	*43.7*	*32.7*	*29.2*	*28.1*	*31.2*	*22.2*	*31.1*	*22.4*	*22.3*	*27.8*
PH	829	699	759	**2,061**	*30.9*	*23.8*	*27.2*	*31.4*	*24.0*	*26.1*	*21.7*	*20.1*	*23.7*
AQ	582	537	497	471	**1,522**	*52.5*	*53.8*	*54.5*	*44.6*	*59.0*	*44.0*	*43.7*	*31.1*
SY	632	581	506	491	799	**3,168**	*30.5*	*58.8*	*48.1*	*35.9*	*44.6*	*41.0*	*19.1*
BS	645	578	488	561	819	967	**4,100**	*70.7*	*56.5*	*33.6*	*51.3*	*42.0*	*16.1*
MG	156	142	146	147	255	275	331	**468**	*50.4*	*62.2*	*57.5*	*52.1*	*40.4*
BB	214	211	189	204	379	409	480	236	**850**	*52.2*	*46.2*	*43.8*	*29.4*
EC	676	598	539	538	898	1,138	1,376	291	444	**4,290**	*77.8*	*49.9*	*17.1*
HI	453	416	384	372	669	766	880	269	393	1335	**1,717**	*41.1*	*28.8*
HP	375	355	354	320	665	652	668	244	372	793	653	**1,590**	*22.2*
SC	555	522	482	488	474	606	659	189	250	735	494	353	**6,296**

The numbers of genes shared (see Methods) between genomes (lower left triangle), the percentage of genes shared between genomes (the total number divided by the number of genes in the smallest genome; upper right triangle) and the numbers of genes per genome (bold). HI, *H. influenzae*[16]; MG, *M. genitalium*[17]; SY, *Synechocystis* sp. PCC 6803 (ref. 18); MJ, *M. jannaschii*[19]; EC, *E. coli*[20]; MT, *M. thermoautotrophicum*[21]; HP, *H. pylori*[22]; AF, *A. fulgidus*[23]; BS, *B. subtilis*[24]; BB, *B. burgdorferi*[25]; SC, *S. cerevisiae*[26]; AQ, *A. aeolicus*[27]; PH, *P. horikoshii*[28].

[1]*European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany.* [2]*Max-Delbrück-Centrum for Molecular Medicine, 13122 Berlin-Buch, Germany.* [3]*Bioinformatics Group, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands. Correspondence should be addressed to P.B. (e-mail: bork@embl-heidelberg.de).*

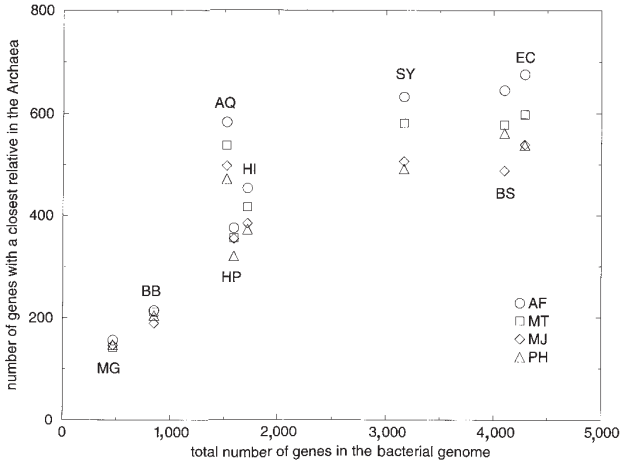

Fig. 1 Relationship between the number of genes in a genome and the number of genes that have a closest relative (Table 1) in another genome. The Archaea are chosen as reference species because they all have the same evolutionary distance to the Bacteria; hence, phylogenetic effects on the number of shared genes are eliminated. The number of shared genes between two genomes correlates with genome size. The exception to the general trend is *A. aeolicus*, which, relative to its genome size, has too many genes with closest relatives in the Archaea.

remainder of the Euryarchaeota topology (Fig. 2*a*) does not correspond with the 16s rRNA phylogeny, but is supported by sequence comparisons of RNA polymerase subunit B (ref. 7) and other proteins shared among the four genomes.

In addition to revealing the topology of the phylogenetic tree, neighbour joining also reveals information about variations in branch lengths. These variations have distinctive causes. Of the Bacteria, *M. genitalium* and *A. aeolicus* have the shortest distance to the center of the tree. In *M. genitalium*, this appears to be due to a secondary loss of genes, given its late branching within the Bacteria. This has left *M. genitalium* with a set of relatively essential genes that have a high probability of being shared with other species. *A. aeolicus* has, compared with other bacteria of a similar size, many genes with orthologues in the Archaea (Table 1, Fig. 1), although it is clearly a bacterium (bootstrap value 100). If one assumes, on the basis of studies of ancient gene duplications[8], that the root of the tree of life lies between the Bacteria and the Archaea, this implies that *A. aeolicus* is not only similar to the last common ancestor of the Bacteria with respect to the sequences of single genes, as has been reported earlier for 16s rRNA (ref. 4), but also with respect to its gene content. *A. aeolicus* can hence be regarded as a primitive species, aside from being a species with primitive genes.

There are a few aspects in which our tree differs from the 16s rRNA tree. These mainly concern the bacterial phylogeny. The spirochete *Borrelia burgdorferi* does not cluster with the purple bacteria, and the cyanobacterium *Synechocystis* appears as a sister species of *A. aeolicus*. The bootstrap value for the position of *B. burgdorferi* is low; however, that of the clustering of *Synechocystis* with *A. aeolicus* is high. In 16s rRNA-based phylogenies (Fig. 2*b*), and also in phylogenies based on proteins involved in replication, transcription and translation[9], the relative phylogenetic positions of the Gram-positive bacteria, purple bacteria, cyanobacteria and spirochetes are ill resolved. With the availability of more genomes, the robustness of the observed patterns should become clearer and we may be able to further clarify the phylogeny of these groups.

In the Archaea, *Methanococcus jannaschii* and *Methanobacterium thermoautotrophicum* cluster together, relative to *Archaeoglobus fulgidus* (bootstrap value of 100). This does not cor-

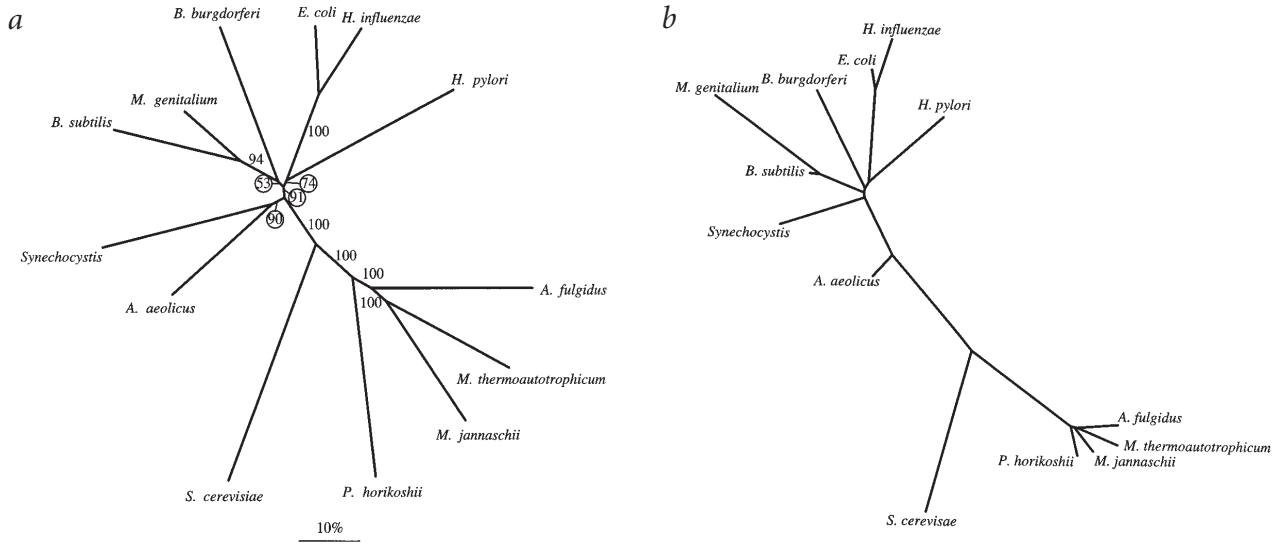

Fig. 2 Genome phylogeny. ***a***, Phylogeny of completely sequenced cellular genomes derived from gene content. The similarity between two genomes is expressed as the fraction of the genes in each of the genomes that have a closest relative gene in the other genome. The fraction is calculated by dividing the number of pairs of closest relatives (Table 1) by the total number of genes in the smallest genome of the two, the latter posing an upper limit to the number of shared genes. The distance between two genomes is then: 1–(number of shared genes/genes in smallest genome). The phylogeny is a neighbour-joining clustering of the resulting distance matrix. To obtain confidence estimates for the tree, a delete-half-jackknife[29] was implemented; that is, bootstrap values were calculated by selecting random subsets of 50% of the genes per genome, reanalysing the fractions of shared genes and recalculating the trees. The values represent the number of times (out of 100) a specific cluster was present. The length of the scale bar corresponds with a 10% difference in gene content. The phylogeny includes the first 14 genomes published, except for *Mycoplasma pneumoniae*. *M. genitalium* and *M. pneumoniae* are close relatives, the gene content of *M. genitalium* being a subset of that of *M. pneumoniae*[30], making the similarity between the two 100% in our measure. *M. genitalium* was chosen of the two because it is the smallest completely sequenced genome; our analysis covers the size range of the published genomes. ***b***, Phylogeny of the species in this paper constructed on the basis of 16s rRNA. The phylogeny is identical to a previously published version[4], and can be extracted from the 16s rRNA database (http://rdp.life.uiuc.edu/). The phylogenetic position of *S. cerevisiae* relative to the prokaryotes is not included in this database; *S. cerevisiae* was added to the tree at its consensus position, and its branch length is not necessarily representative. The phylogenetic positions of the cyanobacteria, Gram-positive bacteria and purple bacteria are ill resolved, as is reflected in the short branch lengths separating these groups.

respond with the 16s rRNA tree of the Archaea, in which *M. thermoautotrophicum* and *A. fulgidus* are more closely related than either is to *M. jannaschii*[4,6] (Fig. 2*b*). Individual protein sequences, however, tend to favour *M. thermoautotrophicum* and *M. jannaschii* as sister groups relative to *A. fulgidus*[7]. In 369 sets of four sequences that were shared among the four Archaea used in this analysis, the level of sequence identity between *M. thermoautotrophicum* and *M. jannaschii* is higher than that of either of them with *A. fulgidus* ($P<0.001$, using Spearman's rank correlation). Furthermore, neighbour-joining trees of the 369 sets most often showed *M. jannaschii* and *M. thermoautotrophicum* as sister species (45%) relative to *A. fulgidus*, with either of the two (22% and 32%, respectively) when *P. horikoshii* was used as outgroup.

Our tree formulated on the basis of gene content does not correlate with phenotype; for example, the pathogenic species in the set, such as *M. genitalium*, *H. influenzae* and *H. pylori*, do not cluster together, neither do the hyperthermophilic species *A. aeolicus*, *P. horikoshii*, *A. fulgidus* and *M. jannaschii*. Genes that are shared between species correlate with phenotypic features, for example, in the case of the genes that are shared between the pathogens *H. influenzae* and *H. pylori* but that are absent in the relatively benign *E. coli*. Of these genes, 70% are involved in the interaction with the host. This set of genes, however, is only small (17 genes) compared with the set that is shared between *H. influenzae* and *E. coli*, but absent in *H. pylori*[10] (508 genes). Thus, although the gene content shared between species qualitatively reflects correlations in phenotype, gene content shared quantitatively depends on genome size and phylogenetic position. A phenotypic feature such as hyperthermophily is, of course, also at least partly due to adaptations in the genes themselves rather than in gene content.

Reports of the horizontal transfer of large sets of genes, for example, into the *E. coli* genome[11], and from Bacteria to Archaea and Eukarya[1], have led to the view that horizontal gene transfer is a "major force"[1], rather than an interesting but anecdotal event. The correspondence of the genome tree with the 16s rRNA tree and the generally high bootstrap values show that gene content still carries a strong phylogenetic signature. Such a phylogenetic pattern is the result of the differential acquisition and loss of genes along the various evolutionary lineages, for example by expansion and shrinkage of gene families. The fact that gene content carries a strong phylogenetic signature implies that either there are relatively few horizontal transfer events, or the events occur mainly between closely related species or affect closely related species in the same manner (for example, when they predate their radiation), or the genes that are transferred generally replace an orthologous gene that is already present in the genome. Given the small number of sequenced genomes, a complete, quantitative model of genome evolution that includes probabilities of horizontal gene transfer, gene duplication and gene loss can not at present be parameterized.

Methods

Genes shared between two genomes were determined using an operational definition of orthology. After a Smith-Waterman comparison[12,13] of all the genes between two genomes, compared at the amino-acid level using a parallel Biocellerator computer (http://www.cgen.com), pairs of homologous sequences were selected using a cutoff value (E=0.01). E values in Smith-Waterman comparisons are reliable indicators of the ratio of false positives to true positives in homology detection[14]. From the resulting lists, we selected pairs of genes that are each other's 'closest relative' in their respective genomes: that is, the level of identity between the two genes is the highest when compared with the level of identity of each of the two genes with all the other genes in the other's genome. To include the possibility of fusion and splitting of genes, multiple genes from one genome can have the same single closest relative in another genome, as long as the alignments with this single gene do not overlap. The closest relative is an operational definition of 'orthology'[5], a concept introduced for genes whose independent evolution reflects a speciation event, rather than a gene duplication event, and who probably perform the same function. Orthology, however, is not an absolute, as it is a statement about the history of genes. The original concept does not include the possibility of horizontal gene transfer, and more elaborate criteria have been proposed for finding orthologous genes[2,15]. Such criteria lead to systematic biases in the number of orthologues that can be identified between species, the size of the bias depending on the evolutionary distance between the species[2]. Hence, they can not be used to construct a phylogenetic tree on the basis of gene content. Variations in the rate of sequence evolution only affect the results when they affect the detection of homology. Decreasing the E-value threshold to E=0.001 led to small changes in the fraction of genes with closest relatives between species (<3%), and did not change the topology of the clustering.

Acknowledgements

This work was supported by BMBF.

Received 19 August; accepted 12 November 1998.

1. Doolittle, W.F. & Logsdon, J.M. Archaeal genomics: do Archaea have a mixed heritage? *Curr. Biol.* **8**, R209–R211 (1998).
2. Huynen, M.A. & Bork, P. Measuring genome evolution. *Proc. Natl Acad. Sci. USA* **95**, 5849–5856 (1998).
3. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
4. Olsen, G.J., Woese, C.R. & Overbeek, R. The winds of (evolutionary) change: breathing new life into microbiology. *J. Bacteriol.* **176**, 1–6 (1994).
5. Fitch, W.M. Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**, 99–110 (1970).
6. Maidak, B.L. *et al.* The RDP (Ribosomal Database Project). *Nucleic Acids Res.* **25**, 109–111 (1997).
7. Klenk, H. & Zillig, W. DNA-dependent RNA polymerase subunit B as a tool for phylogenetic reconstructions: branching topology of the archaeal domain. *J. Mol. Evol.* **38**, 420–432 (1994).
8. Baldauf, S., Palmer, J.D. & Doolittle, W.F. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proc. Natl Acad. Sci. USA* **93**, 7749–7754 (1996).
9. Gruber, T.M. & Bryant, D.A. Molecular systematic studies of eubacteria, using σ^{70}-type σ factors of group 1 and group 2. *J. Bacteriol.* **179**, 1734–1747 (1997).
10. Huynen, M.A., Dandekar, T. & Bork, P. Differential genome analysis applied to the species-specific features of *Helicobacter pylori*. *FEBS Lett.* **426**, 1–5 (1998).
11. Lawrence, J.G. & Ochman, H. Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl Acad. Sci. USA* **95**, 9413–9417 (1998).
12. Smith, T. & Waterman, M.S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).
13. Pearson, W. Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276**, 71–84 (1998).
14. Brenner, S., Chotia, C. & Hubbard, T.J. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA* **95**, 6073–6078 (1998).
15. Tatusov, R.L., Koonin, E.V. & Lipman, D.J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).
16. Fleishmann, R. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae*. *Science* **269**, 496–512 (1995).
17. Fraser, C.M. *et al.* The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403 (1995).
18. Kaneko, T. *et al.* 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 (1996).
19. Bult, C.J. *et al.* Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**, 1058–1072 (1996).
20. Blattner, F.E. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462 (1997).
21. Smith, D.R. *et al.* Complete genome sequence of *Methanobacterium thermoautotrophicum* δH: functional analysis and comparative genomics. *J. Bacteriol.* **17**, 7135–7155 (1997).
22. Tomb, J.-F. *et al.* The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**, 539–547 (1997).
23. Klenk, H.P. *et al.* The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390**, 364–370 (1997).
24. Kunst, F. *et al.* The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249–256 (1997).
25. Fraser, C.M. *et al.* Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390**, 580–586 (1997).
26. Mewes, H.W. *et al.* Overview of the yeast genome. *Nature* **387**, 7–65 (1997).
27. Deckert, G. *et al.* The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **392**, 353–358 (1998).
28. Kawarabayasi, Y. *et al.* Complete sequence and gene organization of the genome of a hyper-thermophylic archaebacterium *Pyrococcus horikoshii* OT3. *DNA Res.* **5**, 55–76 (1998).
29. Wu, C.F.J. Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Stat.* **14**, 1261–1295 (1986).
30. Himmelreich, R., Plagens, H., Hilbert, H., Reiner, B. & Herrmann, R. Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. *Nucleic Acids Res.* **24**, 4420–4449 (1996).