# Whole genome-based phylogenetic analysis of free-living microorganisms

**Sorel T. Fitz-Gibbon* and Christopher H. House[1]**

Department of Microbiology and Molecular Genetics, University of California, 1602 Molecular Sciences Building, 405 Hilgard Avenue, Los Angeles, CA 90095-1489, USA and [1]Department of Earth and Space Sciences and the IGPP Center for Astrobiology, University of California, Los Angeles, CA 90095-1567, USA

Received July 7, 1999; Revised and Accepted September 10, 1999

## ABSTRACT

**A phylogenetic 'tree of life' has been constructed based on the observed presence and absence of families of protein-encoding genes observed in 11 complete genomes of free-living microorganisms. Past attempts to reconstruct the evolutionary relationships of microorganisms have been limited to sets of genes rather than complete genomes. Despite apparent rampant lateral gene transfer among microorganisms, these results indicate a single robust underlying evolutionary history for these organisms. Broadly, the tree produced is very similar to the small subunit rRNA tree although several additional phylogenetic relationships appear to be resolved, including the relationship of *Archaeoglobus* to the methanogens studied. This result is in contrast to notions that a robust phylogenetic reconstruction of microorganisms is impossible due to their genomes being composed of an incomprehensible amalgam of genes with complicated histories and suggests that this style of genome-wide phylogenetic analysis could become an important method for studying the ancient diversification of life on Earth. Analyses using informational and operational subsets of the genes showed that this 'tree of life' is not dependent on the phylogenetically more consistent informational genes.**

## INTRODUCTION

Historically, determining the phylogenetic relationships of microorganisms was difficult due to the lack of discernable morphological characters. The phylogenetic analysis of universally conserved nucleic acid or protein sequences (in particular the small subunit rRNA gene; 1) subsequently became a powerful tool for microbial taxonomy, allowing the taxonomic identification of microorganisms with only a single gene sequence. However, in spite of the success of rRNA microbial taxonomy, the evolutionary relationships between major groups of prokaryotes is still unclear because phylogenetic analyses of single gene sequences lack the information to resolve these deep branches. Further, misalignment and differing evolutionary rates can result in phylogenetic trees with the wrong topology (2). Also, the horizontal transfer of genes from one species to another provides a means by which each gene may tell of an independent history.

The recently completed sequences of several microbial genomes provide an enormous amount of data with which to address these problems, however, the task of interpreting this genome data is difficult. Careful alignments and phylogenetic analyses of large numbers of conserved proteins give inconsistent phylogenetic results (3–5). The extent of these inconsistencies has led to speculation that there may not be a single tree that can be used to represent the history of life on Earth. Here we present a phylogenetic tree based on the observed presence and absence of protein-encoding gene families found in 11 genomes of free-living microorganisms. This method of phylogenetic analysis is analogous to using the distribution of morphological features observed in a set of organisms to determine their phylogenetic relationships. However, unlike most morphological characters, amino acid sequences are unlikely to be highly similar unless they share a common ancestor, eliminating the problems associated with convergent evolution. Here, using the observed presence and absence of protein encoding genes as the basis of a phylogenetic reconstruction, we show that there is a strong signal within the genomes reflecting the evolutionary histories of the organisms despite horizontal gene transfer, gene duplication and gene loss.

## MATERIALS AND METHODS

For this analysis, we used all of the published complete genome data for free-living microorganisms (6–15), plus the soon to be published genome data for the free-living crenarchaeon *Pyrobaculum aerophilum* (16).

### Construction of data matrices

The first step in this analysis groups proteins based on pairwise sequence similarity. Comparisons were done using the FASTA3 (17) software, comparing each protein sequence in turn to each of the 11 databases of all protein sequences for each organism. The proteins were grouped if any pairwise similarity score was greater than a preset *z*-score (17,18) regardless of the length of the matching region or the relative lengths of the proteins. FASTA3 *z*-scores are based on an

*To whom correspondence should be addressed. Tel: +1 310 825 8924; Fax: +1 310 206 3088; Email: sorel@mbi.ucla.eu

The authors wish it to be known that, in their opinion, both authors should be regarded as joint First Authors

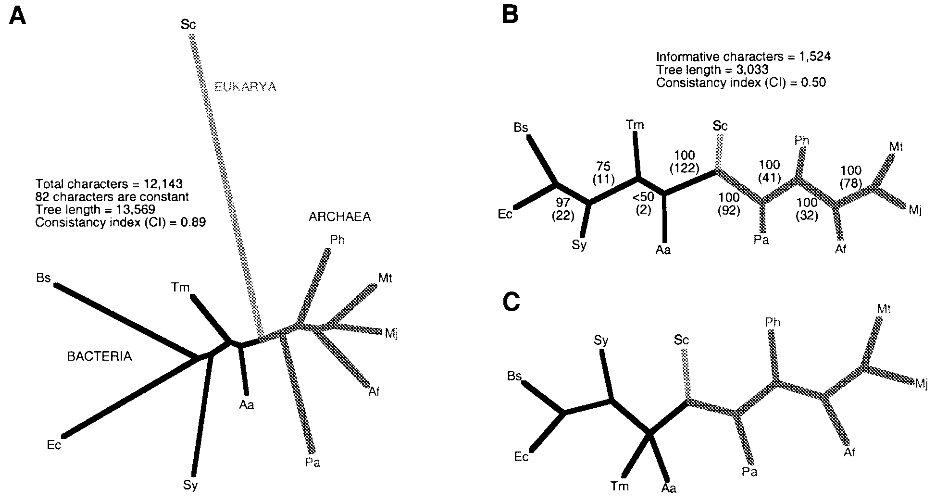

**Figure 1.** Phylogenetic analysis of 11 free-living microorganisms, *A.aeolicus* (6) (Aa), *A.fulgidus* (7) (Af), *B.subtilis* (8) (Bs), *E.coli* (9) (Ec), *M.jannaschii* (10) (Mj), *M.thermoautotrophicum* (11) (Mt), *P.aerophilum* (S.T.Fitz-Gibbon *et al*., in preparation) (Pa), *P.horikoshii* (12) (Ph), *S.cerevisiae* (13) (Sc), *Synechocystis* sp. (14) (Sy) and *T.maritima* (15) (Tm) using the presence/absence of protein groups in the each genome as characters. (**A**) The single most parsimonious phylogram (unrooted) produced when using a *z*-score cut-off of 170 as the criterion for the identification of homologous protein groups. (**B**) The same phylogram shown with only the parsimony informative characters. Because 10 619 of the 12 143 protein groups identified at this *z*-score cut-off are unique to only one taxon (and therefore phylogenetically uninformative), the long terminal branches evident in (A) are not present in (B). Values obtained from 1000 bootstrap replicates are indicated along with the decay index (in parentheses) (21). (**C**) The consensus topology for all of the tested *z*-score cut-offs in the range 140–500 for both maximum parsimony and distance (neighbor joining) methods.

extreme value distribution and scaled to have a mean of 50 and a standard deviation of 10. The presence or absence of each protein group was scored for each genome to construct the data matrix for phylogenetic analysis. By grouping all recognizable members of gene families into the same group (even protein sequences linked via another intermediate protein sequence or via a fused multidomain protein), protein families of varying sizes among the genomes do not influence our phylogenetic analysis. The data matrices are available on the World Wide Web at http://www.astrobiology.ucla.edu/data/nar1/treedata.html

### Phylogenetic analysis and statistics

Parsimony and distance analyses were performed using PAUP v.4.0b1 (Sinauer Associates) for a series of data matrices derived using the following *z*-score cut-offs: 140, 150, 160, 170, 180, 190, 200, 300, 500, 1000 and 2000. Bootstrap scores and consistency indices were calculated using PAUP v.4.0b1. The character consistency index is equal to $m_i/s_i$, where $m_i$ is the minimum conceivable number of steps for character *i* on any tree (always equal to 1 for our binary data), and $s_i$ is the number of reconstructed steps for character *i* on this tree (19). The consistency index for all characters on a tree is the minimum possible tree length divided by the observed tree length (20). The decay index is defined as the number of additional steps required to collapse the branch in question (21) and was calculated using AUTODECAY v.3.0 and PAUP v.4.0b1.

### Character subsets

Characters were selected for inclusion into 'informational' or 'operational' subsets on the basis of published gene annotations for all organisms except *Aquifex aeolicus*, *P.aerophilum* and *Pyrococcus horikoshii.* Characters labeled as 'informational' or 'operational' had at least one member categorized as 'informational' or 'operational' respectively (see below). Any category which included proteins involved in 'replication, transcription and translation' was labeled as 'informational'. Other categories (excluding 'hypotheticals' and 'unknowns') were labeled as operational. Characters that had members in both 'informational' and 'operational' subsets were excluded from the analysis.

## RESULTS

### Phylogenetic analysis using whole genomes

Maximum parsimony and distance analysis of the distribution of protein coding genes results in a tree topology (Fig. 1) which is very similar to that predicted by phylogenetic analysis of small subunit rRNA sequences (1). The tree is extremely well supported by the data, as indicated by the bootstrap values (22), consistency indices (19), decay indices (21; AUTODECAY v.3.0), and the consistency across differing *z*-score cut-offs.

Figure 1A and B show the parsimony results using a *z*-score cut-off of 170. The phylogenetic relationships between the members of the bacterial domain are identical to those predicted by the small subunit rRNA in spite of the weak resolution of these groups in small subunit rRNA analyses (23–25) and the independent nature of these two types of phylogenetic analysis.

Figure 1C shows the consistency across a wide range of *z*-score cut-offs using both tree building algorithms. The only unresolved position on this consensus tree is the relative branching order of *A.aeolicus* and *Thermotoga maritima.* Despite the ambiguity of the relative positions of these taxa, these results increase our

**Table 1.** Results of 'alternative trees' analysis

| Alternative trees | Number of informative characters (consistency index = 1) | Tree topology |
|---|---|---|
| Figure 1 | 456 (30%) | ((((ec,bs),sy),tm),aa),(sc,(pa,(ph,(af,(mj,mt))))) |
| 1 | 113 (7%) | (((ec,sy),aa),(tm,bs)),(sc,(pa,(mt,(mj,(af,ph))))) |
| 2 | 92 (6%) | (((aa,tm),(ec,bs)),sy),(sc,(mt,((mj,ph),(af,pa)))) |
| 3 | 86 (6%) | ((((ec,sc),(sy,bs)),aa),tm),(af,((ph,pa),(mj,mt))) |
| 4 | 41 (3%) | ((((ec,bs),tm),sy),aa),(sc,(pa,(ph,(mj,(mt,af))))) |
| 5 | 10 (1%) | (((ec,bs),aa),(tm,sy)),(sc,(pa,(ph,(af,(mj,mt))))) |
| 6 | 23 (2%) | (((ec,bs),(sy,aa)),tm),(sc,(pa,(ph,(af,(mj,mt))))) |
| 7 | 0 | (((sy,ec),aa),(bs,tm)),(sc,(pa,(ph,(af,(mj,mt))))) |

Alternative phylogenetic trees constructed (using maximum parsimony) after sequentially removing all informative characters with a consistency index of 1 (all characters entirely consistent with the former tree topology) starting with the tree in Figure 1A. The number of characters with a consistency index of 1 is also shown for each tree (along with the percentage these characters represent of the original 1524 informative characters).

confidence that both of these organisms are 'early diverging', as suggested by small subunit rRNA sequences.

Surprisingly, within the archaeal domain the two methanogens form a monophyletic group with the exclusion of *Archaeoglobus fulgidus*, whereas in most small subunit rRNA trees the methanogens are paraphyletic, with *Methanobacterium thermoautotrophicum* as the sister taxon to *A.fulgidus*. The support for the pairing we observed is dominated by the presence of 81 gene groups (at a *z*-score cut-off of 170) unique to the methanogen genomes, which include 64 protein groups of unknown function, 11 protein groups involved in methanogenesis (merABCDG and mtrABCDEG), and a seryl-tRNA synthetase group.

## 'Alternative trees' analysis

In the analysis shown in Figure 1, 30% (456/1524) of the informative characters have a consistency index of 1, i.e. a distribution entirely consistent with the tree topology. Since this represents a minority of the data, we tested for strong alternative topologies not detected by the bootstrap analysis. Starting with the tree shown in Figure 1, we removed all gene groups (characters) with a consistency index of 1 and created a phylogenetic tree with the remaining (non-consistent) characters. This process favors the formation of an alternative tree topology. The procedure was then repeated until the resulting alternative tree had no characters with a consistency index of 1. A total of seven alternative trees were formed (Table 1). In general, the results failed to show any single strong alternative hypothesis. One of the alternative trees constructed places yeast well within the bacterial domain, as a sister taxon to *Escherichia coli*, reflecting the presence of bacterial genes in this eukaryote. None of the alternative trees have yeast within the Archaea, and none have archaeal and bacterial taxa mixed together.

## Analysis without 'bacterial' *Saccharomyces cerevisiae* genes

In Figure 1, *S.cerevisiae* (yeast) falls between the Bacteria and Archaea as in small subunit rRNA. Because this result may be an artifact of the chimeric nature of the eukaryotic genome, we repeated the analysis after removing all yeast genes in which the best match to a bacterial gene was stronger than the best match to an archaeal gene. The resulting tree is identical to the tree produced with the total data set. This suggests that the placement of this taxon is not biased by its chimeric nature. However, we are not able to rule out the possibility that unequal rates of evolution are affecting yeast's position.

## Analysis of 'informational' and 'operational' subsets of genes

As archaeal genome sequences became available, it became clear that archaeal proteins involved in informational processes (replication, transcription and especially translation) are more frequently eukaryal-like than are proteins involved in other cellular processes (4,7,10,11,26). Jain *et al.* documented that individual protein phylogenetic trees for informational molecules are more likely to be congruent with the small subunit rRNA tree, while trees built for other types of proteins ('operational') showed greater variations in their topologies (27). In order to test whether our tree topology was dependant upon the phylogenetically more consistent informational proteins we repeated our analysis on two subsets of characters identified as 'informational' or 'operational' (3). The resulting trees (Fig. 2) both have similar topologies to the tree derived from the complete data set (Fig. 1) with minor branching differences (weakly supported and probably due to the smaller number of characters) in fact found only for the informational subset. Thus even the phylogenetically dissonant set of operational genes still yield an overall pattern which is consistent with the small subunit rRNA tree. Interestingly, the 'informational' tree has a striking increase in the length of the branch separating the Bacteria from the Archaea and yeast, further reflecting the intriguing fundamental difference between these two sets of genes.

## Analysis including non-free-living microorganisms

Although the results of this phylogenetic reconstruction of free-living microorganisms is well supported, the inclusion of genomes of non-free-living taxa [i.e. *Helicobacter pylori* (28), *Mycobacterium tuberculosis* (29) and *Haemophilus influenzae* (30)] yields mixed results. Although the expected pairing of *H.influenzae* with its close relative *E.coli* is easily resolved, the inclusion of the other two pathogenic bacteria greatly lowered

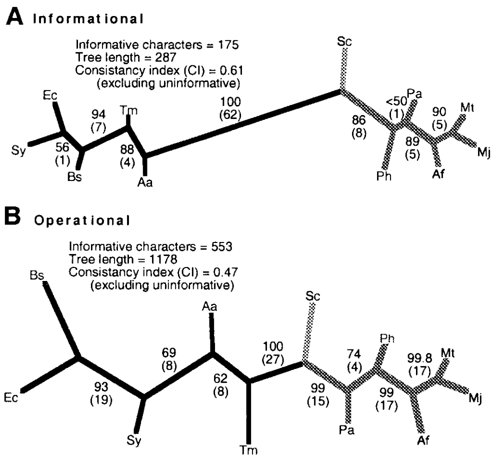

**Figure 2.** Phylogenetic analysis of the 11 microorganisms shown in Figure 1 using only 'informational' genes (**A**) or only 'operational' genes (**B**). The *z*-score cut-off was 170. Values obtained from 1000 bootstrap replicates (of informative characters only) are indicated along with the decay index (in parentheses). Both the 'informational' maximum parsimony tree and the 'operational' maximum parsimony tree have a similar topology to the tree produced with the full complement of genes (Fig. 1).

the statistical support within the bacterial domain, implying that the analysis could not clearly resolve their phylogenetic position.

The most parsimonious tree (at a *z*-score cut-off of 170) using the 11 free-living microorganisms and the three non-free-living taxa does not have a tree topology consistent with the tree shown in Figure 1. The differences are that in the new topology, *T.maritima* is a sister group to *Bacillus subtilis* and *A.aeolicus* is a sister group to the rest of the Proteobacteria (*H.pylori*, *H.influenzae* and *E.coli*). These changes to the topology seem to be an artifact of a complex interaction principally involving *H.pylori* and *T.maritima*. Exclusion of either of these two restores the tree to a topology consistent with the one shown in Figure 1 whereas the exclusion of any other single taxa does not. We believe that the 1.6 Mb genome of *H.pylori* is not representative of the early Proteobacteria and therefore creates an artifact when included in this phylogenetic reconstruction.

## DISCUSSION

This analysis shows that a robust tree topology can be produced based on the taxonomic distribution of the gene groups found in the genomes of free-living organisms. The results are remarkably similar to results from phylogenetic analysis of the small subunit rRNA gene, increasing confidence in both methods and suggesting that a tree of life can indeed be constructed and used to understand early microbial evolution.

Similar to small subunit rRNA analysis, the deepest branches can be difficult to resolve. In this analysis, the relative positions of *A.aeolicus* and *T.maritima* are uncertain. Satisfactory resolution of these branches may be possible in the future by inclusion of data from additional genomes, particularly from other early diverging bacteria. However, in the meantime, despite the uncertainty of the branching order, our results do support the position of these two lineages as the deepest branches of the bacterial domain, a notion that has been challenged (15,31).

Within the archaeal domain, all branches have 100% bootstrap support and relatively high decay indices. This uniformly strong support includes the non-traditional pairing of the two methanogens, *M.thermoautotrophicum* and *Methanococcus jannaschii*. This topology implies a different evolutionary history for these taxa than is currently accepted, although it is possible that the small subunit rRNA tree is correct and gene groups were lost en masse from the *A.fulgidus* lineage. We suspect our topology is correct, as the support for the traditional pairing of *M.thermoautotrophicum* with *A.fulgidus* is not conclusive. Analysis of other genes, including large subunit rRNA (32) and radA (33), support, albeit weakly, the pairing of these methanogens.

The 'alternative trees' analysis tested for strong alternative topologies missed by the bootstrap tests. None were found in which the yeast fell within the Archaea, or in which archaeal and bacterial taxa mixed together. The 'alternative' tree analysis did, however, pick up an affinity of yeast to *E.coli*, perhaps reflecting components of the yeast genome derived from the proto-mitochondrion.

The only significant difference between the analysis of the two subsets of genes, 'informational' and 'operational', was in the relative length of the branch separating the bacterial domain from the yeast and Archaea, which was longer in the informational gene subset. This difference is consistent with the observation by Rivera *et al.* (3) of generally shorter distances (substitutions/position) between *M.jannaschii* and bacterial ortholog pairs for 'operational' versus 'informational' genes. It has been proposed that the shorter distances between the operational orthologs may be due to more frequent horizontal transfers (3), or at least more recent horizontal transfers (34), perhaps due to more stringent requirements for precise interactions between the informational molecules (27,34). Alternatively, when fundamental differences in informational processing between the Bacteria and Archaea arose in evolution, there may have been rapid rates of amino acid replacement in the entire suite of informational genes, i.e. widespread co-evolution of informational genes.

Analysis of eukaryotes by this method is currently limited since a reasonably complete set of predicted proteins from whole genome data is available for only one eukaryote (yeast). The placement of yeast in Figure 1 is consistent with its position in the small subunit rRNA tree, allowing there to be three distinct domains of life. However, this placement will remain tenuous pending both further eukaryotic representatives and a better ability to control for the effects of unequal rates of evolution.

## ACKNOWLEDGEMENTS

We thank Charles Marshall (University of California Los Angeles) for critical reading of an earlier version of the manuscript. This work was supported by a grant to Jeffrey H. Miller from the National Institutes of Health (GM57917) and by the UCLA Center for Astrobiology, NASA Astrobiology Institute. IGPP publication no. 5259

## REFERENCES

1. Woese,C.R., Kandler,O. and Wheelis,M.L. (1990) *Proc. Natl Acad. Sci. USA*, **87**, 4576–4579.
2. Marshall,C.R. (1997) *Comput. Sci. Statist.*, **29**, 218–226.
3. Rivera,M.C., Jain,R., Moore,J.E. and Lake,J.A. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 6239–6244.
4. Ribeiro,S. and Golding,G.B. (1998) *Mol. Biol. Evol.*, **15**, 779–788.
5. Feng,D.F., Cho,G. and Doolittle,R.F. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 13028–13033.
6. Deckert,G., Warren,P.V., Gaasterland,T., Young,W.G., Lenox,A.L., Graham,D.E., Overbeek,R., Snead,M.A., Keller,M., Aujay,M. *et al*. (1998) *Nature*, **392**, 353–358.
7. Klenk,H.P., Clayton,R.A., Tomb,J.F., White,O., Nelson,K.E., Ketchum,K.A., Dodson,R.J., Gwinn,M., Hickey,E.K., Peterson,J.D. *et al*. (1997) *Nature*, **390**, 364–370.
8. Kunst,F., Ogasawara,N., Moszer,I., Albertini,A.M., Alloni,G., Azevedo,V., Bertero,M.G., Bessiaeres,P., Bolotin,A., Borchert,S. *et al*. (1997) *Nature*, **390**, 249–256.
9. Blattner,F.R., Plunkett,G.R., Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al*. (1997) *Science*, **277**, 1453–1474.
10. Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleischmann,R.D., Sutton,G.G., Blake,J.A., FitzGerald,L.M., Clayton,R.A., Gocayne,J.D. *et al*. (1996) *Science*, **273**, 1058–1073.
11. Smith,D.R., Doucette-Stamm,L.A., Deloughery,C., Lee,H., Dubois,J., Aldredge,T., Bashirzadeh,R., Blakely,D., Cook,R., Gilbert,K. *et al*. (1997) *J. Bacteriol*., **179**, 7135–7155.
12. Kawarabayasi,Y., Sawada,M., Horikawa,H., Haikawa,Y., Hino,Y., Yamamoto,S., Sekine,M., Baba,S., Kosugi,H., Hosoyama,A. *et al*. (1998) *DNA Res.*, **5**, 55–76.
13. Goffeau,A., Authora,A. and Authorb,B. (1997) *Nature*, **387**, 5.
14. Kaneko,T., Sato,S., Kotani,H., Tanaka,A., Asamizu,E., Nakamura,Y., Miyajima,N., Hirosawa,M., Sugiura,M., Sasamoto,S. *et al*. (1996) *DNA Res.*, **3**, 109–136.
15. Nelson,K.E., Clayton,R.A., Gill,S.R., Gwinn,M.L., Dodson,R.J., Haft,D.H., Hickey,E.K., Peterson,L.D., Nelson,W.C., Ketchum,K.A. *et al*. (1999) *Nature*, **399**, 323–329.
16. Fitz-Gibbon,S., Choi,A., Miller,J.H., Stetter,K.O., Simon,M., Swanson,R. and Kim,U.-J. (1997) *Extremophiles*, **1**, 36–51.
17. Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
18. Pearson,W.R. (1995) *Protein Sci.*, **4**, 1145–1160.
19. Kluge,A.G. and Farris,J.S. (1969) *Syst. Zool.*, **18**, 1–32.
20. Farris,J.S. (1989) *Cladistics*, **5**, 417–419.
21. Bremer,K. (1988) *Evolution*, **42**, 795–803.
22. Felsenstein,J. (1985) *Evolution*, **39**, 783–791.
23. Pace,N.R. (1997) *Science*, **276**, 734–740.
24. Hugenholtz,P., Pitulle,C., Hershberger,K.L. and Pace,N.R. (1998) *J. Bacteriol*., **180**, 366–376.
25. Woese,C.R. (1987) *Microbiol. Rev.*, **51**, 221–271.
26. Koonin,E.V., Mushegian,A.R., Galperin,M.Y. and Walker,D.R. (1997) *Mol. Microbiol.*, **25**, 619–637.
27. Jain,R., Rivera,M.C. and Lake,J.A. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 3801–3806.
28. Tomb,J.F., White,O., Kerlavage,A.R., Clayton,R.A., Sutton,G.G., Fleischmann,R.D., Ketchum,K.A., Klenk,H.P., Gill,S., Dougherty,B.A. *et al*. (1997) *Nature*, **388**, 539–547.
29. Cole,S.T., Brosch,R., Parkhill,J., Garnier,T., Churcher,C., Harris,D., Gordon,S.V., Eiglmeier,K., Gas,S., Barry,C.E. *et al*. (1998) *Nature*, **393**, 537–544.
30. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. *et al*. (1995) *Science*, **269**, 496–512.
31. Brown,J.R. and Doolittle,W.F. (1997) *Microbiol. Mol. Biol. Rev.*, **61**, 456–502.
32. Maidak,B.L., Cole,J.R., Parker,C.T.,Jr, Garrity,G.M., Larsen,N., Li,B., Lilburn,T.G., McCaughey,M.J., Olsen,G.J., Overbeek,R. *et al*. (1999) *Nucleic Acids Res.*, **27**, 171–173.
33. Sandler,S.J., Hugenholtz,P., Schleper,C., DeLong,E.F., Pace,N.R. and Clark,A.J. (1999) *J. Bacteriol*., **181**, 907–915.
34. Woese,C. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 6854–6859.