Phylogeny of *Firmicutes* with special reference to *Mycoplasma* (*Mollicutes*) as inferred from phosphoglycerate kinase amino acid sequence data

Matthias Wolf,[1] Tobias Müller,[1] Thomas Dandekar[1] and J. Dennis Pollack[2]

[1]University of Würzburg, Department of Bioinformatics, Biocenter, Am Hubland, 97074 Würzburg, Germany

[2]Department of Molecular Virology, Immunology and Medical Genetics, Ohio State University, 2066 Graves Hall, 333 West 10th Avenue, Columbus, OH 43210, USA

Correspondence
Thomas Dandekar
dandekar@biozentrum.uni-wuerzburg.de

J. Dennis Pollack
pollack.1@osu.edu

The phylogenetic position of the *Mollicutes* has been re-examined by using phosphoglycerate kinase (Pgk) amino acid sequences. Hitherto unpublished sequences from *Mycoplasma mycoides* subsp. *mycoides*, *Mycoplasma hyopneumoniae* and *Spiroplasma citri* were included in the analysis. Phylogenetic trees based on Pgk data indicated a monophyletic origin for the *Mollicutes* within the *Firmicutes*, whereas *Bacilli* (*Firmicutes*) and *Clostridia* (*Firmicutes*) appeared to be paraphyletic. With two exceptions, i.e. *Thermotoga* (*Thermotogae*) and *Fusobacterium* (*Fusobacteria*), which clustered within the *Firmicutes*, comparative analyses show that at a low taxonomic level, the resolved phylogenetic relationships that were inferred from both the Pgk protein and 16S rRNA gene sequence data are congruent.

INTRODUCTION

Morphologically and microbiologically, *Mollicutes* are classified as *Bacteria* that were probably derived from lactobacilli, bacilli or streptococci by regressive evolution and genome reduction, to produce the smallest and simplest free-living and self-replicating cells (Razin *et al.*, 1998). Their lifestyle is, in general, parasitic. Structurally, *Mollicutes* are characterized by the complete lack of cell wall and the presence of an internal cytoskeleton (Balish & Krause, 2002; Dandekar *et al.*, 2002).

Based on 16S rRNA data, the taxonomy, as well as the phylogeny and evolution, of *Mollicutes* have recently been discussed (Johansson & Pettersson, 2002; Maniloff, 2002). By phylogenetic analysis, low-G+C, Gram-positive *Bacteria* (*Firmicutes*) comprise three groups: *Bacilli*, *Clostridia* and *Mollicutes*. However, based on 16S rRNA gene sequence data, only the *Mollicutes* are well-supported as being monophyletic (Ludwig & Klenk, 2001).

Published online ahead of print on 12 December 2003 as DOI 10.1099/ijs.0.02868-0.

Abbreviations: ML, maximum-likelihood; MP, maximum-parsimony; NJ, neighbour-joining; Pgk, phosphoglycerate kinase.

The GenBank/EMBL/DDBJ accession numbers for the Pgk sequences of *Mycoplasma mycoides* subsp. *mycoides*, *Spiroplasma citri* and *Mycoplasma hyopneumoniae* are BX293980, AJ580006 and AY319328, respectively.

In this study, we present the results of our analysis that used phosphoglycerate kinase (Pgk) amino acid sequences as a molecular marker, instead of 16S rRNA, to examine the phylogeny of *Firmicutes* taxa. Pgk is one of the oldest 'housekeeping' enzymes; its evolutionary time has been estimated to be about 40 million years, which is about twice as long as was required for 1% mutation to occur in cytochrome *c* or glyceraldehyde-3-phosphate dehydrogenase (Ciccarese *et al.*, 1989). Other reports consider that Pgk is evolving at a linear rate of four to six accepted point mutations in 100 million years, i.e. about the same rate as for cytochrome *c* (Fothergill-Gilmore, 1986). Even for a 'housekeeping' enzyme, this is a very conserved sequence (Fothergill-Gilmore, 1986; Fothergill-Gilmore & Michels, 1993). The *pgk* gene may be an example of a 'core' household gene (Daubin *et al.*, 2002). The metabolic role of Pgk, especially in *Mollicutes*, has recently been discussed (Pollack *et al.*, 2002). The role of Pgk is particularly consequential in *Mollicutes*, as these bacteria lack cytochrome pigments and the citric acid cycle and are thought to synthesize most of their ATP by substrate phosphorylation during glycolysis, mediated by the presumably essential action of Pgk and pyruvate kinase (Pollack, 2002).

The focus of this study is on the phylogenetic position of the amino acid sequences of *Mollicutes* Pgks and their relationships to the 16S rRNA *Mollicutes* subgroups that were established by Johansson *et al.* (1998). We included unpublished Pgk sequences from *Mycoplasma mycoides*

subsp. *mycoides*, *Mycoplasma hyopneumoniae* and *Spiroplasma citri*. Furthermore, the utility of Pgk as a phylogenetic marker to analyse the phylogeny and evolution of *Firmicutes* is discussed.

In initial studies, Pollack (2002) hypothesized that Pgk should be an attractive marker for phylogenetic analyses. Similarly, our own preliminary analyses of the *pgk* gene were also performed by using a computational neighbour-joining (NJ) method to analyse 100 complete *pgk* gene sequences from different life forms (*Bacteria*, *Archaea* and *Eukarya*). We found relationships and clustering that were similar or identical to those already established by microbiological and phenotypical criteria. Individual and entire groupings, e.g. *Crenarchaeota*, *Euarchaeota*, ciliates, fungi, plants, mammals and the α, β and most of the γ divisions of the *Eubacteria* were entirely separable. Furthermore, the NJ branching order of the *Mollicutes* exactly followed the predicted 16S rRNA groupings that were described by Johansson *et al.* (1998). Of additional interest was the closer relationship of the *Mollicutes* to the low-G+C non-spore-formers *Staphylococcus aureus* and *Lactobacillus delbrueckii* subsp. *bulgaricus*, and their greater distance from the low-G+C spore-formers *Bacillus* and *Clostridium* spp. The NJ-derived Pgk tree suggested that *Mycoplasma* spp. are related more closely to the *Streptococcus*/*Lactobacillus* subgroups and less closely to *Bacillus* and *Clostridium* spp. The relationship of the *Mollicutes* to *Streptococcus*/*Lactobacillus* was first reported by Neimark (1979), who concluded that the *Mollicutes* (acholeplasmas) descended from this group. His view was based on both the similarity of their fructose 1,6-bisphosphate-activated lactate dehydrogenases and the immunological homology of their aldolases.

In this study, the Pgk tree was re-examined and tested by more sensitive and rigorous computational means than the NJ techniques. Here, we have used distance, maximum-parsimony (MP) and maximum-likelihood (ML) methods to calculate a Pgk-based phylogenetic tree of the *Firmicutes*.

METHODS

We studied all available complete amino acid Pgk sequences of *Firmicutes*, plus related sequences (*Fusobacterium* and *Thermotoga*), firstly by using BLAST (basic local alignment search tool), i.e. iterative sequence alignment procedures (Altschul *et al.*, 1997). We included three previously unpublished *Mollicutes* sequences. Alignment and direct comparison of amino acid sequences (GenBank accession numbers are given in Fig. 1) were performed with CLUSTALX (Thompson *et al.*, 1994) and the Windows-based multi-sequence alignment editor of Hepperle (2002).

An MP analysis of aligned amino acid sequences was conducted by using PAUP* version 4.0b10 win32 (Swofford, 2002). Heuristic searches with 10 random taxon addition replicates and tree bisection–reconnection swapping were applied. The MulTrees and Collapse options of PAUP* were used and character changes were interpreted with ACCTRAN optimization. Characters were weighted equally and coded as unordered; gaps were treated as missing data. Bootstrap support was estimated, based on 100 replicates.

For an ML analysis that used the WAG model (Whelan & Goldman, 2001), TREEPUZZLE (Strimmer & von Haeseler, 1996) was used with 10 000 quartet-puzzling steps. Further, using the JTT model (Jones *et al.*, 1992), an ML topology was obtained by using PROML, as implemented in PHYLIP 3.6 (Felsenstein, 1993).

Additionally, by using default settings, two NJ trees (Saitou & Nei, 1987) were generated by using TREECON for Windows (Van de Peer & De Wachter, 1994) and PHYLIP 3.6. Bootstrap support was estimated, based on 500 and 100 replicates, respectively. All trees were rooted by using *Nostoc* sp. (Q8YPR1) and displayed by using TREEVIEW (Page, 1996).

RESULTS AND DISCUSSION

From comparison of different substitution models and different tree phylogeny algorithms, molecular phylogenetic analyses of Pgk sequences resulted in identical tree topologies with high bootstrap support values (Fig. 1).

The monophyletic *Mollicutes* (bootstrap support of up to 95 %) are the sister group of *Bacilli* II plus *Fusobacterium* (*Fusobacteria*) (Fig. 1). This large clade (*Mollicutes*, *Bacilli* II and *Fusobacterium*) is the sister group to *Thermotoga* (*Thermotogae*), which, like *Fusobacterium*, does not belong to the *Firmicutes*. Because of the relative positions of *Fusobacterium* and *Thermotoga*, the *Firmicutes* appear to be polyphyletic. *Bacilli* I (Fig. 1), followed by *Clostridia* I and II (Fig. 1), clustered at basal positions within this polyphyletic assemblage that comprises low-G+C, Gram-positive *Bacteria* (*Firmicutes*).

It is of note that, regardless of the algorithm applied, *Bacilli* and *Clostridia* appeared to be paraphyletic. Within the *Mollicutes*, the pneumoniae group, the spiroplasma group and the hominis group (the latter represented by *Mycoplasma pulmonis* and *M. hyopneumoniae*) are supported strongly (Fig. 1). The grouping of *M. pulmonis* within the pneumoniae group and the respective positions of *Bacillus subtilis*, *Bacillus halodurans*, *Streptococcus pyogenes* and *Lactococcus lactis* also reflect and support genome trees that were established by gene content or gene order (Dandekar *et al.*, 2002). Hitherto unpublished sequences clustered as follows: *M. hyopneumoniae*, as mentioned above, is the sister group to *M. pulmonis*. *M. mycoides* subsp. *mycoides* is the sister group to *Mycoplasma capricolum* subsp. *capricolum* and *Spiroplasma citri* is the sister group to the *M. mycoides*/*M. capricolum* cluster.

At a low taxonomic level, especially within the *Mollicutes*, the resolved phylogenetic relationships inferred from both protein and 16S rRNA gene sequence data are congruent. The nine *Mollicutes* clustered into three Pgk subclades (Fig. 1). The members of each of these Pgk subclades were, without exception, the same as previously grouped by 16S rRNA analyses (Johansson *et al.*, 1998; Johansson & Pettersson, 2002): 16S rRNA group III (*Mycoplasma pneumoniae*, *Mycoplasma genitalium*, *Ureaplasma urealyticum* and *Mycoplasma penetrans*), 16S rRNA group IV

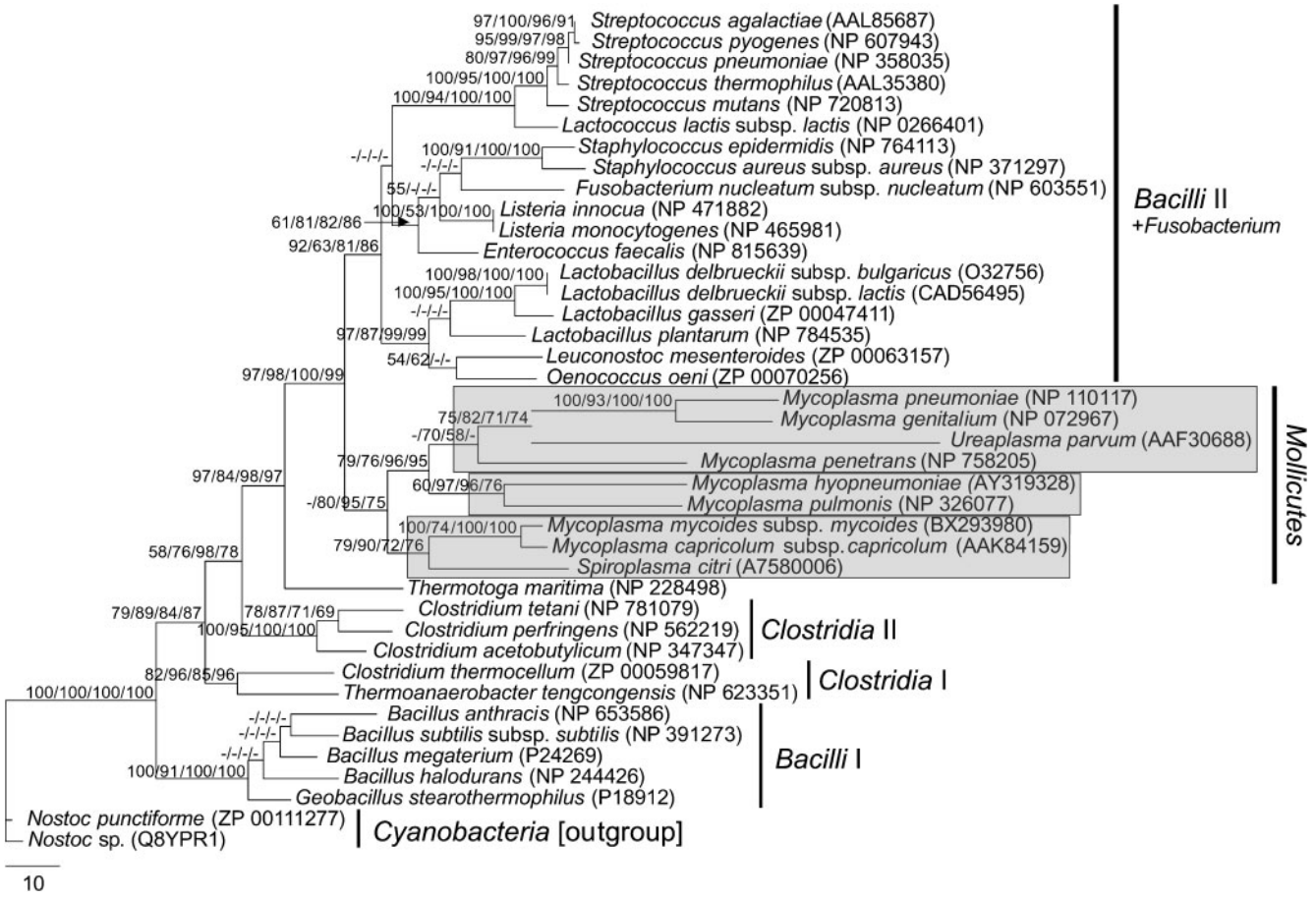

Fig. 1. ML tree, as derived from PROML analysis of Pgk sequences. The outgroup is forced. Numbers at nodes indicate bootstrap support values (>50 %) for clusters to the right of them, as calculated by MP/ML (TREEPUZZLE)/NJ (PHYLIP)/ NJ (TREECON). Boxed areas enclose the following 16S rRNA subgroups of *Mollicutes* (Johansson *et al*., 1998): pneumoniae group III, hominis group IV and spiroplasma group II.

(*M. hyopneumoniae* and *M. pulmonis*) and 16S rRNA group II (*M. mycoides* subsp. *mycoides*, *M. capricolum* subsp. *capricolum* and *S. citri*).

Furthermore, the non-spore-forming genera *Streptococcus*, *Staphylococcus*, *Listeria* and *Lactobacillus* (clustering within *Bacilli* II) and the spore-forming genera *Geobacillus* and *Bacillus* (*Bacilli* I) are well-supported as being monophyletic (Fig. 1). The genus *Mycoplasma* appeared to be paraphyletic (because of *Ureaplasma* and *Spiroplasma*) and *Clostridium* is also paraphyletic in this analysis (Fig. 1). A more schematic (unrooted) representation of relationships within the *Firmicutes* is given in Fig. 2.

We believe that Pgk may be used as an appropriate marker, yielding high bootstrap support values, for phylogenetic analysis of *Firmicutes* taxa. In most cases, we found that support is even higher than that found for 16S rRNA analyses (data not shown).

Although *Fusobacterium nucleatum* is an anaerobic, Gram-negative bacterium that is not classified among the low-G+C, Gram-positive *Firmicutes*, it was studied because of sequence similarities obtained by BLAST and because several of its core metabolic features have been reported to be similar to those of *Clostridium* spp., *Enterococcus* spp. and *Lactococcus* spp. (Kapatral *et al*., 2002). However, like other members of the *Firmicutes*, it has a low DNA G+C content of 27 %. Its cell-wall structure results in its Gram-negative character. Our analyses placed it in the *Bacilli* II group. As reported by Kapatral *et al*. (2002), it possesses distinguishing genomic features that are found in some members of the *Bacilli* II group: clustering of ORFs, 16S rRNA gene sequence similarities, uracil monophosphate biosynthesis, Rho factor, elongation factor EF-G, subunits of glutamine tRNA, its peptidases, absence of fatty acid desaturase and certain acyltransferases and transposase content. Interestingly, *F. nucleatum* lacks nucleoside diphosphate kinase (NdK), a supposedly ubiquitous gene that expresses an essential metabolic activity. The gene was not found in any *Clostridium* spp. (*Firmicutes*) nor any *Mollicutes* (*Mycoplasma* or *Ureaplasma* species) sequenced so far (Pollack *et al*., 2002).

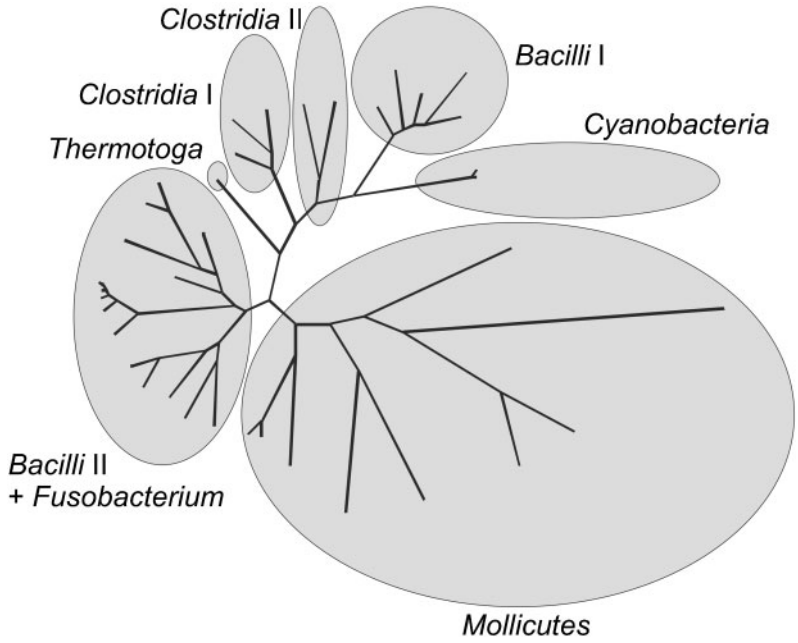

Fig. 2. Schematic (unrooted) representation of relationships within the *Firmicutes*.

The phylogenetic position of *Thermotoga* within the *Firmicutes* was unexpected. The Pgk from *Thermotoga* was included in the analysis only because of its high sequence similarity, obtained by BLAST search, to *Firmicutes* Pgk sequences. *Thermotoga* has an affinity to the low-G+C, Gram-positive *Bacteria* (Nelson *et al.*, 1999). Daubin *et al.* (2002) suggested that the phylogenetic position of hyperthermophilic and radioresistant species is close to that of mesophilic species, like *Bacilli*. We believe that these opinions tend to minimize the concept of the 'hyperthermophilic origin of life'.

The exact position of *Thermotoga* within the 'tree of life' still remains an open question, as different markers have yielded varying results; they either place *Thermotoga* close to the root of the 'tree of life' (e.g. Brown *et al.*, 2001) or further 'up' from the root (Brochier & Philippe, 2002; Daubin *et al.*, 2002). Also, the position of *Thermotoga* with respect to other species is in part compromised by an unknown but important degree of horizontal transfer of genes from other, in particular archaean, species. Similarly, adaptation to a hyperthermophilic environment is a confounding possibility. These factors complicate predictions that involve the genomic content of molecules from organisms that are adapted to high temperature (there are specific adaptations in enzymes such as RNA polymerases, as well as to increase the stability of nucleic acids at higher temperatures).

The value of Pgk as a phylogenetic marker was investigated by Chattopadhyay & Chakrabarti (2003). In that study, evolutionary conclusions were derived solely from a basic statistic (the mean second moment of the codon base distribution) of the coding sequence of Pgk. This statistic was used to position taxonomic groups or organisms in the 'vertical position' on the evolutionary tree. The results for Pgk were convincing, whereas those for other genes (e.g. the glycolytic enzyme glyceraldehyde-3-phosphate dehydrogenase) could not be used as acceptable phylogenetic markers in these computational analyses. Therefore, these and our results strengthen the usefulness of Pgk as a phylogenetic marker.

The Pgk tree of the *Mollicutes* is in agreement with Neimark (1979), in suggesting that *Mycoplasma* spp. are related more closely to the *Streptococcus*/*Lactobacillus* subgroup than to *Bacillus* and *Clostridium* spp.

Pgk shows several advantages for use as a marker enzyme, because of its widespread distribution in central metabolic activity. Further research will extend this analysis to include enzyme-structure data and more biochemical data, to study in more detail the relationships that have been examined here, including other clades and species.

ACKNOWLEDGEMENTS

We gratefully acknowledge receipt of the unpublished *Mollicutes* Pgk gene sequences of *Mycoplasma mycoides* subsp. *mycoides* SC from Joakim Westberg and Karl-Erik Johansson (National Veterinary Institute, Uppsala, Sweden), *Mycoplasma hyopneumoniae* from F. Chris Minion (Iowa State University, Ames, Iowa, USA) and Steven P. Djordjevic (Elizabeth McArthur Agricultural Institute, Cambden, NSW, Australia) and *Spiroplasma citri* from Frédéric Laigret and colleagues (INRA, Villenave d'Ornon, France). We thank the students of the Bioinformatics/Phylogenetics courses at the University of Würzburg for valuable discussions.

REFERENCES

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new

generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.

Balish, M. F. & Krause, D. C. (2002). Cytadherence and the cytoskeleton. In *Molecular Biology and Pathogenicity of Mycoplasmas*, pp. 491–519. Edited by S. Razin & R. Herrmann. New York: Kluwer.

Brochier, C. & Philippe, H. (2002). Phylogeny: a non-hyperthermophilic ancestor for *Bacteria*. *Nature* **417**, 244.

Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E. & Stanhope, M. J. (2001). Universal trees based on large combined protein sequence data sets. *Nat Genet* **28**, 281–285.

Chattopadhyay, S. & Chakrabarti, J. (2003). Temporal changes in phosphoglycerate kinase coding sequences: a quantative measure. *J Comput Biol* **10**, 83–93.

Ciccarese, S., Tommasi, S. & Vonghia, G. (1989). Cloning and cDNA sequence of the rat X-chromosome linked phosphoglycerate kinase. *Biochem Biophys Res Commun* **165**, 1337–1344.

Dandekar, T., Snel, B., Schmidt, S., Lathe, W., Suyama, M., Huynen, M. & Bork, P. (2002). Comparative genome analysis of the *Mollicutes*. In *Molecular Biology and Pathogenicity of Mycoplasmas*, pp. 255–279. Edited by S. Razin & R. Herrmann. New York: Kluwer.

Daubin, V., Gouy, M. & Perrière, G. (2002). A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res* **12**, 1080–1090.

Felsenstein, J. (1993). PHYLIP (phylogeny inference package), version 3.6. Department of Genetics, University of Washington, Seattle, USA.

Fothergill-Gilmore, L. A. (1986). Domains of glycolytic enzymes. In *Multidomain Proteins: Structure and Function*. Edited by D. G. Hardie & J. R. Coggins. Amsterdam: Elsevier.

Fothergill-Gilmore, L. A. & Michels, P. A. (1993). Evolution of glycolysis. *Prog Biophys Mol Biol* **59**, 105–235.

Hepperle, D. (2002). Align: a multicolor sequence alignment editor (available at http://www.gwdg.de/~dhepper/software.htm).

Johansson, K.-E. & Pettersson, B. (2002). Taxonomy of *Mollicutes*. In *Molecular Biology and Pathogenicity of Mycoplasmas*, pp. 1–31. Edited by S. Razin & R. Herrmann. New York: Kluwer.

Johansson, K.-E., Heldtander, M. U. K. & Pettersson, B. (1998). Characterization of mycoplasmas by PCR and sequence analysis with universal 16S rDNA primers. *Methods Mol Biol* **104**, 145–165.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**, 275–282.

Kapatral, V., Anderson, I., Ivanova, N. & 22 other authors (2002). Genome sequence and analysis of the oral bacterium *Fusobacterium nucleatum* strain ATCC 25586. *J Bacteriol* **184**, 2005–2018.

Ludwig, W. & Klenk, H.-P. (2001). Overview: a phylogenetic backbone and taxonomic framework for procaryotic systematics. In *Bergey's Manual of Systematic Bacteriology*, 2nd edn, vol. 1, pp. 49–65. Edited by D. R. Boone, R. W. Castenholz & G. M. Garrity. New York: Springer.

Maniloff, J. (2002). Phylogeny and evolution. In *Molecular Biology and Pathogenicity of Mycoplasmas*, pp. 31–45. Edited by S. Razin & R. Herrmann, R. New York: Kluwer Academic/Plenum Publishers.

Neimark, H. (1979). Phylogenetic relationships between mycoplasmas and other prokaryotes. In *The Mycoplasmas*, vol. 1, pp. 43–61. Edited by M. F. Barile & S. Razin. New York: Academic Press.

Nelson, K. E., Clayton, R. A., Gill, S. R. & 26 other authors (1999). Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399**, 323–329.

Page, R. D. M. (1996). TREEVIEW: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci* **12**, 357–358.

Pollack, J. D. (2002). Central carbohydrate pathways: metabolic flexibility and the extra role of some ''housekeeping'' enzymes. In *Molecular Biology and Pathogenicity of Mycoplasmas*, pp. 163–201. Edited by S. Razin & R. Herrmann. New York: Kluwer.

Pollack, J. D., Myers, M. A., Dandekar, T. & Herrmann, R. (2002). Suspected utility of enzymes with multiple activities in the small genome *Mycoplasma* species: the replacement of the missing ''household'' nucleoside diphosphate kinase gene and activity by glycolytic kinases. *OMICS* **6**, 247–258.

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

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406–425.

Strimmer, K. & von Haeseler, A. (1996). Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol Biol Evol* **13**, 964–969.

Swofford, D. L. (2002). PAUP*: Phylogenetic Analysis Using Parsimony (*and other methods), version 4.0b10 win32. Sunderland, MA: Sinauer Associates.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (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.

Van de Peer, Y. & De Wachter, R. (1994). TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. *Comput Appl Biosci* **10**, 569–570.

Whelan, S. & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* **18**, 691–699.