NOTE

Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis

Sabine Hansmann[1] and William Martin[2]

Author for correspondence: William Martin. Tel: +49 211 811 3011. Fax: +49 211 811 3554.
e-mail: w.martin@uni-duesseldorf.de

[1] Institut für Genetik, Technische Universität Braunschweig, Spielmannstr. 7, D-38023 Braunschweig, Germany

[2] Institut für Botanik III, Universität Düsseldorf, Universitätsstr. 1, D-40225 Düsseldorf, Germany

Thirty-nine proteins encoded in a large gene cluster that is well-conserved in gene content and gene order across 18 sequenced prokaryotic genomes were extracted, aligned and subjected to phylogenetic analysis. In individual analyses of the alignments, only two probable examples of lateral gene transfer between archaea and eubacteria were detected, involving the genes for ribosomal protein Rpl23 and adenylate kinase. Amino acid sequences for 35 of the 39 proteins were concatenated to yield a data set of 9087 amino acid positions per genome. Many of these proteins, 33 of which are ribosomal proteins, are not highly conserved across distantly related organisms and thus contain many regions that are difficult to align. Phylogenetic analyses were performed with subsets of the concatenated data from which the most highly variable sites had been iteratively removed, using the number of different amino acids that occur at a given site as a criterion of variability. Glycine, which has a strong influence on protein structure, tended to be more frequent at the most conserved (least polymorphic) sites. With most subsets of the data, the proteins from the cyanobacterium *Synechocystis* tended to branch with their homologues from Gram-positive bacteria. The results indicate that excluding only a few percentage of poorly alignable sites from phylogenetic analysis can have a severe impact upon the phylogeny inferred and that bootstrap support for branches can fluctuate substantially, depending upon which sites are excluded.

Keywords: comparative genomics, evolution, gene cluster, lateral transfer, substitution rate

One of the hopes that biologists have vested in genomics is that the copious amounts of data emerging from genomes, particularly prokaryotic ones, will help us to better understand evolutionary history. One of the most significant insights emerging from prokaryotic genomes is the degree of horizontal transfer that has occurred during prokaryotic evolution (Doolittle, 1999a, b). Yet despite a now-evident flux of genes through prokaryotic genomes through time, there also appears to be functional classes of genes that are less freely exchanged than others, possibly because of co-evolutionary constraints due to physical interactions between proteins, for example within ribosomes (Jain *et al.*, 1999). In line with this view, Wächtershäuser (1998) recently identified a conserved cluster of gene order encompassing about 50 genes found to exist in 19 completely sequenced (or nearly so) prokaryotic genomes. The cluster encodes primarily proteins of translation (ribosomal proteins), transcription (RNA polymerase), and related functions. However, a few proteins belonging to other functional classes, notably adenylate kinase (Sánchez & Müller, 1998) and enolase (Hannaert *et al.*, 2000) involved in energy metabolism, are also found in the cluster in some genomes (Wächtershäuser, 1998). This cluster has proven useful from the standpoint of plastid phylogeny by virtue of its tendency to preserve gene order, revealing evolutionary relationships among plastids (Stoebe & Kowallik, 1999). We have investigated the molecular phylogeny of the proteins encoded

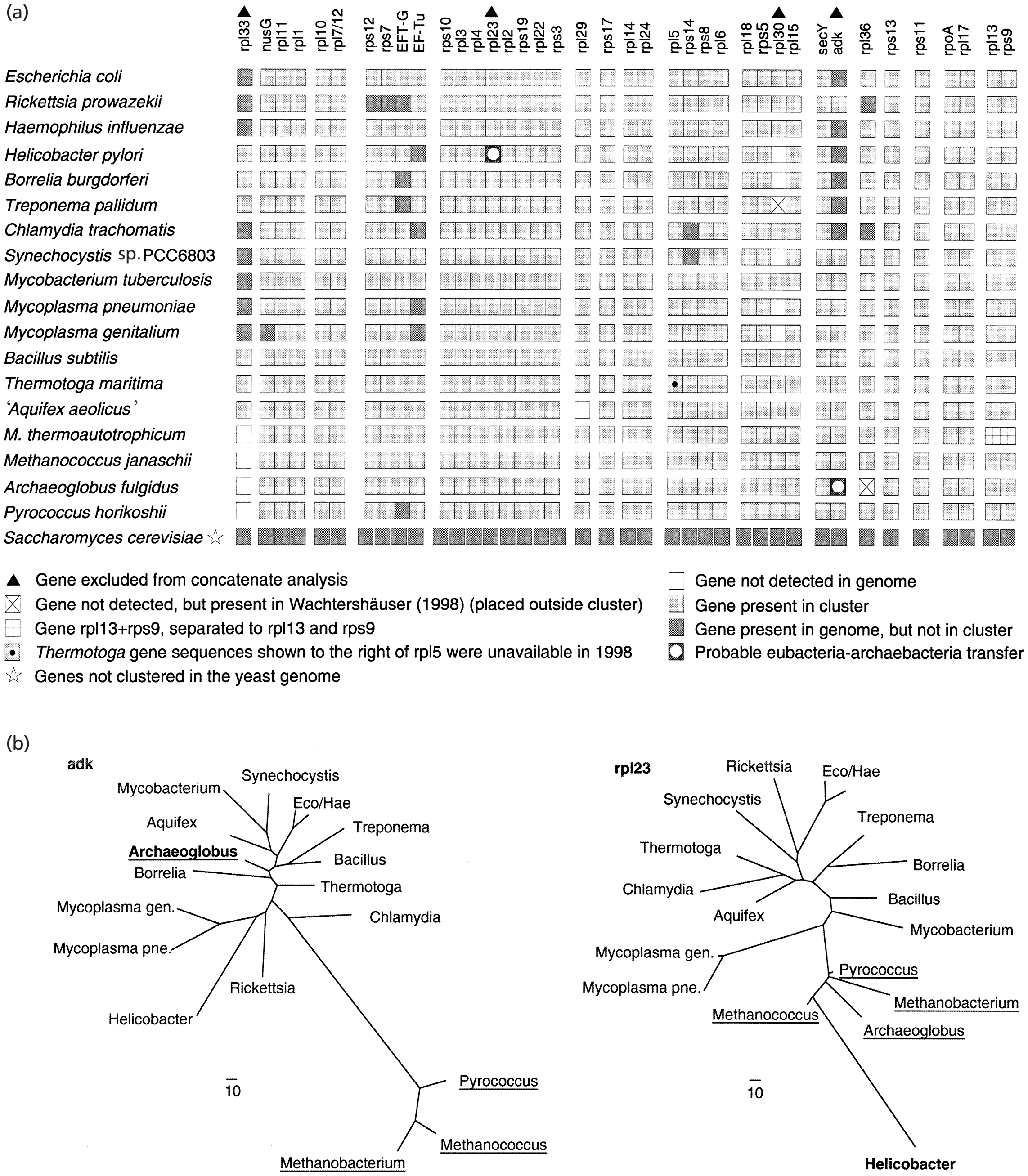
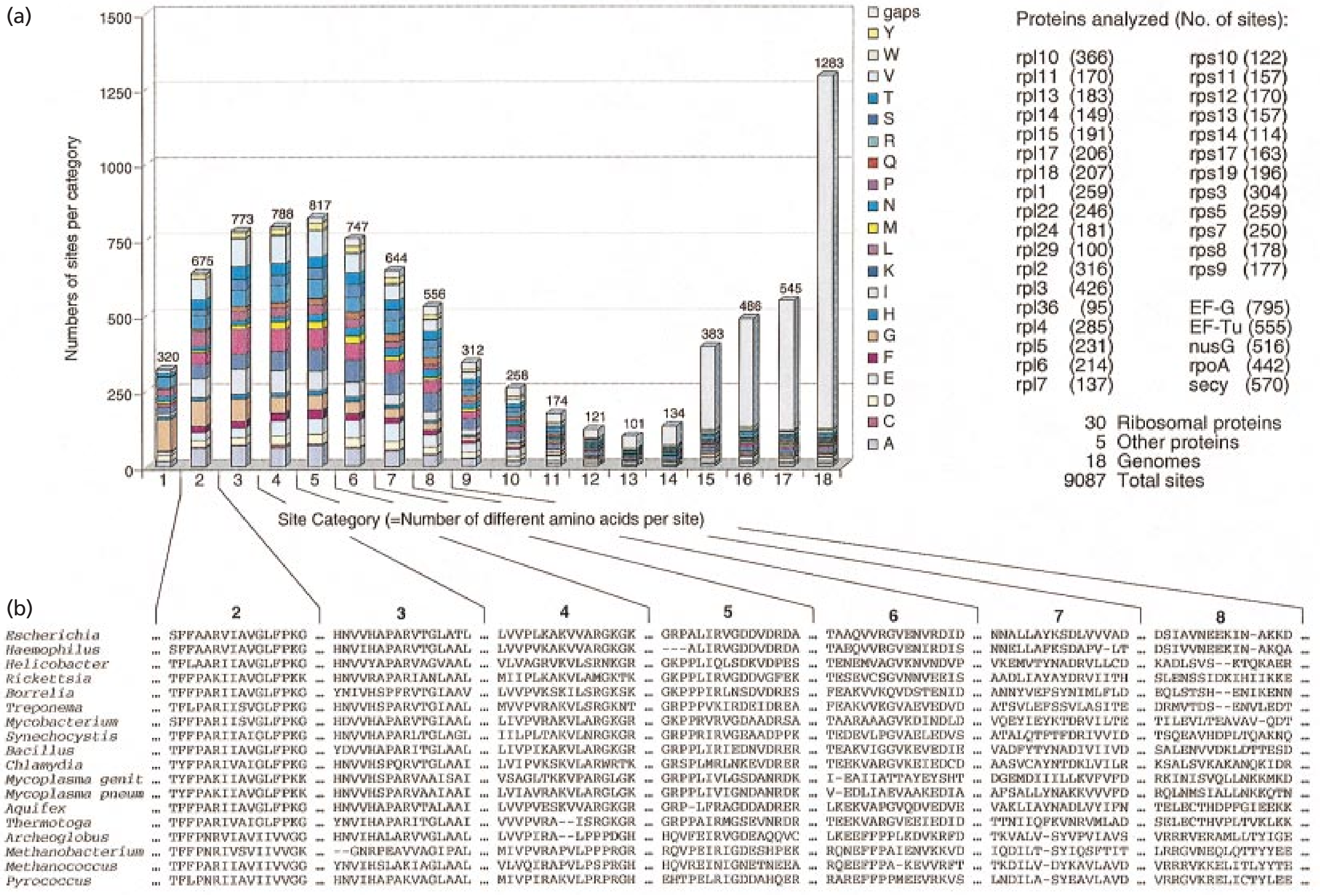


Fig. 1. (a) Distribution of genes in a portion of the conserved gene cluster recently identified (Wächtershäuser, 1998). The depiction of the cluster is restricted to those genes that are prevalent among many genomes hence amenable to phylogenetic analysis. Symbols and minor discrepancies to the tabulation in Wächtershäuser (1998) are indicated. *M. thermoautotrophicum*, *Methanobacterium thermoautotrophicum*. Proteins that were excluded from concatenated analysis are indicated by a triangle. (b) PROTML topologies obtained for Adk and Rpl23, which suggest the occurrence of lateral gene transfers (indicated in boldface type) between archaea (underlined) and eubacteria. A more comprehensive phylogeny of Adk genes, including eukaryotic but not archaeal homologues was recently reported by Sánchez & Müller (1998).

in that conserved gene cluster across several sequenced genomes.

The published genome sequences of 18 prokaryotes were retrieved: *Haemophilus influenzae* Rd (Fleischmann *et al.*, 1995), *Mycoplasma genitalium* (Fraser *et al.*, 1995), *Methanococcus jannaschii* (Bult *et al.*, 1996), *Synechocystis* sp. PCC 6803 (Kaneko *et al.*, 1996), *Mycoplasma pneumoniae* (Himmelreich *et al.*, 1996), *Helicobacter pylori* (Tomb *et al.*, 1997), *Escherichia coli* (Blattner *et al.*, 1997), *Methanobacterium thermoautotrophicum* (recently reclassified



	2	3	4	5	6	7	8
Escherichia	... SFFAARVIAVGLFPKG ...	HNVVHAPARVTGLATL ...	LVVPLKAKVVARGKGK ...	GRPALIRVGDDVDRDA ...	TAAQVVRGVENVRDID ...	NNALLAYKSDLVVVAD ...	DSIAVNEEKIN-AKKD ...
Haemophilus	... SFFAARVIAVGLFPKG ...	HNVIHAPARVTGLAAL ...	LVVPVKAKVVARGKGK ...	---ALIRVGDDVDRDA ...	TAEQVVRGVENIRDIS ...	NNELLAFKSDAPV-LT ...	DSIVVNEEKIN-AKQA ...
Helicobacter	... TFLAARIIAVGLFPKG ...	HNVVYAPARVAGVAAL ...	VLVAGRVKVLSRNKGR ...	GKPPLIQLSDKVDPES ...	TENEMVAGVKNVNDVP ...	VKEMVTYNADRVLLCD ...	KADLSVS--KTQKAER ...
Rickettsia	... TFFPAKIIAVGLFPKK ...	HNVVRAPARIANLAAL ...	MIIPLKAKVLAMGKTK ...	GKPPLIRVGDDVGFEK ...	TESEVCSGVNNVEEIS ...	AADLIAYAYDRVIITH ...	SLENSSIDKIHIIKKE ...
Borrelia	... TFFPARIIAVGLFPKG ...	YNIVHSPFRVTGIAAV ...	LVVPVKSKILSRGKSK ...	GKPPPIRLNSDVDRES ...	FEAKVVKQVDSTENID ...	ANNYVEFSYNIMLFLD ...	EQLSTSH--ENIKENN ...
Treponema	... TFLPARIISVGLFPKG ...	HNVVHSPARVTGIAAL ...	MVVPVRAKVLSRGKNT ...	GRPPPVKIRDEIDREA ...	FEAKVVKGVAEVEDVD ...	ATSVLEFSSVLASITE ...	DRMVTDS--ENVLEDT ...
Mycobacterium	... SFFPARIISVGLFPKG ...	HDVVHAPARVTGIAAL ...	LIVPVRAKVLARGKGR ...	GKPPRVRVGDAADRSA ...	TAARAAAGVKDINDLD ...	VQEYIEYKTDRVILTE ...	TILEVLTEAVAV-QDT ...
Synechocystis	... TFFPARIIAIGLFPKG ...	HNVVHAPARLTGLAGL ...	IILPLTAKVLNRGKGR ...	GRPPRIRVGEAADPPK ...	TEDEVLPGVAELEDVS ...	ATALQTPTFDRIVVID ...	TSQEAVHDPLTQAKNQ ...
Bacillus	... TFFPARIIAVGLFPKG ...	YDVVHAPARITGLAAL ...	LIVPIKAKVLARGKGR ...	GRPPLIRIEDNVDRER ...	TEAKVIGGVKEVEDIE ...	VADFYTYNADIVIIVD ...	SALENVVDKLDTTESD ...
Chlamydia	... TYFPARIVAIGLFPKG ...	HNVVHSPQRVTGLAAI ...	LVIPVKSKVLARWRTK ...	GRSPLMRLNKEVDRER ...	TEEKVARGVKEIEDCD ...	AASVCAYNTDKLVILR ...	KSALSVKAKANQKIDR ...
Mycoplasma genit	... TYFPAKIIAVGLFPKK ...	HNVVHSPARVAAISAI ...	VSAGLTKKVPARGLGK ...	GRPPLIVLGSDANRDK ...	I-EAIIATTAYEYSHT ...	DGEMDIIILLKVFVFD ...	RKINISVQLLNKKMKD ...
Mycoplasma pneum	... TYFPAKIIAVGLFPKK ...	HNVVHSPARVAAIAAI ...	LVIAVRAKVLARGLGK ...	GRPPLIVIGNDANRDK ...	V-EDLIAEVAAKEDIA ...	AFSALLYNAKKVVVFD ...	RQLNMSIALLNKKQTN ...
Aquifex	... TFFPARIIAVGLFPKG ...	HNVVHAPARVTALAAI ...	LVVPVESKVVARGKGR ...	GRP-LFRAGDDADRER ...	LKEKVAPGVQDVEDVE ...	VAKLIAYNADLVYIPN ...	TELECTHDPFGIEEKK ...
Thermotoga	... TFFPARIVAIGLFPKG ...	YNVIHAPARITGLAAI ...	VVVPVRA--ISRGKGR ...	GRPPAIRMGSEVNRDR ...	TEEKVARGVEEIEDID ...	TTNIIQFKVNRVMLAD ...	SELECTHVPLTVKLKK ...
Archeoglobus	... TFFPNRVIAVIIVVGG ...	HNVIHALARVVGLAAL ...	LVVPIRA--LPPPDGH ...	HQVFEIRVGDEAQQVC ...	LKEEFFPPLKDVKRFD ...	TKVALV-SYVPVIAVS ...	VRRRVERAMLLTYIGE ...
Methanobacterium	... TFFPNRIVSVIIVVGK ...	--GNRPEAVVAGIPAL ...	MIVPVRAPVLPPPRGR ...	RQVPEIRIGDESHPEK ...	RQNEFFPAIENVKKVD ...	IQDILT-SYIQSPTIT ...	LRRGVNEQLQTTYYEE ...
Methanococcus	... TFFPARIIAVIIVVGG ...	YNVIHSLAKIAGLAAL ...	VLVQIRAPVLPSPRGH ...	HQVREINIGNETNEEA ...	RQEEFPPA-KEVVRFT ...	TKDILV-DYKAVLAVD ...	VRRRVKKELITLYYTE ...
Pyrococcus	... TFLPNRIIAVIIVVGG ...	HNVIHAPAKVAGLAAL ...	MIVPIRAKVLPRPRGH ...	EHTPELRIGDDAHQER ...	RAREFFPPMEEVRKVS ...	LNDILA-SYEAVLAVD ...	VRRGVKRELICTYLEE ...

Fig. 2. Distribution of polymorphic sites in the concatenated alignment of proteins indicated in Fig. 1. (a) Columns indicate the number of amino acids that occur at a given category of site, whereby the category is defined by the number of different amino acids that occur at the site (see text). Scale at left and numbers at the top of columns indicate the number of sites per category. For example, there are 675 sites in the alignment at which two different amino acids (or one amino acid only and one gap) occur. Within columns, the proportions of amino acids and gaps that occur in the given category are shown. The proteins included in the concatenated data and number of sites in the individual protein alignments are shown at right. (b) Randomly chosen segments of the sorted alignment illustrating patterns of sequence similarity at positions in categories of sites 2–8. Gaps are indicated by dashes.

as *Methanothermobacter thermoautotrophicus*; Wasserfallen *et al*., 2000) (Smith *et al*., 1997), *Bacillus subtilis* (Kunst *et al*., 1997), *Archaeoglobus fulgidus* (Klenk *et al*., 1997) *Borrelia burgdorferi* (Fraser *et al*., 1997), '*Aquifex aeolicus*' (Deckert *et al*., 1998), *Pyrococcus horikoshii* (Kawarabayasi *et al*., 1998), *Mycobacterium tuberculosis* (Cole *et al*., 1998), *Treponema pallidum* (Fraser *et al*., 1998), *Chlamydia trachomatis* (Stephens *et al*., 1998), *Rickettsia prowazekii* (Andersson *et al*., 1998), and *Thermotoga maritima* (Nelson *et al*., 1999). Using the tabulation provided by Wächtershäuser (1998) and using BLAST searches (Altschul *et al*., 1997) the presence or absence of genes in the cluster and in the respective genome was reexamined. As shown in Fig. 1(a), only very minor discrepancies to the earlier analysis were found.

Notably, only about half of the genes in the cluster identified by Wächtershäuser (1998) are shown in Fig. 1(a), since we have only indicated those genes that are present in enough of the genomes to justify inclusion in phylogenetic analysis. Since not all of the genes in the cluster are present in all genomes, there is a trade-off between the number of sequences per genome (increasing the amount of data for analysis) and the number of genomes (increasing the number of taxa for analysis) one can investigate. In this study, we have analysed those proteins shown in the figure.

The identified protein sequences were extracted and aligned using the program PILEUP of the Genetics Computer Group (1994) package. In total, 39 proteins were extracted (Fig. 1). Phylogenetic analyses of the 39 individual alignments were performed by constructing neighbour-joining (Saitou & Nei, 1987) trees from matrices of maximum-likelihood distances between amino acid sequences as a starting topology for local rearrangement using PROTML with the JTT-matrix as implemented in the MOLPHY package (Adachi & Hasegawa, 1996a). The individual PROTML trees were inspected by eye for possible cases of lateral transfer between archaea and eubacteria. For three proteins, adenylate kinase (Adk) and the ribosomal proteins L23 (Rpl23) and L33 (Rpl33), the monophyly of archaea and eubacteria was violated. *Archaeoglobus* Adk branched within the eubacteria, *Helicobacter* Rpl23 branched within the archaea, suggesting the occurrence of lateral transfers for these proteins (Fig. 1b). Rpl33 is only about 50 amino acids long in most species and is very poorly conserved, such that the interleaving of archaeal and eubacterial Rpl33 sequences (not shown) is at least as likely to be due to these factors as it is to be due to lateral transfer. Rpl30 is similarly short and poorly conserved as Rpl33 (alignments and trees available via ftp from 134.169.70.80/ftp/pub/incoming/baktFeb00). These four proteins were excluded from further analysis. This left 35 alignments (gene designations indicated in the right portion of Fig. 2a) from 18 prokaryotic genomes, amounting to 9087 amino acid sites (including gaps) per genome whereby, as in the cluster itself, one or the other protein is missing from one or the other genome (see Fig. 1a).

Overall, many of these remaining 35 proteins are not dramatically well conserved across distantly related prokaryotes and many of the alignments possess conspicuously large gapped regions. But all 35 do possess one or the other highly conserved region that might contain a phylogenetic signal and all do appear to be alignable, albeit not always over their entire length. Molecular phylogeneticists, when confronted with a 'poor' alignment – a judgement usually made by inspecting the data by eye to see if the sites appear to be conserved – exclude regions of uncertain alignment from analysis. This widely practiced procedure is intuitively justifiable, because highly variable (poorly conserved) sites may contain misleading information, this being due to stochastic similarity or bias attributable to any number of factors (Lockhart *et al.*, 1999; Philippe & Laurent, 1998). Although much attention has been given to the importance of different computer models of phylogenetic inference (Nei, 1996) and to the problem of how to take into account the variability of substitution rates across different sites (Yang *et al.*, 1998), the problem of how to distinguish between which sites in a given data set should be included in a phylogenetic analysis and which should be excluded from analysis has been less widely dealt with.

One approach to this problem is to assume *a priori* that the most polymorphic sites (the ones that are least conserved and hence most difficult to align) have the highest probability of containing misleading information. If, for the purposes of this paper, we assume this to be true (it is also possible that the least polymorphic sites might be the most misleading), then systematically excluding the most variable sites from the alignment until only a 'conserved core' of positions is left would, in principle, remove potentially misleading information. Yet there are currently no objective criteria for defining what might constitute a conserved core of an amino acid alignment. It must lie somewhere between all and none of the sites in an alignment, but the problem of how to define a hierarchy of variability for excluding sites from analysis is not trivial. One solution to this, suggested by M. Hasegawa (personal communication) is to assume a star phylogeny (that is, to make no assumption about topology) and to simply count the number of different amino acids that occur at a given position in the alignment. The greater the number of different amino acids that occur at the position, the more variable (less conserved) it should be. Using that hierarchy of variability, one could systematically exclude the most variable sites from an alignment, generating subsets of the data with increasing sequence conservation, and subject each subset to phylogenetic analysis. A program (SORTAL) written in C++ for UNIX environments has been written (Hansmann, 1997) that performs these steps automatically (available from the ftp site).

SORTAL accepts a standard PHYLIP (Felsenstein, 1993) infile as input. For each site of the pre-aligned data, the number of different amino acids that occur at a given site are counted and scored. Strictly conserved sites are assigned the score '1', sites at which only two different amino acids occur (for example either alanine or arginine) obtain the score '2', sites at which only three different amino acids (for example either alanine, arginine or glycine) occur obtain the score '3', etc., irrespective of how many times each different amino acid occurs at the given site. Sites with an identical score are grouped together as a class. The positions of the alignment are then written into an output file that has the same format as an interleaved PHYLIP infile, but the positions are sorted into classes of ascending positional variability. That is, the first positions written into in the sorted alignment are the invariant sites, followed by the positions at which only two amino acids occur, followed by those at which three different amino acids occur, etc. Gaps are counted as being different from every other amino acid at the site (including other gaps), such that the highly gapped sites are scored as being highly variable. This file can be inspected by eye and edited by hand. Alternatively, subsets of the data can be iteratively generated from this file starting with all of the sites, then progressively excluding the most variable class of sites from the previous subset.

Fig. 2(a) (left panel) shows the distribution of sites with different numbers of amino acids per site in the present 9087-site data and the proportions of each amino acid for the different classes of sites. Glycine, an amino acid with a strong influence on protein structure, is more common among the least polymorphic sites and comprises almost a third (103/320 residues) of the non-polymorphic sites. Otherwise, no striking preference for sites of particular variability to possess particular types of amino acids was observed. A non-uniform distribution of amino acids across sites of differing substitution rate is not incorporated into current models of amino acid substitution, but this parameter might deserve future inspection for generating more realistic models for phylogeny inference (Adachi & Hasegawa, 1996b).

Fig. 2(b) shows randomly selected segments of the sorted alignment to provide an impression of the degree of sequence conservation that is contained in sites of differing variability in this data. Sites with only two, three or four different amino acids appear well-conserved and represent data that molecular phylogeneticists might feel comfortable analysing. Sites with seven or eight different amino acids (Fig. 2b) are not well-conserved and belong to the type of sites that many phylogeneticists might exclude from analysis. Sites with five or six different amino acids lie somewhere in between in this data. Sites with more than eight different amino acids (not shown) are very poorly conserved in this data.

As mentioned above, four proteins were excluded from analysis because of possible archaeal–eubacterial gene transfers or because of insufficient length. Within eubacteria and within archaea the possibility of lateral transfer is difficult to demonstrate or to exclude due to the overall mean low degree of conservation in most of these proteins. Thus, by concatenating the 35-protein data there is still a chance that we are grouping different proteins encoded in the conserved cluster that may not share the same evolutionary history. Nonetheless, it was of interest to examine the phylogenies obtained with subsets of this concatenated data.

We constructed maximum-likelihood (ML) phylogenies for the whole data set of concatenated sequences and for each subset of the data from which the most variable fraction of sites had been iteratively removed (Fig. 3a). PROTML (Adachi & Hasegawa, 1996a) was employed with the JTT-F model using local rearrangements starting from the neighbour-joining tree of maximum-likelihood distances. For 18 OTUs, this yields 17 trees (Fig. 3), because the class of invariant sites does not produce a tree. Only five branches were common to all 17 trees: (i) the branch uniting the two *Mycoplasma* species, (ii) that uniting *Escherichia* and *Haemophilus*, (iii) that uniting *Treponema* and *Borrelia*, (iv) that separating *Archaeoglobus* from the other archaea, and (v) that separating eubacteria and archaea.

In addition to these five branches, several others were observed for most of the subsets of the data. *Mycobacterium*, *Synechocystis* and *Bacillus* shared a common branch in trees 3–18, whereby 'tree *18*' designates the tree that was obtained from the data set that contained all sites, including those where *18* different amino acids (or gaps) per site occur, 'tree *3*' designates the tree that was obtained from the highly conserved data set that consisted only of sites that contained 1, 2 or a maximum of *3* different amino acids per site, etc. In trees 4–18 *Rickettsia* shared a common branch with *Escherichia* and *Haemophilus*. The most basally branching eubacterium was *Aquifex* for trees inferred from the more conservatively evolving sites, *Aquifex* and *Thermotoga* shared a common branch in trees 7–18. Data sets containing 11–18 different amino acids per site state gave an identical topology, shown in Fig. 3(b). Overall, *Helicobacter* and *Chlamydia* were the most prone to branching at different positions across trees constructed from subsets of differing degree of conservation.

The concatenated sequences from the cyanobacterium *Synechocystis* did not tend to branch with homologues from other Gram-negative bacteria in these analyses, rather they always remained close to Gram-positive homologues. This is in contrast to a recent analysis of gene content (presence or absence of gene) across prokaryotic genomes (Huynen *et al*., 1999), in which *Synechocystis* tended to branch with *Aquifex*. Resolving deeper branches of prokaryotic phylogeny with rRNA is notoriously difficult (Pace, 1997). Affinity between cyanobacteria and high-G + C Gram-positives has been observed in the case of RecA, though not for rRNA from the same species (Eisen, 1995). Affinity between cyanobacteria and Gram-positives is also observed in phylogenies of Hsp70, though not specifically to either high- or low-G+C groups (Gupta, 1998). It has been suggested that the two photosystems of cyanobacteria might reflect an ancient lateral acquisition from photosynthetic Gram-positives, which possess a homologue of photosystem I (Stackebrandt *et al*., 1996; Xiong *et al*., 1998). In the 39 individual analyses, *Synechocystis* branched with *Mycobacterium* in the case of seven proteins: Adk, Rpl3, Rpl11, Rpl22, Rpl29, Rps8 and NusG. Thus, there was an overall tendency for *Synechocystis* to branch with *Mycobacterium* with these data. When comparable data from many prokaryotic genomes become available for analysis, it can be anticipated that this tendency may change, and it is too early to assess whether or not this affinity may relate to lateral transfer between these groups.

Notably, some branches with high RELL bootstrap proportions (Adachi & Hasegawa, 1996a) for some subsets of the data were not detected or were contradicted with high bootstrap proportions (BP) in other subsets. This is particularly clear in the case of *Chlamydia*. All branches with a BP greater than 0·95 are indicated by dots in Fig. 3. In trees 11–18, *Chlamydia* branched consistently and with a high BP with the two spirochaetes (Fig. 3b) (BP 1·0 in all eight trees). When the 174 most variable sites of the 6024 positions in data set 11 are excluded (a mere 2·9 % of the sites, the most variable ones), the position of *Chlamydia* switches to branch with the mycoplasmas and *Helicobacter* (tree 10) but also with a high BP (0·97) (Fig. 3c). When the 258 most variable sites of the 5859 positions in data set 10 are excluded (4·4 %), the position of *Chlamydia* changes to the most basal among eubacteria (BP 0·69) (tree 9). Excluding the next 312 most variable sites (5·9 %) places *Chlamydia* with *Helicobacter* (BP 0·59) (tree 8). Excluding the 556

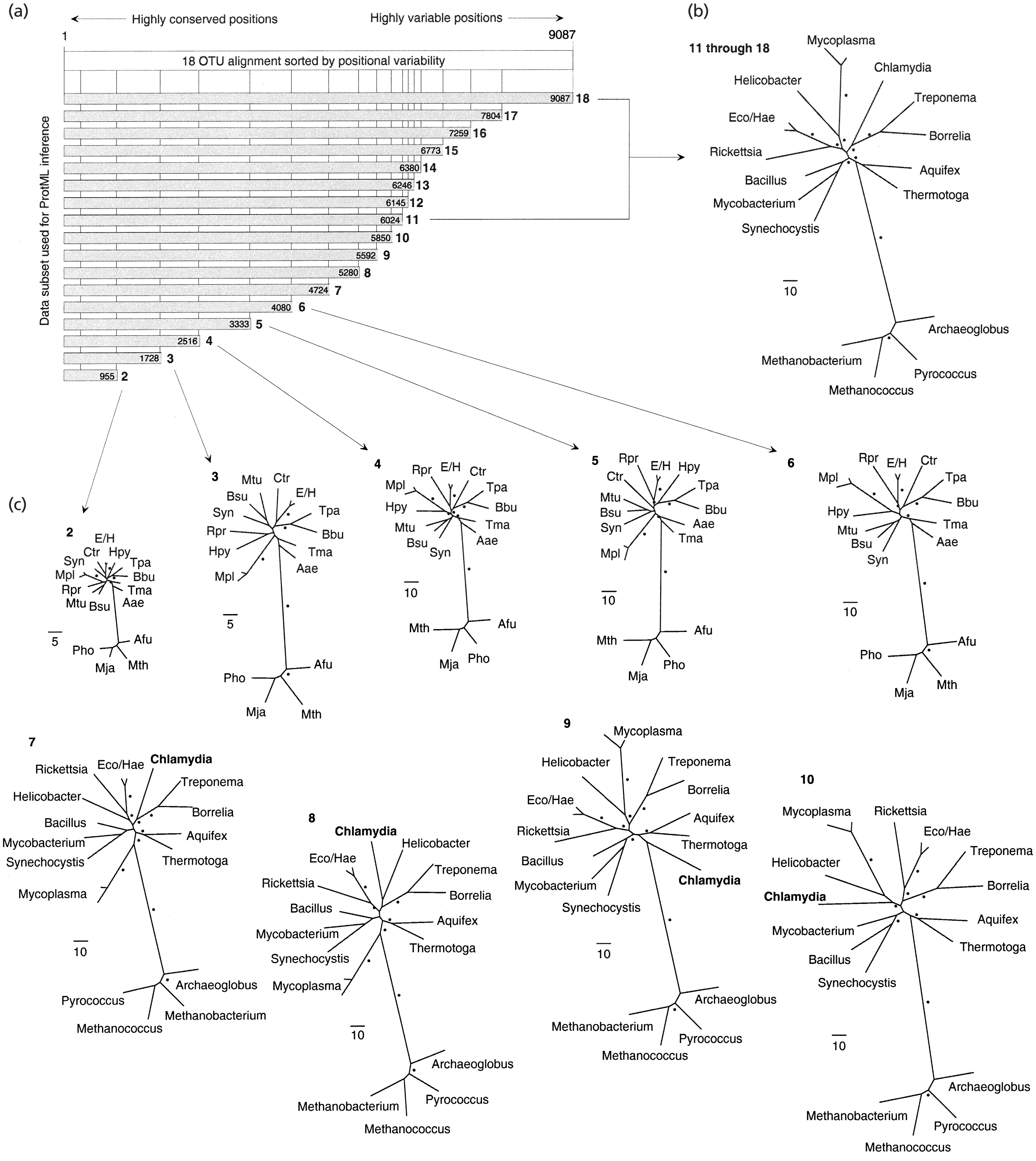


Fig. 3. Data sets used for phylogenetic inference and trees. (a) Schematic representation of the concatenated alignment sorted by positional variability and 17 subsets of the data that contain sites with up to a maximum of two, three or four, etc., different amino acids. Numbers in bars indicate the number of sites in the respective subset. Numbers to the right of bars designate the subset and the corresponding tree inferred from it. (b) Tree inferred from subset 11. The scale bar indicates 10 substitutions. Branches with RELL (Adachi & Hasegawa, 1996a) bootstrap proportions >0·95 are indicated by dots. Data subsets 12–18 gave the same topology but with differing BPs and scale (not shown). Eco/Hae, *Escherichia* and *Haemophilus*. Mycoplasma designates the two species sampled, *M. genitalium* and *M. pneumoniae*. (c) Trees inferred from subsets 2–10. Trees are drawn to the same scale as in (b) except trees 2 and 3. In trees 2–6, taxon designations are abbreviated. In trees 7–10, the position of *Chlamydia* (see text) is highlighted.

sites with eight states per position brings *Chlamydia* back to the spirochaetes, again with a high BP (0·96) (tree 7). These results indicate that excluding only a small percentage of poorly alignable sites from analysis can have a severe impact upon the phylogeny inferred and that bootstrap proportions for branches can depend heavily upon which sites are excluded. These results were obtained with concatenated data, but we have observed very similar results with various individual proteins (unpublished).

Previous studies have shown that eukaryotes, generally speaking, possess an archaeal-like genetic apparatus (Zillig, 1991; Langer *et al.*, 1995). In particular, they possess archaeal-like ribosomes in the cytosol, with eubacterial-type ribosomes in chloroplasts and mitochondria, a circumstance that is generally interpreted as reflecting an archaeal ancestry of the host that acquired mitochondria (Doolittle, 1996; Martin & Müller, 1998; Roger, 1999) and a eubacterial ancestry of mitochondria and chloroplasts (Andersson *et al.*, 1998; Gray *et al.*, 1999; Martin *et al.*, 1998). In yeast, the genes for the ribosomal and other proteins studied here are not clustered as in prokaryotes, rather they are strewn about the chromosomes. We retrieved yeast cytosolic (i.e. not mitochondrial) homologues of these 35 proteins for this data set to examine the phylogeny of the eukaryotic proteins in the context of their prokaryotic homologues. The resulting data set contained quite a few more gaps, there were 10225 sites. The position of baker's yeast homologues of these proteins split the branch between archaea and eubacteria such that its node was, on average, about four times closer to the former than to the latter, a position that did not change dramatically across the 18 different ML trees inferred, with the exception of the subsets containing only two or three different amino acids, in which the branch to eubacteria became extremely long (alignments and results available from the ftp site above).

On the whole, the present data very clearly indicate distinctness of archaea and eubacteria, common branching for two species of *Mycoplasma*, for two methanogens and for two spirochaetes. As far as the other branches go, there is no more certainty that these 35 proteins and 9087 amino acid sites come close to reflecting the historical relatedness of these prokaryotes (and a eukaryote) than there is with any other data set, including rRNA. Furthermore, we have concatenated these sequences for analysis, such that the various trees may depict patterns of similarity that consist of unknown proportions of common descent, possible lateral gene transfer, reconstruction artefact, and any number of biasses in the data. But the very same uncertainties exist with rRNA and other phylogenetic markers. There is doubt concerning the ability of rRNA to accurately reflect phylogenetic relationships among eukaryotes (Philippe & Laurent, 1998; Embley & Hirt, 1998; Roger, 1999) and the problem among prokaryotes is probably much more severe (Doolittle, 1999a, b; Jain *et al.*, 1999). It is possible that many of the proteins encoded in the conserved gene cluster investigated here have co-evolved with their cognate rRNA. Concatenated sequence data have recently been employed to gain insights into the phylogeny of plastid genomes and mitochondrial genomes (Martin *et al.*, 1998; Lockhart *et al.*, 1999; Gray *et al.*, 1999). These proteins from the conserved gene cluster identified by Wächtershäuser (1998) may supplement bacterial systematics as genome data accumulate for prokaryotes and eukaryotes.

There can be little doubt that the data containing two, three and four amino acids per site appear more amenable to phylogenetic analysis than the data containing seven or more amino acids per site in this sample (see Fig. 2b). About half of the sites contained within the data analysed here appear as if they contain evidence for sequence conservation across this taxon sample. Molecular phylogeneticists intuitively remove the questionable regions of alignments prior to analysis. The sorting program used here does the same thing, but it does so systematically. For phylogenetic inference models that assume sites to evolve independently, the order of sites in the data are irrelevant. Deleting certain parts of the data may violate the assumptions of many models of phylogenetic inference, but that is true regardless of whether difficult-to-align regions are excluded intuitively by hand, as is currently common practice, or systematically by computer. Models of phylogenetic inference are important, but these findings indicate that the issue of which sites in a data set should be excluded from analysis seems to be important as well.

Acknowledgements

We thank the Deutsche Forschungsgemeinschaft (W.M.) and the Heinrich-Böll-Stiftung (S.H.) for financial support, Carmen Rotte for critical reading of the manuscript, and two referees for their very helpful suggestions.

References

Adachi, J. & Hasegawa, M. (1996a). *Computer Science Monographs, no.* 28. *MOLPHY version* 2.3: *programs for Molecular Phylogenetics Based on Maximum Likelihood.* Tokyo: Institute of Statistical Mathematics.

Adachi, J. & Hasegawa, M. (1996b). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J Mol Evol* **42**, 459–468.

Altschul, S. F., Madden, T. L., Schaffer, 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.

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

Blattner, F. R., Plunkett, G., III, Bloch, C. A. & 14 other authors (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1474.

Brown, J. R. & Doolittle, W. F. (1997). Archaea and the prokaryote-to-eukaryote transition. *Microbiol Mol Biol Rev* **61**, 456–502.

Bult, C. J., White, O., Olsen, G. J. & 37 other authors (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**, 1058–1073.

Cole, S. T., Brosch, R., Parkhill, J. & 39 other authors (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544.

Deckert, G., Warren, P. V., Gaasterland, T. & 12 other authors (1998). The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **392**, 353–358.

Doolittle, W. F. (1996). Some aspects of the biology of cells and their possible evolutionary significance. In *Evolution of Microbial Life* (Society for General Microbiology Symposium 54), pp. 1–21. Edited by D. Roberts, P. Sharp, G. Alderson & M. Collins. Cambridge: Cambridge University Press.

Doolittle, W. F. (1999a). Phylogenetic classification and the universal tree. *Science* **284**, 2124–2128.

Doolittle, W. F. (1999b). Lateral genomics. *Trends Biochem Sci* **24**, M5–M8.

Eisen, J. A. (1995). The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16S rRNAs from the same species. *J Mol Evol* **41**, 1105–1123.

Embley, T. M. & Hirt, R. P. (1998). Early branching eukaryotes? *Curr Opin Genet Dev* **8**, 624–629.

Felsenstein, J. (1993). PHYLIP (Phylogeny inference package) manual, version 3.5c. Distributed by the author. University of Washington, Department of Genetics, Seattle, Washington.

Fleischmann, R. D., Adams, M. D., White, O. & 37 other authors (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.

Fraser, C. M., Gocayne, J. D., White, O. & 26 other authors (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403.

Fraser, C. M., Casjens, S., Huang, W. M. & 22 other authors (1997). Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390**, 580–586.

Fraser, C. M., Norris, S. J., Weinstock, G. M. & 22 other authors (1998). Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281**, 375–388.

Genetics Computer Group (1994). Program manual for version 8, 575 Science Drive, Madison, WI 53711, USA.

Gray, M. W., Burger, G. & Lang, B. F. (1999). Mitochondrial evolution. *Science* **283**, 1476–1481.

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.

Hannaert, V., Brinkmann, H., Nowitzki, U. & 7 other authors (2000). Enolase from *Trypanosoma brucei*, from the amitochondriate protist *Mastigamoeba balamuthi*, and from the chloroplast and cytosol of *Euglena gracilis*: pieces in the evolutionary puzzle of the eukaryotic glycolytic pathway. *Mol Biol Evol* **17**, 989–1000.

Hansmann, S. (1997). *Einfluss der Variabilität von Aminosäurepositionen auf die Rekonstruktion phylogenetischer Bäueme*. Studienarbeit, University of Braunschweig.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B. C. & Herrmann, R. (1996). Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res* **24**, 4420–4449.

Huynen, M., Snel, B. & Bork, P. (1999). Lateral gene transfer, genome surveys, and the phylogeny of prokaryotes. *Science* **286**, 1443a.

Jain, R., Rivera, M. C. & Lake, J. A. (1999). Horizontal transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci USA* **96**, 3801–3806.

Kaneko, T., Sato, S., Kotani, H. & 21 other authors (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., Sawada, M., Horikawa, H. & 22 other authors (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., Clayton, R. A., Tomb, J. F. & 48 other authors (1997). The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390**, 364–370.

Kunst, F., Ogasawara, N., Moszer, I. & 148 other authors (1997). The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249–256.

Langer, D., Hain, J., Thuriaux, P. & Zillig, W. (1995). Transcription in archaea: similarity to that in eukarya. *Proc Natl Acad Sci USA* **92**, 5768–5772.

Lockhart, P. J., Howe, C. J., Barbrook, A. C., Larkum, A. W. D. & Penny, D. (1999). Spectral analysis, systematic bias and the evolution of chloroplasts. *Mol Biol Evol* **16**, 573–576.

Martin, W. & Müller, M. (1998). The hydrogen hypothesis of the first eukaryote. *Nature* **392**, 37–41.

Martin, W., Stoebe, B., Goremykin, V., Hansmann, S., Hasegawa, M. & Kowallik, K. V. (1998). Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* **393**, 162–165.

Nei, M. (1996). Phylogenetic analysis in molecular evolutionary genetics. *Annu Rev Genet* **30**, 371–403.

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.

Pace, N. (1997). A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–730.

Philippe, H. & Laurent, J. (1998). How good are deep phylogenetic trees? *Curr Opin Genet Dev* **8**, 616–623.

Roger, A. J. (1999). Reconstructing early events in eukaryotic evolution. *Am Nat* **154**, S146–S163.

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

Sánchez, L. B. & Müller, M. (1998). Cloning and heterologous expression of *Entamoeba histolytica* adenylate kinase and uridylate/cytidylate kinase. *Gene* **209**, 219–228.

Smith, D. R., Doucette-Stamm, L. A., Deloughery, C. & 34 other authors (1997). Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: functional analysis and comparative genomics. *J Bacteriol* **179**, 7135–7155.

Stackebrandt, E., Rainey, F. A. & Ward-Rainey, N. (1996). Anoxygenic phototrophy across the phylogenetic spectrum: current understanding and future perspectives. *Arch Microbiol* **166**, 211–223.

Stephens, R. S., Kalman, S., Lammel, C. J. & 9 other authors (1998). Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* **282**, 754–759.

Stoebe, B. & Kowallik, K. V. (1999). Gene-cluster analysis in chloroplast genomics. *Trends Genet* **15**, 244–347.

Tomb, J. F., White, O., Kerlavage, A. R. & 41 other authors (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**, 539–547.

Wächtershäuser, G. (1998). Towards a reconstruction of ancestral genomes by gene cluster alignment. *Syst Appl Microbiol* **21**, 473–477.

Wasserfallen, A., Nölling, J., Pfister, P., Reeve, J. & Conway de Macario, E. (2000). Phylogenetic analysis of 18 thermophilic *Methanobacterium* isolates supports the proposals to create a new genus, *Methanothermobacter* gen. nov., and to reclassify several isolates in three species, *Methanothermobacter thermautotrophicus* comb. nov., *Methanothermobacter wolfeii* comb. nov. and *Methanothermobacter marburgensis* sp. nov. *Int J Syst Evol Microbiol* **50**, 43–53.

Xiong, J., Inoue, K. & Bauer, C. E. (1998). Tracking molecular evolution of photosynthesis by characterization of a major photosynthesis gene cluster from *Heliobacillus mobilis*. *Proc Natl Acad Sci USA* **95**, 14851–14856.

Yang, Z., Nielsen, R. & Hasegawa, M. (1998). Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol Biol Evol* **15**, 1600–1611.

Zillig, W. (1991). Comparative biochemistry of Archaea and Bacteria. *Curr Opin Genet Dev* **1**, 544–551.