CMLS, Cell. mol. life sci. 53 (1997) 34–50
1420-682X/97/010034-17 $ 1.50 + 0.20/0

CMLS Cellular and Molecular Life Sciences

Phylogenetic relationship of organisms obtained by ribosomal protein comparison

E.-C. Müller* and B. Wittmann-Liebold

Max-Delbrück-Centrum für Molekulare Medizin, Robert-Rössle-Str. 10, D-13122 Berlin (Germany),
Fax +49 30 94063869, e-mail: emu@mdc-berlin.de
Received 24 October 1996; accepted 30 October 1996

Abstract. The evolutionary relationships of ribosomal proteins from eubacteria, archaea, eukaryotes, chloroplasts and mitochondria were examined by their degree of conservation, their structural and functional properties and by the occurrence of conserved structural elements. The structural domains formed by the different protein families were studied. The occurrence of monophyletic groups was investigated for each protein family within the archaea. Phylogenetic trees were constructed between these organisms and together with the homologous sequences of the other phylogenetic domains. All organisms belonging to the archaea clearly formed a monophyletic group. The conserved sequence motifs were checked for the potential to form similar secondary structural elements.
Key words. Evolution; phylogenetic tree; eubacteria; archaea; eukarya; proteins.

Abbreviations. *A. castellanii = Acanthamoeba castellanii*; *A. polyphaga = Acanthamoeba polyphaga*; *A. nidulans = Aspergillus nidulans*; *A. longa = Astasia longa*; *B. stearothermophilus = Bacillus stearothermophilus*; *B. subtilis = Bacillus subtilis*; *C. elegans = Caenorhabditis elegans*; *C. trachomatis = Chlamydia trachomatis*; *C. reinhardtii = Chlamydomonas reinhardtii*; *C. ellipsoidea = Chlorella ellipsoidea*; *Cgdab = Citrus greening disease-associated bacterium*; *C. phi = Cryptomonas phi*; *C. paradoxa = Cyanophora paradoxa*; *D. mobilis = Desulfurococcus mobilis*; *E. virginia = Epifagus virginia*; *E. coli = Escherichia coli*; *E. gracilis = Euglena gracilis*; *H. influenzae = Haemophilus influenzae*; *H. marismortui = Haloarcula marismortui*; *H. halobium = Halobacterium halobium*; *H. morrhuae = Halobacterium morrhuae*; *H. volcanii = Halobacterium volcanii*; *L. biflexa = Leptospira biflexa*; *M. vannielii = Methanococcus vannielii*; *M. luteus = Micrococcus luteus*; *M. capricolum = Mycoplasma capricolum*; *M. genitalis = Mycoplasma genitalis*; *M. leprae = Mycobacterium leprae*; *M. smegmatis = Mycoplasma smegmatis*; *O. bertiana = Oenothera bertiana*; *P. tetraurelia = Paramecium tetraurelia*; *P. aphid = Pea aphid*; *P. thunbergii = Pinus thunbergii*; *P. sativum = Pisum sativum*; *P. purpurea = Porphyra purpurea*; *P. vulgaris = Proteus vulgaris*; *S. cerevisiae = Saccharomyces cerevisiae*; *S. typhimurium = Salmonella typhimurium*; *S. marcescens = Serratia marcescens*; *S. platensis = Spirulina platensis*; *S. carnosus = Staphyloccoccus carnosus*; *S. griseus = Streptomyces griseus*; *S. virginae = Streptomyces virginiae*; *S. acidocaldarius = Sulfolobus acidocaldarius*; *S. solfataricus = Sulfolobus solfataricus*; *T. pyriformis = Tetrahymena pyriformis*; *T. celer = Thermococcus celer*; *T. maritima = Thermotoga maritima*; *T. aquaticus = Thermus aquaticus*; *T. thermophilus = Thermus thermophilus*; *V. faba = Vicia faba.*

The study of macromolecules provides an important way of investigating the evolution of life. By comparing the 16S rRNA sequences from various organisms, three phylogenetic domains have been obtained: the archaea, the bacteria and the eukarya [1, 2]. Each of these domains has been proposed to be monophyletic and different from the others. The distinction both between archaea and bacteria and between these and the eukarya was confirmed by the group-specific shaping of homologous features [3]. The archaea seem to have two subdivisions based on 16S rRNA sequences [2]. One branch comprises a rather homogeneous group of often sulphur-dependent, extremely thermophilic organisms, and has been named the crenarchaeota (called eocytes by Lake [4]). The second branch comprises the euryarchaeota. This group contains methanogens and extreme halophiles. Lake [5] has argued from the construction of evolutionary trees of rRNA sequences that archaea are not monophyletic. He constructed a tree consisting of two main branches: the eukarya and crenarchaeota (which form one group) and the euryarchaeota and eubacteria (the other group).

* Corresponding author.

Wettach et al. [6] and Reiter et al. [7] found gene control sequences in DNA of the euryarchaeota *Methanococcus vannielii* and the crenarchaeota *Sulfolobus* sp. B12 that resemble the TATA box, a regulatory sequence stretch found in eukaryotic genes, but not in eubacterial ones. Creti et al. [8] and Rowlands et al. [9] had shown that *Pyrococcus woesei* expresses a protein with structural and functional similarities to eukaryotic TATA-binding protein (TBP) molecules. Marsh et al. [10] reported the identification of sequences of the archaebacterium *Thermococcus celer* that would produce a protein very similar to the TATA-binding protein found in eukaryotes. Both these results in conjunction with observations of similarity in RNA polymerase subunit composition [11] support the idea that the major features of the eukaryotic transcription apparatus were well established before the origin of eukaryotic cellular organization.

Gupta et al. [12, 13] investigated the 70-kD heat shock proteins (HSP70) and could not identify the three monophyletic branches proposed by Woese. Analyses of the HSP70 sequences indicate a close evolutionary relationship between the gram-positive group of bacte-

ria and the archaebacterial species on one hand, and the gram-negative bacteria and eukaryotic homologues on the other.

Since the substitutional bias is minimized in highly conserved proteins, inferences based on comparisons of their amino acid sequences have been proposed to be more reliable than those based on the corresponding nucleotide sequences. To obtain additional information about the phylogenetic relations of organisms of the three phylogenetic domains, we have investigated sequence data of the ribosomal proteins (rproteins). The ribosomes and their components provide an appropriate means for the study of evolutionary aspects, since this organelle performs the translation of the genetic information into proteins and occurs in all organisms [14]. More than 1900 complete rprotein sequences presently stored in databases are the foundation for constructing evolutionary trees with more reliability than was previously possible and to perform comparisons of their multiple alignments and predictions of common conserved secondary structural elements. The goal of the investigations described in this paper was to make an extensive comparison of r-protein sequences based on the huge and still increasing data available, to obtain clues on the phylogenetic relationship between such distantly related organisms as eubacteria, archaea and eukaryotes, and to find hints for the grouping of rproteins that are derived from mitochondria.

We have taken into consideration five groups of ribosomal proteins: eubacterial and archaebacterial proteins, and proteins encoded by chloroplast, mitochondrial or nuclear genomes. Only complete sequences are suitable for phylogenetic considerations.

Escherichia coli and *Bacillus stearothermophilus* are the best characterized representatives of the eubacteria. For the *E. coli* ribosome all proteins have been established [15], and for the *B. stearothermophilus* organelle complete sequences have been determined for 47 proteins contained in the NBRF, SwissProt and RIBO databases [RIBO is a database of the Max Planck Institute, Berlin (B. Wittmann-Liebold, A. Köpke, M. Dzionara et al., unpublished)], complemented by the sequences of BstS14 [16] and of proteins S2, S4, S6, S8, S16, L11 (M. Kimura, personal communication).

The chloroplast equivalents of 44 out of the 55 ribosomal proteins that constitute the *E. coli* ribosome have been identified by sequencing the corresponding nuclear or chloroplast genes from land plants, cyanobacteria and algae. The chloroplast-encoded proteins show somewhat greater sequence identities to their counterparts from *E. coli* than the nuclear-encoded ones [17].

Mitochondria and other plastids contain ribosomes whose constituent proteins are partly encoded by the organelle genome, while the others are specified by the nuclear genome and imported into the organelle post-translationally. The databases hold about 30 complete mitochondrial sequences from yeast, mitochondrial L3 sequences of human and *Rattus norwegicus* (rat), and also some mitochondrial sequences of land plants. Takemura et al. [18] determined the complete sequences of liverwort mitochondrial DNA and identified genes encoding 16 different ribosomal proteins (S1, S2, S3, S4, S7, S8, S10, S11, S12, S13, S14, S19, L2, L5, L6, L16). The sequences of the ciliates and the amoeboid protozoon are of interest for their relationship to the sequences of the land plants. The complete primary sequence of the mitochondrial DNA of *Acanthamoeba castellanii* was determined by Burger et al. [19]. This mtDNA encodes 16 ribosomal proteins similar to the mtDNA from liverwort: liverwort has proteins S1 and S10, but *Acanthamoeba* has the genes of proteins L11 and L14. Vodkin et al. [20] found an rpl14 gene in mtDNA of *A. polyphaga*. The ciliated protozoon *Paramecium tetraurelia* retained four genes encoding proteins L2, L14, S12 and S14 [21].

On the basis of the 16S rRNA the phylogenetic tree of the archaea consists of two subdivisions [22]: the crenarchaeota with *Sulfolobus solfataricus* and *S. acidocaldarius*, and the euryarchaeota with *Haloarcula marismortui*. Forty-nine ribosomal sequences of *H. marismortui* have been published (October 1996). Most of them show similarities to eukaryotic and eubacterial rproteins, but for some of them no counterpart could be detected. Furthermore, 24 sequences of *S. solfataricus* and *S. acidocaldarius* have been recorded in the databases.

Seventy-five of 80 rproteins of rat and human have been determined from the eukaryotic group of organisms; 65 yeast proteins are known to be related to the rat proteins [23].

Materials and methods

The amino acid sequence data and the nucleic acid sequences were obtained from the protein and nucleic acid databases SWISSPIR and GENEMBL. SWISSPIR comprises the entries of SWISSPROT, the PIR entries not included in SWISSPROT and weekly updates of SWISSPROT SWnew.

To get an evolutionary tree for each protein family from S1 to S21 (small subunit), and from L1 to L36 (large subunit), we performed a FASTA search, looking for the most related ribosomal sequences of *E. coli*. FASTA is a program of the GCG program package of the University of Wisconsin Genetics Computer Group [24]. A multiple global alignment of the sequences was performed with the GCG program PILEUP [25]. The resulting multiple sequence files were the basis for the construction of phylogenetic trees with the PROPTREE program [26, 27]. In this new approach each amino acid is represented by a vector of 11 steric and physicochemical properties [28], either present or not. Such properties (hydrophobic, polar, charged, aromatic, aliphatic

Table 1. The most conserved ribosomal proteins with respect to *E. coli*. Homology to the other species is given in percentage of identical residues, length of the homologous region (computed by means of FASTA) and length of the whole sequence (SwissProt, April 1995).

	E. coli	*B. stearothermophilus*	Liverwort (chloroplast)	Liverwort (mitochondrial)	*H. marismortui*	Rat
S11	128	68.8/121/128	52.5/120/130	52.7/110/125	45.4/119/128	41.2/119/151
S12	123	67.9/137/139	70.7/123/123	61.8/123/126	34.4/90/147[b]	29.0/112/142
S19	91	66.3/90/91	62.0/92/92	46.1/76/93	40.5/84/140	33.3/78/145
L14	123	64.5/121/122	53.7/121/122	-	27.3/121/132	35.9/103/140
L2	272	60.4/275/275	49.5/277/277	50.7/142/501	38.1/235/239	30.3/241/257
L5	178	59.3/177/179	43.2/176/179[c]	26.5/181/188	39.7/131/176	31.3/134/178
S3	232	58.4/219/218	41.9/217/217	32.5/120/430	27.3/183/304	25.8/186/243
S7	178	57.7/156/155	43.2/155/155	42.1/140/230	26.2/141/205	29.0/148/204
S5	166	54.9/164/166	-	-	27.5/160/212	27.7/148/293
S17	83	52.4/82/85	-	-	45.0/80/111	33.3/84/157
S4	206	49.8/205/198	36.4/206/202	29.3/164/196	29.0/69/171	30.0/90/193
L6	176	49.4/174/177	-	38.7/93/101	24.6/179/177	22.8/192/192
S8	129	49.2/128/130[a]	44.2/129/132	42.4/92/152	26.7/131/129	25.4/126/129

[a]M. Kimura, personal communication, [b]*M. vannielii*, [c]*Euglena gracilis*.

etc.) are well known to be essential for the protein structure formation resulting from the underlying sequences. No assumptions concerning the evolutionary pathway from one individual to another are needed to construct the trees. One of the main advantages of tree construction with PROPTREE is that determination of the branching order and the calculation of the branch lengths are simultaneous steps and negative branches are always excluded by the construction principle.

To construct phylogenetic trees, we also used the programs TREE [25] and CLUSTREE [29] which are part of the HUSAR package [30]. The multiple alignments done by TREE are obtained by pairwise alignments of the sequences. The closer two sequences resemble each other, the more confidence one has in the alignment. The gaps in the alignments of two closely related sequences are not ignored merely because an alignment with some distantly related sequence might be improved. The branching order of the sequences in the tree is determined according to the method of Fitch and Margoliash [31]. The branch length is computed with a least-squares approach. The method used in CLUSTREE is the Neighbor-Joining (NJ) method of Saitou and Nei [32]. First, percent divergence figures are calculated between all pairs of sequences. These divergence figures are then used by the NJ method to give the tree. The NJ method does not explicitly assume a constant rate of evolution and is known to give a correct topology at a high rate in a wide variety of situations [33]. A bootstrap algorithm can be used to show confidence levels for groupings.

All trees computed with CLUSTREE, TREE and PROPTREE are unrooted trees.

To predict the secondary structure, we used the service of the European Molecular Biology Laboratory (EMBL) Heidelberg World Wide Web (WWW) server [34]. For homologous proteins, alignment procedures predict the secondary structure more accurately than any method using the sequence information only [35]. Two amino acid sequences evolved in nature are almost sure to have identical space structure if they share 30% amino acids [36]. Rost and Sander [37] use this evolutionary information in multiple sequence alignments as input to neural networks. Using a position-specific conservation weight as part of the input increases the accuracy of prediction. The networks were trained on different sets of proteins with known three-dimensional structures of the Brookhaven Protein Data Bank. The average accuracy for all sequence-unique chains is above 72%.

Results

Most conserved ribosomal proteins. Table 1 lists the most conserved rproteins with respect to *E. coli* in decreasing order and contains the percentage identities, the lengths of the homologous regions and the lengths of the entire sequences. The representatives of *B. stearothermophilus* (eubacteria), liverwort (chloroplasts), liverwort (mitochondria), *H. marismortui* (archaea) and rat (eukarya) were chosen to allow for a general representation of the data.

High sequence similarities were found for some eubacterial ribosomal proteins of functional importance, such as proteins S7, S11, S12, S19 and L2. *E. coli* S12 is essential for maintaining the accurate message translation by the ribosomes [38] and is related to the S12 counterparts, the human and rat S23 proteins [39]. A consensus sequence of 17 amino acids (between residues 40 and 56) is present in the alignment of eubacteria, chloroplasts and mitochondria. The rat sequence matches at 10 positions. In this region the lysine at position 42 of *E. coli* causes resistance to streptomycin and leads to increased accuracy of translation by the ribosomes [40]. This lysine is conserved in rat S23 at position 60, *H. halobium* S12 (62), *H. morrhuae* (61) and

M. vannielii (63), but in *S. acidocaldarius* (30) an arginine is aligned. Spiramycin binds to the S12 protein and to proteins L27, L35, L17 and L18 of the large subunit inhibiting the growth of the polypeptide chain [41].

Together with S4 and S5, protein S12 participates in a region designated as the recognition complex by Oakes et al. [42]. In *E. coli*, S4 is one of the primary rRNA-binding proteins that initiates assembly of the 30S subunit and is known to protect several 16S rRNA regions from chemical modification during assembly [43]. The similarity of protein S4 to other eubacteria is high, but it is limited to a relatively short region of residues 100 to 153 (rat S9) and 98 to 143 (*H. marismortui*).

The proteins of the S3 family are also highly conserved phylogenetically. The *E. coli* S3 interacts with the 16S rRNA, and it resides in the head of the small subunit near the site to which the polypeptide release factor RF-2 binds [44]. The so-called KH motif exists in the S3 proteins of all domains [45]. This motif was first identified in human hnRNP K protein [46]. Cross-link studies [47] demonstrate a direct interaction between the KH motif of *E. coli* S3 and the 16S RNA. The cross-linking position 88 (lysine) was found near the core sequence of the KH motif.

In *E. coli*, protein S7 is a primary binding protein to 16S rRNA, its binding site has been established [47], and this binding is a prerequisite for the assembly of S9 and S19 [43].

The proteins S8 and S11 interact with a highly conserved site in the central domain of the 16S rRNA named the platform ring [42]. The same is true for S6 and S18, but no relatives to *E. coli* S6 from chloroplasts or from eukarya have so far been found in the databases. The homology with other eubacteria (*B. subtilis* and *Thermus thermophilus*) is only about 30%. Eubacteria and chloroplasts with sequences related to S18 of *E. coli* exist, but no homologues from eukarya and archaea are known.

S17 is one of the primary assembly proteins in *E. coli*. It binds to the 16S rRNA in the 5′ domain [43]. S19 of *E. coli* interacts with proteins S7, S9 and S14. It was found in close proximity to protein S13 [48].

In *E. coli*, the highly conserved proteins L1, L2, L6 and L11 belong to the early-assembly proteins of the large subunit [49]. The L1 protein forms the ridge on the large subunit. Protein L2 is part of the peptidyltransferase centre, and it is the most conserved protein from the large subunit. It can even be substituted in *E. coli* ribosomes by the halophilic protein, HmaL2, and the human L8 equivalent (M. Ühlein, unpublished). L18 binds to 5S rRNA and is also associated with the peptidyltransferase centre [50]. If one considers these proteins, the high conservation coincides with an important functional role in the translational process. However, the highly conserved L14 protein is among the late-assembly proteins and does not bind directly to 23S rRNA [49]. The similarity of the rat L1 to *E. coli* is very low (about 10%), and there are no eukaryotic relatives to *E. coli* L11 (see below). Furthermore, proteins L16, L17, L20 and L24 – all functionally important proteins – show no high conservation among the phylogenetic domains.

The sequences of S7, S10 and S12 of *E. coli*, *Haemophilus influenzae*, *Mycobacterum leprae* and *Th. thermophilus* (E–eubacteria), human and *S. cerevisiae* (N–eukarya), *Halobacterum halobium*, *S. solfataricus*, and *M. vannielii* (A–archaea) were added to get a longer sequence for the phylogenetic computations. The multiple alignment (fig. 1) represents – in bold letters and in the consensus line – the high conservation of these proteins and the greater similarity of archaea and eukarya against eubacteria.

Archaeal phylogenies. Conflicting data in the literature describe the grouping of organisms combined to form the phylogenetic domain of archaea. The question arises whether the ribosomal proteins can shed light on this much discussed field in evolution.

The crenarchaeota representative with the largest number of known ribosomal sequences is *S. solfataricus*; the corresponding euryarchaeota is *H. marismortui*. Table 2 shows the protein families with the largest number of known archaebacterial sequences and the lengths of these proteins. We compared these sequences with the others to see whether or not the archaea form a monophyletic group.

The following *S. solfataricus* amino acid sequences so far exist in the databases: S7, S10, S12, L1, L5, L10, L11, L12, L46, LX. Gene organization of the archaeal rproteins in general is in clusters as found for the eubacterial genome, although differences occur for gene replacements, gene transfer, changes in promotor location and in differences in the intervening sequences [51, 14]. The genes for the *S. solfataricus* S12, S7, S10 and the elongation factor 1α protein are arranged analogously as in the *E. coli* str operon [52]. The genes for the *S. solfataricus*, *S. acidocaldarius* and *Halobacterium cutirubrum* L11, L1, L10, and L12 proteins occur in the same order as the equivalent genes in *E. coli* [53, 54].

Sequence similarity compared by phylogenetic tree relation shows that the tree of the S7 family is monophyletic with respect to the archaea (see fig. 2). The archaebacterial proteins share nearly 50% with the eukaryotic rat sequence, *S. solfataricus* with *E. coli* 31%, *M. vannielii* with *E. coli* 34% (calculated with BESTFIT from the GCG package). Like the tree of S7, the tree of S12 consists of a monophyletic archaebacterial branch joined with the eukaryotic sequences, and these two form the whole tree with the group of eubacterial, chloroplast and mitochondrial sequences.

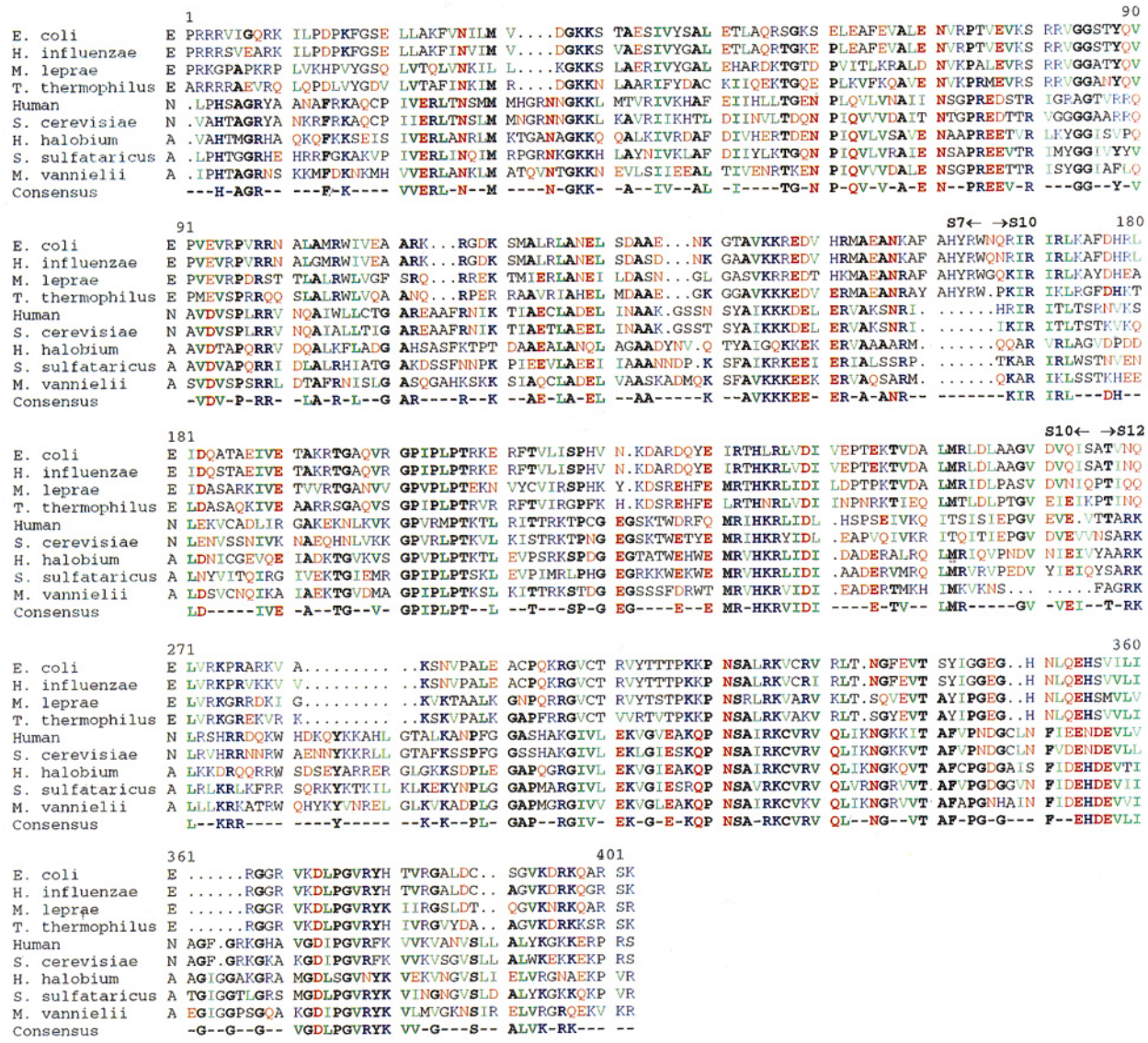

Figure 1. Multiple alignment of a combined sequence from S7, S10 and S12. N-terminal and C-terminal parts without similarity are removed. Conserved amino acids are shown in bold letters, acidic amino acids D, E, N and Q in red; basic H, K and R in blue; very similar I, L and V in green. The greater similarity of archaeal (A) and eukaryotic (N) sequences as compared to the eubacterial ones (E) is recognizable.

The archaebacterial S10 proteins show sequence similarity to the eubacterial S10 proteins and to the eukaryotic S20 proteins. All three groups are monophyletic, but the distance between the archaebacterial branch and the eubacterial branch seems to be smaller than that between the archaebacterial branch and the eukaryotic one. Some significant changes in the S10 sequences are found: the conserved sequence GPIPLPTK in archaea and eubacteria corresponds to GPIRMPTK in eukarya. The most important difference is the length of the eukaryotic sequences: they are about 15 aa longer than the other proteins. Removing these residues at the N-terminal end, the tree is changed to the branching order of the other families (E, (A, N)).

The archaebacterial and the eubacterial L1 proteins each form a monophyletic group. A weak, statistically insignificant similarity to L3 eukaryotic proteins is found. On the other hand, the L5 archaebacterial proteins are more related to the eukaryotic L11 proteins than to the eubacterial group. The situation is the same for the L10 family. The problem in this family is due to the different lengths: eubacteria from 165 to 179 aa, archaea from 335 to 352 aa, and eukarya from 305 to 322 aa. There is only one conserved region at the N-terminal end with five amino acids in a row in the three phylogenetic domains and some smaller regions or single conserved amino acids in the sequences. In many cases inserts or deletions are necessary to obtain the alignment. Comparisons of the percent identity in the protein alignments indicate that a higher identity can be found between the archaea and the eukaryotic P0 proteins than between the archaeal L10 and the eubacterial L10 counterparts.

Table 2. Protein families with the largest number of known archaebacterial sequences and their lengths.

	S7	S10	S12	L1	L5	L10	L11	L12
S. solfataricus	193	102	147	221		335	170	105
S. acidocaldarius	195	102	151		178			105
H. cutirubrum				212		352	163	114
H. halobium	210		142	212		352		114
H. marismortui	205	99		212	177	348	162	115
Haloferax volcanii				210		350	159	113
Halococcus morrhuae	203		142					
M. vannielii	194	91	147	222	181	336		99
Th. celer	215		147					
Th. acidophilum		104						
Pyrococcus woesei		102						

Compared with the respective *H. marismortui* and *S. solfataricus* sequences, the *E. coli* L11 sequence exhibits 33 and 35% identities. There is a weak similarity of 21% to the rat and *S. cerevisiae* L12 sequences. The alignment shows strongly conserved regions of eubacterial and archaebacterial proteins; the eukaryotic proteins are scarcely involved. Within the aligned region, *E. coli*, *S. solfataricus* and *H. marismortui* L11 proteins contain nine, ten and 11 proline residues. At seven positions, the proline residues are conserved in all three sequences. The rat L12 protein contains eight proline residues, only two of which are conserved. In this case the similarity of eubacterial and archaebacterial sequences is also reflected in the predicted secondary structure and the phylogenetic tree (fig. 3). The secondary structures of the eukayotic L12 and archaebacterial L11 sequences are quite different. Against the background of these data we cannot consider the eukaryotic L12 proteins to be relatives of the archaebacterial L11 proteins.

The archaebacterial L12 proteins form a monophyletic group and are most related to the P2 protein group of rat, human and yeast. The branch of archaea and P2 proteins is joined with the P1 group of eukaryotic proteins.

The result of the computation of all phylogenetic trees with representatives of crenarchaeota and euryarcheota suggests that the archaea are in fact a monophyletic group more similar to each other than to eukarya or eubacteria. This corresponds to the higher identity within the archaea than to the other phylogenetic domains.

Tree comparison of ribosomal proteins with respect to mitochondrial rproteins. While the branching of the ribosomal chloroplasts and eubacteria ensues according to their taxonomic affiliations, the branching of the mitochondrial proteins is not uniform. Many mitochondrial proteins are considerably longer than the corresponding *E. coli* proteins, which makes it difficult to yield accurate alignments. The lengths and percentage identities of the *E. coli*, *A. castellanii*, and the mitochondrial liverwort are shown in table 3. Proteins S12, L2 and L14 of *A. castellanii* are the most conserved proteins, whereas S2, S3, S11, L5 , L6 and L11 are not significantly similar to the *E. coli* counterparts.

Within the trees we find neither a monophyletic group of mitochondrial proteins nor a similar branching order in all trees. Furthermore, different trees were computed with PROPTREE and TREE.

The S3 family comprises ribosomal proteins of eubacteria, archaea and eukarya from chloroplast, nuclear and mitochondrial genomes (E, A, C, N, M). The archaea *H. marismortui* and *H. halobium* form a monophyletic group; likewise, the eukaryotic proteins are encoded by the nuclear genes. The eubacterial proteins are closely related to the proteins of chloroplast genomes derived from land plants, algae and *Cyanophora paradoxa*. The distances between the species are small, and they might change as new sequences become available. The S3 chloroplast rproteins of *Chlamydomonas reinhardtii*, *C. humicola*, *C. peterfii* and *C. frankii* are considerably longer than those found in chloroplasts of land plants (682 and 807 versus 218). They are quite different from the other eubacterial and all other chloroplast sequences within the aligned region. The mitochondrial sequences of land plants form a monophyletic group. The mitochondrial S3 of *A. castellanii* consists of 294 aa. It seems to be more related to the group of eubacteria and chloroplasts than to the mitochondrial group. This sequence is too distant from the sequences of the archaea and eukarya (N) to compute a tree. We can describe the complete tree computed without *A. castellanii* as: (((E, C), M), (A, N)).

The *rps3* gene encoding the S3 proteins of chloroplasts has been found between *rpl22* and *rpl16*, as is true for the genes in the S10 operon of *E. coli*. In most algae and land plants the *rps3* gene is uninterrupted, but in *Euglena gracilis* it contains two introns [55]. In the chloroplast genome of the *Chlamydomonas* species there is no gene equivalent to *rsp3* in the expected region between *rpl22* and *rpl16* [56].

Two or more representatives of each type (A, N, E, C, M) have been sequenced from the L14 family. The phylogenetic tree constructed with TREE (fig. 4) shows the grouping of the organisms according to their taxonomic affiliations except for *Th. thermophilus*. The archaea and eukarya (nuclear genes) are robustly separated

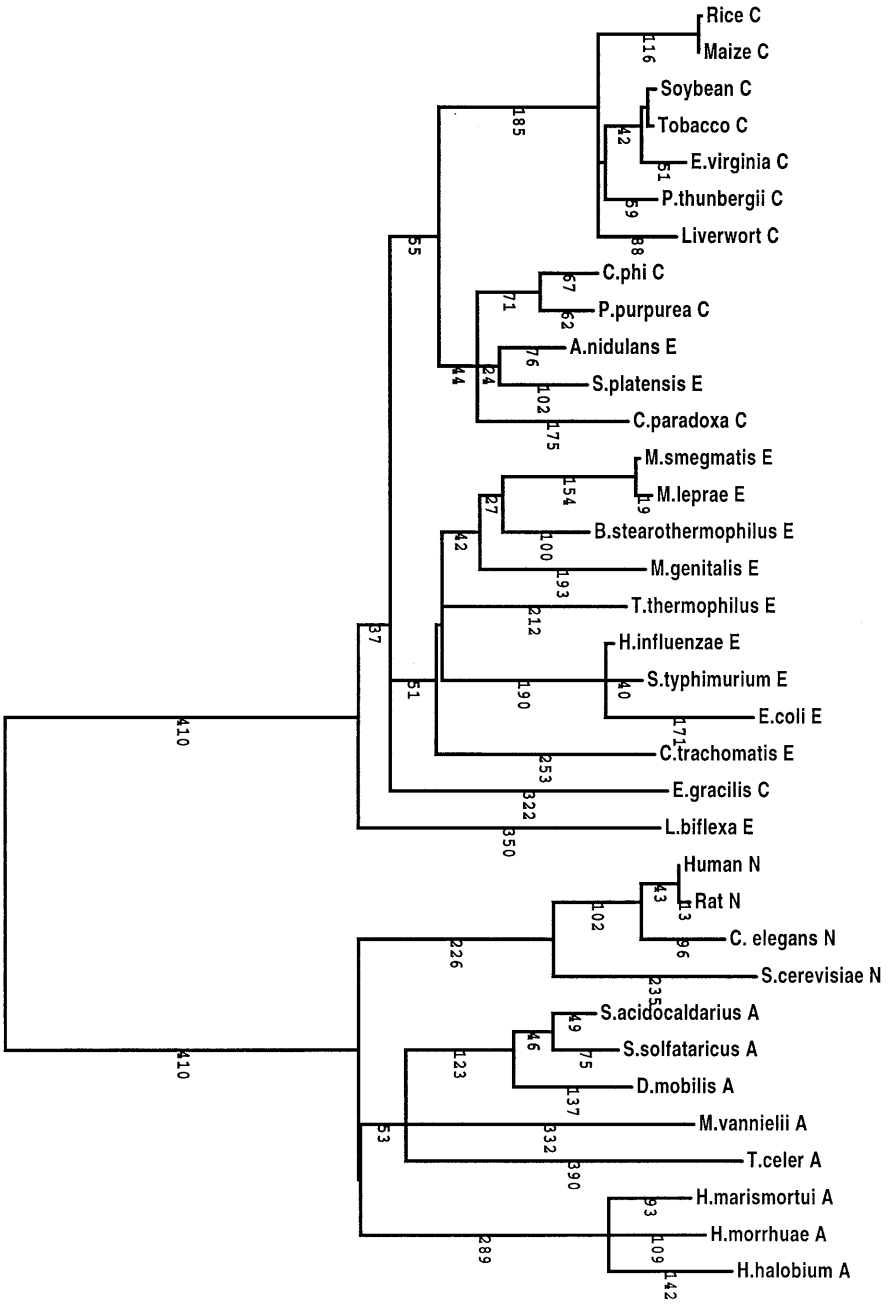

Figure 2. Phylogenetic tree of the S7 rproteins with the monophyletic archaeal (A) group, the nuclear encoded (N) group, the eubacterial (E) group and the chloroplast (C).

from the eubacteria and eukarya (chloroplasts). The mitochondrial sequences were found to be most related to the branch of eubacteria and chloroplasts. This result does not correspond to the tree of Vodkin et al. [20] computed with the PAUP package [57], where the mitochondrial representatives *Acanthamoeba polyphaga*, *Tetrahymena pyriformis* and *Paramecium tetraurelia* are the outermost ones. Because of the properties of the amino acids the tree computed with PROPTREE differs from the result obtained with TREE. It looks like the tree of Vodkin et al. [20]; however, the proteins from *A. polyphaga* and the ciliates are exchanged. Calculations of the tree with these parts of the sequences occurring in all species did not change the tree. Alignments reveal that the proteins of *A. polyphaga* and *A. castellanii* have an insert of six amino acids (near aa 52) that are not found in the homologous proteins of the other organisms. On the other hand, in *A. polyphaga* the sequence KKNI

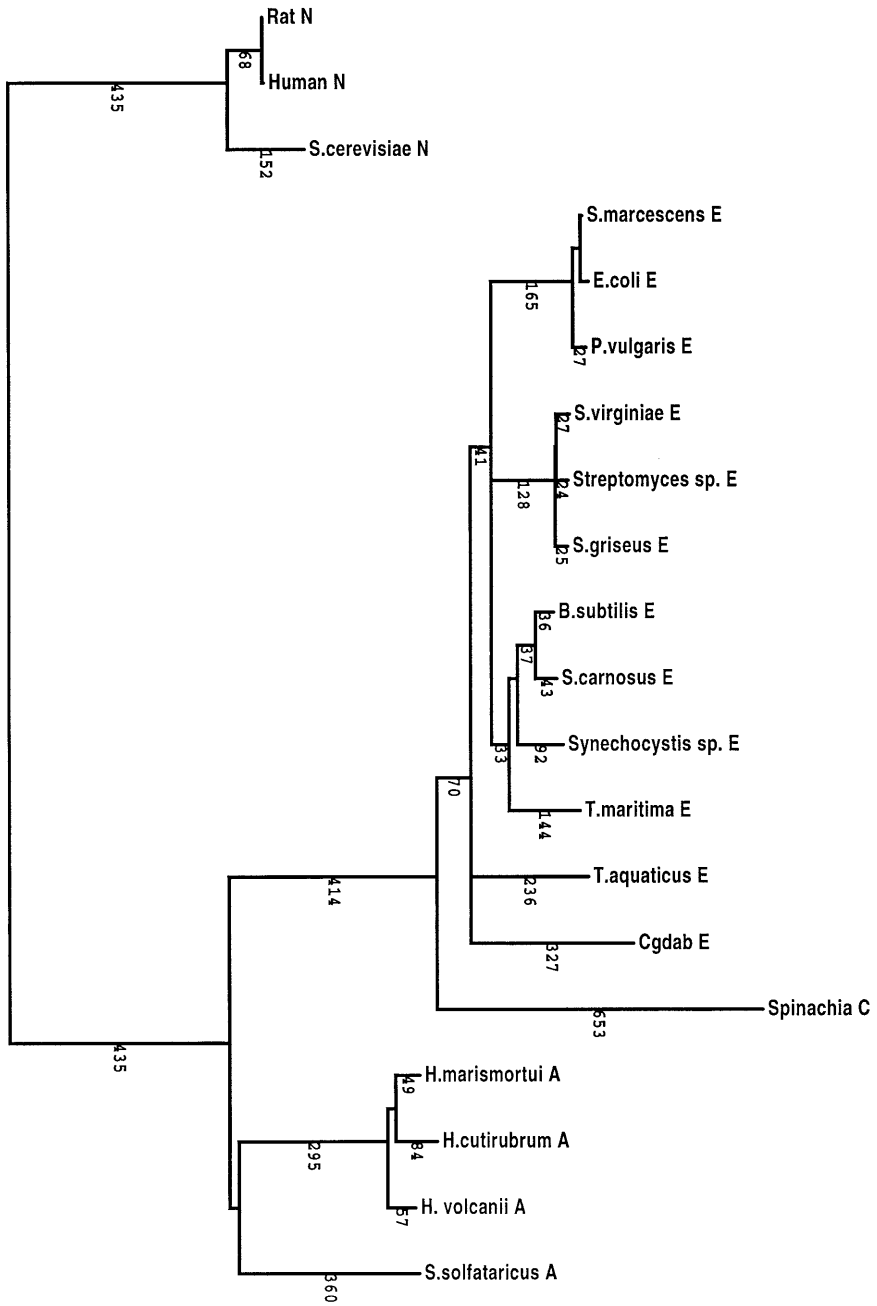

Figure 3. Tree of the L11 rproteins. The archaea are more related to the L11 eubacterial and chloroplast species than to the L12 eukarya.

occurs twice; in *A. castellanii* the sequence KKSI is repeated as KKNI. However, this insert does not influence the position in the tree.

The tree of the L2 proteins computed with PROPTREE can be described as ((((A, N), (E, C)), (*A. castellanii*, Yeast)), *P. tetraurelia*); the TREE result is ((A, N), (((E, C), *A. castellanii*), Yeast), *P. tetraurelia*). The L2 of *E. coli* (272 aa) and *P. tetraurelia* (259 aa) correspond in 27% of 212 aa; L2 *E. coli* and *A. castellanii* are identical in 47% of the sequences. The prediction of the secondary structure of *P. tetraurelia* differs from that computed for *E. coli* and other L2 proteins. In contrast, the percentage identity of L14 of *P. tetraurelia* (119 aa) and *E. coli* (123 aa) is 36% in a region of 113 aa, and the prediction of the secondary structure of both proteins corresponds.

The proteins of the S14 family are quite different in their lengths. Chloroplasts, mitochondria and the eubacteria *E. coli* and *Pea aphid* are about 100 aa, other eubacteria, archaea and eukarya are about 60 aa, and yeast mitochondrial is 115 aa. The multiple alignment shows a gap of 42 aa near the N-terminal end for the

Table 3. The *A. castellanii* proteins: their lengths and the percentage identity in comparison with *E. coli* and mitochondrial Liverwort.

	E. coli	*A. cast.*	Liverwort mt	FASTA *A. cast.*/ *E. coli*	FASTA *A. cast.*/ Liverwort
S2	240	312	237	29.2/72	25.4/118
S3	232	294	430	19.1/110	21.7/263
S4	206	374	196	21.7/166	25.3/182
S7	178	337	230	24.8/141	
S8	129	127	152	29.5/122	36.0/100
S11	128	173	125		
S12	123	127	126	48.4/124	50.4/127
S13	117	119	120	28.6/112	33.0/103
S14	98	99	99	26.3/80	36.5/96
S19	91	78	93	32.8/58	30.6/85
L2	272	253	501	47.4/249	45.0/249
L5	178	177	188		
L6	176	181	101		29.0/100
L11	141	339		24.4/90	
L14	123	129		41.6/125	
L16	136	140	135	34.9/120	26.8/127

Table 4. Homology between eubacterial and eukaryotic sequences and the tree representation (according to Felsenstein [61]) of archaea (A), eubacteria (E), chloroplast-encoded (C), mitochondrial-encoded (M) and nuclear-encoded (N) sequences in phylogenetic trees.

Eubacteria, archaea, chloroplasts, mitochondria	Eukarya	Tree representation	
S3	S3	(((E, C), M),	(A, N))
S4	S9	(((E, C), M),	(A, N))**
S5	S2	(E,	(A, N))
S7	S5	(((E, C), M),	(A, N))**
S8	S15a	(((E, C), M),	(A, N))
S9	S16	(E,	N)
S10	S20	(E,	(A, N))
S11	S14	((((E, C), Liverwort),	(A, N), *A. cast*)
S12	S23, S28 (Yeast)	(((E, C), M),	(A, N))
S13	S18	(((E, C), M),	(A, N))
S14	S29	(((E, C), M),	(A, N))*
S17	S11	((E, C),	(A, N))
S19	S15	(((E, C), M),	(A, N))
L2	L8	(((((E, C), *A. cast*),	(A, N)), *P. tetr.*), (Liv., Primrose))
L3, L9 (mt Yeast)	L3	((E, M),	(A, N))
L5	L11	(((E, C, Liverwort),	(A, N)), *A. cast*)
L6	L9	(((E, Yeast), (A, N)	(*A. cast.*, Liverwort))
L10	P0	(E,	(A, N))
L12	P1, P2	(	A, N)
L13	L13a	((E, C),	(A, N))
L14	L23	(((E, C), M)	(A, N))*
L15	L27	(E,	(A, N))
L18	L5	(E,	(A, N))
L22	L17	((E, C),	(A, N))
L23	L23a	((E, C),	(A, N))
L24	L26	((E, C),	(A, N))
L30	L7		(A, N)

*See figures, **the sequences of *A. castellanii* were shortened at the N-terminal end where no alignment exists with other proteins.

short eubacteria. The archaea and eukarya are slightly similar (*E. coli* – rat: 30.8% in 26 aa) to all the other S14 proteins. This fact is reflected in the prediction of the secondary structure, which is similar only in the C-terminal region for all S14 proteins. The tree constructed with TREE (fig. 5a) differs from other trees because of the protein lengths. The short eubacteria are joined with the branch of eukarya and archaea. The mitochondria form a branch which is connected with chloroplasts and long eubacteria. With PROPTREE (fig. 5b) we obtained a correct branching order with respect to eubacteria and chloroplasts. However, the mitochondria including *A. castellanii* are the outermost group.

Table 4 summarizes these eubacterial r-proteins which show similarity to the eukaryotic ones and the branching order within the phylogenetic tree.

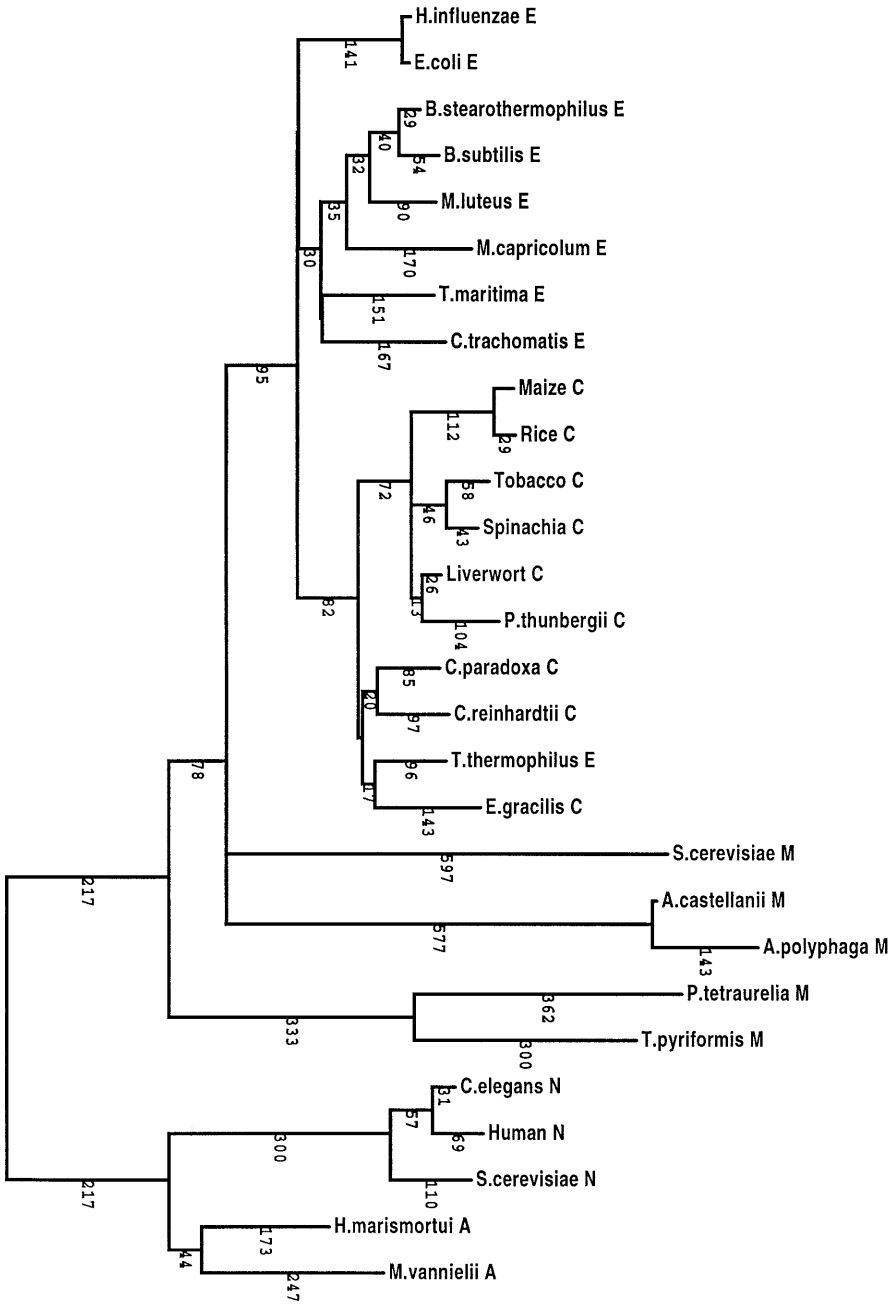

Figure 4. Tree computed with TREE of the L14 rproteins.

Comparison of the tree computional methods. The CLUSTREE, PROPTREE and TREE programs were tested with the multiple sequence file shown in figure 1. The length is 402 amino acids. The unrooted trees computed by CLUSTREE (fig. 6a), PROPTREE (fig. 6b) and TREE (fig. 6c) are almost the same with respect to the eubacterial branch. The second branch, consisting of the eukaryotic species at the one side and the archaea at the other, is very similar in PROPTREE and TREE computation. The CLUSTREE result does not show a clear monophyletic group of the archaea, but the branch lengths of all archaea are shorter to the eukaryotic than to the eubacterial species. The S7 result of CLUSTREE with eight archaea is equivalent to the result of TREE (fig. 2) with a monophyletic branch of archaea.

Conserved regions and secondary structure. We performed calculations for predicting the secondary structure with one representative each of archaea, eubacteria, chloroplasts, mitochondria and eukarya to examine any correlation between conservation and structural features. We wanted to find out whether or not the conserved amino acids are situated in helices, β-sheets or loops. The predicted classes (method of Rost and Sander [37]) of the most conserved *E. coli* proteins are shown in table 5.

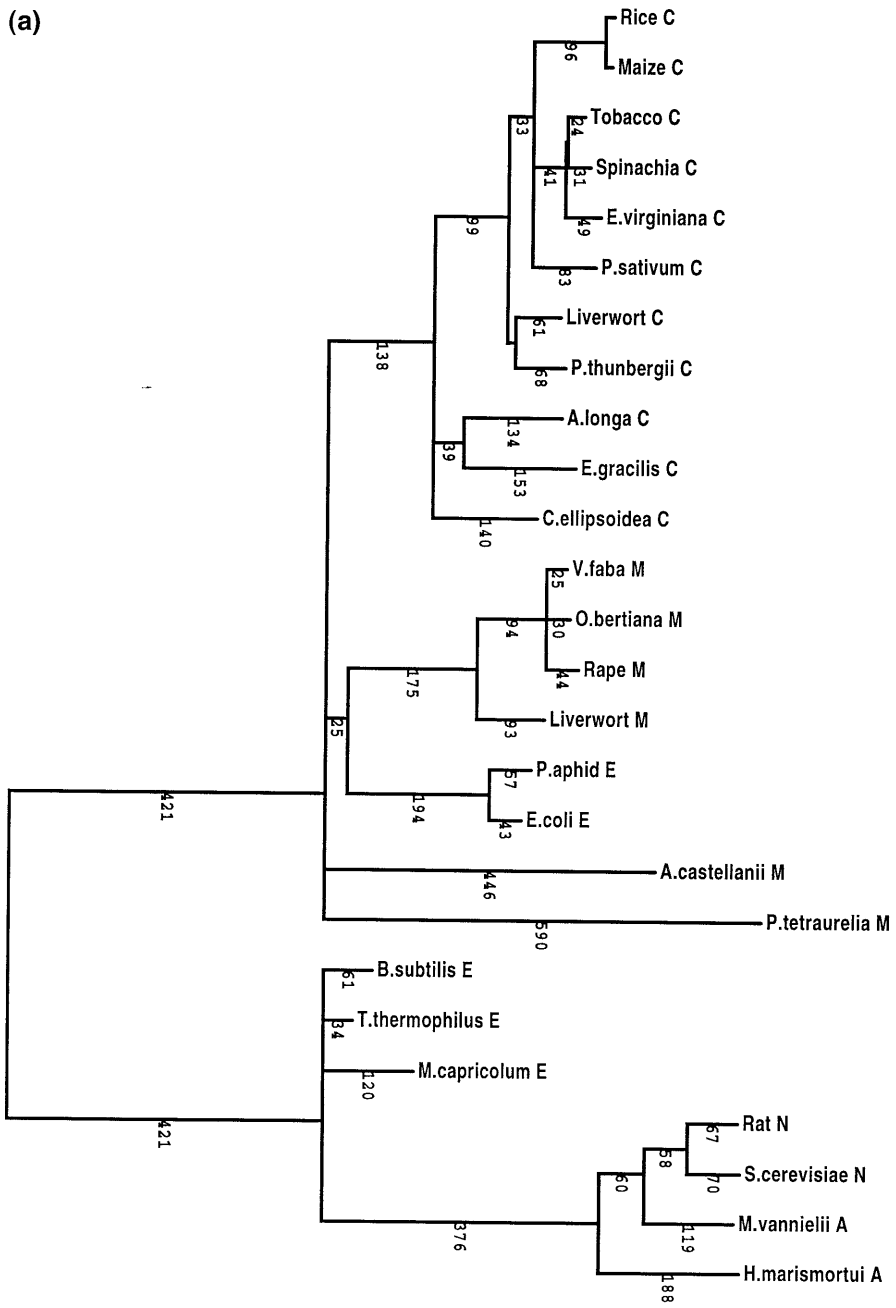

Figure 5. (a).

The S3 sequences differ in the predicted secondary structure of eubacteria versus rat and *H. marismortui* in the area following the KH sequence. Helical and non-helical regions are predicted within the conserved regions.

The region with the highest similarities in the S4 sequences extends from aa 94 to 138 (*E. coli*) and was predicted as α-helical and β-strand elements for all domains. A highly conserved region exists at the C-terminal end of eubacteria and chloroplasts predicted as β-strand.

The 3D structure of the S5 protein of *B. stearothermophilus* was reported by Ramakrishnan and White [58]. The loop2 region (aa 21 to 31) contains conserved arginine and lysine residues which are expected to interact with the 16S rRNA, and this was confirmed by direct sequencing of the contact region between the peptide and oligonucleotide binding site [47]. The structurally important residues are particularly well conserved within all domains and may also contain helices.

A similar secondary structure is predicted for archaea, eukarya and eubacteria of the S7 proteins, and these are

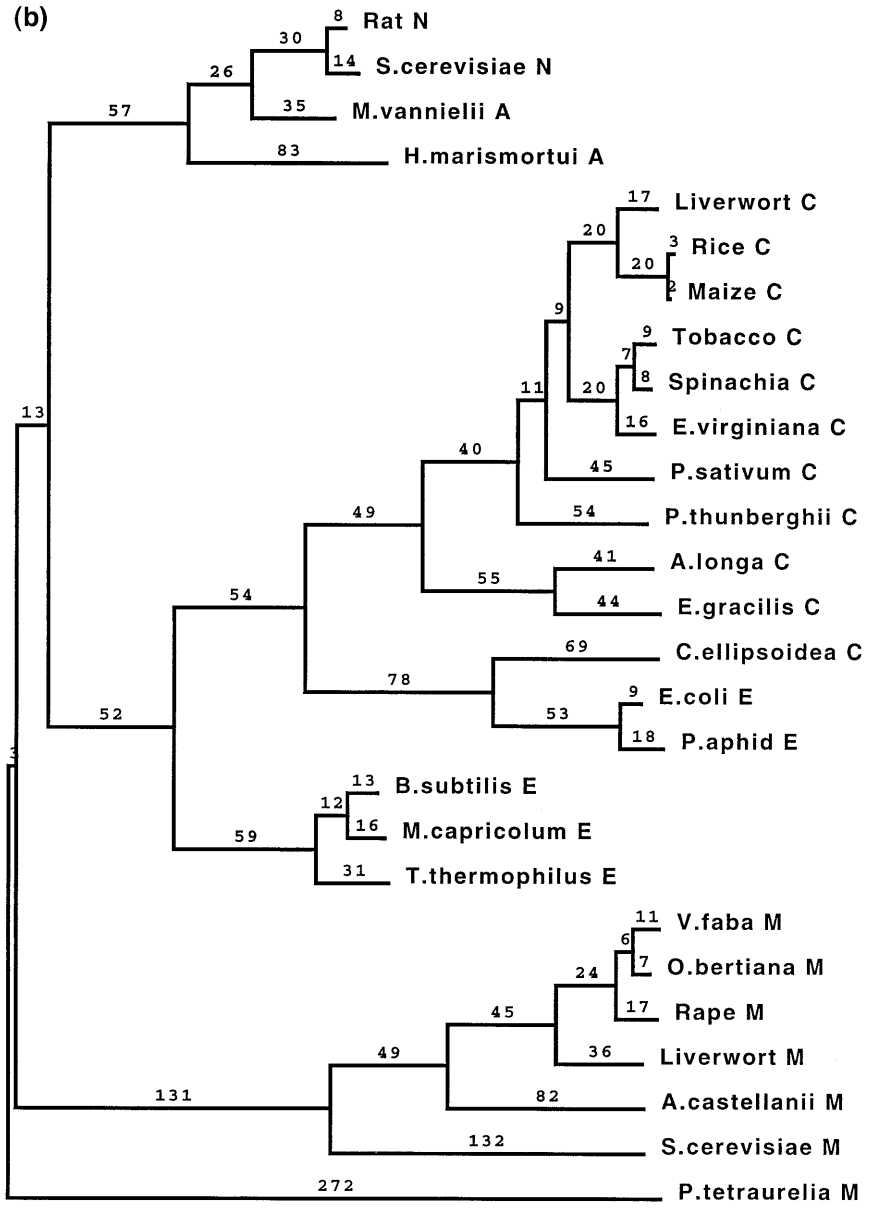

Figure 5. Tree of the S14 rproteins computed with TREE. (a) Because of their sequence lengths, the eubacteria *B. subtilis*, *T. thermophilus* and *M. capricolum* seem to form a group with the eukarya and the archaea. (b) Branching order according to the taxonomic affiliations of the eubacterial sequences computed with PROPTREE. With TREE, the mitochondrial sequences form the outermost group because of their amino acid alterations in the aligned sequences.

predicted to contain more than 40% α-helical areas and only 14% as β-strands. One helical structure is situated in a conserved region with the consensus sequence GKKxxAxxIxxAxxxIxxxT (x various amino acids). Two of three peptides of the S7 protein of *E. coli* which bind to the 16S rRNA [47] are predicted in α-helices.

The S12 proteins are predicted to be nearly 10% as α-helical, 32% as β-strand and 58% in loops. Eubacteria and chloroplasts share the consensus sequence TxxQLxRxx. In *E. coli* this sequence is ATVNQLVRKP (aa 1 to 10), and this is an RNA binding motif, with the sequence VNQLVR calculated as α-helical. More similarities between all phylogenetic domains of the protein family S12 exist in the region shown in figure 1.

Unlike the S11 proteins, the predicted secondary structure of the eubacterial S8 equivalent proteins is quite different from the archaeal and eukaryotic proteins. The conserved regions are predicted as α-, β- and loop structures for all domains.

The 3D structure of S17 from *B. stearothermophilus* was investigated by NMR spectroscopy [59]. The protein consists of β strands connected by several extended loops. Urlaub et al. [47] found the peptides cross-linked to the rRNA by a lysine in *E. coli* (position 29) and *B.*

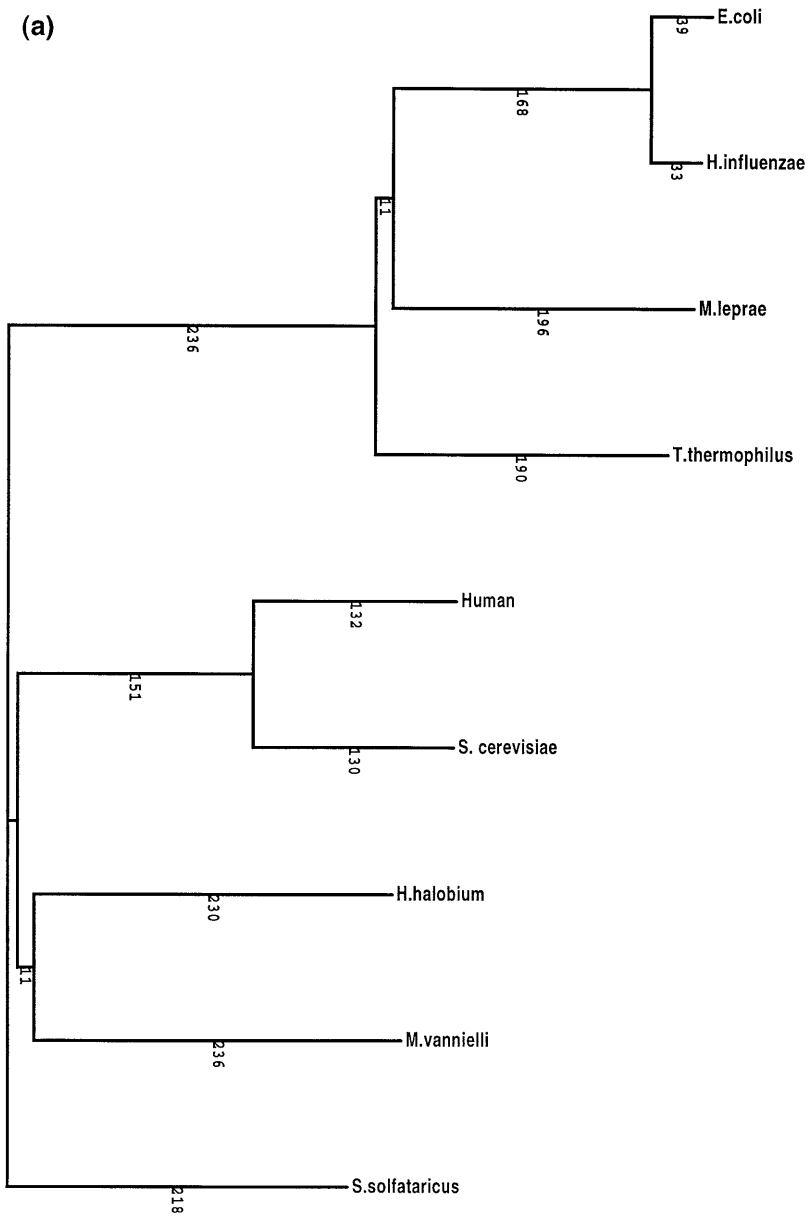

Figure 6. (a). For legend see p. 48.

stearothermophilus (position 31). The consensus sequence of eubacteria and eukarya in this region is K 10x KxPxYxK 6x K, with the underlined lysine as the cross-linking position. The archaebacterial sequences are short and do not match in this region.

The L2 proteins of *E. coli*, *B. stearothermophilus*, rat and *H. marismortui* are similar with respect to the predicted secondary structures, all sharing β-strands and loops. The *P. tetraurelia* protein is less conserved, a fact that is reflected in the secondary structure with α-helical regions and an insert of 12 aa in the N-terminal half of the protein.

The crystal structure of L6 from *B. stearothermophilus* [60] reveals that the protein contains two domains with nearly identical topology. The similarity of the two structural domains suggests that the protein evolved from a single protein by gene duplication early in the evolution. However, the primary sequences of the two domains are not similar. The cross-link experiments made by Urlaub et al. [47] confirm the supposed binding site of the 23S rRNA at the C-terminal end. The amino acid tyrosine (position 156) bound to the 23S RNA is situated in a highly conserved loop region. The eubacterial consensus sequence is RxPEPYKGKG. Only the lysine (position 157) is conserved in yeast (RL9_YEAST), rat and *H. marismortui*.

A relatively large percentage of helical structure has been predicted for the L5 and L11 proteins; it is not

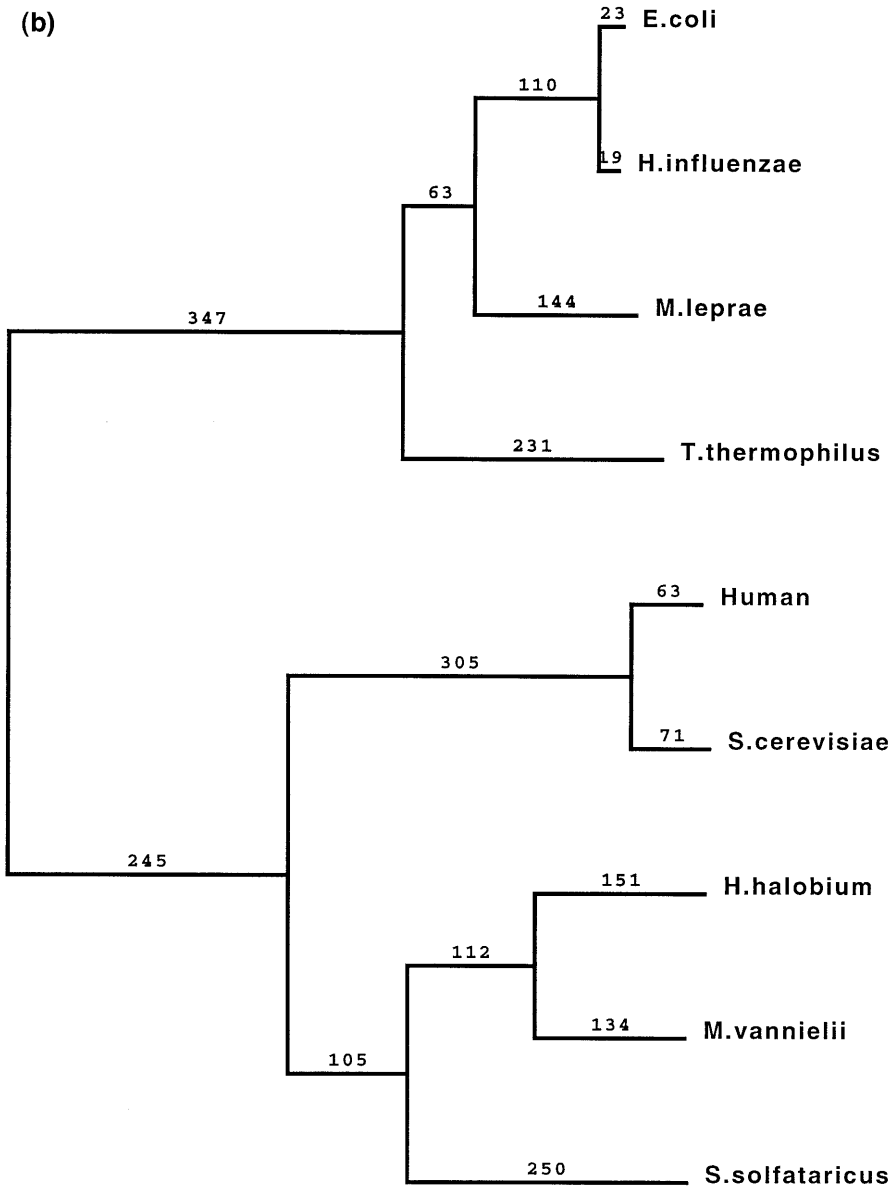

Figure 6. (b). For legend see p. 48.

limited to conserved parts of the sequence. Otherwise, the L14 proteins seem to have only one helix, at the C-terminal end.

Discussion

We have used the huge pool of more than 1900 complete amino acid sequences of ribosomal proteins to reinvestigate the phylogenetic relationship of the organisms of the distant domains of eubacteria, archaea and eukarya by (1) direct sequence comparison, (2) computation of phylogenetics, and (3) searching for conserved sequence motifs and common secondary structural elements.

Three different algorithms (CLUSTREE, PROPTREE and TREE) were employed to investigate the descent of the organisms. The new method for computing phylogenetic trees is based on the properties of the amino acids themselves. There are no assumptions or models used to describe mutation rates. In principle, other properties (e.g. presence or absence of secondary structure motifs) may be introduced into the analysis. PROPTREE produces results in accordance with the taxonomic affiliations of the organisms, often better than TREE. But different lengths of proteins and gaps are unfavourable and may more or less influence the trees. Of course, to some extent this is valid for all phylogenetic calculation methods. In our approach a gap is interpreted as an 'amino acid' without properties. In future we will test the method with a gap penalty and a gap length penalty (work in progress).

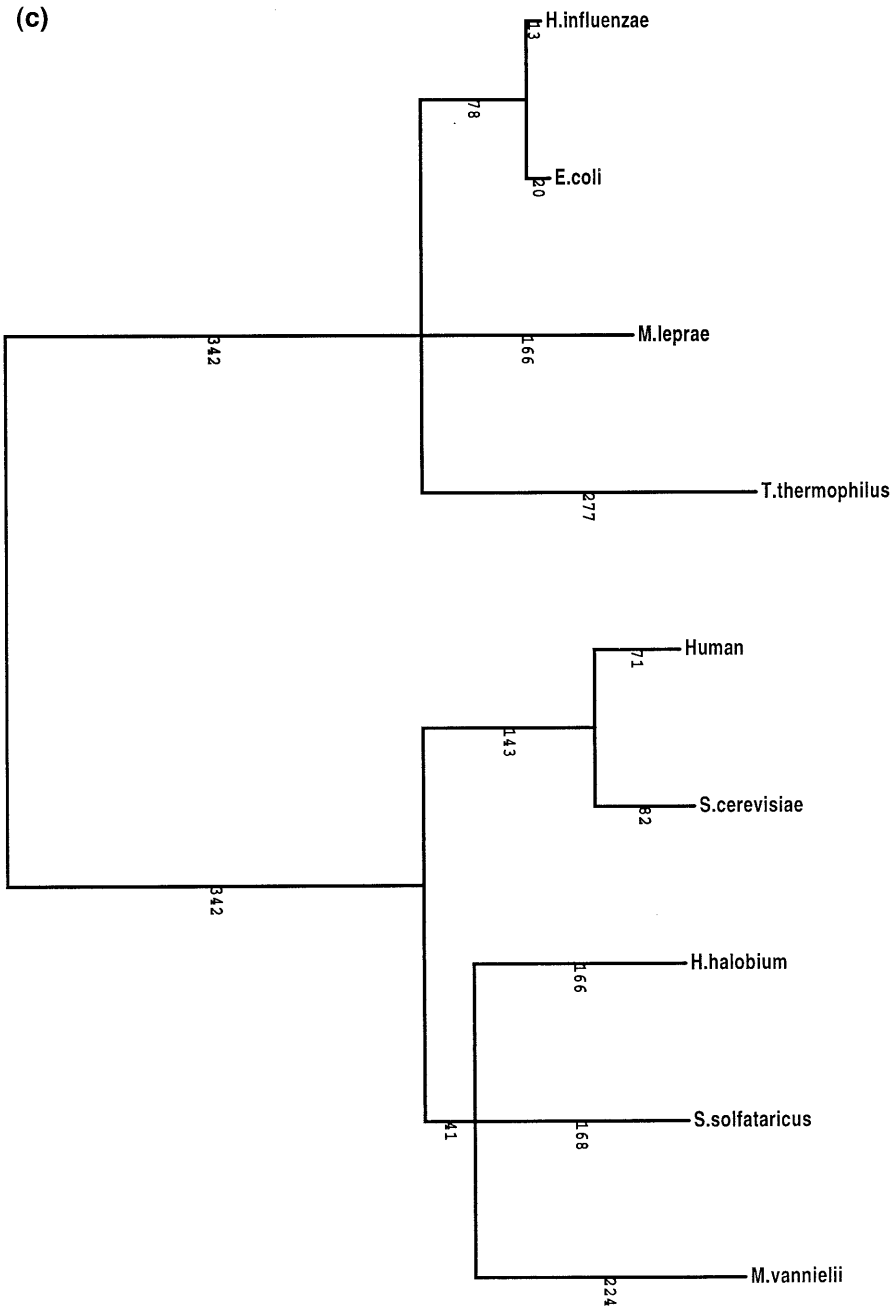

Figure 6. Phylogenetic analysis based on the sequences S7, S10 and S12. The branch lengths are arbitrary values. (a) Neighbor-joining tree (CLUSTREE); (b) tree derived from the amino acid properties (PROPTREE); (c) tree computed with its own alignment (TREE).

Concerning the branches, the methods used yield the same trees. The results of the longer sequences from S7, S10 and S12 do not differ from the shorter ones with respect to the branches.

In our systematic comparison of all known ribosomal proteins of archaea, eubacteria and eukarya, the archaea are a monophyletic group in all trees and more similar to the eukarya. Our data support neither a strong relationship of the archaea only to the gram-positive bacteria (figure 7a) as found by Gupta and Singh [13] for heat shock proteins of Hsp 70 genes, nor the results of rRNA sequences ([5], fig. 7b).

The general tree as derived from the ribosomal proteins is described as follows (fig. 7c): The archaea are more closely related to the eukarya than to the eubacteria. Proteins of chloroplasts and mitochondria often have different lengths, shorter or longer than most of the rest of the protein family. Nevertheless, the chloroplasts are

Table 5. Percentage of α-, β- and loop structures of the best-conserved *E. coli* proteins, method of Rost and Sander [37].

E. coli	α [%]	β [%]	loop [%]
S3	35.8	25.4	38.8
S4	39.5	11.7	48.8
S5 pred.	30.7	33.1	36.2
*	23.4	31.4	45.2
S7	56.7	9.0	34.3
S8	39.2	21.5	39.3
S11	21.1	28.9	50.0
S12	9.8	31.7	58.5
S17 pred.	0.0	57.8	42.2
*	0.0	47.6	52.4
S19	13.2	25.3	61.5
L2	0.0	34.9	65.1
L5	37.6	16.9	45.5
L6 pred.	17.6	38.1	44.3
*	19.9	41.5	38.6
L11	45.8	16.2	38.0
L14	7.3	45.4	47.2

*Results of *B. stearothermophilus* of the 3D structure by crystallography or nuclear magnetic resonance spectroscopy.

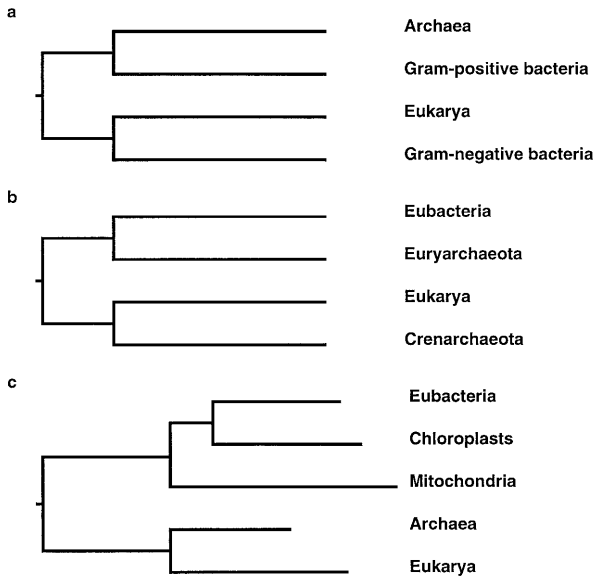

Figure 7. (a) Schematic dendrogram found by Gupta and Singh [13] for heat shock proteins; (b) schematic dendrogram by Lake [5] for RNA sequences; (c) our schematic tree resulting of ribosomal proteins.

most related to the eubacteria. Often both groups are not clearly separated. Most of the mitochondrial proteins form a group together with eubacteria and chloroplasts. However, some of them differ so much that they look like an outgroup of the tree.

Acknowledgement. This work was supported by a grant to Dr. B. Wittmann-Liebold from the Deutsche Forschungsgemeinschaft (SFB 344, YE6).

1 Woese C. R. and Fox G. E. (1977) The concept of cellular evolution. Molec. Evol. **10:** 1–6
2 Woese C. R., Kandler O. and Wheelis M. L. (1990) Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eukarya. Proc. Natl. Acad. Sci USA **87:** 4576–4579
3 Zillig W. (1991) Comparative biochemistry of archaea and bacteria. Curr. Opin. Genet. Dev. **1:** 544–551
4 Lake J. A. (1990) Origin of the eucaryotic nucleus: rRNA sequences genotypically relate eocytes and eucaryotes. In: The Ribosome: Structure, Function and Evolution, pp. 579–588, Hill W. E. (ed.), Amer. Soc. Microbiol., Washington, DC
5 Lake J. A. (1987) Prokaryotes and archaebacteria are not monophyletic: rate invariant analysis of rRNA genes indicates that eukaryotes and eocytes form a monophyletic taxon. Cold Spring Harbor Symposia on Quantitative Biology **52:** 839–846
6 Wettach J., Gohl H. P., Tschochner H. and Thomm M. (1995) Functional interaction of yeast and human TATA-binding proteins with an archaeal RNA polymerase and promoter. Proc. Natl. Acad. Sci. USA **92:** 472–476
7 Reiter W. D., Hudepohl U. and Zillig W. (1990) Mutational analysis of an archaebacterial promoter: essential role of a TATA box for transcription efficiency and start-site selection in vitro. Proc. Natl. Acad. Sci. USA **87:** 9509–9513
8 Creti R., Londei P. and Cammarano P. (1993) Complete nucleotide sequence of an archaeal (*Pyrococcus woesei*) gene encoding a homolog of eukaryotic transcription factor IIB (TFIIB). Nucleic Acids Res. **21:** 2942
9 Rowlands T., Baumann P. and Jackson S. P. (1994) The TATA-binding protein: a general description factor in eukaryotes and archaebacteria. Science **264:** 1251
10 Marsh T. L., Reich C. I., Whitelock R. B. and Olsen G. J. (1994) Transcription factor IID in the Archaea: sequences in the *Thermococcus celer* genome would encode a product closely related to the TATA-binding protein of eukaryotes. Proc. Natl. Acad. Sci. USA **91:** 4180–4184
11 Klenk H. P. and Doolittle W. F. (1994) Archaea and eukaryotes versus bacteria? Curr. Biol. **4:** 920–922
12 Gupta R. S. and Golding G. B. (1993) Evolution of HSP70 gene and its implications regarding relationships between archaebacteria, eubacteria and eukaryotes. J. Mol. Evol. **37:** 573–582
13 Gupta R. S. and Singh B. (1994) Phylogenetic analysis of 70 kD heat shock protein sequences suggests a chimeric origin for the eukaryotic cell nucleus. Curr. Biol. **4:** 1104–1114
14 Wittmann-Liebold B., Köpke A. K. E., Arndt E., Krömer W., Hatakeyama T. and Wittmann H. G. (1990) Sequence comparison and evolution of ribosomal proteins and their genes. In: The Ribosome: Structure, Function and Evolution, pp. 598–616, Hill WE (ed.), Amer. Soc. Microbiol., Washington, D.C.
15 Wittmann-Liebold B. (1986) Ribosomal proteins: their structure and evolution. In: Structure, Function and Genetics of Ribosomes, pp. 326–378, Hardesty B. and Kramer G. (eds), Springer Verlag, Berlin, Heidelberg, New York
16 Herfurth E., Briesemeister U. and Wittmann-Liebold B. (1994) Complete amino acid sequence analysis of ribosomal protein S14 from *Bacillus stearothermophilus* and homology studies to other ribosomal proteins. FEBS Lett. **351:** 114–118
17 Harris E. H., Boynton J. E. and Gillham N. W. (1994) Chloroplast ribosomes and protein synthesis. Microbiol. Rev. **58:** 700–754
18 Takemura M., Oda K., Yamato K., Ohta E., Nakamura Y., Nozato N. et al. (1992) Gene clusters for ribosomal proteins in the mitochondrial genome of a liverwort, *Marchantia polymorpha*. Nucleic Acids Res. **20:** 3199–3205
19 Burger G., Plante I., Lonergan K. M. and Gray M. W. (1995) The mitochondrial DNA of the amoebid protozoon, *Acanthamoeba castellanii*: complete sequence, gene content and genome organization. J. Mol. Biol. **245:** 522–537
20 Vodkin M. H., Gordon V. R. and McLaughlin G. L. (1993) A ribosomal protein in *Acanthamoeba polyphaga* is conserved in eukaryotic nuclei, organelles and bacteria. Gene **131:** 141–144
21 Pritchard A. E., Seilhamer J. J., Mahalingam R., Sable C. L., Venuti S. E. and Cummings D. J. (1990) Nucleotide sequence of the mitochondrial genome of *Paramecium*. Nucleic Acids Res. **18:** 173–180

22 Sogin M. L. (1991) Early evolution and the origin of eukaryotes. Curr. Opin. Genet. Dev. **1:** 457–463
23 Wool I.G., Chan Y.-L. and Glück A. (1995) Structure and evolution of mammalian ribosomal proteins. Biochem. Cell Biol. **73:** 933–947
24 Pearson W. R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol. **183:** 63–98
25 Feng D. F. and Doolittle R. F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol. **25:** 351–360
26 Schmidt W. (1995) Phylogeny reconstruction for protein sequences based on amino acid properties. J. Mol. Evol. **41:** 522–530
27 Schmidt W. and Müller E.-C. (1996) A distance measure based on binary properties of the individuals and its application to molecular phylogeny reconstruction. Bull. Math. Biol. **58:** 449–469
28 Taylor W. R. (1986) Identification of protein-sequence homology by consensus template. J. Mol. Biol. **188:** 233–258
29 Thompson J. D., Higgins D. G. and Gibson T. J. (1994) CLUSTALW: 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
30 HUSAR (1995) Heidelberg Unix Sequence Analysis Resources, Rel. 3.0
31 Fitch W. M. and Margoliash E. (1967) Construction of phylogenetic trees. Science **15:** 279–284
32 Saitou N. and Nei M. (1987) The neighbor-joining method. A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425
33 Saitou N. and Imanishi T. (1989) Relative efficiencies of the Fitch-Margoliash, maximum parsimony, maximum-likelihood, minimum-evolution and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. Mol. Biol. Evol. **6:** 514–525
34 Rost B., Sander C. and Schneider R. (1994) PHD – an automatic mail server for protein secondary structure prediction. CABIOS **10:** 53–60
35 Rost B. and Sander C. (1993) Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Evol. **232:** 584–599
36 Schneider R. and Sander C. (1991) Database of homology-derived structures and the structural meaning of sequence alignment. Proteins **9:** 56–68
37 Rost B. and Sander C. (1994) Combining evolutionary information and neural networks to predict protein secondary structure. Proteins **19:** 55–77
38 Gorini L. (1974) Streptomycin and misreading of the genetic code. In: Ribosomes, pp. 791–803, Nomura M, Tissières A and Lengyel P (eds), Cold Spring Harbor
39 Kitaoka Y., Olvera J. and Wool I. G. (1994) The primary structure of rat ribosomal protein S23. Biochem. and Biophys. Res. Comm. **202:** 314–320
40 Funatsu G., and Wittmann H. G. (1972) Ribosomal proteins. XXXVIII. Location of amino acid replacements in protein S12 isolated from *Escherichia coli* mutants resistant to streptomycin. J. Mol. Biol. **68:** 547–550
41 Bischof O., Urlaub H., Kruft V. and Wittmann-Liebold B. (1995) Peptide environment of the peptidyl transferase center from *Escherichia coli* 70 S ribosomes as determined by thermoaffinity labeling with dihydrospiramycin. J. Biol. Chem. **270:** 23060–23064
42 Oakes M. I., Scheinman A., Atha T., Shankweiler G. and Lake J. A. (1990) Ribosome structure: three-dimensional locations of rRNA and proteins. In: The Ribosome: Structure, Function and Evolution, pp. 180–193, Hill W. E. (ed.), Amer. Soc. Microbiol., Washington, DC
43 Noller H. F., Moazed D., Stern S., Powers T., Allen P. N., Robertson J. M. et al. (1990) Structure of rRNA and its functional interactions in translation. In: The Ribosome: Structure, Function and Evolution, pp. 73–92, Hill W. E. (ed.), Amer. Soc. Microbiol., Washington, DC
44 Tate W. P., Brown C. M. and Kastner B. (1990) Codon recognition by the polypeptide release factor. In: The Ribosome: Structure, Function and Evolution, pp. 393–401, Hill W. E. (ed.), Amer. Soc. Microbiol., Washington, DC
45 Gibson T. J., Thompson J. D. and Heringa J. (1993) The KH domain occurs in a diverse set of RNA-binding proteins that include the antiterminator NusA and is probably involved in binding to nucleic acids. FEBS Lett. **73:** 12–17
46 Siomi H., Matunis M. J., Michael W. M. and Dreyfuss G. (1993) The pre-mRNA binding K protein contains a novel evolutionarily conserved motif. Nucleic Acids Res. **21:** 1193–1198
47 Urlaub H., Kruft V., Bischof O., Müller E.-C. and Wittmann-Liebold B. (1995) Protein-rRNA binding features and their structural and functional implications in ribosomes as determined by cross-linking studies. EMBO Journal **14:** 4578–4588
48 Pohl T., and Wittmann-Liebold B. (1988) Identification of a crosslink in the *Escherichia coli* ribosomal protein pair S13–S19 at the amino acid level. J. Biol. Chem. **263:** 4293–4301
49 Nierhaus K. H. (1991) The assembly of prokaryotic ribosomes. Biochimie **73:** 739–755
50 Cooperman B. S., Weitzman C. J. and Concito C. (1990) Antibiotic probes of *Escherichia coli* peptidyltranferasecenter. In: The Ribosome: Structure, Function and Evolution, pp. 123–133, Hill W. E. (ed.), Amer. Soc. Microbiol., Washington, DC
51 Matheson A. T., Auer J., Ramirez C. and Böck A. (1990) Structure and evolution of archaebacterial ribosomal proteins. In: The Ribosome: Structure, Function and Evolution, pp. 617–635, Hill W. E. (ed.), Amer. Soc. Microbiol., Washington, DC
52 Ianniciello G., Gallo M. and Arcari P. Bocchini V. (1994) Organization of a *Sulfolobus solfataricus* gene cluster homologous to the *Escherichia coli str* operon. Biochem. Mol. Biol. Internat. **33:** 927–937
53 Ramirez C., Shimmin L. C., Newton C. H., Matheson A. T. and Dennis P. P. (1989) Structure and evolution of the L11, L1, L10, and L12 equivalent ribosomal proteins in eubacteria, archaebacteria, and eukaryotes. Can. J. Microbiol. **35:** 234–244
54 Ramirez C., Shimmin L. C., Leggatt P. and Matheson A. T. (1994) Structure and transcription of the L11-L1-L10-L12 ribosomal protein gene operon from the extreme thermophilic archaeon Sulfolobus acidocaldarius. J. Mol. Biol. **244:** 242–249
55 Christopher D. A., Cushman J. C., Price C. A. and Hallick R. B. (1988) Organization of ribosomal protein genes *rpl23*, *rpl2*, *rps19*, *rpl22* and *rps3* on the *Euglena gracilis* chloroplast genome. Curr. Genet. **14:** 275–286
56 Liu X.-Q., Huang C. and Xu H. (1993) The unusual *rps3*-like orf712 is functionally essential and structurally conserved in *Chlamydomonas*. FEBS Lett. **336:** 225–230
57 Swofford D. L. (1991) PAUP: Phylogenetic Analysis using Parsimony. Version 3.0s
58 Ramakrishnan V. and White S. W. (1993) The structure of ribosomal protein S5 reveals sites of interaction with 16S rRNA. Nature **358:** 768–771
59 Golden B., Hoffman D. W., Ramakrishnan V. and White S. W. (1993a) Ribosomal protein S17: characterization of the three-dimensional structure by ^{1}H and ^{15}N NMR. Biochemistry **32:** 12812–12820
60 Golden B. L., Ramakrishnan V. and White S. W. (1993b) Ribosomal protein L6: structural evidence of gene duplication from a primitive RNA binding protein. EMBO J. **12:** 4901–4908
61 Felsenstein J. (1993) PHYLIP (Phylogeny Interference Package) version 3.5p