J Mol Evol (2000) 51:173–181
DOI: 10.1007/s002390010078

JOURNAL OF MOLECULAR EVOLUTION

© Springer-Verlag New York Inc. 2000

Determining the Relative Rates of Change for Prokaryotic and Eukaryotic Proteins with Anciently Duplicated Paralogs

Justin M. Kollman, Russell F. Doolittle

Center for Molecular Genetics, University of California San Diego, 9500 Gilman Dr., La Jolla, CA 92093-0634, USA

Received: 30 December 1999 / Accepted: 5 April 2000

Abstract. The relative rates of change for eight sets of ubiquitous proteins were determined by a test in which anciently duplicated paralogs are used to root the universal tree and distances are calculated between each taxonomic group and the last common ancestor. The sets included ATPase subunits, elongation factors, signal recognition particle and its receptor, three sets of tRNA synthetases, transcarbamoylases, and an internal duplication in carbamoyl phosphate synthase. In each case phylogenetic trees were constructed and the distances determined for all pairs. Taken over the period of time since their last common ancestor, average evolutionary rates are remarkably similar for Bacteria and Eukarya, but Archaea exhibit a significantly slower average rate.

Key words: Universal tree — Evolutionary rate — Paralogous rooting — Protein sequence evolution

Introduction

It is well known that different proteins change at different but characteristic rates (for example, Dickerson 1971; McLaughlin and Dayhoff 1972). Differences in rates along different evolutionary lineages can be assessed by the so-called relative-rate test (Sarich and Wilson 1973). Thus, if sequences are available for three or more taxa (A, B, C, etc.), and C is more different from the other sequences than they are from each other (i.e., C is the outlier), the distance AC can be compared with the distance BC to determine the relative evolutionary rates of A and B (Li 1997). The shortcoming of the test is that it is not possible to make a judgment about the rate of change leading to the outlier because the position of the root of such a tree is unknown.

This limitation has proven especially vexing in attempts to determine divergence times between eukaryotes and prokaryotes, and in some past studies similar rates of change had to be assumed for the three urkingdoms: Bacteria, Archaea, and Eukarya (Doolittle et al. 1996; Feng et al. 1997). Attempts to determine rates of change in various bacterial taxa have been reported in the past, most often based on host–parasite relationships and/or the timing of major ecological events. For example, Ochman and Wilson (1987) used such methods to examine evolutionary rates among bacterial lineages; indeed, they found evidence for uniform substitution rates in Eukarya and Bacteria. More recently, Ochman et al. (1999) used the host–parasite method to establish recent rates of 16S rRNA evolution in bacterial endosymbionts of insects, in which case the bacterial divergence times were correlated with the divergence times of their hosts inferred from the fossil record.

Following the leads of others who have used paralogous pairs of proteins to root the "universal tree" (Gogarten et al. 1989; Iwabe et al. 1989), we have extended the paralogous approach to determining rates of change. This method employs paralogous proteins which were duplicated prior to the last universal common ancestor and which have representatives in each of the three extant urkingdoms. The proteins used were: ATPase subunit α (60kDa) and β (70kDa), signal recognition particle (SRP) and its receptor (SRα), elongation factors (EF)

Correspondence to: Russell F. Doolittle

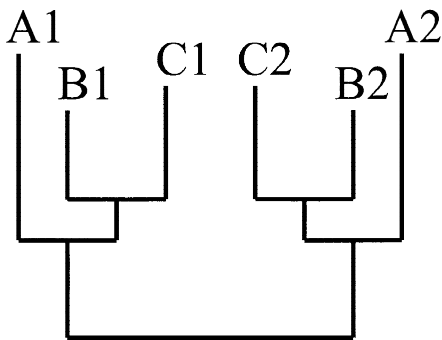


Fig. 1. Model for paralogous rooting method. A, B, and C represent different taxa, and 1 and 2 paralogous sequences. The paralogs are the result of gene duplication prior to the last common ancestor of A, B, and C. Sequence 2 is used as an outgroup to root the tree for sequence 1, and vice versa.

1α/Tu and 2/G, tyrosine and tryptophan tRNA synthetases (TyrRS and TrpRS), valine and isoleucine tRNA synthetases (ValRS and IleRS), threonine and serine tRNA synthetases (ThrRS and SerRS), N- and C-terminal repeats of carbamoyl phosphate synthase (nCPS and cCPS), and aspartate and ornithine transcarbamoylases (ATC and OTC).

In phylogenetic trees constructed with these protein sets, each paralog serves as the outgroup with which to root the other (Fig. 1). These outgroups enabled us to calculate the distance to the LCA of each pair of urkingdoms, the values of which can be used in a relative rate test to compare the pace of evolution for the three urkingdoms since the time of their divergence. We have found that in the time since the LCA, Eukarya and Bacteria have changed with nearly identical average rates. Archaea, however, exhibit a reduced average rate of sequence change.

Materials and Methods

Multiple Alignments and Phylogenetic Trees. Protein sequences were found by BLAST searching (Altschul et al. 1997) with a seed sequence for each data set (Table 1). The results were examined and sequences chosen for a broad representation of Archaea, Bacteria, and Eukarya. Multiple alignments were generated by the progressive method of Feng and Doolittle (1987) using all sequences from both paralogs (see Table 1 for the number of sequences used in each alignment). Phylogenetic trees based on these alignments were constructed by a least-squares distance method with the program TREE (Feng and Doolittle 1996), as well as by a character-based parsimony method with the program PAPA (Doolittle and Feng 1990). All multiple alignments and phylogenetic trees referred to in this paper are available by anonymous FTP at juno.ucsd.edu in the directory labeled rates.

Distance Measurement. Percent identities (the fraction of identical residues in alignment between two proteins) were calculated for every pair of proteins in each data set. Average percent identity was then determined between urkingdoms and between each urkingdom and the outgroup (Table 2). Although percent identity is only a crude inverse measure of difference, it does offer a common perspective on relatedness and is intuitively useful as a first approximation of relative rates of change.

Two different measures of evolutionary distance were employed. The first of these depends on a simple Poisson relationship combined with a substitution matrix which corrects for the nature of amino acid exchanges (Feng et al. 1985). This method, which we will refer to as a modified Poisson method, assigns a similarity score for each pair of proteins based on an amino acid substitution table. The similarity score is then converted to a measure of evolutionary distance:

$$d = -\ln(S_{obs} - S_{rand}/S_{id} - S_{rand}) \times 100 \qquad \text{(Eq. 1)}$$

in which d is the evolutionary distance and ($S_{obs} - S_{rand}/S_{id} - S_{rand}$) is the similarity score corrected for amino acid composition (Feng et al. 1985). The Blosum62 substitution table (Henikoff and Henikoff 1993) was used as the scoring matrix.

Because the method described above does not adequately account for varying evolutionary rates at different amino acid sites in a protein, a second measurement of distance was employed which corrects for such variation. The method of Grishin (1995) corrects for both among-site variation and the nature of the amino acid interchange:

$$q = \ln(1 + 2d)/2d \qquad \text{(Eq. 2)}$$

where d is the measure of evolutionary distance and q is the fraction of unchanged residues (Grishin 1995). We will refer to this method as the Grishin method. A comparison of these two methods by computer simulation (Feng and Doolittle 1997) suggests that for very diverged sequences the modified Poisson method tends to underestimate and the Grishin distance to exaggerate the true distance.

The Relative Rate Test. The relative rates of change of protein sequences can be estimated by calculating distances between the taxa of interest and the last common ancestor (Sarich and Wilson 1973). Comparison of the distance from two taxa to their last common ancestor can reveal differences in evolutionary rates along the lineages leading to these taxa. The TREE program automatically generates the distances between all sequence pairs; all that was needed was to average them appropriately.

The relative rate test is designed to find the distance from the taxa of interest to the last common ancestor according to the relationship:

$$AL = (AO + AB - BO)/2 \qquad \text{(Eq. 3)}$$

where AO and BO are the average distances from two taxa to the outgroup, AB is the average distance between these two taxa, and AL is the distance from taxa A to the LCA.

Average modified Poisson and Grishin method distances were compiled between each pair of urkingdoms and between each urkingdom and the paralogous outgroup. This information was then used to determine distances to the LCA of each pair of urkingdoms, taken separately. By considering distances for only two urkingdoms at a time, we avoid potential complications related to branching order of the urkingdoms. This was useful if only because it has been suggested that rooting of universal trees by paralogous outgroups may suffer from "long-branch attraction" (Phillipe and Forterre 1999).

The significance of differences in rate was assessed with Student's t test (Triola 1997). For each pair of urkingdoms, the ratio of distance to the LCA was calculated, as well as the mean (x) and sample standard deviation (s) for all proteins compared. The t values were then calculated to determine whether the mean was significantly different from unity:

$$t = |x - 1.0|/(s/\sqrt{n}) \qquad \text{(Eq. 4)}$$

Table 1. Seed sequences used for database searches

Paralog	Accession #	Organism	Residues used in alignment[a]	# Sequences used in alignment		
				Archaea	Bacteria	Eukarya
ATPase α/60kDa	P55987	*Helicobacter pylori*	147–347	5	5	5
ATPase β/70kDa	P00824	*Escherichia coli*	125–325	4	5	5
SRα	P51835	*Bacillus subtilis*	39–329	5	5	4
SRP	P37105	*B. subtilis*	17–300	5	5	5
EF-1α/Tu	P02990	*E. coli*	entire sequence	5	6	5
EF-2/G	P02996	*E. coli*	1–400	5	8	5
TyrRS	P25151	*B. subtilis*	54–342	5	5	5
TrpRS	P21656	*B. subtilis*	7–286	5	5	5
IleRS	P00956	*E. coli*	94–606	6	9	6
ValRS	P07118	*E. coli*	78–558	6	7	7
ThrRS	P56071	*H. pylori*	214–498	4	6	4
SerRS	P56458	*H. pylori*	141–415	3	6	6
nCPS	AE001514	*H. pylori*	11–333	3	8	9
cCPS	AE001514	*H. pylori*	563–857	3	8	9
ATC II	P00479	*E. coli*	entire sequence	5	7	6
OTC β	P04391	*E. coli*	entire sequence	5	7	7

[a] Residue numbers for representative protein only. Other sequences in the alignments were manually trimmed to align with the given segment.

Table 2. Average percent identity within and across paralogous pairs

	Average percent identity with outgroup[a]			Average percent identity within lineages[b]		
	Archaea	Bacteria	Eukarya	AB	BK	AK
ATPase α/60kDa	28.2 ± 2.1	26.1 ± 1.6	27.3 ± 1.7	32.4 ± 1.6	33.8 ± 1.9	67.1 ± 1.9
ATPase β/70kDa	27.3 ± 2.1	25.9 ± 1.7	28.4 ± 1.5	29.0 ± 1.6	28.2 ± 1.3	69.0 ± 2.5
SRα	33.3 ± 2.5	32.3 ± 3.1	26.6 ± 2.3	40.4 ± 4.2	29.5 ± 1.6	37.2 ± 2.3
SRP	31.1 ± 4.3	32.1 ± 3.7	29.9 ± 3.4	38.6 ± 2.9	33.2 ± 2.2	44.5 ± 2.5
EF-1α/Tu	25.5 ± 2.2	24.9 ± 2.1	25.8 ± 1.6	40.8 ± 3.9	37.0 ± 1.9	53.3 ± 3.3
EF-2/G	25.5 ± 1.7	25.0 ± 2.2	25.9 ± 1.2	33.4 ± 2.2	29.9 ± 2.1	42.5 ± 1.8
TyrRS	22.1 ± 2.1	18.8 ± 2.2	20.9 ± 1.6	25.5 ± 2.3	21.7 ± 2.0	35.3 ± 3.8
TrpRS	21.3 ± 2.3	20.7 ± 1.9	19.7 ± 2.6	23.8 ± 2.4	24.4 ± 2.1	33.6 ± 7.7
IleRS	28.0 ± 4.0	26.7 ± 3.6	25.2 ± 2.1	39.8 ± 3.2	32.2 ± 1.3	37.3 ± 1.7
ValRS	30.3 ± 3.5	26.2 ± 2.1	—[c]	31.6 ± 2.8	—[c]	—[c]
ThrRS	16.1 ± 1.4	16.4 ± 1.7	—[c]	24. ± 1.6	—[c]	—[c]
SerRS	16.9 ± 1.6	16.6 ± 1.5	16.4 ± 1.6	41.3 ± 3.5	38.3 ± 3.3	45.0 ± 2.4
nCPS	—[d]	33.6 ± 3.6	32.3 ± 3.2	—[d]	48.3 ± 3.4	—[d]
cCPS	—[d]	35.5 ± 2.7	30.2 ± 1.7	—[d]	42.2 ± 1.5	—[d]
ATC	28.6 ± 2.7	23.9 ± 2.9	26.4 ± 2.4	34.2 ± 2.2	33.0 ± 1.9	45.5 ± 2.4
OTC	—[e]	25.7 ± 2.5	25.5 ± 2.9	—[e]	37.2 ± 3.5	—[e]

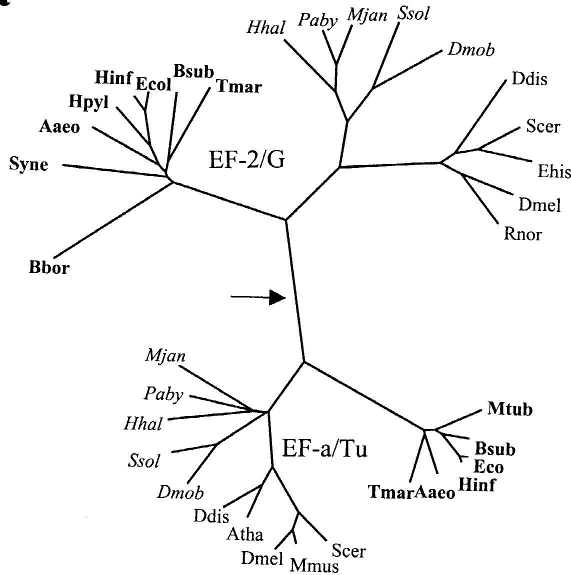
[a] Calculated as the mean of pairwise identities between the specified lineage and the last common ancestor.
[b] Calculated as the mean of the pairwise identities between the designated urkingdoms within the same lineage.
[c] Measurements omitted from calculations due to eucaryotic acquisition of the mitochondrial sequence.
[d] Measurements omitted from calculations due to unresolved archaeal phylogeny.
[e] Measurements omitted from calculations due to unresolved archaeal and bacterial phylogeny.

where n is the number of protein sets compared. The level of confidence (based on one-tailed alpha value) was then determined for the t value and the appropriate degrees of freedom ($n - 1$).

Results

Phylogenetic Trees

Phylogenetic trees constructed with the least-squares modified Poisson distance method are presented in Figs. 2–5. Except where noted, the relationships among the three urkingdoms are in agreement with those obtained by character-based parsimony, as well as with results from previously published analyses (ATPase subunits—Gogarten et al. 1989; Iwabe et al. 1989; elongation factors—Iwabe et al. 1989; ValRS/IleRS—Brown and Doolittle 1995; CPS—Lawson et al. 1996; TyrRS/TrpRS—Brown et al. 1997; SRP/SRα—Gribaldo and Cammarano 1998; ATC/OTC—Labedan et al. 1999). The majority of the trees have a similar topology, with Archaea and Eukarya as sister groups. For convenience,



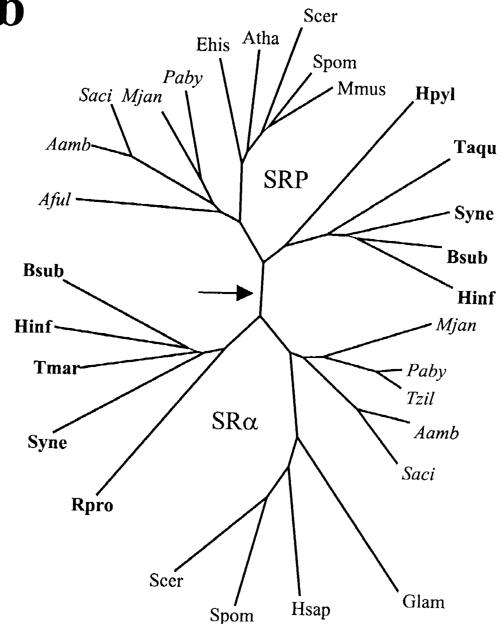


Fig. 2. Phylogenetic trees constructed from two paralogous sets of sequences: signal recognition particle (SRP) and its receptor (SRα) **(a)** and elongation factors (EF-2/G and EF1α/Tu) **(b)**. Trees were constructed using the least-squares distance method. The arrow indicates the branch separating paralogs. See Table 3 for abbreviations.

this will be referred to as the standard topology. There were occasional anomalies that violated the standard topology, mostly the result of apparent horizontal gene transfer or organellar import. As detailed in the following paragraphs, most of these were dispensed with by omitting the groups violating the standard topology from the calculations.

Elongation Factors

The phylogenetic tree of the elongation factors conformed to the "standard topology" with no exceptions (Fig. 2a).

SRP/SRα

The phylogenetic tree of the SRP/SRα gave the standard topology for both paralogs (Fig. 2b). A similar SRP/SRα tree has been reported by others (Gribaldo and Cammarano 1998). However, in the parsimony tree constructed for SRP/SRα (not shown) the SRα part of the tree was unresolved for the deepest branching order.

ValRS/IleRS

This tree had several anomalies (Fig. 3a). The IleRS sequences conformed to the "standard topology" with the exception of a small group of bacterial pathogens that clusters within the Eukarya. It has been postulated that these pathogens acquired the IleRS gene from a eukaryotic host (Sassanfar et al. 1996), and those sequences were therefore omitted from rate calculations. The ValRS half of the tree has all of the eukaryotic sequences very close to the bacterial sequences. Evidence has been presented indicating that eukaryotic ValRS is an import of the mitochondrial gene into the eukaryotic host (Hashimoto et al. 1998). Accordingly, the eukaryotic sequences of ValRS were not included in rate calculations. Furthermore, the *Rickettsia* ValRS sequence clustered within the Archaea and was also removed from the calculations.

SerRS/ThrRS

The SerRS sequences have the standard topology, whereas the ThrRS topology has Eukarya and Bacteria clustered together (Fig. 3b). Like ValRS, eukaryotic ThrRS seems to be a case of organellar import (Doolittle and Handy 1998) and was also omitted from rate calculations.

ATC/OTC

The tree constructed from the transcarbamoylase sequences also has a few anomalies (Fig. 4a). Thus, the ATC sequences exhibit the "standard topology" with the exception of a group of γ-proteobacteria that cluster with the Archaea. The situation suggests horizontal transfer, and the γ-proteobacteria sequences were not used for rate calculations. The OTC sequences fall into three distinct clusters: one entirely eukaryotic, one entirely bacterial, and a third that is a mixture of archaeal and bacterial sequences. The latter, which may be the result of one or more horizontal transfer events, were omitted from distance calculations, only the well-grouped eukaryotic and bacterial lineages being used.

nCPS/cCPS

The phylogenetic tree made from the N-terminal and C-terminal regions of carbamoyl phosphate synthase is

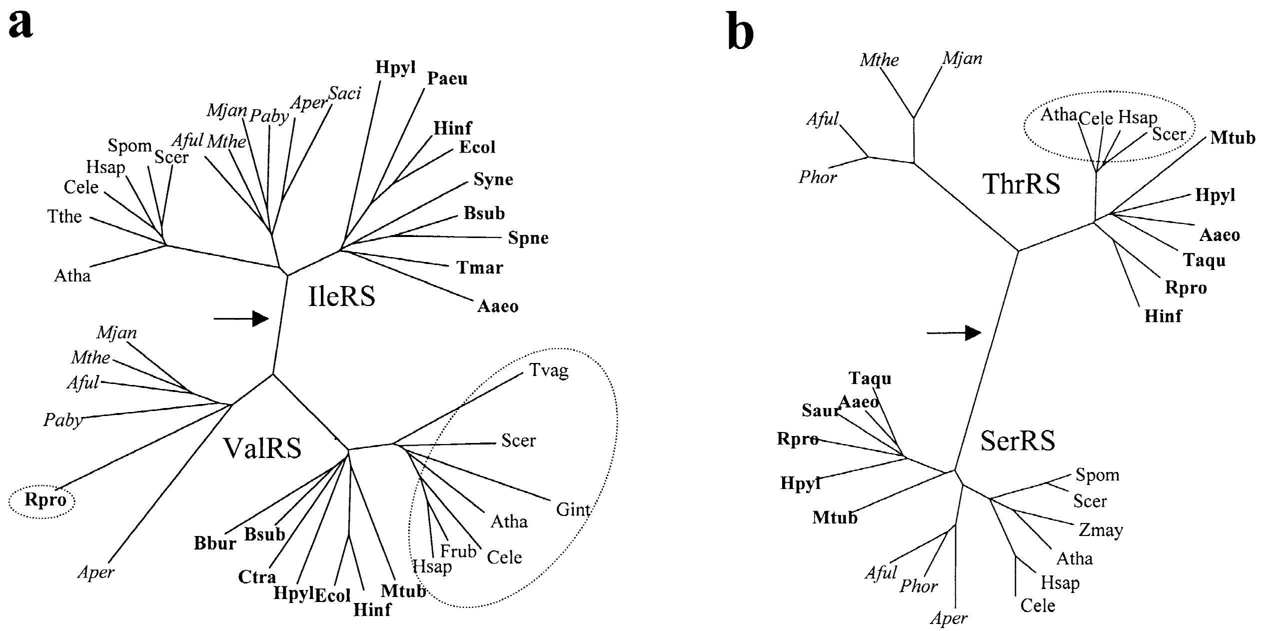


Fig. 3. Phylogenetic trees constructed from two paralogous sets of sequences: threonine and serine aminoacyl tRNA synthetases (ThrRS and SerRS) **(a)** and valine and isoleucine aminoacyl tRNA synthetases (ValRS and IleRS) **(b)**. Trees were constructed using the least-squares distance method. Dotted lines indicate sequences omitted from rates calculations. The arrow indicates the branch separating paralogs. See Table 3 for abbreviations.

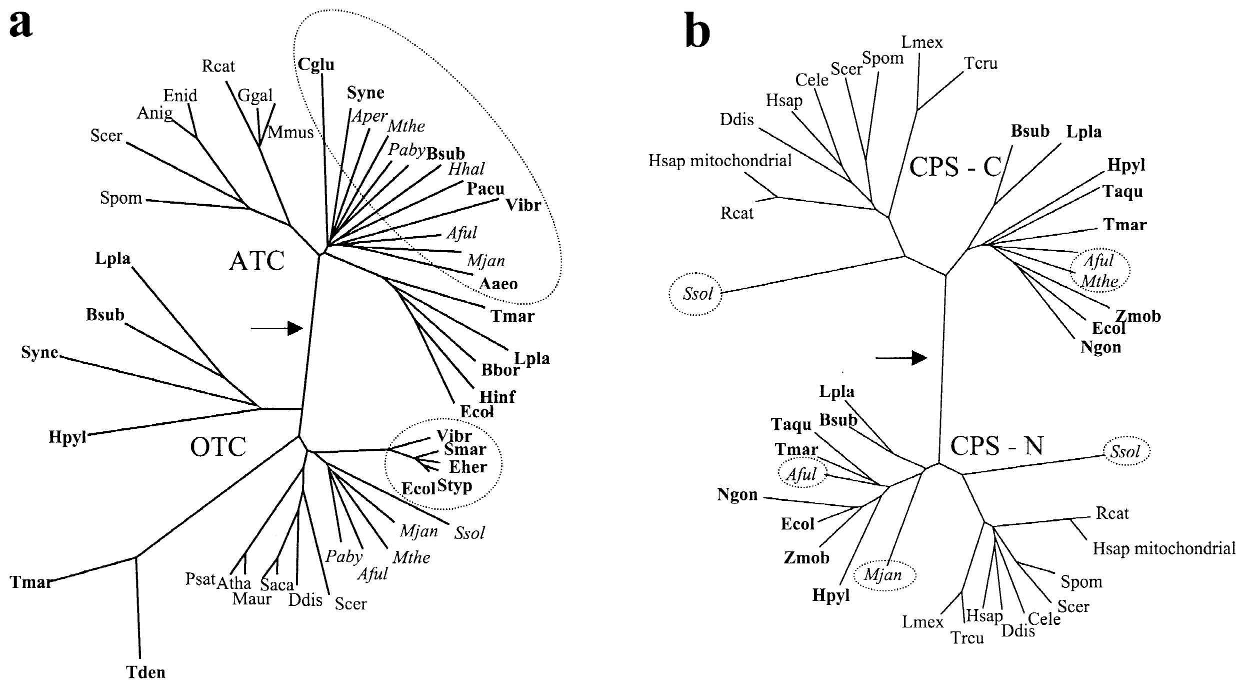


Fig. 4. Phylogenetic trees constructed from two paralogous sets of sequences: aspartate and ornithine transcarbamoylases (ATC and OTC) **(a)** and N- and C-terminal regions of carbamoyl phosphate synthase (nCPS and cCPS) **(b)**. Trees were constructed using the least-squares distance method. Dotted lines indicate sequences omitted from rate calculations. The arrow indicates the branch separating paralogs. See Table 3 for abbreviations.

ambiguous with respect to the position of Archaea (Fig. 4b). In a previous study of CPS phylogeny (Lawson et al. 1996), the *Sulfolobus* sequence—the only available archaeal sequence at the time—was grouped with the eukaryotic sequences in both the N- and C-terminal regions. This same topology was obtained in the present study; now, however, newly available archaeal sequences—*Archaeoglobus, Methanococcus,* and *Methonobacterium*—are scattered among the bacterial group. Because of the disparate phylogenetic positions of the archaeal sequences, none were used in the rate calculations.

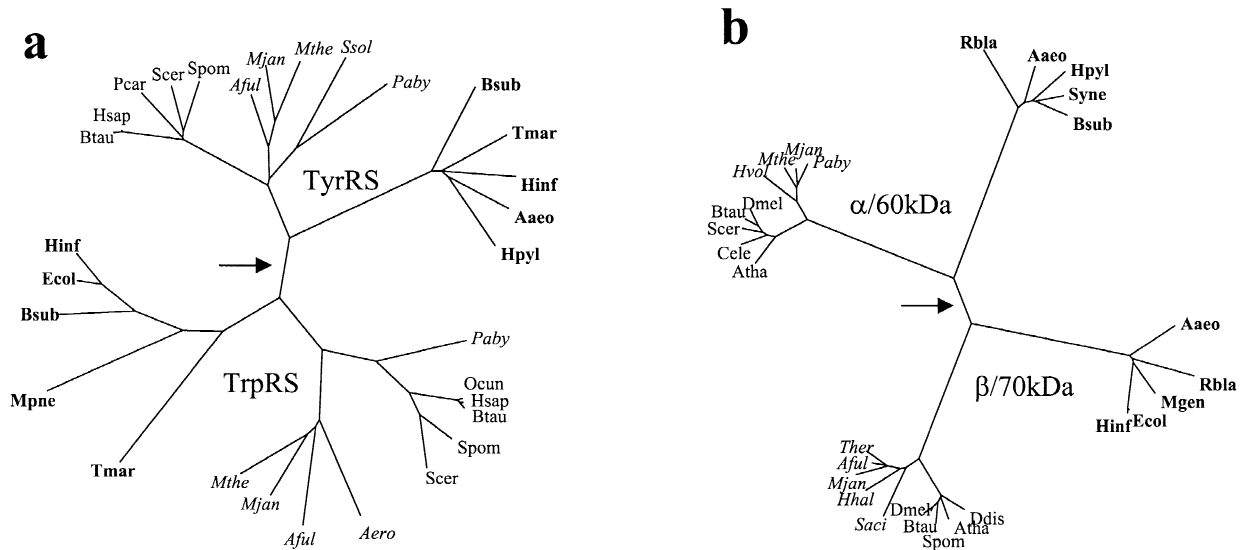


Fig. 5. Phylogenetic trees constructed from two paralogous sets of sequences: tyrosine and tryptophan aminoacyl tRNA synthetases (TyrRS and TrpRS) **(a)** and ATPase subunits (α/60 kDa and β/70 kDa) **(b)**. Trees were constructed using the least-squares distance method. The arrow indicates the branch separating paralogs. See Table 3 for abbreviations.

TyrRS/TrpRS

The TyrRS/TrpRS tree gives the standard topology for both paralogs with the single exception of the *Pyrococcus* TrpRS sequence, which falls at the base of the eukaryotic cluster (Fig. 5a).

ATPase Subunits

The tree of ATPase subunits (Fig. 5b), although conforming to the standard topology, is somewhat unusual in having very long internal branch lengths and very short terminal branches. Taken at face value, the suggestion is for a very rapid rate of sequence change immediately after the LCA, followed by dramatic and coincidental slow-downs in the more recent past. Whatever the case, there should be no effect on the calculation of average rate in the interval since the LCA, although we must remain cautious about the results in this case.

Determination of Relative Rates

The degree of divergence of Archaea, Bacteria, and Eukarya from their paralogous outgroup was estimated for each data set in three different ways: average percent identity, average modified Poisson, and average Grishin distance. The average percent identities between each urkingdom and the paralogous outgroup, as well as percent identities between the orthologs from each urkingdom, are presented in Table 2. For the most part, the data do not suggest much difference in divergence among the three urkingdoms for a given protein. Of the 16 proteins studied, only in the case of SRα does there appear to be a significant difference among the urkingdoms in the average percent identity with the outgroup, the eukaryotic sequences having a lower percent identity than the archaeal or bacterial sequences, suggesting an increase in evolutionary rate for the eukaryotic protein. As noted above, percent identity is only a rough guide to the extent of divergence.

A graphic comparison of evolutionary rates was obtained by plotting the distance between pairs of urkingdoms and their LCA against each other (Fig. 6). If the rates were identical for any pair of urkingdoms, that data point would fall on the diagonal. When the modified Poisson method was used, the data typically fell close to the diagonal (Fig. 6a). When the Grishin method was used, the deviation from perfectly equal rates was greater (Fig. 6b).

Both methods indicate that, on average, Eukarya and Bacteria are changing with the same rate. In contrast, Archaea are changing at about 90% the rate of Eukarya and Bacteria when calculated with the modified Poisson, and only 75–80% the rate of Bacteria and Eukarya when calcuated with the Grishin method. As noted, the “modified Poisson” method tends to underestimate and the Grishin method exaggerate extreme distances.

When the Grishin method is used, there are three cases where a set of sequences in one urkingdom is changing at a much greater rate than another urkingdom. In the most obvious case, eukaryotic SRα sequences are changing at least twice as fast as bacterial and archaeal SRα sequences, as was apparent from comparisons of percent identity alone. In a second case, the eukaryotic C-terminal region of CPS is changing at about three times the rate of bacterial counterparts. In the third case, the bacterial ATC sequences are changing at about three times the rate of the Archaeal sequences. In the remain-

Table 3. Abbreviations used in phylogenetic trees

Abbreviation	Organism	Abbreviation	Organism
Archaea		Saca	*Squalus acanthias*
Aamb	*Acidianus ambivalens*	Scer	*Saccharomyces cerevisiae*
Aful	*Archaeoglobus fulgidus*	Spom	*Schizosaccharomyces pombe*
Aper	*Aeropyrum pernix*	Tcru	*Trypanosoma cruzi*
Dmob	*Desulfurococcus mobilis*	Tthe	*Tetrahymena thermophila*
Hhal	*Halobacterium halobium*	Zmay	*Zea mays*
Hsal	*Halobacterium salinarium*		
Hvol	*Haloferax volcanii*	**Bacteria**	
Mjan	*Methanococcus janaschii*	Aaeo	*Aquifex aeolicus*
Mthe	*Methanobacterium thermoautotrophicum*	Bbur	*Borrelia burgdorferi*
Paby	*Pyrococcus abyssi*	Bsub	*Bacillus subtilis*
Pfur	*Pyrococcus furiosus*	Cglu	*Cornybacterium glutamicum*
Phor	*Pyrococcus horikoshii*	Cper	*Clostridium perfringens*
Saci	*Sulfolobus acidocaldarius*	Ctra	*Chlamydia trachomatis*
Ssol	*Sulfolobus solfactaricus*	Ecol	*Escherichia coli*
Ther	*Thermococcus* sp.	Eher	*Erwinia herbicola*
Tzil	*Thermococcus zilligii*	Hinf	*Haemophilus influenzae*
Eukarya		Hpyl	*Helicobacter pylori*
Anig	*Aspergillus niger*	Lpla	*Lactobacillus plantarum*
Atha	*Arabidopsis thaliana*	Mgen	*Mycoplasma genitalium*
Btau	*Bos taurus*	Mpne	*Mycobacterium pneumoniae*
Cele	*Caenorhabditis elegans*	Mtub	*Mycobacterium tuberculosis*
Ddis	*Dictyostelium discoidium*	Ngon	*Neisseria gonorrhoeae*
Dmel	*Drosophila melanogaster*	Paeu	*Pseudomonas aeruginosa*
Ehis	*Entamoebae histolytica*	Pflu	*Pseudomonas fluorescens*
Enid	*Emericella nidulans*	Rbla	*Rhodobacter blasticus*
Frub	*Fugu rubripes*	Rpro	*Rickettsia prowazekii*
Ggal	*Gallus gallus*	Saur	*Staphylococcus aureus*
Gint	*Giardia intestinalis*	Smar	*Serratia marcescens*
Hsap	*Homo sapiens*	Spne	*Streptococcus pneumoniae*
Lmex	*Leishmania mexicana*	Styp	*Salmonella typhimurium*
Maur	*Mesocricetus auratus*	Syne	*Synechocystis PCC6803*
Mmus	*Mus musculus*	Taqu	*Thermus aquaticus*
Ocun	*Oryctolagus cunniculus*	Tden	*Treponema denticola*
Pcar	*Pneumocystis carinii*	Tmar	*Thermatoga maritima*
Psat	*Pisum sativum*	Tvag	*Trichomonas vaginalis*
Rcat	*Rana catesbeiana*	Vibr	*Vibrio* sp. 1693
Rnor	*Rattus norvegicus*	Zmob	*Zymomonas mobilis*

ing protein sets, there is a clear but lesser tendency for archaeal sequences to be changing more slowly.

Student's *t* test provides confirmation that Archaea are changing more slowly than Eukarya or Bacteria. The test gives 90% confidence (modified Poisson method) or 99% confidence (Grishin method) that the bacterial sequences have changed more rapidly than archaeal sequences. Similarly, Student's *t* test gives 99% confidence (modified Poisson method) and 95% confidence (Grishin method) that the eukaryotic sequences have changed faster than archaeal sequences. The slight differences between eukaryotic and bacterial rates found with both methods were not significantly different by this test.

Discussion

Knowing the relative rates of change of proteins from the three major groups of living organisms bears heavily on studies aimed at establishing the "real tree of life." Recently, the paralogous method of rooting the universal tree has been questioned. Brinkmann and Philippe (1999) and Philippe and Forterre (1999) conducted a careful analysis that employed many of the same protein sets used in our study and concluded that the tree of life has been misconstrued because bacterial sequences have been changing much faster than archaeal and eukaryotic sequences. The kinship of the Archaea and Eukarya, they feel, is merely the result of "long branch attraction." Our results challenge that conclusion; bacterial sequences are not changing faster than eukaryotic ones.

Beyond that, in recent years there has been an enormous amount of controversial comment on the matter of molecular clocks. Much of the dispute has centered around differences in expectation and exactitude. No one argues that protein sequence clocks always run at uniform rates, and hardly anyone thinks that a single protein will provide a reliable clock. But there is some basis for

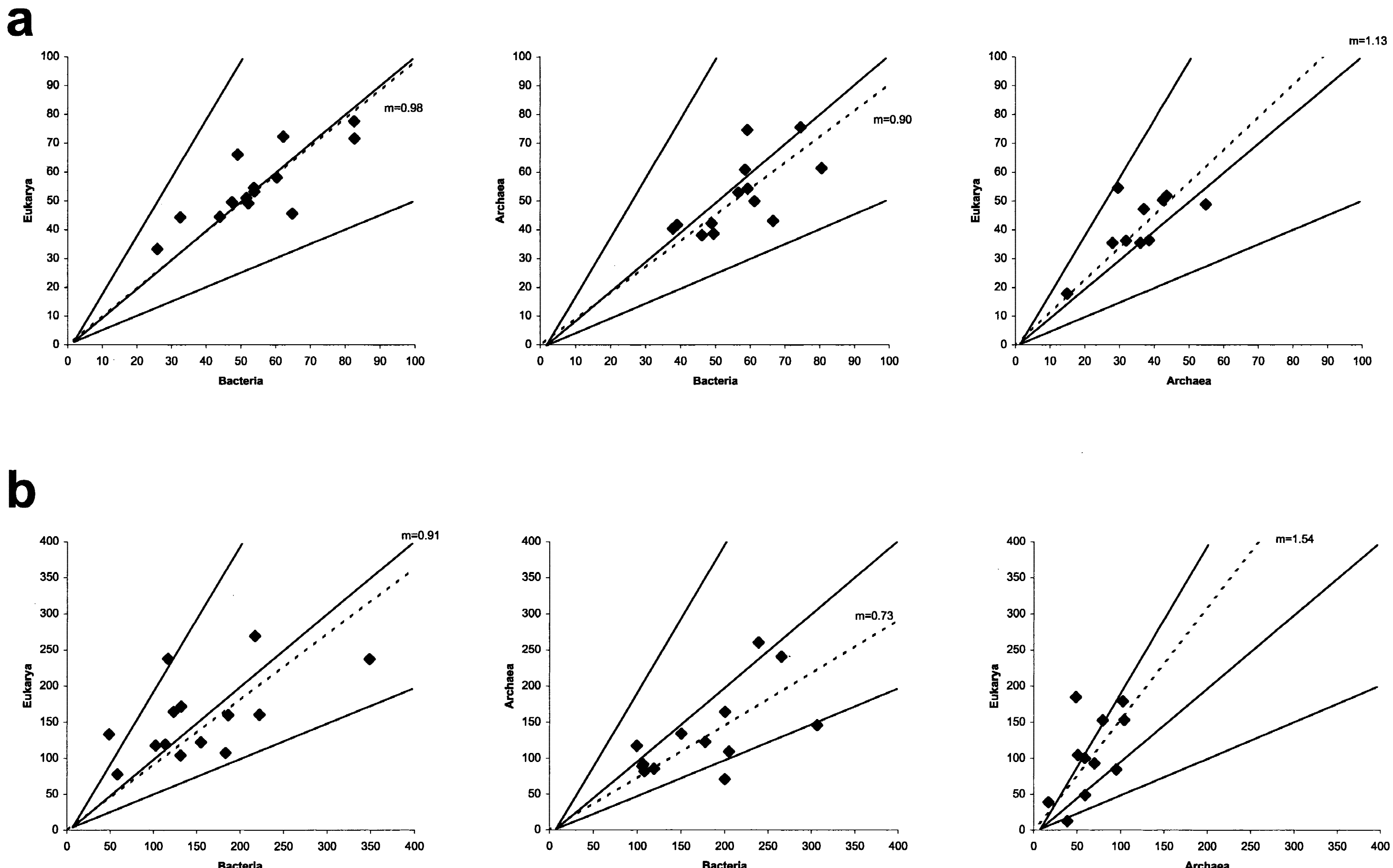


Fig. 6. Comparison of distance to the LCA in different urkingdoms. For each set of proteins, the average distance between the LCA and each urkingdom are plotted against the same distance for the other urkingdoms. Distances were calculated with the "modified Poisson" method **(a)** or the "Grishin method" **(b)**. Solid lines indicate equal rates (the central line) or cases where one urkingdom is changing twice as quickly as the other (upper and lower lines). The dotted line is the best-fit curve for the data points, constrained to pass through the origin; *m* equals the slope of the best-fit curve.

combining carefully considered groups of sequences and using them to determine divergence times. Common pitfalls of the method include horizontal transfers (both individual and organelle-mediated) and unsuspected paralogy. If such anomalies can be identified and culled from comparisons, then not unreasonable results can be achieved.

In this regard, we have been able to use paralogous outgroups to calculate the average distance to the LCA for each pair of urkingdoms. For the protein sets studied, bacterial and eukaryotic sequences have changed at virtually the same average rate in the interval since their LCA, but both groups have changed significantly faster than the archaeal sequences. It is not possible to know whether this difference is the result of a decreased archaeal rate or an increase in the rates of Bacteria and Eukarya.

The question arises: Why are the archaeal proteins changing more slowly? Perhaps it has to do with adaptation to extreme environments, many proteins being constrained against further change. Or, perhaps the difference reflects a more unbridled change among Eukarya and aerobic bacteria in an oxygen-rich world. Whatever the case, for the sequences used in this study there has been a generally slower rate of sequence evolution among the Archaea.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller A, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Brinkmann H, Phillipe H (1999) Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol 16:817–825

Brown JR, Doolittle WF (1995) Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. Proc Natl Acad Sci USA 92:2441–2445

Brown JR, Robb FT, Weiss R, Doolittle WF (1997) Evidence for the early divergence of tryptophanyl- and tyrosyl-tRNA synthetases. J Mol Evol 45:9–16

Dickerson RE (1971) The structure of cytochrome c and the rates of molecular evolution. J Mol Evol 1:26–45

Doolittle RF, Feng D-F (1990) Nearest neighbor procedure for relating progressively aligned amino acid sequences. Meth Enzymol 183: 659–669

Doolittle RF, Handy J (1998) Evolutionary anomalies among the aminoacyl-tRNA synthetases. Curr Opin Genet Dev 8:630–636

Doolittle RF, Feng D-F, Tsang T, Cho G, Little E (1996) Determining the divergence times of the major kingdoms of living organisms with a protein clock. Science 271:470–477

Feng D-F, Doolittle RF (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol 25:351–360

Feng D-F, Doolittle RF (1996) Progressive alignment of amino acid sequences and the construction of phylogenetic trees from them. Meth Enzymol 266:368–382

Feng D-F, Doolittle RF (1997) Converting amino acid alignment scores into measures of evolutionary time. J Mol Evol 44:361–370

Feng D-F, Johnson MS, Doolittle RF (1985) Aligning amino acid sequences: comparison of commonly used methods. J Mol Evol 21: 112–125

Feng D-F, Cho G, Doolittle RF (1997) Determining divergence times with a protein clock: update and reevaluation. Proc Natl Acad Sci USA 94:13028–13033

Gogarten JP, Kibak H, Dittrich H, Taiz L, Bowman EJ, Bowman EJ, Manolson EJ, Poole RJ, Date T, Oshima T, Konishi J, Denda K, Yoshida M (1989) Evolution of vacuolar H^+-ATPase: implications for the origin of eukarya. Proc Natl Acad Sci USA 86:6661–6665

Gribaldo S, Cammarano P (1998) The root of the universal tree of life inferred from anciently duplicated genes encoding components of protein-targeting machinery. J Mol Evol 47:508–516

Grishin NV (1995) Estimation of the number of amino acid substitutions per site when substitution rate varies among sites. J Mol Evol 41:675–679

Hashimoto T, Sánchez LB, Shirakura T, Müller M, Hasegawa M (1998) Secondary absence of mitochondria in *Giardia lamblia* and *Trichimonas vaginalis* revealed by valyl-tRNA synthetase phylogeny. Proc Natl Acad Sci USA 95:6860–6865

Henikoff S, Henikoff JG (1993) Performance evaluation of amino acid substitution matrices. Prot Struct Funct Genet 17:49–61

Iwabe N, Kuma K-I, Hasegawa M, Osawa S, Miyata T (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukarya inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci USA 86:9355–9359

Labedan B, Boyen A, Baetans M, Charlier D, Chen P, Cunin R, Durbeco V, Glansdorff N, Herve G, Legrain C, Liang Z, Purcarea C, Roovers M, Sanchez R, Toong T-L, Van De Casteele M, van Vliet G, Xu Y, Zhang Y-F (1999) The evolutionary history of transcarbamoylases: a complex set of paralogous genes was already present in the last universal common ancestor. J Mol Evol 49:461–473

Lawson FS, Charlebois RL, Dillon J-AR (1996) Phylogenetic analysis of carbamoylphosphate synthetase genes: complex evolutionary history includes an internal duplication within a gene which can root the tree of life. Mol Biol Evol 13:970–977

Li W-H (1997) Molecular evolution. Sinauer Associates, Sunderland, MA, pp 216–218

McLaughlin PJ, Dayhoff MO (1972) Evolution of species and proteins: a time scale. *In* Atlas of protein sequence and structure, vol. 5 National Biomedical Research Foundation, Washington, DC, pp 47–52

Ochman H, Wilson AC (1987) Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. J Mol Evol 26:74–86

Ochman H, Elwyn S, Moran NA (1999) Calibrating bacterial evolution. Proc Natl Acad Sci USA 96:12638–12643

Philippe H, Forterre P (1999) The rooting of the universal tree of life is not reliable. J Mol Evol 49:509–523

Sarich VM, Wilson AC (1973) Generation time and genomic evolution in primates. Science 179:1144–1147

Sassanfar M, Kranz JE, Gallant P, Schimmel P, Kiyotaka S (1996) A eubacterial *Mycoplasma tuberculosis* tRNA synthetase is eukaryote-like and resistant to a eubacterial-specific antisynthetase drug. Biochemistry 35:9995–10003

Triola MF (1997) Elementary statistics, 7th ed. Addison-Wesley, Menlo Park, CA, pp 377–381