J Mol Evol (1999) 48:218–235

JOURNAL OF MOLECULAR EVOLUTION

Phylogeny of Organisms Investigated by the Base-Pair Changes in the Stem Regions of Small and Large Ribosomal Subunit RNAs

Jinya Otsuka, Goro Terai, Takao Nakano

Department of Applied Biological Science, Faculty of Science and Technology, Science University of Tokyo, Noda 278, Japan

Received: 23 October 1997 / Accepted: 12 August 1998

Abstract. In order to obtain the evolutionary distance data that are as purely additive as possible, we have developed a novel method for evaluating the evolutionary distances from the base-pair changes in stem regions of ribosomal RNAs (rRNAs). The application of this method to small-subunit (SSU) and large-subunit (LSU) rRNAs provides the distance data, with which both the unweighted pair group method of analysis and the neighbor-joining method give almost the same tree topology of most organisms except for some Protoctista, thermophilic bacteria, parasitic organisms, and endosymbionts. Although the evolutionary distances calculated with LSU rRNAs are somewhat longer than those with SSU rRNAs, the difference, probably due to a slight difference in functional constraint, is substantially decreased when the distances are converted into the divergence times of organisms by the measure of the time scale estimated in each type of rRNAs. The divergence times of main branches agree fairly well with the geological record of organisms, at least after the appearance of oxygen-releasing photosynthesis, although the divergence times of Eukaryota, Archaebacteria, and Eubacteria are somewhat overestimated in comparison with the geological record of Earth formation. This result is explained by considering that the mutation rate is determined by the accumulation of misrepairs for DNA damage caused by radiation and that the effect of radiation had been stronger before the oxygen molecules became abundant in the atmosphere of the Earth.

Correspondence to: J. Otsuka

Key words: Base pairs — Divergence time — Functional constraint — Molecular clock — Mutation rate — Phylogenetic tree — Ribosomal RNA

Introduction

As the nucleotide sequence data of ribosomal RNAs (rRNAs) have increased, investigations of phylogeny have been expanded to organisms more anciently diverged than those investigated by the analysis of proteins and/or protein genes; the first attempt concentrated mainly on animals, plants, and fungi using the nucleotide base changes observed in 5S rRNAs (Hori and Osawa 1979; Kumazaki et al. 1983) and then expanded to both eukaryotes and prokaryotes using the base changes in the small-subunit (SSU) rRNAs (Woese 1987) and those in the large-subunit (LSU) rRNAs (De Rijk et al. 1995). This is reasonable because these SSU and LSU rRNAs are ubiquitous, i.e., present in mitochondria and photosynthetic plastids as well as in prokaryotes and host eukaryotes. In these trials of investigating the phylogeny, however, the base-change probabilities are estimated from counting the base changes observed at all the sites except for those in the variable regions, without considering the effect of selection that would have been different depending on sites.

The secondary structure of stem and loop regions is fairly well characterized in each type of rRNA, suggesting its importance for function, and, in practice, the sequence fragments, which play the essential roles in the manifestation of ribosomal function, are beginning to be identified experimentally in SSU and LSU rRNAs (Raue et al. 1989). The nucleotide bases in these sequence frag-

ments are highly conserved and most of them are centered on loop or single-stranded regions, although some of the them involved in the elongation and ribosomal subunit association are contained in the edges of some stem regions. Moreover, the kingdom-specific sequence fragments are also indicated on LSU rRNAs (Egebjerg et al. 1989).

Probably because of the mixing of base changes under different influences of selection, the previous analyses on SSU and LSU rRNAs result in the base change rates varying from evolutionary lineage to lineage, and they could not help constructing unrooted trees (Woese 1987) and tree topology (De Rijk et al. 1995), apart from the ''molecular clock'' hypothesis initially proposed for investigating the phylogeny of organisms with the estimation of their divergence times (Zuckerkandl and Pauling 1962; Margoliash 1963). Although many kinds of algorithms (Fitch and Margoliash 1967; Farris 1972; Sattath and Tversky 1977; Fitch 1981; Tateno et al. 1982; Efron 1982; Faith 1985; Saitou and Nei 1987) have been proposed for reconstructing phylogenic trees from evolutionary distance data including nonadditive ones, the problem of how the evolutionary distance data faithfully reflecting the ''true'' phylogeny can be deduced from the observed base changes remains unresolved. For the purpose of challenging the latter problem, it is necessary to look for the base changes that have ticked away fairly regularly under a definite functional constraint.

Recently, authors have proposed a theoretical method to evaluate the nucleotide base changes under the functional constraint of maintaining the matched base pairs G:C, C:G, A:U, and U:A (Otsuka et al. 1997a). This is a theoretical expression for the previous indication that the nucleotide bases in the stem regions of 5S rRNAs are under the influence of selection to maintain the matched base pairs (Curtiss and Vournakis 1984; Horimoto et al. 1989). According to this theory, the rate of changes between the matched base pairs is expected to be slower than the mutation rate of individual nucleotide bases by the elimination of mismatched base pairs, and the functional constraint acting on each base-pairing site can be evaluated by examining the fractions of mismatched base pairs appearing at the site. In the present paper, this theoretical method is applied to SSU and LSU rRNAs, because the situations indicated on 5S rRNAs is similarly seen in most stem regions of SSU and LSU rRNAs. Moreover, the comparison of homologous rRNA sequences shows a remarkable feature that the nucleotide base changes have occurred more frequently in stem regions than in loop regions. If the observed base changes are under a definite functional constraint of maintaining the matched base pairs, they may be suitable for investigating the phylogeny of organisms. First, we examine the base-pair contents in every stem region of each type of rRNAs and choose the stem regions where the base changes seem to have occurred under almost the same functional constraint of base-pairing. The evolutionary distance between different organisms is then calculated in terms of the base-pair change probabilities estimated from the base-pair changes observed in stem regions. The evolutionary distances thus obtained are nearly additive, and the phylogenetic tree of most organisms can be drawn with a measure of time, which is tolerable for the comparison with the geological record of organisms.

Method for Evaluating Evolutionary Distance from Observed Base-Pair Changes

In the stem regions of rRNAs, the base-pairings of G:C and C:G are most outstanding, the pairings of A:U and U:A are second, and other mispairings are scarcely observed. Moreover, these four types of base pairs are changeable at most base-pairing sites in stem regions. This characteristic feature indicates the following elementary process of base changes in stem regions. Because it is of a very low probability that the nucleotide bases at two sites are substituted simultaneously, it is natural to consider that any counterpart of the pair bases is only substituted within a short interval of time. Thus, any of the four favorable base pairs, G:C, C:G, A:U, and U:A, would be converted into less favorable or unfavorable base pairs by a substitution, but the latter base pairs also have a chance to return to the favorable ones by successive substitutions. On the other hand, a pair of nucleotide bases at the positions suitable for forming a stable base pair would have been exposed to selection according to the strength of its base-pairing; G:C and C:G pairs are most favorable, A:U and U:A pairs are second, G:U and U:G are less favorable, and others are unfavorable.

Because the difference in frequency observed between favorable base pairs and unfavorable ones is decisive as shown in the following section, the time changes of 16 possible base pairs can be split into two types of equations; one is the equation for the time change of favorable base pairs and the other concerns the ratios of unfavorable base pairs to favorable base pairs (Otsuka et al. 1997a). According to this theory, the former type of equation is symbolically expressed in the following form, with the use of the probability $P(X{:}Y,t)$ that a favorable base-pair $X{:}Y$ occupies the base-pairing sites at time t:

$$\frac{d}{dt}P(X{:}Y,t) = \frac{\partial}{\partial t}P(X{:}Y,t) + (S_{XY} - \bar{S})P(X{:}Y,t) \tag{1}$$

Here $X{:}Y$ stands for A:U, U:A, G:C, and C:G. The first term on the right side of Eq. (1) represents the time change between the favorable base pairs, and its explcit form is expressed as follows.

$$\frac{\partial}{\partial t}P(A{:}U,t) = -\{(\alpha^2)+(\beta^2)+(\gamma^2)\}P(A{:}U,t) + (\alpha^2)P(G{:}C,t) + (\beta^2)P(U{:}A,t) + (\gamma^2)P(C{:}G,t) \tag{2-1}$$

$$\frac{\partial}{\partial t}P(U{:}A,t) = -\{(\alpha^2)+(\beta^2)+(\gamma^2)\}P(U{:}A,t) + (\alpha^2)P(C{:}G,t) + (\beta^2)P(A{:}U,t) + (\gamma^2)P(G{:}C,t) \tag{2-2}$$

$$\frac{\partial}{\partial t}P(G{:}C,t) = -\{(\alpha^2)+(\beta^2)+(\gamma^2)\}P(G{:}C,t) + (\alpha^2)P(A{:}U,t) + (\beta^2)P(C{:}G,t) + (\gamma^2)P(U{:}A,t) \tag{2-3}$$

$$\frac{\partial}{\partial t}P(C{:}G,t) = -\{(\alpha^2)+(\beta^2)+(\gamma^2)\}P(C{:}G,t) + (\alpha^2)P(U{:}A,t) + (\beta^2)P(G{:}C,t) + (\gamma^2)P(A{:}U,t) \tag{2-4}$$

where (α^2), (β^2), and (γ^2) are the base-pair change rates, each expressed by

$$(\alpha^2) = \left\{ \frac{1}{2(\alpha+\beta+\gamma) - S_{(GU)}} + \frac{1}{2(\alpha+\beta+\gamma) - S_{(AC)}} \right\} \alpha^2 \tag{3-1}$$

$$(\beta^2) = \frac{2}{2(\alpha+\beta+\gamma) - S} \beta^2 \tag{3-2}$$

$$(\gamma^2) = \left\{ \frac{1}{2(\alpha+\beta+\gamma) - S_{(CU)}} + \frac{1}{2(\alpha+\beta+\gamma) - S_{(GA)}} \right\} \gamma^2 \tag{3-3}$$

The three parameters, α, β, and γ, are the substitution rates formally adopted from the three-parameter model of substitutions. Although the substitution rates in the original three-parameter model (Kimura 1981) are regarded as the change rates of nucleotide bases under the assumption of neutral changes, the substitution rates α, β, and γ in the present formulation are considered to be the ''true'' rates of mutations, which further experience the selection (Otsuka et al. 1997a, b). $S_{(GU)}$, $S_{(AC)}$, $S_{(CU)}$, and $S_{(GA)}$ denote the rates of elimination (negative selection) of mismatched pairs, G:U or U:G, A:C or C:A, C:U or U:C, and G:A or A:G, respectively, and the elimination rates of other mismatched pairs A:A, U:U, G:G, and C:C, which appear in the stem region at much lower frequencies, are bundled into one rate denoted by *S*, for simplicity. The second term on the right side of Eq. (1) represents the influence of selection for favorable base pairs; S_{XY} is the rate of selection for the base-pair *X:Y* and $\overline{S}$ is an average of the selective rates S_{XY}'s for all four types of favorable base pairs. If the contents of the four favorable base pairs are equal, the selective term vanishes and the changes between the favorable base pairs are well represented by Eq. (2-1) to (2-4). As shown in the following section, the content of base pairs G:C and C:G is observed to be approximately double that of A:U and U:A in the rRNAs from most organisms, suggesting the presence of base-pair change flow streaming from G:C and C:G to A:U and U:A. If the contents of these base pairs are almost-constant in homologous rRNAs from different organisms, however, the flow intensity may be common to these rRNAs, and the base-pair changes observed between the organisms may be regarded as those due to the mutual changes of favorable base pairs, which are represented by Eqs. (2-1) to (2-4).

Equations (2-1) to (2-4) take forms similar to the equations of individual nucleotide base changes in the three-parameter model. If we tentatively consider a correspondence of base pairs A:U, U:A, G:C, and C:G to the single nucleotide bases A, U, G, and C, we can see that the base-pair change rates, (α^2) (β^2), and (γ^2), formally correspond to the substitution rates α, β, and γ, respectively, in the usual three-parameter model. This means that a new evolutionary distance defined in terms of (α^2), (β^2), and (γ^2) can be evaluated from counting the base-pair changes observed in the comparison of homologous rRNAs derived from different organisms by the following procedure. In the comparison of homologous rRNA sequences I and II, three types of base-pair change probabilities, *P*, *Q*, and *R*, are estimated by counting the base-pair changes according to the following three categories:

Category for *P*				
I	A:U	G:C	U:A	C:G
II	G:C	A:U	C:G	U:A
Category for *Q*				
I	U:A	A:U	C:G	G:C
II	A:U	U:A	G:C	C:G
Category for *R*				
I	U:A	G:C	A:U	C:G
II	G:C	U:A	C:G	A:U

That is, the base-pair change probability *P* is estimated from counting the fraction of base-pairing sites where base-pair changes from A:U to G:C and from U:A to C:G, and vice versa, are observed between the sequences I and II. Similarly, the change probabilities *Q* and *R* are estimated as the ratio of base-pairing sites showing the corresponding categories of base-pair changes to the total number of base-pairing sites. The relation connecting the base-pair change rates with the base-pair change probabilities is then obtained in the following form:

$$(\alpha^2)t = \frac{1}{8} \ln \frac{1-2(Q+R)}{\{1-2(P+Q)\}\{1-2(P+R)\}} \tag{4-1}$$

$$(\beta^2)t = \frac{1}{8} \ln \frac{1-2(P+R)}{\{1-2(P+Q)\}\{1-2(Q+R)\}} \tag{4-2}$$

$$(\gamma^2)t = \frac{1}{8} \ln \frac{1-2(P+Q)}{\{1-2(P+R)\}\{1-2(Q+R)\}} \tag{4-3}$$

where *t* is the divergence time of compared sequences I and II or of the organisms from which the compared sequences I and II are derived, respectively. If a new evolutionary distance is defined by

$$K(t) = 2\{(\alpha^2) + (\beta^2) + (\gamma^2)\}t \tag{5}$$

this distance is then evaluated from the estimated values of the base-pair change probabilities *P*, *Q*, and *R* with the use of Eq. (4-1) to (4-3). Assuming the binomial distribution for each type of change probability, we can also derive the variance σ_K^2 for the evolutionary distance (5) in the following form.

$$\sigma_K^2 = \{a^2P + b^2Q + c^2R - (aP + bQ + cR)^2\}/n \tag{6}$$

where

$$a = \frac{1}{4} \left\{ \frac{1}{1-2(P+Q)} + \frac{1}{1-2(P+R)} \right\} \tag{7-1}$$

$$b = \frac{1}{4} \left\{ \frac{1}{1-2(P+Q)} + \frac{1}{1-2(Q+R)} \right\} \tag{7-2}$$

$$c = \frac{1}{4} \left\{ \frac{1}{1-2(Q+R)} + \frac{1}{1-2(P+R)} \right\} \tag{7-3}$$

and *n* is the total number of base-pairing sites for which base-pair changes are counted.

Together with the derivation of Eq. (1), the ratio of an unfavorable base pair to favorable base pairs is also derived under the assumption that the elimination rates of unfavorable or mismatched base pairs are much faster than the substitution rates. Two examples are given here:

$$P(G{:}U, t) \approx \frac{\alpha}{2(\alpha+\beta+\gamma) - S_{(GU)}} \{P(A{:}U, t) + P(G{:}C, t)\} \tag{8-1}$$

$$P(A{:}G, t) \approx \frac{\gamma}{2(\alpha+\beta+\gamma) - S_{(GA)}} \{P(C{:}G, t) + P(A{:}U, t)\} \tag{8-2}$$

These relations are useful for estimating the elimination rates of unfavorable base pairs, $S_{(GU)}$ and $S_{(GA)}$, as well as the ''true'' mutation rates, α and γ, when the divergence time *t* of compared sequences is known and the values of (α^2), (β^2), and (γ^2) are estimated by Eq. (4-1), (4-2), and (4-3), respectively.

Preliminary Investigation of Nucleotide Base Sequences

At first, all the nucleotide base sequences of SSU rRNAs and those of LSU rRNAs stored in the databases by Van de Peer et al. (1997) and De Rijk et al. (1997) are examined in the preliminary investigation. The nomenclature for the stem regions is also used according to these databases. Although the homologous alignment of SSU rRNAs and that of LSU rRNAs have been carried out for the sequence data stored in the databases, the base pairs in each stem region are checked again by every nucleotide sequence, and then the correspondence of base-pairing sites in each stem region is reexamined among the homologous rRNAs from different organisms. At this stage of examination, rRNAs from mitochondria in animals are excluded from the present study. These rRNAs are much shorter than the corresponding type of rRNAs from host eukaryotic genomes and show much faster rates of base changes even in the stem regions.

The contents of mismatched base pairs as well as of matched base pairs are then counted in every stem region of available rRNAs. By this examination, the following stem regions are excluded from the present study, because of the high frequency of occurrence of mismatched base-pairs and/or of insertions of deletions.

SSU rRNAs
- Eukaryota; 1, 2, 3, 6, 8, 10, E10-1, 11, E23-1~6, E23-8, E23-10, 31, 35, 40, 41, 44, 49, 50
- Eubacteria; 1, 2, 3, 6, 10, 11, 17, 18, 25, 29, 32, 37, P37-2, 42, 45, 46, 49, 50
- Archaebacteria; 3, 6, 11, 16, 18, 22, P23-1, 29, 35, P37-2, 39, 42, 49, 50

LSU rRNAs
- Eukaryota; A1, B1, B2, B3, B4, B5, B6, B7, B8, B9, B14-1, C1-1, C1-2, C1-3, D4-1, D10, D21-1, E9-1, E20-1, E20-2, E24, G1-1, G5-2, H1-1, H1-2, H1-3, I1
- Eubacteria; A1, B2, B9, B13-1, B14-1, C1-1, C1-2, C1-3, D4-1, D5, D5-1, D14-1, D21-1, E9-1, E11-1, E15, E20-1, E20-2, G1-1, G5-1, G5-2, G12, H1-1, H1-2, H1-3, I1
- Archaebacteria; A1, B2, B8, B14-1, B15, C1-1, C1-2, C1-3, D4-1, D14-1, D20, D21-1, E9-1, E11-1, E15, E20-1, E20-2, G1-1, G5-1, G5-2, H1-1, H1-2, H1-3

Besides the incomplete stem regions mentioned above, the stem regions that carry stable base-pairs highly conserved in each kingdom are also excluded from the present calculation. These stem regions are as follows.

SSU rRNAs
- Eukaryota; 5, 7, 20, 21, 23, 26
- Eubacteria; 20, 39,
- Archaebacteria; 2, 15, 20, 21, 32

LSU rRNAs
- Eukaryota; E1, E16, E22, E26, E27, some part of G3, G9, G17, G20
- Eubacteria; D19, E5, E22
- Archaebacteria; B5, E5, E16, E22, E27, G9

The inclusion of these stem regions leads not only to an underestimation of base-pair change probabilities within the same kingdom but also to an overestimation of evolutionary distance between different kingdoms, because most of the conserved base-pairs are kingdom-specific. The SSU and LSU rRNAs of plant mitochondria and photosynthetic plastids are similar to the SSU and LSU rRNAs in Eubacteria, respectively, with respect to the loci of both the incomplete stem regions and the conserved base pairs. Thus, the inclusion and exclusion of stem regions in SSU and LSU rRNAs from these organelles are treated in the same way as for the respective types of rRNAs in Eubacteria.

Contents of all possible base-pairs appearing in the stem regions adopted in the present study are listed in Table 1. These contents are those obtained by taking an average over several phyla which show almost the same ratio of (A:U) to (G:C). Some Protoctista (*Plasmodium*, *Dictyostelium*, *Crithidia*, *Trypanosoma*, *Staurastrum*, and *Entamoeba*) show an abnormally high (A:U) content and the average of their base-pair contents is shown separately from those of many other Protoctista. The base-pair contents of Euryarchaeota are those averaged over Methanococcus, Methanobacter, Methanomicrobium, Halobacteria, and Thermoplasma, while the contents of Thermococcus are shown separately from other Euryarchaeota because of their lower (A:U) content. Most Eubacteria (Proteobacteria alpha, beta, gamma, delta, and epsilon, Spirochetes, Cyanobacteria, Fibrobacter, Green sulfur bacteria, Chlamydiae, Fusobacteria, Flavobacteria, Planctomyces, and Mycoplasmas) show the standard ratio of (A:U) to (G:C). Even in the Gram-positive high G + C and low G + C groups, the contents of base pairs are not much different from those in other groups. On the other hand, both SSU and LSU rRNAs from Radioresistant micrococci, *Thermatoga* and Green nonsulfur bacteria show a higher content of (G:C), and their average contents of base pairs are listed in Table 1 as those of some other Eubacteria. Because the optimal growth temperatures of these Eubacteria as well as of Crenarchaeota are high, the higher content of (G:C) in these organisms may be the result of adaptation to retain the secondary structure of rRNAs in a high-temperature environment. Thus, the evolutionary distances of these organisms from the other organisms might be calculated to be longer than those in the true phylogeny. Both SSU and LSU rRNAs from fungal mitochondria are irregular, showing a high content of (A:U), which is almost-equal to the content of (G:C). Thus, the rRNAs from fungal mitochondria are also excluded from the present study, and the phylogenic relation of mitochondria to the other

Table 1. Contents of all possible base-pairs observed in the stem regions used in the present study[a]

Main branches		(A:U)	(G:C)	(G:U)	(A:G)	(C:A)	(C:U)	(A:A)	(U:U)	(G:G)	(C:C)
SSU rRNA											
Eukaryota	Animals	0.360	0.566	0.065	0.001	0.004	0.001	0.000	0.000	0.000	0.003
	Plants	0.407	0.515	0.065	0.000	0.004	0.000	0.000	0.000	0.000	0.009
	Fungi	0.389	0.526	0.074	0.000	0.007	0.000	0.000	0.000	0.000	0.004
	Most Protoctista	0.405	0.513	0.062	0.003	0.008	0.001	0.000	0.003	0.000	0.005
	Some other Protoctista	0.491	0.416	0.071	0.005	0.005	0.003	0.0003	0.000	0.001	0.005
Archaebacteria	Crenarchaeota	0.148	0.791	0.044	0.009	0.008	0.000	0.000	0.000	0.000	0.000
	Most Euryarchaeota	0.267	0.675	0.051	0.003	0.004	0.000	0.000	0.000	0.000	0.000
	Thermococcus group	0.133	0.823	0.037	0.007	0.000	0.000	0.000	0.000	0.000	0.000
Eubacteria	Most Eubacteria	0.314	0.621	0.062	0.000	0.001	0.001	0.000	0.001	0.000	0.000
	Gram Positives Low G + C	0.294	0.650	0.049	0.001	0.001	0.000	0.000	0.005	0.000	0.000
	Gram Positives High G + C	0.236	0.677	0.087	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Some other Eubacteria	0.195	0.757	0.048	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Mitochondria	In Plants	0.295	0.619	0.076	0.000	0.001	0.009	0.000	0.000	0.000	0.000
Plastids	In Plants and Protoctista	0.343	0.593	0.059	0.000	0.000	0.000	0.000	0.000	0.000	0.001
LSU rRNA											
Eukaryota	Animals	0.303	0.620	0.060	0.001	0.006	0.002	0.001	0.003	0.002	0.002
	Plants	0.302	0.631	0.056	0.001	0.005	0.004	0.000	0.001	0.000	0.000
	Fungi	0.396	0.520	0.072	0.000	0.003	0.005	0.001	0.000	0.001	0.002
	Most Protoctista	0.380	0.545	0.062	0.000	0.003	0.005	0.001	0.000	0.001	0.003
	Some other Protoctista	0.413	0.491	0.082	0.001	0.005	0.005	0.000	0.001	0.000	0.002
Archaebacteria	Crenarchaeota	0.145	0.783	0.071	0.000	0.001	0.000	0.000	0.000	0.000	0.000
	Most Euryarchaeota	0.259	0.634	0.103	0.000	0.000	0.001	0.000	0.000	0.000	0.003
	Thermococcus group	0.127	0.816	0.056	0.000	0.000	0.000	0.000	0.000	0.000	0.001
Eubacteria	Most Eubacteria	0.285	0.627	0.079	0.000	0.001	0.001	0.003	0.000	0.000	0.004
	Gram Positives Low G + C	0.302	0.623	0.067	0.001	0.002	0.002	0.002	0.000	0.000	0.001
	Gram Positives High G + C	0.231	0.670	0.090	0.002	0.000	0.001	0.000	0.001	0.000	0.005
	Some other Eubacteria	0.145	0.797	0.054	0.002	0.001	0.000	0.000	0.000	0.000	0.001
Mitochondoria	In Plants	0.333	0.587	0.057	0.002	0.006	0.004	0.002	0.003	0.004	0.002
Plastids	In Plants and Protoctista	0.303	0.603	0.081	0.003	0.004	0.003	0.000	0.000	0.000	0.003

[a] (*X:Y*) means *X:Y* + *Y:X*.

organisms will be investigated with the use of rRNAs from plant mitochondria. The rRNAs from photosynthetic plastides in plants and Protoctista show almost the same content of base-pairs, and the average content of them is listed in Table 1.

Results

The base-pair change probabilities P, Q, and R are estimated by counting the corresponding base-pair changes observed in the stem regions by a pairwise comparison of homologous rRNAs. If, in the homologous alignment, a matched base pair in one source correspponds to a mismatched base pair in the counterpart, such a matched base-pair change is omitted from the count of base-pair changes. The distance between the homologous rRNAs from different sources is evaluated in terms of the evolutionary distance defined by Eq. (5). This procedure is carried out separately for SSU rRNAs and LSU rRNAs.

Construction of a Phylogenetic Tree

The construction of a phylogenetic tree is carried out by essentially the same procedure as the unweighted pair group method of analysis (UPGMA). Because the sequence data from a great number of organisms are available especially in the case of SSU rRNAs, the tree construction is carried out in the following way. The distance matrix, whose elements are evolutionary distances calculated from the base-pair changes in stem regions, is decisively divided into the three submatrices, which correspond to those of Eukaryota, Archaebacteria, and Eubacteria, respectively. Moreover, each of the submatrices is further divided into the smaller parts corresponding to those of phyla or subdivisions, mostly consistently with the proposal for Eukaryota by Whittaker and Margulis (1978) and that for Archaebacteria and Eubacteria by Woese (1987), when the organisms giving the shorter distances are arranged at the nearer position in the row and column. Thus, we first cluster the species in a same phyla or subdivision. Among all possible pairs of species, we choose the one that gives the shortest distance K_{12}. This pair of species, 1 and 2, is then regarded as a combined taxonomic unit (1–2), and the third species 3 is chosen as that giving the shortest distance $K_{(1-2)3}$, where $K_{(1-2)3}$ means the average distance of K_{13}, between species 1 and 3, and K_{23}, between species 2 and 3. This procedure is continued until all species in a

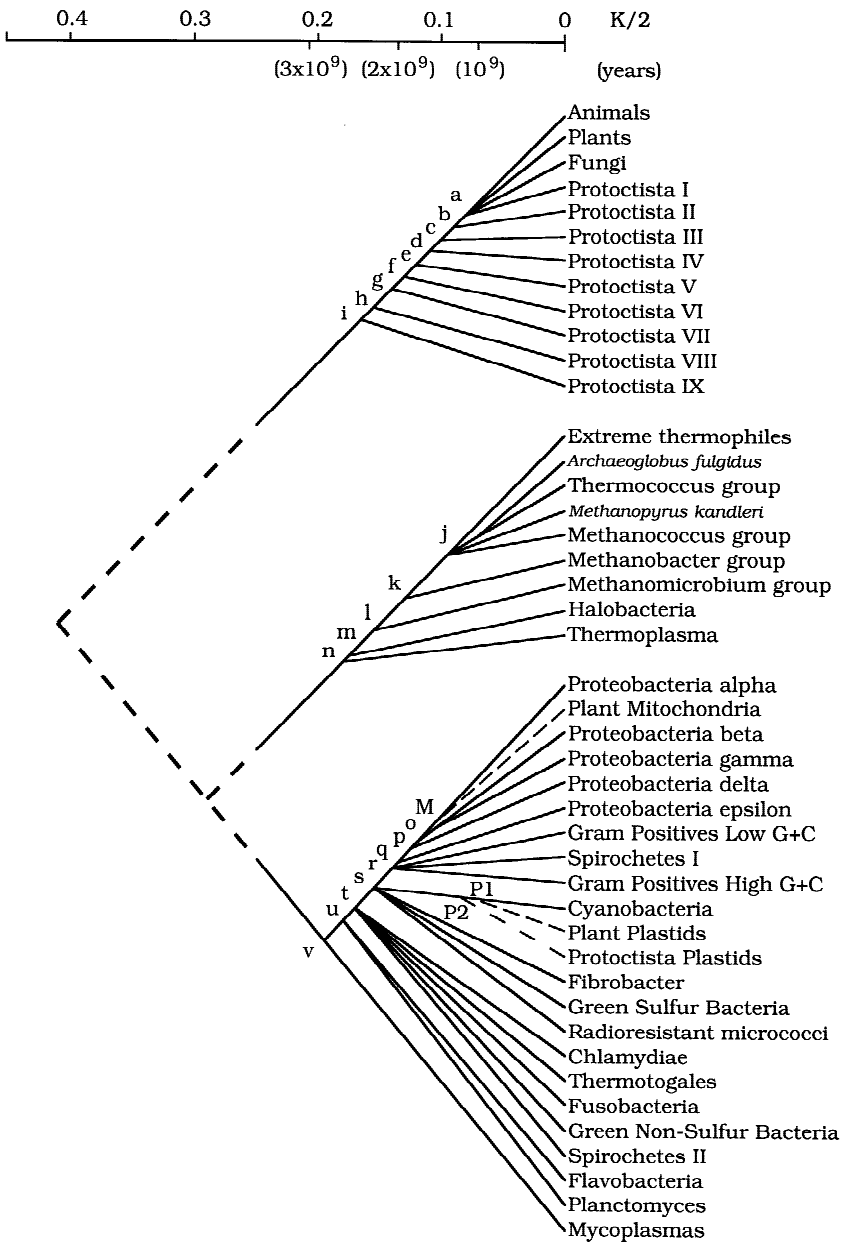

Fig. 1. The phylogenetic tree of Eukaryota, Archaebacteria and Eubacteria constructed by the unweighted pair-group method of analysis. Although the evolutionary distances between animals, plants and fungi are calculated to be almost the same, Protoctista are divided into nine groups (I–IX) according to their evolutionary distances. These groups are as follows. I: Dinoflagellatta, Rizopoda (e.g., *Hartmannella*), Chrysophyta, Haptophyta, Cryptophyta, Zoomastigina, Xanthophyta, Eustigmatophyta, Bacillariophyta, Phaeophyta, Rhodophyta (e.g., *Bangia*), Gamophyta, Chlorophyta, Actinopoda, Ciliophora (e.g., *Rhodomonas*), Apicomplexa (e.g., *Toxoplasma*), Hypochytridiomycota, Chytridiomycota, Oomycota, II: Ciliophora (e.g., *Paramecium*), III: Rizopoda (e.g., *Paulinella*), IV: Labyrinthulomycota, V: Ciliophora (e.g., *Tetrahymena*), VI: Rhodophyta (e.g., *Gelidium*), VII: Apicomplex (e.g., *Plasmodium*), VIII: Acrasiomycota, IX: Rizopoda (e.g., *Entamoeba*). The divergence of Eukaryota, Archaebacteria and Eubacteria is shown by *thick broken lines* because the rate constancy seems to be broken in the comparison between the three kingdoms. Plant mitochondria and photosynthetic plastids are assigned to have diverged from the Proteobacteria alpha subdivision and the Cyanobacteria, respectively, solely by their shortest distances, and their phylogenic relations are shown by *thin broken lines*. The base-pair change probabilities, evolutionary distances and standard deviations at the main branching points, *a* to *v*, and *M*, *P1*, and *P2*, are listed in Table 2. For the time scale, see the text.

phlum or subdivision are combined. In parallel to the clustering of species within each phylum, the submatrix of each kingdom is reduced to the interphylum matrix which consists of the elements, each representing an average of evolutionary distances obtained by all interphylum comparison of species. Then, we proceed to cluster the phyla that show the shorter interphylum distance. This procedure is carried out by the same way as for clustering species, each phylum being regarded as a new taxonomic operational unit at this step. Because Eubacteria consist of many phyla and subdivisions, such clustering of phyla is started in several groups of phyla or

Table 2. Three types of base-pair change probabilities, evolutionary distances, numbers of base-pairing sites, and standard deviations calculated at the main branching points in the tree drawn in Fig. 1 (the case of SSU rRNAs)

Main branching points		Change probability			Evolutionary distance ($K/2$)	Number of base-pairing sites	Standard deviation
		P	Q	R			
Eukaryota	a	0.085	0.035	0.028	0.083	275	0.014
	b	0.086	0.033	0.042	0.091	243	0.015
	c	0.105	0.032	0.038	0.101	237	0.017
	d	0.097	0.044	0.048	0.110	242	0.017
	e	0.095	0.052	0.060	0.121	240	0.018
	f	0.107	0.077	0.037	0.132	250	0.019
	g	0.126	0.047	0.062	0.142	243	0.020
	h	0.155	0.050	0.042	0.154	235	0.022
	i	0.143	0.063	0.055	0.163	224	0.023
Archaebacteria	j	0.077	0.069	0.021	0.095	255	0.015
	k	0.128	0.056	0.031	0.129	243	0.019
	l	0.143	0.051	0.053	0.153	241	0.022
	m	0.136	0.084	0.056	0.173	250	0.023
	n	0.153	0.073	0.059	0.182	239	0.025
Eubacteria	o	0.091	0.065	0.047	0.119	212	0.019
	p	0.103	0.067	0.042	0.125	201	0.020
	q	0.120	0.057	0.048	0.135	209	0.021
	r	0.116	0.067	0.048	0.139	205	0.022
	s	0.122	0.071	0.059	0.155	203	0.023
	t	0.130	0.077	0.067	0.172	202	0.025
	u	0.124	0.098	0.065	0.182	201	0.026
	v	0.155	0.078	0.070	0.197	202	0.028
Plant mitochondria vs Proteobacteria alpha	M	0.095	0.056	0.038	0.110	198	0.019
Plant plastids vs Cyanobacteria	P1	0.091	0.020	0.019	0.073	197	0.015
Protoctista plastids vs Cyanobacteria	P2	0.099	0.024	0.025	0.084	196	0.016
Eubacteria vs Archaebacteria		0.153	0.137	0.113	0.290	88	0.057
Eukaryota vs Archaebacteria		0.195	0.137	0.105	0.334	75	0.073
Eukaryota vs Eubacteria		0.220	0.185	0.145	0.506	72	0.121

subdivisions showing shorter interphylum distances. For example, the alpha, beta, and gamma subdivisions of Proteobacteria (or Purple bacteria in Woese's nonmenclature) are first clustered and then clustered with episilon subdivision, while Gram-positive low G + C and high G + C are clustered with some Spirochetes. These groups, each obtained by clustering some phyla or subdivision, and the remaining phyla are then subject to the next step of clustering. Such procedure of clustering phyla is continued in each kingdom until the final two groups or phyla are clustered with the average evolutionary distance between them. Finally, the three kingdoms are compared with the evolutionary distances each calculated as the mean interkingdom distance.

SSU rRNAs. The phylogenetic tree of organisms constructed by above procedure is shown in Fig. 1, where the projection of the branching point to the abscissa indicates the evolutionary distance between the branches. Although the nucleotide base sequences of SSU rRNAs are available from a large number of organisms, the functional constraint acting on this type of rRNAs is so strong that the evolutionary distances calculated between the species in the same phylum is too short to be distinguished from each other in most cases. Thus, the construction of tree is started from the taxonomic unit of a phylum or of a higher category. In particular, the evolutionary distances among different phyla are negligibly small in animals, plants, and fungi, and the evolutionary distances among animals, plants, and fungi are calculated to be almost the same, although the sequence data of SSU rRNAs from 165 species of Ascomycota, 63 species of Basidiomycota, and 25 species of Zygomycota are used. On the other hand, distinctive evolutionary distances are calculated among the organisms assembled under the name of Protoctista. A considerable number of Protoctista diverge at almost the same period as animals do from plants, but others diverge earlier at several points in time. These Protoctista are divided into nine groups (I–IX) according to their evolutionary distances. Although most of the organisms are denoted only by the names of phyla for simplicity, some Protoctista, which are assembled under a same phylum in taxonomy but divided into different groups, are exemplified by their genus names in the figure legend. However, it should be also noted that groups VII, VIII, and IX, which are assigned to have diverged at earlier times, correspond to those showing a high content of (A:U), and the evolutionary distances of these Protoctista might be somewhat overestimated. Archaebacteria are usually divided into

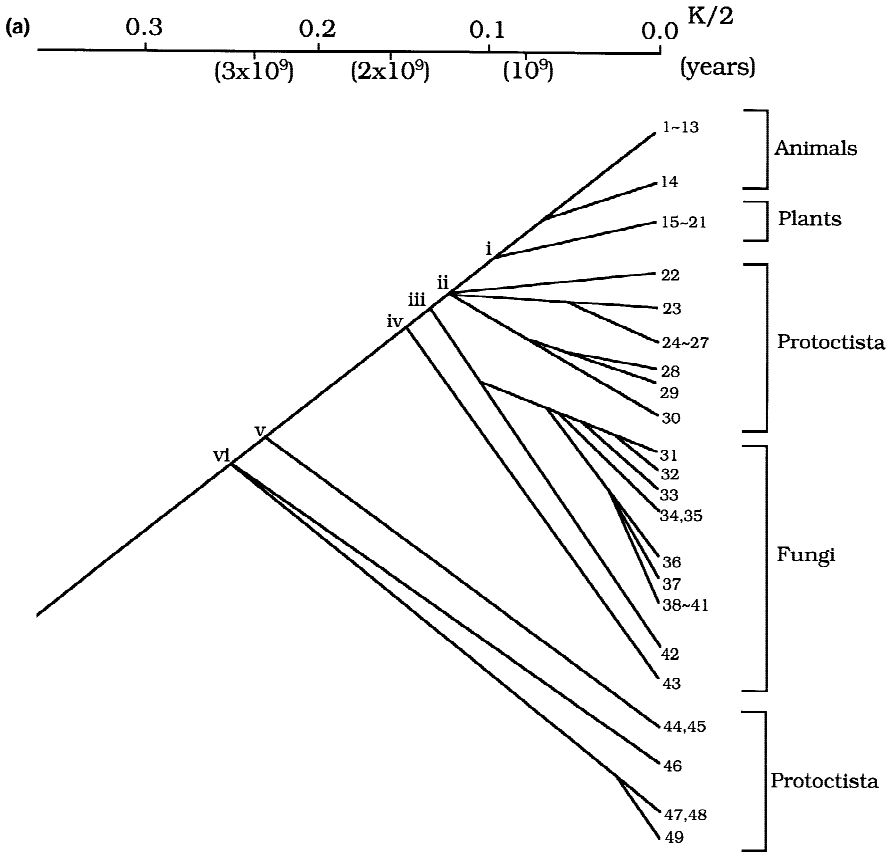

Fig. 2. The phylogenetic trees of **(a)** Eukaryota, **(b)** Archaebacteria, and **(c)** Eubacteria, each constructed by the unweighted pair group method of analysis. The organisms denoted by *numbers* and the abbreviated names of phyla or subdivisions in Archaebacteria and Eubacteria are as follows. **(a)** 1 *Homo,* 2 *Mus,* 3 *Rattus,* 4–6 *Xenopus,* 7 *Acipenser,* 8 *Anguilla,* 9 *Latimeria,* 10 *Lepidosiren,* 11 *Neoceratodus,* 12 *Oncorhynchus,* 13 *Protopterus,* 14 *Herdmania,* 15 *Oryza,* 16 *Arabidopsis,* 17 *Brassica,* 18 *Citrus,* 19 *Fragaria,* 20 *Lycopersicon,* 21 *Sinapis,* 22 *Chlorella,* 23 *Prorocentrum,* 24–27 *Toxoplasma,* 28 *Hyphochytrium,* 29 *Phytophthora,* 30 *Scytosiphon,* 31–32 *Schizosaccharomyces,* 33 *Pneumocystis,* 34–35 *Cryptococcus,* 36 *Arxula,* 37 *Candida,* 38–41 *Saccharomyces,* 42 *Mucor,* 43 *Entomophaga,* 44–45 *Plasmodium,* 46 *Dictyostelium,* 47–48 *Trypanosoma,* 49 *Crithidia.* **(b)** ETR (Extreme thermophiles): 50 *Thermoproteus,* 51 *Pyrobaculum,* 52 *Thermofilum,* 53 *Desulfurococcus,* 54–58 *Sulfolobus,* 59 *Stygiolobus,* 60–61 *Acidianus,* ARF: 62 *Archaeoglobus fulgidus,* TRC (Thermococcus): 63 *Thermococcus,* MTC (Methanococcus): 64 *Methanococcus,* MTB (Methanobacter): 65 *Methanobacterium,* MTM (Methanomicrobium): 66 *Methanospirillium,* HLB (Halobacteria), 67–68 *Halobacterium,* 69 *Halococcus,* 70 *Haloferax,* 71 *Natronobacterium,* TRP (Thermoplasma): 72 *Thermoplasma.* **(c)** PTB (Proteobacteria): 73–74 *Agrobacterium,* 75 *Bartonella,* 76–77 *Bradyrhizobium,* 78 *Rhodopseudomonus,* 79–82 *Rhodobacter,* 83–85 *Richettsia,* 86–87 *Acetobacter,* 88 *Walbachia,* 89–92 *Bordetella,* 93 *Pseudomonas capacia,* 94 *Thiobacillus cuprinus,* 95–96 *Neisseria,* 97–103 *Escherichia,* 104 *Salmonella,* 105 *Plesiomonas,* 106 Buchnera, 107 *Aeromonas,* 108–115 *Haemophilus,* 116 Coxiella, 117 *Pseudomonas aeruginosa,* 118 *Pseudomonas perfectomarina,* 119 *Ruminobacter,* 120 *Thiobacillus ferrooxidans,* 121–122 *Campylobacter,* 123 *Helicobacter,* GPB-1 (Gram positive low G + C): 124–141 *Bacillus,* 142–150 *Listeria,* 151–154 *Staphylococcus,* 155–158 *Leuconostoc,* 162–165 *Streptococcus,* 166–167 *Lactobacillus,* 168–172 *Clostridium,* 173 *Pectinatus,* 174 *Peptococcus,* SRC-I (Spirochetes I): 175 *Leptospira,* GPB-h (Gram positive high G + C), 176–177 *Frankia,* 178–181 *Streptomyces,* 182 *Micrococcus,* 183–189 *Mycobacterium,* CAB (Cyanobacteria): 190 *Anacystis,* 191 *Synechocystis,* GSB (Green sulfur bacteria), 192 *Chlorobium,* FLB (Flavobacteria): 193 *Flavobacterium,* 194 *Flexibacter,* PLM (Plancytomyces), 195 *Pirellula,* TRG (Thermotogales): 196 *Thermotoga,* RRM (Radioresistant micrococcus): 197 *Thermus,* SRC-II (Sprirochates II): 198–200 *Borrelia,* MYP (Mycoplasmas): 201–205 *Mycoplasma.* The assignment of plant mitochondria and photosynthetic plastids is carried out by the same procedure as for the case of SSU rRNAs. The base-pair change probabilities, evolutionary distances, and standard deviations at the main branching points, *i* to *xvi*, and *M, P1,* and *P2,* are listed in Table 3. For the time scale, see the text.

two large groups, i.e., Crenarchaeota and Euryarchaeota, but the distances among the subdivisions (Thermococcus, Methanococcus, Methanobacter, Methanomicrobium, Halobacteria, Thermoplasma) of Euryarchaeota are much longer than the distance of Crenarchaeota (Extreme thermophiles) from the Thermococcus group belonging to Euryarchaeota. *Archaeoglobus fulgidus* and *Methanopyrus kandleri* are explicitly referred to by their species names in the figure, because their positioning is undetermined in the classification by Woese (1987) as well as by Fox et al. (1977). In the distance matrix of Eubacteria, we have found several examples in which the organisms assembled under the same genus are divided into two or more phyla or subdivisions. Such examples are already furnished by *Pseudomonas,* some being clustered into the beta subdivision and others into the gamma subdivision (Woese 1987). In addition to these examples, some of the organisms called Mycoplasmas (*Myco-*

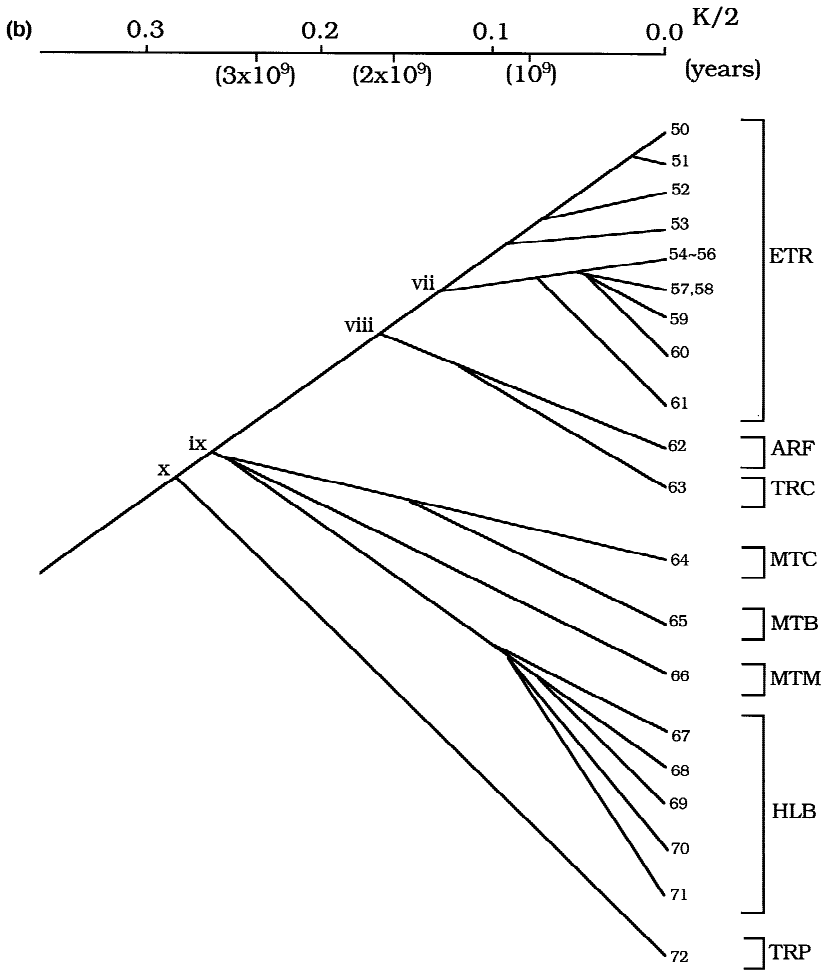

Fig. 2. Continued.

plasma carpricolum, M. ellychnium, M. feliminutum, M. mycoides, and *M. putretaciens*) are clustered together with the organisms in Gram-positive low G + C, while the other organisms also called Mycoplasmas show the longest distance from the other Eubacteria. The organisms called the Spirochetes are also divided into two groups, I and II: group I contains *Ancona, Canela, Jequitaia, Leptonema,* and *Leptospira,* while group II consists of *Borrelia, Brachyspira, Brevinema, Serpula, Spirochaeta,* and *Treponema.* A more detailed list of organisms under the categories shown in Fig. 1 is available, if it is requested, from the authors.

The phylogenetic origins of plant mitochondria and photosynthetic plastids are assigned by looking for the phyla that show the shortest distances from these endosymbionts. By this procedure, the mitochondria are uniquely assigned to have diverged from Proteobacteria alpha subdivision, and the photosynthetic plastids in higher plants and Protoctista are assigned to have diverged from Cyanobacteria. This is consistent with the previous indications (Dickerson 1980; Margulis 1981; Yang et al. 1985; Van den Eynde et al. 1988). However, the evolutionary distances of these endosymbionts are not incorporated into the evaluation of evolutionary distances at the branching points in the earlier periods. The reason for this treatment is discussed in the next subsection.

The evolutionary distance between kingdoms is estimated by the average value of evolutionary distances calculated for all pairs of organisms taken from the kingdoms compared. This calculation leads us to the result that the evolutionary distance between Archaebacteria and Eubacteria is the shortest. However, the distances among the three kingdoms are not additive but the distance between Eukaryota and Archaebacteria is calculated to be shorter than the distance between Eukaryota and Eubacteria. Thus, the distance of Eukaryota from Prokaryota is drawn by the thick broken lines in Fig. 1 with the use of the average value of the distances between Eukaryota and Archaebacteria and between Eukaryota and Eubacteria.

The evolutionary distances calculated at the main branching points, *a* to *v,* and *M, P1,* and *P2,* shown in Fig. 1 are listed in Table 2, together with the values of each type of change probabilities. Each value of the change probabilities in this table is the one recalculated

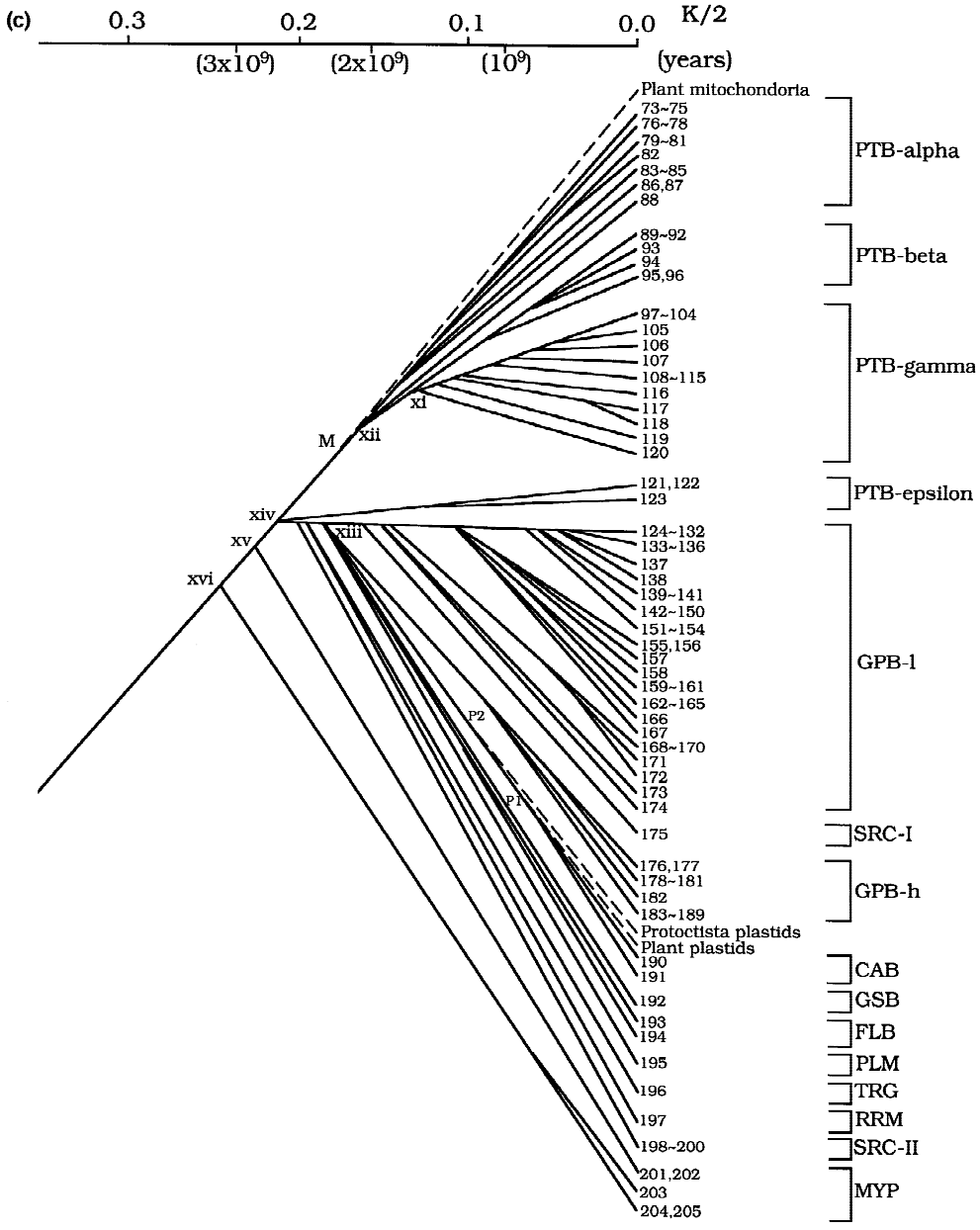

Fig. 2. Continued.

as an average of the probabilities estimated for all the pairs of rRNAs sequences, each taken from different phyla or kingdoms at their branching point. The numbers of base-pairing sites for counting base-pair changes are different depending on which phyla or kingdoms are compared, and they are listed in the sixth column in Table 2. As seen in this table, the three types of change probabilities estimated at most branching points retain the relation $P > Q > R$, reconfirming that the preferential selection of G:C and C:G base pairs has little influence on the base-pair changes observed even between Eukaryota and Prokaryota. The slight deviation seen at some branching points (*c*, *d*, *e*, *g*, and *l*) is due to the comparison of (G:C)-rich thermophilic organisms with the normal (G:C) content of organisms. At any rate, the phylogenetic tree constructed on the basis of the base-pair changes in SSU rRNAs well represents an overall feature of three kingdoms, indicating how different phyla or subdivisions have diverged in the respective kingdoms.

LSU rRNAs. In comparison with SSU rRNAs, LSU rRNAs seem to be under weaker functional constraint. The slightly higher fractions of mismatched base pairs are counted in LSU rRNAs, as shown in Table 1, and the higher base-pair change probabilities are estimated for the stem regions of LSU rRNAs. This tendency is seen in both eukaryotic and prokaryotic LSU rRNAs, and the base-pair changes in LSU rRNAs are suitable for investigating the phylogenetic relations of organisms within

Table 3. Three types of base-pair change probabilities, evolutionary distances, numbers of base-pairing sites, and standard deviations calculated at the main branching points in the tree drawn in Figs. 2a–c (the case of LSU rRNAs)

Main branching points		Change probability			Evolutionary distance ($K/2$)	Number of base-pairing sites	Standard deviation
		P	Q	R			
Eukaryota	i	0.088	0.065	0.013	0.095	503	0.011
	ii	0.111	0.050	0.041	0.119	500	0.013
	iii	0.121	0.053	0.045	0.131	507	0.013
	iv	0.146	0.047	0.045	0.146	499	0.015
	v	0.168	0.069	0.085	0.215	483	0.020
	vi	0.166	0.090	0.078	0.224	436	0.021
Archaebacteria	vii	0.088	0.094	0.037	0.131	593	0.012
	viii	0.092	0.130	0.043	0.166	607	0.014
	ix	0.162	0.138	0.074	0.263	581	0.021
	x	0.166	0.144	0.084	0.284	568	0.023
Eubacteria	xi	0.102	0.074	0.045	0.132	529	0.013
	xii	0.113	0.086	0.072	0.168	511	0.016
	xiii	0.142	0.085	0.066	0.188	538	0.017
	xiv	0.140	0.102	0.080	0.212	522	0.018
	xv	0.155	0.106	0.074	0.225	538	0.019
	xvi	0.159	0.110	0.089	0.246	525	0.021
Plant mitochondria vs Proteobacteria alpha	M	0.138	0.075	0.068	0.178	491	0.017
Plant plastids vs Cyanobacteria	P1	0.075	0.032	0.029	0.075	529	0.009
Protoctista plastids vs Cyanobacteria	P2	0.103	0.039	0.031	0.100	531	0.011
Eubacteria vs Archaebacteria		0.173	0.163	0.103	0.336	333	0.034
Eukaryota vs Archaebacteria		0.229	0.164	0.118	0.444	315	0.051
Eukaryota vs Eubacteria		0.222	0.164	0.151	0.481	300	0.054

the same kingdom. Thus, the phylogenetic trees of Eukaryota, Archaebacteria and Eubacteria are separately shown in Figs. 2a–c, respectively. Because the available nucleotide base sequence data of LSU rRNAs are still limited to those from a relatively small number of organisms, the phylogenetic clustering is started from the comparison of evolutionary distances, each calculated between individual organisms. Although most of the organisms are indicated by their genus names in the figures for simplicity, some organisms, which are assembled under the same genus in taxonomy but are clearly separated in the present analysis, are distinguished by denoting their species names.

The calculation of evolutionary distance between eukaryotic LSU rRNAs clearly shows that the distance between animals and plants is shorter than the distances of fungi from animals and plants. According to this result, the divergence of animals and plants occurred more recently than they had diverged from fungi, as shown in Fig. 2a. Moreover, some Protoctista, which correspond to those designated as the group I in Fig. 1, are also allocated to the middle position between fungi and the animal–plant group. This characteristic feature is essentially the same as that obtained by the application of the present method to 5S rRNAs of Eukaryota (Otsuka et al. 1997a), although the current method of counting the individual nucleotide base changes observed in stem and loop regions could not resolve the divergence of animals, plants and fungi (Hori and Osawa 1979). The second group of Protoctista are assigned to have diverged earlier, but this might be due to the higher content of (A:U) in the stem regions of these organisms. The nucleotide sequences of LSU rRNAs from other Protoctista such as *Didymium iridis, Physarum polycephalum, Tetrahymena pyriformis, Tetrahymena thermophila, Euglena gracilis, Giardia ardeae, G. intestinalis, G. muris,* and *Entamoeba histolytica* are also available, but they show much more irregularity and are excluded from the present analysis.

The distances among the divisions of Archaebacteria are also expanded in comparison with the result of SSU rRNAs, as shown in Fig. 2b. This figure also shows the phylogenetic problem of the Crenarchaeota posed by the fact that, based on SSU rRNAs, the distance between the Crenarchaeota and some of the Euryarchaeota is shorter than the distance between the other divisions of Euryarchaeota. *Archaeoglobus fulgidus,* whose positioning is undetermined in taxonomy, is also assigned to be an organism near Thermococcus consistently with the result of SSU rRNAs.

The evolutionary distances among different phyla of Eubacteria are also more expanded, as shown in Fig. 2c, than those calculated with SSU rRNAs, both retaining almost the same branching orders, although LSU rRNA sequences from some phyla or subdivisions, e.g., Proteobacteria subdivision delta, are not available. The Mycoplasmas, from which LSU rRNA sequences are available, are those corresponding to the species showing the longer distances in SSU rRNAs. The phylogenetic origins of mitochondria and photosynthetic plastids are also ascribed to Proteobacteria alpha subdivision and Cyanobacteria, respectively, by their shortest distances from

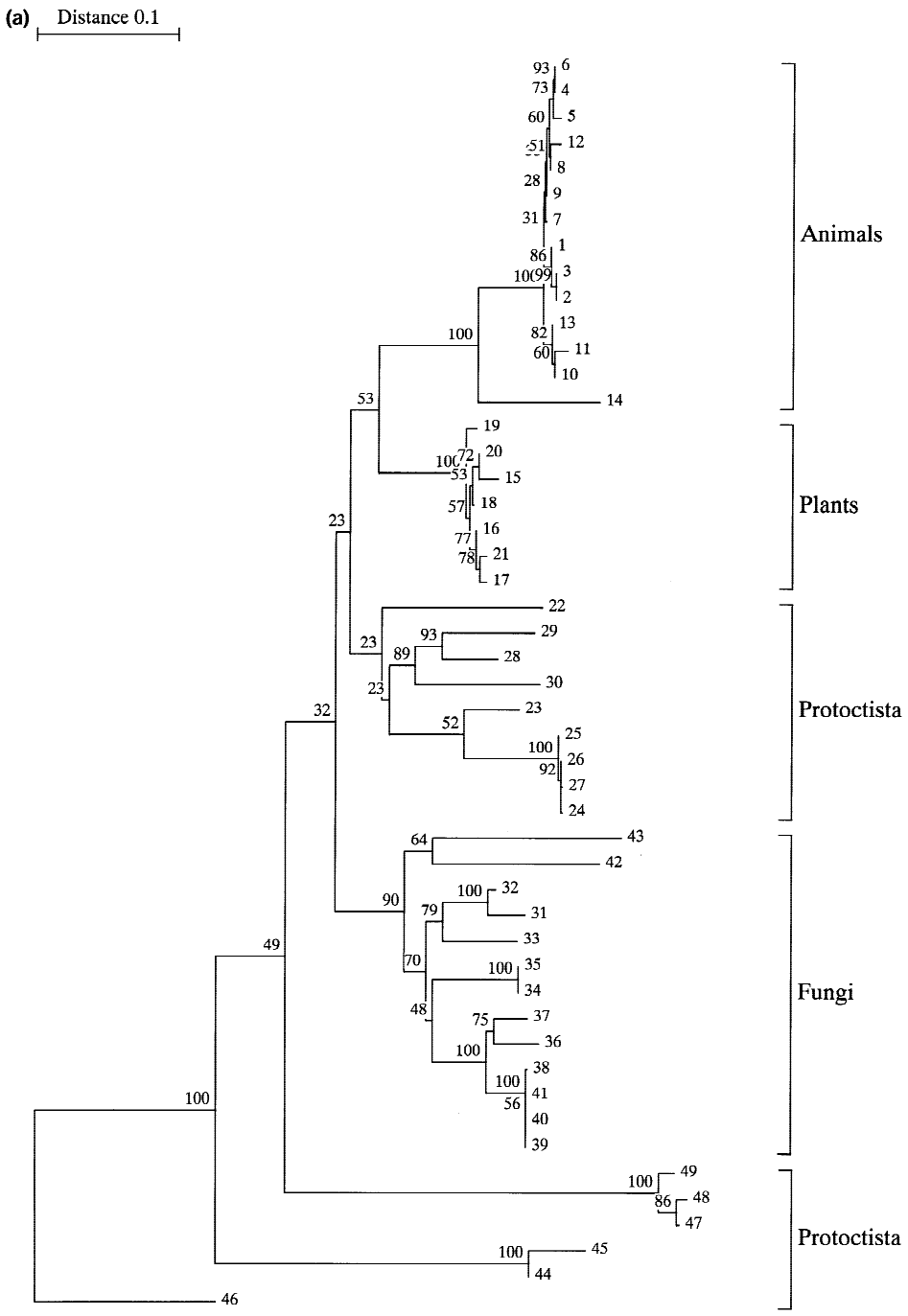

Fig. 3. The trees constructed by the neighbor-joining method. The abbreviated phylum names and the *numbers* for organisms are the same as those in the legend to Fig. 2. In this representation of tree topology, the evolutionary distance between two organisms corresponds to the sum of the branch lengths of the respective organisms, each of which is drawn horizontally. **(a)** Tree of Eukaryota constructed by the boot strap resampling. 46, *Dictyosterium* (Protoctista), is used as an outgroup. **(b)** Tree of Archaebacteria constructed by the bootstrap resampling. 142, *Listeria* (Eubacteria), is used as an outgroup. **(c)** Tree of Eubacteria constructed by the bootstrap resampling. 70, *Haloferax* (Archaebacteria), is used as an outgroup. **(d)** Tree of Eubacteria constructed without bootstrap resampling.

the above Eubacteria. Consistently with the result of SSU rRNAs, this result of LSU rRNAs also shows that the divergence of photosynthetic plastids in Protoctista from Cyanobacteria occurred prior to the endosymbiosis of the Cyanobacteria in higher plants.

The base-pair change probabilities, evolutionary distances, and the numbers of base-pairing sites calculated at the main branching points *i* to *xvi*, and *M*, *P1*, and *P2* in Figs. 2a–c are listed in Table 3. In addition to the weaker functional constraint, the variance around the expectation value of each evolutionary distance is evaluated to be much smaller in LSU rRNAs, because the

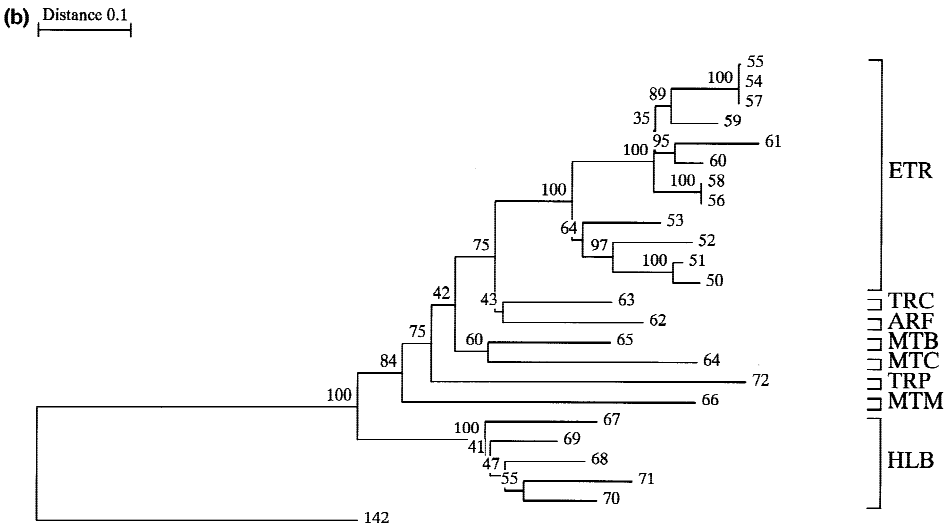

Fig. 3. Continued.

number of base-pairing sites is much larger in LSU rRNAs than that in SSU rRNAs. In this sense, the base-pair changes observed in LSU rRNAs may be more suitable for investigating the phylogeny of closely related organisms than those in SSU rRNAs, if the base-pair changes under a definite functional constraint are chosen from suitable stem regions. The evolutionary distances calculated among the three kingdoms are also listed in Table 3, showing that the evolutionary distance between Archaebacteria and Eubacteria is the shortest, the distance between Archaebacteria and Eukaryota is middle, and the distance between Eukaryota and Eubacteria is the longest. This result is consistent with the result of SSU rRNAs, although this triangle relation of evolutionary distances among the three kingdoms is probably due to a reflection of the functional constraint specific to each kingdom. At any rate, the present result does not easily accept the previous proposal that Archaebacteria are phylogenetically closer to Eukaryota than to Eubacteria (Hori et al. 1982; De Rijk 1995; Langer et al. 1995).

Comparison with the Trees Constructed by Some Other Kinds of Algorithm

If the rate of base changes is constant in all evolutionary lineages, the branch lengths in the lineages are additive and the same tree topology should be obtained by any procedure of constructing the tree. In practice, however, the distance data evaluated by the current method contain nonadditivity, and various kinds of algorithm have been proposed to reconstruct a tree from the nonadditive distance data. Thus, it may be necessary to check the additivity of our data by drawing trees with the use of other kinds of algorithm. This checking is carried out by the aid of the software package TREECON for Windows (Van de Peer and De Wachter 1994). Among the tree-constructing softwares stored in this package, the neighbor-joining (NJ) algorithm produces the tree topology most similar to that represented in the preceding subsection. This will be illustrated by the trees constructed on the basis of the base-pair changes observed in LSU rRNAs from Eukaryota, Archaebacteria and Eubacteria. According to our formulation, the bootstrap analysis is also applicable to the base-pair changes in stem regions by the current software, because the base-pair sequence can be converted into a virtual series of bases by making the corresppondence of A:U, U:A, G:C, and C:G to A, U, G, and C.

The tree of Eukaryota, which is obtained by the NJ algorithm for bootstrap resampling of base-pair changes in stem regions, is shown in Fig. 3a. As shown in this figure, the tree topology of animals, plants, fungi and Protoctista is consistent with that shown in Fig. 2a. However, it should be noted here that *Chlorella*, number 22 in the figure, is calculated to be of a shorter distance from plants than the distance between plants and animals while the other five organisms (*Prorocentrum*, *Toxoplasma*, *Hyphochytrium*, *Phytophthora*, and *Scytosiphon*) of Protoctista show longer distances from both plants and animals. By the NJ algorithm, animals and plants are first combined, and then *Chlorella* and the five organisms of Protoctista are combined by giving the smallest sum of branch lengths at this stage. In this procedure, the least-squares estimation of branch lengths makes the branch lengths of plants shorter than those in other phyla as seen in Fig. 3a, although there is no evidence that the change rate in plants is slower than those in other phyla. On the contrary, the evolutionary distance of *Chlorella* is evaluated as an average of its distances from animals and plants in Fig. 2a.

The tree shown in Fig. 3b is constructed by the bootstrap-NJ software for Archaebacteria. In this case, the divergence of divisions in Euryarchaeota also precedes the divergence of Crenarchaeota from Euryarchaeota,

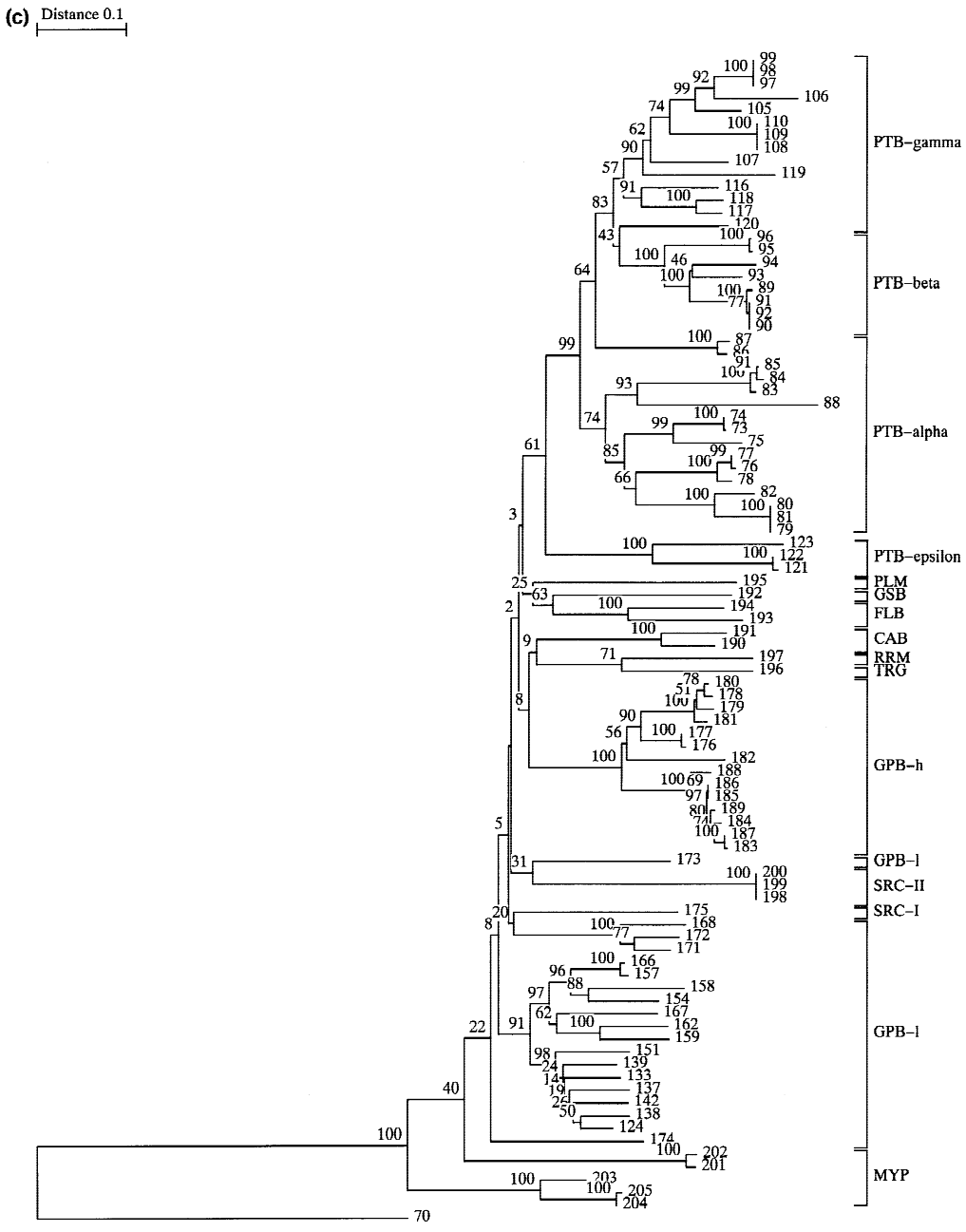

Fig. 3. Continued.

just like the feature shown in Fig. 2b. Although the topological relation of Halobacteria to Methanobacter, Methanomicrobium, and Thermoplasma is somewhat different from the relation shown in the preceding subsection, the difference in evolutionary distance between these divisions is very slight, falling in the range of standard deviation of their evolutionary distances.

The tree shown in Fig. 3c is constructed for Eubacteria by the boostrap-NJ software. In this tree, Proteobacteria are clearly separated from Gram-positive low G + C and high G + C, consistently with the result shown in Fig. 2c, although the arrangement of phyla and subdivisions is somewhat different from that in Fig. 2c. At any rate, the distances between these branching points are generally shorter and bootstrap values are lower. For reference, instead of the bootstrap resampling, the tree of Eubacteria is also drawn by applying the NJ algorithm to the distance data obtained from base-pair changes in all stem regions adopted in the present study, and it is shown in Fig. 3d. As is easily seen, this tree is also similar to the tree shown in Fig. 2c. These results indicate that different phyla of Eubacteria diverged during a relatively short period of evolution.

On the other hand, the photosynthetic plastids and plant mitochondria show somewhat troublesome behavior in the NJ algoritm. If the evolutionary distance data of photosynthetic plastids are incorporated into the data of Eubacteria at the start of the NJ procedure, the photo-

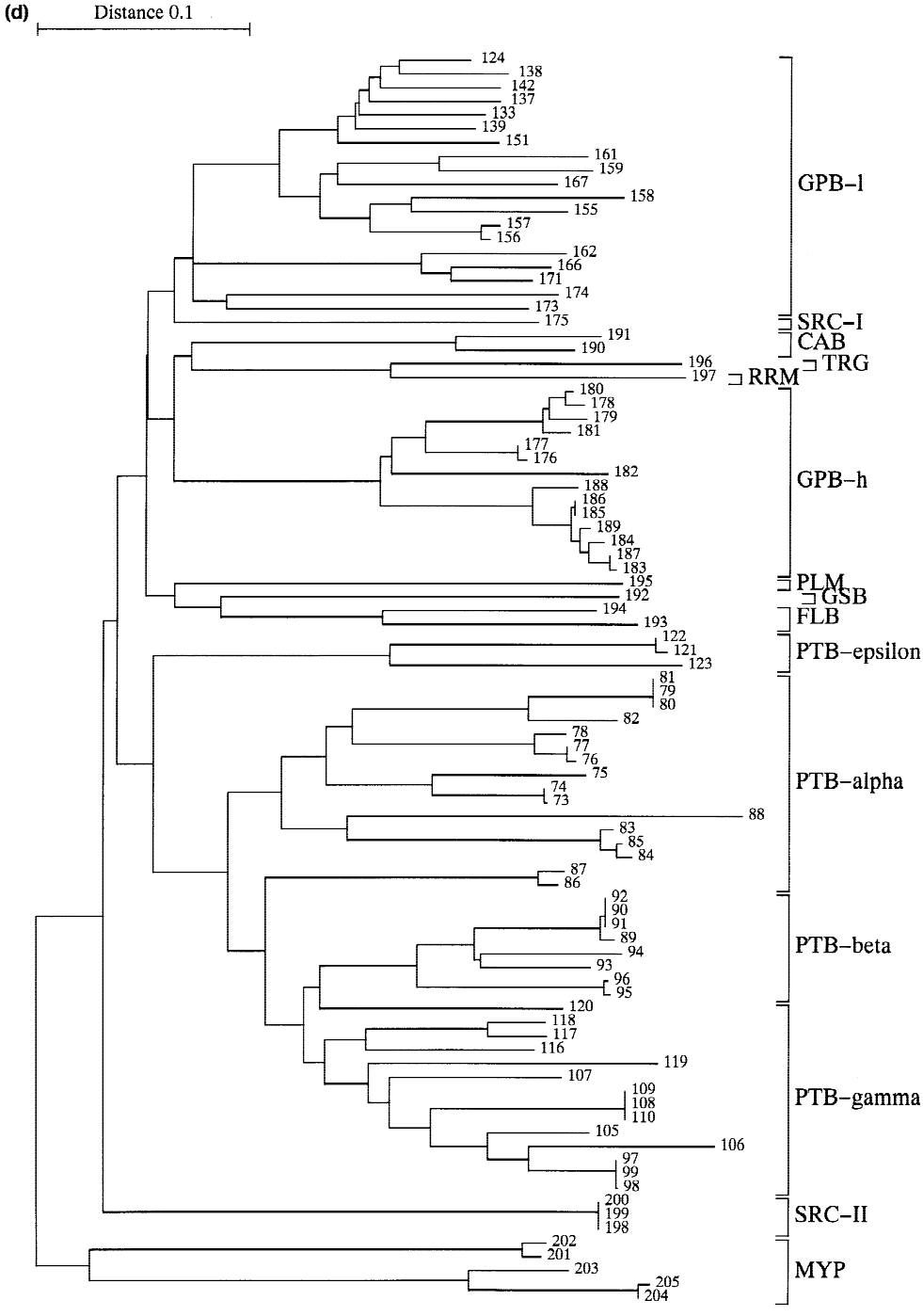

Fig. 3. Continued.

synthetic plastids are first combined with Cyanobacteria, because of the shorter distances between them, and their positions are moved to the outside of Spirochetes group II (denoted as SRC-II in the figure) and Mycoplasmas. This is due to the fact that photosynthetic plastids show a considerably longer distance from Eubacteria other than Cyanobacteria. In this case, the branch length of Cyanobacteria is expressed as much shortened as if the change rate in Cyanobacteria had been much slower than that in the other Eubacteria. However, we have no other evidence that the change rate in Cyanobacteria is very different from the rate in other Eubacteria. This is the reason why we have assigned the phylogeny of photosynthetic plastid only by their evolutionary distance from Cyanobacteria in the preceding subsection. A similar situation also occurs with the data of mitochondria. Probably, rRNAs in these endosymbionts retain the sequence properties similar to those of their original Eubacteria in some parts but have been specialized in other parts differently from those in free living organisms. For this

reason, we have treated the rRNAs in these endosymbionts separately from those in free living organisms.

Except for such endosymbionts, the present distance data of carefully selected sequences and regions are nearly additive, reproducing almost the same tree topology by different procedures, i.e., the UPGMA and the NJ algorithm. Thus, we will proceed with measuring the divergence times at the main branching points in the phylogenetic trees constructed in the preceding subsection.

Estimation of Divergence Time and Comparison with Geological Record

According to Eq. (5), the evolutionary distance is porportional to the divergence time, under the assumption of rate constancy of base-pair changes, in each type of rRNAs. The proportional constant, $(\alpha^2) + (\beta^2) + (\gamma^2)$, is evaluated to be 6.92×10^{-11} year^{-1} in SSU rRNAs and 7.92×10^{-11} year^{-1} in LSU rRNAs, if the animal-plant divergence is assumed to have occurred 1.2×10^9 years ago from the previous estimation (Dickerson 1971). With the use of these values of proportional constants, the evolutionary distances can be converted into divergence times, and the time scales thus obtained for SSU rRNAs and LSU rRNAs are plotted along the distance axes in Figs. 1 and 2, respectively. According to the time scales, the divergence of different phyla or subdivisions leading to the contemporary organisms in Eubacteria is predicted to have begun around 3×10^9 years ago, the divergence of different groups in Archaebacteria occurred slightly later and the divergence of different phyla in Eukaryota more recently. The branching point (*s* in Fig. 1 and *xiv* in Fig. 2c) of Proteobacteria, Green sulfur bacteria, and Cyanobacteria is estimated to have been 2.2×10^9 years ago by the measure of SSU rRNAs and to have been $2.4–2.7 \times 10^9$ years ago by LSU rRNAs. These estimated times are fairly well in agreement with those of the geological records showing the appearance and existence of microbial photosynthesis; the abundance of oxygen molecules in the atmosphere by 2000 million years ago (Cloud 1974), the coccoid cyanobacterium *Eosynochoccus,* which is very similar to the modern coccoid *Gloeothece,* from the 2300-million-year-old rocks of Belcher Island (Golubic and Cambell 1979), and the microbial photosynthesis functioned during the Archean about 3400 million years ago (Reimer et al. 1979). In connection with the beginning of oxygen releasing photosynthesis, the present result also predicts that the divergence of sulfate-releasing photosynthetic bacteria, Green sulfur bacteria and Purple sulfur bacteria (Proteobacteria gamma), occurred in almost the same period as Cyanobacteria diverged from other Eubacteria. This is consistent with the earliest geological records of $^{12}C/^{13}C$ and $^{34}S/^{32}S$ enrichments and of biogenic iron and sulfur deposits that are estimated to have been 2.5×10^9 years ago (Goodwin et al. 1976). In contrast to sulfate-releasing photosynthesis, sulfate respiration or chemosynthesis utilizing sulfate is seen in a wider range of organisms in both Archaebacteria and Eubacteria. Thus, a detailed comparison of chemosynthetic metabolism and of associated protein genes between these organisms may be needed to ascertain whether these sulfur chemosyntheses would have evolved independently in different lineages of Archaebacteria and Eubacteria.

The divergence time of mitochondria and Proteobacteria subdivision alpha is estimated to have been 1.6×10^9 years ago by the measure of SSU rRNAs and to 2.2×10^9 years ago by the measure of LSU rRNAs. These times are consistent with the period during which the oxygen molecules became abundunt in the atmosphere. The divergence times of photosynthetic plastids and the Cyanobacteria show an interesting feature; the chloroplasts in higher plants became endosymbionts $0.9–1.1 \times 10^9$ years ago, more recently than the divergence of higher plants from animals, but the photosynthetic plastids in Protoctista became endosymbionts $1.2–1.3 \times 10^9$ years ago, prior to the plant–animal divergence. Thus, the photosynthetic Protoctista and higher plants seem to have independently acquired cyanobacteria-like organisms as endosymbionts, although some reservation is necessary for the estimation of evolutionary distances of these endosymbionts.

In contrast to the consistency of the above examples with the geological record, the divergence times of the three kingdoms seem to be somewhat overestimated. Even the evolutionary distance between Archaebacteria and Eubacteria corresponds to the divergence occurred about 4.2×10^9 years ago, and the divergence time of Prokaryota and Eukaryota is calculated to have been more ancient than the formation of the Earth, which is believed to have begun 4600 million years ago (Cloud 1988). This discrepancy might partly arise from the kingdom-specificity still remaining in the adopted stem regions, but probably suggests another possibility that the substitution rates were faster before the appearance of oxygen-releasing photosynthesis. On the other hand, a recent attempt of estimating the divergence times by amino acid sequence comparison has reported that Cyanobacteria and Gram-positive and -negative bacteria diverged about 2 billion years ago, and that Archaebacteria and Eubacteria diverged between 3 and 4 billion years ago (Feng et al. 1997). However, this set of divergence times seems to be underestimated in comparison with the geological records. This is probably due to the situation that amino acid changes are much restricted to a narrow range by the strong effect of functional constraint on the protein molecules.

Discussion and Conclusions

In the present formulation, it is also possible to estimate the ‘‘true’’ substitution rate from the base-pair change

rate and the fraction of mismatched base-pairs. With the use of Eqs. (3-1) and (8-1), for example the ''true'' substitution rate α is calculated to be in a range from 1.73×10^{-9} to 0.77×10^{-9} year^{-1} for SSU rRNAs and from 1.58×10^{-9} to 0.79×10^{-9} year^{-1} for LSU rRNAs, because the base-pair change rate $(\alpha^2) + (\beta^2) + (\gamma^2)$ is evaluated to be 6.92 by 10^{-11} year^{-1} in SSU rRNAs and 7.92×10^{-11} year^{-1} in SSU rRNAs, as mentioned in the preceding section, and because the ratio of mismatched base-pairs G:U and U:G to the matched base-pairs G:C, C:G, A:U, and U:A mostly ranges from 0.04 to 0.09 in SSU rRNAs and from 0.05 to 0.10 in LSU rRNAs, as seen in Table 1. Interestingly, this value of substitution rate is almost equal to that estimated previously from the synonymous base changes observed in the comparison of the hemoglobin genes between mouse and human (Kimura 1980; Otsuka et al. 1997b). Moreover, this result also indicates that the base-pair changes in stem regions have occurred at a slower rate than the ''true'' substitution rate by one order of magnitude, providing the reason why the base-pair changes is suitable for resolving the phylogeny of organisms diverged a few thousand million years ago.

Why does the constancy of substitution rate hold commonly for different generation lengths of organisms? It is shown by a mathematical model in population genetics that the probability of random fixation of a selectively neutral mutant is equal to the occurrence probability of mutation during a particular generation independently of population size (Kimura 1968, 1969). This indication is adequate for claiming that observed base changes are selectively neutral but is still insufficient to explain the rate constancy. The most persuasive interpretation of the rate constancy comes from the consideration that mutations occur during the repetition of DNA damage and repair. The DNA molecule not only suffers spontaneous damage like loss of bases but also is assaulted by natural chemicals and radiation that break its backbone and chemically alter the bases. The damaged DNA molecule is continuously repaired by repair enzymes. However, such repairs may not be necessarily complete but sometimes replace nucleotide bases other than the original ones. If the replication accompanied by the proof-reading process takes place with much higher accuracy than the repair, the occurrence frequency of mutations by single-base changes only depends on the occurrence probability of misrepairs and may be proportional to the time during which the DNA molecule has been subject to radiation and chemicals. Recently, considerable homologies of DNA polymerases and repair systems are indicated between eukaryotes and prokaryotes (Bernad et al. 1989; Ito and Braithwaite 1991; Sakumi et al. 1993; Ishino et al. 1994), and the DNA molecules in most organisms would show almost the same ''true'' mutation rate, regardless of the difference in replication frequency or generation length between organisms. This consideration of rate constancy also provides a reason why the evolutionary distances between Eukaryota, Archaebacteria and Eubacteria are estimated to be longer than those expected from the geological record. Probably, the occurrence frequency of DNA damage has been influenced by the content of oxygen molecules in the atmosphere on the Earth, and the mutation rate was higher at the ancient time before the appearance of oxygen-releasing photosynthesis. The faster mutation rate in animal mitochondria might be due to the incompletion of repair systems. In addition to the effect of selection on base changes, the estimation of ''true'' mutation rate in various organisms and organelles may be an important problem in the future study of molecular evolution, getting an insight into the molecular mechanism underlying the processes of repair and replication.

References

Bernad A, Blanco L, Lazaro JM, Martin G, Salas M (1989) A conserved 3′→5′ exonuclease active site in prokaryotic and eukaryotic DNA polymerases. Cell 59:219–228

Cloud PE Jr (1974) Evolution of ecosystems. Am Sci 62:54–66

Cloud PE Jr (1988) Oasis in space: Earth history from the beginning. W.W. Norton, New York

Curtiss WC, Vournakis JN (1984) Quantitation of base substitutions in eukaryotic 5S rRNA: Selection for the maintenance of RNA secondary structure. J Mol Evol 20:351–361

De Rijk P, Van de Peer Y, Van den Broeck I, De Watchter R (1995) Evolution according to large ribosomal subunit RNA. J Mol Evol 41:366–375

De Rijk P, Van de Peer Y, De Wachter R (1997) Database on the structure of large ribosomal subunit RNA. Nucleic Acids Res 25: 117–123

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

Efron B (1982) The jackknife, the bootstrap, and other resampling plans. CBMS-NSF Regional Conference Series in Applied Mathematics No 38. Society for Industrial and Applied Mathematics, Philadelphia, PA

Egebjerg J, Larsen N, Garrett RA (1989) Structural map of 23S rRNA. In: Hill WE, Moore PB, Dahlberg A, Schlessinger D, Garrett RA, Warner JR (eds) The ribosome, structure, function, and evolution. American Society for Microbiology, pp 168–179

Faith DP (1985) Distance methods and the approximation of most-parsimonious trees. Syst Zool 34:312–325

Farris JS (1972) Estimating phylogenic trees from distance matrices. Am Nat 106:645–668

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.

Fitch WM, Margoliash E (1967) Construction of phylogenic tree. Science 155:279–284

Fitch WM (1981) A non-sequential method for constructing trees and hierarchial classification. J Mol Evol 18:30–37

Fox GE, Peckman KJ, Woese CR (1977) Comparative cataloging of 16S ribosomal ribonucleic acid: Molecular approach to prokaryotic systematics. Int J Syst Bacteriol 27:44–57

Golubic S, Campbell SE (1979) Analogous microbial forms in recent subaerial habitats and in Precambrian chert: Gloeothece coerulea (Geitler) and Eosynechococcus moorei (Hofmann). Precambrian Res 8:201–217

Goodwin AM, Monster J, Thode HG (1976) Carbon and sulfur isotope

abundance in Archean iron formations and early Precambrian life. Eco Geol 71:870–891

Hori H, Itoh T, Osawa S (1982) The phylogenic structure of the metabacteria. Zbl Bakt Hyg 1 Abt Orig C3:18–30

Hori H, Osawa S (1979) Evolutionary change in 5S RNA secondary structure and a phylogenic tree of 54 5S RNA species. Proc Natl Acad Sci USA 76:381–385

Horimoto K, Otsuka J, Kunisawa T (1989) Rapid evolutionary repair of base mispairings in stem regions of eukaryotic 5S rRNA. Protein Seq Data Anal 2:93–99

Ishino Y, Iwasaki H, Kato I, Shinagawa H (1994) Amino acid sequence motifs essential to 3′→5′ exonuclease activity of *Escherichia coli* DNA polymerase II. J Biol Chem 269:14655–14660

Ito J, Braithwaite DK (1991) Compilation and alignment of DNA polymerase sequences. Nucleic Acids Res 19:4045–4057

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

Kimura M (1969) The rate of molecular evolution considered from the standpoint of population genetics. Proc Natl Acad Sci USA 63: 1181–1188

Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Kimura M (1981) Estimation of evolutionary distances between homologous nucleotide sequences. Proc Natl Acad Sci USA 78:454–458

Kumazaki T, Hori H, Osawa S (1983) Phylogeny of Protozoa deduced from 5S rRNA sequences. J Mol Evol 91:411–419

Langer D, Hain J, Huriaux P, Zillig W (1995) Transcription in Archaea: Similarity to that in Eucarya. Proc Natl Acad Sci USA 92:5768–5772

Margoliash E (1963) Primary structure and evolution of cytochrome c. Proc Natl Acad Sci USA 50:672–679

Margulis L (1981) Symbiosis in cell evolution: Life and its environment on the early earth. W.H. Freeman, San Francisco

Otsuka J, Nakano T, Terai G (1997a) A theoretical study on the nucleotide changes under a definite functional constraint of forming stable base-pairs in the stem regions of ribosomal RNAs; Its application to the phylogeny of eukaryotes. J Theor Biol 184:171–186

Otsuka J, Fukuchi S, Kikuchi N (1997b) A theoretical method for evaluating the relative importance of positive selection and neutral drift from observed base changes. J Mol Evol 45:178–192

Raue HA, Muster W, Rutgers CA, Riet JV, Planta RJ (1989) rRNA: from structure to function. In: Hill WE, Moore PB, Dahlberg A, Schlessinger D, Garrett RA, Warner JR (eds) The ribosome, structure, function, and evolution. American Society for Microbiology, Washington DC

Reimer TO, Barghoorn ES, Margulis L (1979) Primary productivity in an early Archaen microbial ecosystem. Precambrian Res 9:93–104

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

Sakumi K, Furuichi M, Tsuzuki T, Kakuma T, Kawabata S-I, Maki H, Sekiguchi M (1993) Cloning and expression of cDNA for a human enzyme that hydrolyzes 8-oxo-dGTP, a mutagenic substrate for DNA synthesis. J Biol Chem 268:23524–23530

Sattath S, Tversky A (1977) Additive similarity trees. Psychometrica 42:319–345

Tateno Y, Nei M, Tajima F (1982) Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. J Mol Evol 18:387–404

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

Van de Peer Y, Jansen J, De Rijk P, De Wachter R (1997) Database on the structure of small ribosomal subunit RNA. Nucleic Acids Res 25:111–116

Van den Eynde H, De Baere R, De Roeck E, Van de Peer Y, Vandenberghe A, Willekens P, De Wachter R (1988) The 5S ribosomal RNA sequences of a red algal rhodoplast and a gymnosperm chloroplast: Implication for the evolution of plastids and cyanobacteria. J Mol Evol 27:126–132

Whittaker RH, Margulis L (1978) Protist classification and the kingdoms of organisms. BioSystems 10:3–18

Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221–271

Yang D, Oyaizu Y, Oyaizu H, Olsen GJ, Woese CR (1985) Mitochondrial origin. Proc Natl Acad Sci UAS 82:4443–4447

Zuckerkandl E, Pauling L (1962) Molecular disease, evolution and genetic heterogeneity. In: Kasha M and Pullman B (eds) Horizons in biochemistry. Academic Press, New York, pp 189–225