J Mol Evol (1994) 38:420–432

Journal of Molecular Evolution

DNA-Dependent RNA Polymerase Subunit B as a Tool for Phylogenetic Reconstructions: Branching Topology of the Archaeal Domain

Hans-Peter Klenk, Wolfram Zillig

Max-Planck-Institut für Biochemie, Am Klopferspitz 18, 82152 Martinsried, Germany

Received: 15 February 1993 / Revised: 8 July 1993

Abstract. The branching topology of the archaeal (archaebacterial) domain was inferred from sequence comparisons of the largest subunit (B) of DNA-dependent RNA polymerases (RNAP). Both the nucleic acid sequences of the genes coding for RNAP subunit B and the amino acid sequences of the derived gene products were used for phylogenetic reconstructions. Individual analysis of the three nucleotide positions of codons revealed significant inequalities with respect to guanosine and cytosine (GC) content and evolutionary rates. Only the nucleotides at the second codon positions were found to be unbiased by varied GC contents and sufficiently conserved for reliable phylogenetic reconstructions. A decision matrix was used for the combination of the results of distance matrix, maximum parsimony, and maximum likelihood methods. For this purpose the original results (sums of squares, steps, and logarithms of likelihoods) were transformed into comparable effective values and analyzed with methods known from the theory of statistical decisions. Phylogenetic invariants and statistical analysis with resampling techniques (bootstrap and jackknife) confirmed the preferred branching topology, which is significantly different from the topology known from phylogenetic trees based on 16S rRNA sequences. The preferred topology reconstructed by this analysis shows a common stem for the *Methanococcales* and *Methanobacteriales* and a separation of the thermophilic sulfur archaea from the methanogens and halophiles. The latter coincides with a unique phylogenetic location of a characteristic splitting event replacing the largest RNAP subunit of thermophilic sulfur archaea by two fragments in methanogens and halophiles. This topology is in good agreement with physiological and structural differences between the various archaea and demonstrates RNAP to be a suitable phylogenetic marker molecule.

Key words: DNA-dependent RNA polymerase — Archaea (Archaeobacteria) — Phylogeny — Sequence comparison — Effective values — Codon position inequality — Branching topology — Decision matrix — Statistical testing — Molecular evolutionary clock

Introduction

The phylogeny of the Archaea (archaeobacteria, Woese et al. 1990) was first determined by comparison of 16S rRNA oligonucleotide catalogs (Woese and Fox 1977). During the last decade the branching order of the archaeal domain was investigated by the application of various further phylogenetic reconstruction methods on a growing set of 16S rRNA and rDNA sequences (Tu et al. 1982; Woese 1987; Woese et al. 1991; Lake 1991; Burggraf et al. 1991). Reconstructing phylogenetic trees from nucleic acid sequence data always runs the risk of branching-order artifacts which occur when the overall compositions (i.e., the GC[1] content and the purine/pyrimidine ratio) of the sequences analyzed differ widely from one another (Woese et al. 1991). Using amino

Correspondence to: H.-P. Klenk

[1] Guanosine and cytosine

acid sequences or only the second codon positions of the genes coding for proteins could provide a chance to avoid the risk of branching-order artifacts caused by compositional differences among nucleic acids. The focus of our study was a detailed investigation of the branching topology of the archaeal domain. Our phylogenetic reconstructions are based on the rpoB gene, coding for the largest component of DNA-dependent RNA polymerase (RNAP) of the thermophilic sulfur archaea, and the homologous rpoB1 and rpoB2 genes, coding for the two RNAP components B′and B″ of the methanogens and extreme halophiles, and their derived gene products. First hints on the phylogeny of the Archaea using RNAP as marker molecule were obtained by methods not allowing quantitative treatment of the data (Schnabel et al. 1983; Gropp et al. 1986). Later, quantitative reconstructions suffered from using restricted data sets (Zillig et al. 1989; Pühler et al. 1989a; Sidow and Wilson 1990; Iwabe et al. 1991) that considered only the RNAP component sequences from *Methanobacterium thermoautotrophicum* (Berghöfer et al. 1988), *Halobacterium halobium* (Leffers et al. 1989), and *Sulfolobus acidocaldarius* (Pühler et al. 1989b) known at that time. The recently determined DNA sequences encoding the largest RNAP subunits from *Thermococcus celer* (Klenk et al. 1992a), *Thermoplasma acidophilum* (Klenk et al. 1992b), and *Methanococcus vannielii* (P. Palm, personal communication) are the basis for the reconstruction of the phylogenetic tree presented in this paper, comprising most of the orders of the archaeal domain.

Different phylogenetic reconstruction methods might be affected in different ways by potential branching-order artifacts which might occur (1) when sequences have evolved with highly different evolutionary rates, (2) when some sequences are only distantly related to others, (3) when the nucleotide compositions of the sequences analyzed differ widely from one another, and (4) when various positions in the compared sequences evolve with vastly different rates (Woese et al. 1991). In the present report we propose an analysis procedure which combines the main phylogenetic reconstruction methods (distance matrix, maximum parsimony, and maximum likelihood methods) and alleviates the potential branching-order artifacts of any single reconstruction method. The decision matrix used here allows both the determination of the most-favored branching topology by a combination of phylogenetic reconstruction methods and an estimation of the advantage of the preferred topology over alternative branching topologies.

Materials and Methods

Multiple Sequence Alignments. Amino acid sequences were first aligned with CLUSTAL (Higgins and Sharp 1989), and after visual inspection they were manually adjusted for obvious similarities and economical gapping. The DNA sequence alignment was adjusted to the amino acid alignment.

Phylogenetic Reconstructions from Amino Acid Sequences. Distance matrices were calculated from the sequence alignment by applying the unitary matrix,[2] Dayhoff's log-odds matrix (termed PAM250, Dayhoff 1978), Feng's structure-genetic matrix (Feng et al. 1985), a genetic code matrix in the version of Feng et al. (1985), and a problem-specific similarity matrix derived from ratios between detected and expected amino acid correspondences in aligned sequences of archaeal RNAP components (data not shown). Only Dayhoff's log-odds matrix was applied for calculations from a multiple sequence alignment done with CLUSTAL (Higgins and Sharp 1989) and from pairs of sequences aligned with PIRALIGN (George et al. 1986). Evolutionary distances were calculated from the effective similarity scores according to Feng et al. (1985). Following a proposal of Feng et al. (1985) the results calculated from the genetic code matrix were used directly as phylogenetic distances. Distance trees were reconstructed according to the method of Fitch and Margoliash (1967) with the programs FITCH (using the power option $p = 1$ for the denominator) and KITSCH (Felsenstein 1991), the latter assuming an evolutionary clock. The neighbor-joining technique (Saitou and Nei 1987) was carried out with the program NEIGHBOR (Felsenstein 1991). Reconstructions according to the maximum parsimony method were done with the program PROTPARS (Felsenstein 1991) under default conditions. The maximum likelihood method was applied with the PROTML program (Adachi and Hasegawa 1992).

Phylogenetic Reconstructions from Nucleic Acid Sequences. Matrices of pairwise distances between aligned DNA sequences were generated by DNADIST (Felsenstein 1991) using the one-parameter model of Jukes and Cantor (1969).[3] Distance trees were reconstructed from the distance matrices with the programs FITCH, KITSCH, and NEIGHBOR (Felsenstein 1991). Most parsimonious trees were reconstructed with the programs DNAPARS, DNAPENNY,[4] DNACOMP[5] (Felsenstein 1991) and PAUP (Swofford 1989). The latter was used for unequal weighting of the positions within the alignment (weighted characters). The log-likelihoods of phylogenetic trees were determined without the assumption of an evolutionary clock with the program DNAML (Felsenstein 1991) and under the constraint of an evolutionary clock with the program DNAMLK (Felsenstein 1991). Lake's method of phylogenetic invariants (Lake 1987) was applied with the appropriate option in PAUP (Swofford 1989).

Statistical Tests with Resampling Methods. Confidence intervals for branching topologies were estimated with the bootstrap and jackknife resampling techniques (Efron 1982) by applying the programs SEQBOOT and CONSENSE in addition to DNADIST/FITCH, NEIGHBOR, DNAPARS, DNAML, and PROTPARS as described by Felsenstein (1991).

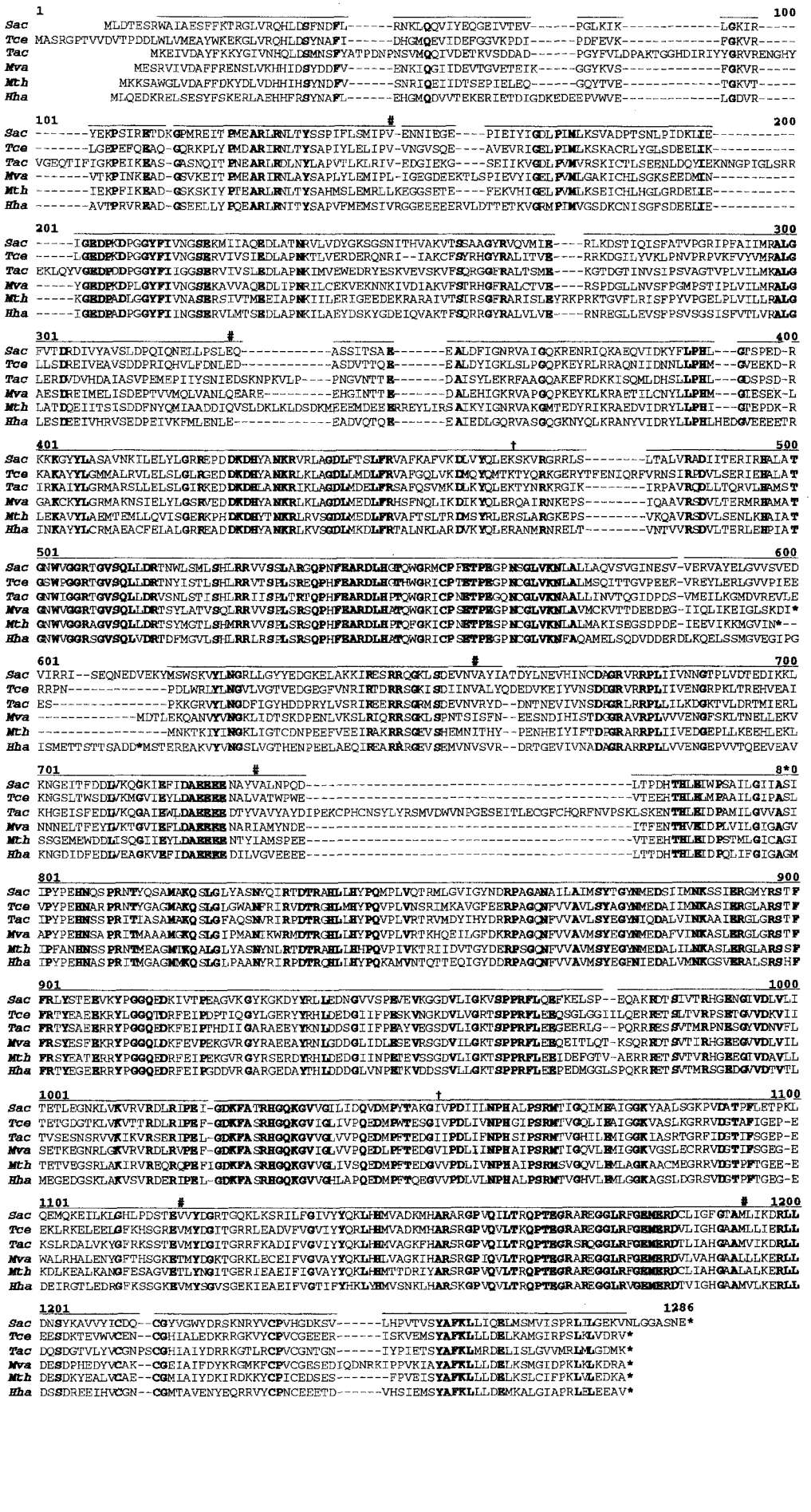
Transformation of Method-Specific Scores into Effective Values. Each of the three main phylogenetic reconstruction methods (distance matrix, maximum parsimony, and maximum likelihood method) measures the quality of a phylogenetic tree with a method-specific score (sums of squares, steps, or logarithms of likelihoods) when applied

[2] Scores only identical amino acids with 1 and all different amino acids with 0

[3] Transitions and transversions were weighted equal since all mutations at the second codon position effect an exchange of the encoded amino acid

[4] Branch and bound algorithm as search strategy

[5] Applying the compatibility method (Le Quesne 1969)



under optimality criteria. A comparison of these method-specific scores for phylogenetic branching topologies obtained by different methods requires the transformation of the method-specific scores into a common scale. When comparing a set of phylogenetic branching topologies in a decision matrix (see Results) there exists for each of the different methods at least one branching topology which shows the best score (least sum of squares or minimum number of steps or highest logarithm of likelihood), termed "optimal topology," and another branching topology which shows the worst score (highest sum of squares or highest number of steps or lowest logarithm of likelihood), termed "worst topology." The branching topologies representing the best or the worst score may differ from one reconstruction method to another reconstruction method. The method-specific scores can be transformed into the effective values (ev_{ij}) of a common scale by means of:

$$ev_{ij} = 100 \times [(\text{score}_{ij} - \text{score}_{wj})/(\text{score}_{oj} - \text{score}_{wj})]$$

in which score_{ij} is the score of branching topology$_i$ calculated with the phylogenetic reconstruction method$_j$. Score_{oj} is the score for the optimal topology and score_{wj} is the score for the worst topology found with reconstruction method$_j$. Effective values (ev_{ij}) of different branching topologies obtained by different phylogenetic reconstruction methods can be compared with each other and are thus suitable as elements of decision matrices.

Results

Amino Acid Sequence Alignment

Figure 1 shows the sequences of RNAP subunit B from *S. acidocaldarius* (Pühler et al. 1989b), *T. celer* (Klenk et al. 1992a), and *T. acidophilum* (Klenk et al. 1992b) aligned with the homologous sequences of RNAP subunits B′ and B″ from *M. thermoautotrophicum* (Berghöfer et al. 1988), *H. halobium* (Leffers et al. 1989), and *M. vannielii* (P. Palm, unpublished). The bar above the *S. acidocaldarius* sequence marks the 1,092 positions used in the phylogenetic reconstructions, comprising 96.2% of the total sequences. Unique inserts in a single sequence or positions without obvious similarities between the different sequences were excluded from the phylogenetic reconstructions. The amino acid sequence of subunit B of RNAP II from *Saccharomyces cerevisiae* (Sweetser et al. 1987, not shown in Fig. 1) was aligned with the archaeal sequences and served as an outgroup for the phylogenetic reconstructions.

Phylogenetic Inequality of the Three Codon Positions

DNA sequences were aligned in accordance with the amino acid alignment shown in Fig. 1. Most parsimonious trees were reconstructed from this data set, considering the 1,092 first or second or third nucleotides in the codons only (labeled in Fig. 1 with a bar). Significantly, the most parsimonious phylogenetic tree reconstructed from the second nucleotides of the codons requires a lower number of steps for the description than those reconstructed from the first or third positions (Fig. 2). Moreover, the branching topologies of these three trees were different from each other.

Figure 3 shows a comparison of the GC contents of the three data sets using the nucleotides occupying each of the three codon positions only. The GC content at the second codon positions shows little variation with a mean value of 37.8 ± 0.7% whereas positions 1 and 3 are more divergent. The second codon positions also show a less variable purine content (52.1 ± 1.2%) than the first (66.2 ± 2.2%) and third codon positions (51.2 ± 4.4%).

The lower evolutionary rate (Fig. 2) and the lower variability of the nucleotide composition at the second codon positions (Fig. 3) suggest the exclusive use of the latter for phylogenetic reconstructions. The nearly constant nucleotide composition of this fraction minimizes possible branching-order artifacts arising when the nucleotide composition of the sequences compared differ widely from one another (Woese et al. 1991). Moreover, the phylogenetic conservation of the nucleotides found at the second codon positions reduces the potential branching-order artifacts which might arise when sequences are only distantly related (Zuckerkandl 1987). Thus, we used only the nucleotides found at the second codon positions for the phylogenetic reconstructions from DNA sequences described below.

Possible Branching Topologies of the Archaea

For the comparison of possible branching orders in a decision matrix it is necessary to determine the relevant branching topologies. The relevant branching topologies (Fig. 4) comprise three classes of topologies, two obligatory classes, and one optional class: (1) All topologies which show the best result with one of the applied reconstruction methods (obligatory); (2) topologies which show the most unfavorable (worst) results with the same methods (obligatory); (3) topologies which are found to be the optimal solutions in phylogenetic analyses based on other marker molecules (optional). Topology A is characterized by a common stem for the two methan-

Fig. 1. Aligned amino acids sequences of archaeal RNAP subunits B, B′, and B″. *Abbreviations* are: **Sac** (*S. acidocaldarius*), **Tce** (*T. celer*), **Tac** (*T. acidophilum*), **Mva** (*M. vannielii*), **Mth** (*M. thermoautotrophicum*), and **Hha** (*H. halobium*). Amino acids occupying the 303 invariant positions were printed in boldface. Informative positions for the common stem of *M. thermoautotrophicum* and *M. vannielii* were marked by # and informative positions for the separation of thermophilic sulfur archaea from methanogens and halophiles were marked by †. Hyphens represent gaps and * indicate termiation codons.

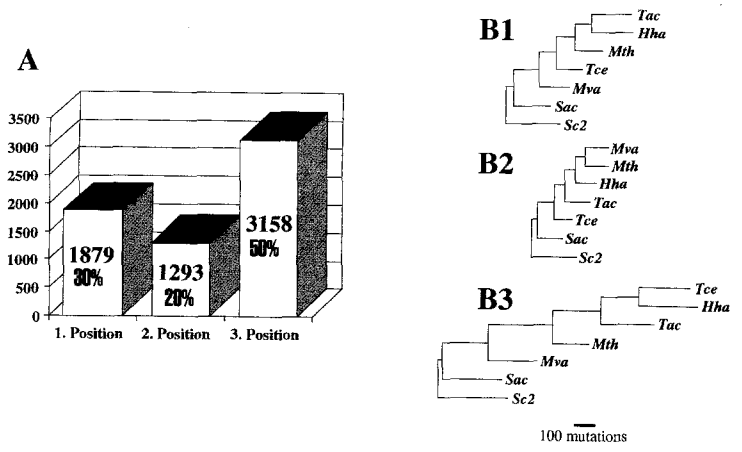


Fig. 2. Phylogenetic inequality of the three codon positions. **A** The numbers of steps necessary for the reconstruction of the most parsimonious phylogenetic trees from the first or second or third nucleotides in the codons only. The percentages give the shares of the total mutations in the rpoB genes. **B** The most parsimonious phylogenetic trees reconstructed form the first (**B1**) or second (**B2**) or third (**B3**) nucleotides of codons only.

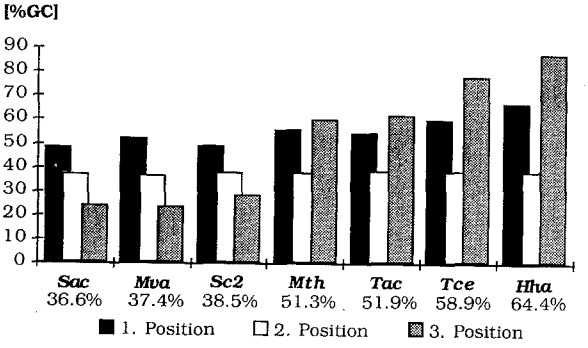


Fig. 3. Codon position inequality with respect to GC content. The *abbreviations* are the same as in Fig. 1. **Sc2** stands for the second-largest subunit of RNAP II from *S. cerevisiae*. The percentages below the abbreviations give the GC content of the genes encoding the RNAP subunits B (B′/B″). The species are ordered according to their GC content.

ogen lineages (*M. vannielii* and *M. thermoautotrophicum*) and by a separation between the thermophilic sulfur archaea (*S. acidocaldarius, T. celer,* and *T. acidophilum*) on the one hand and the two methanogens and *H. halobium* on the other hand. Topologies B, C, and G differ from topology A only in the location of the *T. acidopilum* lineage, showing an uncertainty in the placement of this lineage. Topologies E and H do not show a common stem for the two methanogen lineages. In the following the 10 topologies A to Z will be compared by a decision matrix (Table 1).

Decision Matrix

The purpose of a decision matrix is the identification of the best solution (branching topology) for a problem considered in various ways (phylogenetic reconstruction methods) by application of objective decision criteria. The quality of the 10 topologies shown in Fig. 4 is compared, considering the results of 10 variations of six phylogenetic reconstruction methods (Table 1). The numbers in the table are the effective values (or averages of effective values) determined for the 10 branching topologies (A to Z) specified in the top line, calculated from the results of the reconstruction methods specified in the first column of the table. No assessment of the superiority of one of the phylogenetic reconstruction methods (distance matrix, maximum parsimony, and maximum likelihood methods for both amino acid and nucleic acid sequences) over the others was made. The numbers under the Amino Acid Sequences heading give the effective values obtained by applying five different similarity matrices [1.1–1.5] in the protein distance matrix method. The average effective values of these five reconstructions [1] and the effective values of the other five applied reconstruction methods [2–6] are shown. These method-specific effective values were used as inputs for the two decision criteria applied for the determination of the preferred topology.

The first criterion applied for the calculation of preferred effective values for the results of all six methods is the arithmetical mean criterion, known in the theory of statistical decisions as the Laplace criterion. According to this criterion the best solution (i.e., in our case the best branching topology) is the alternative with the highest arithmetical mean. The calculation procedure is simply adding up the method-specific effective values and dividing the sum by 6. The optimal value for this criterion is 100; the worst value is 0. The second criterion is the minimax criterion, known in the theory of statistical decisions as the Savage-Niehans rule (Niehans 1948; Savage 1951). According to the minimax criterion the best solution is the alternative the most unfavorable result of which does better than the most unfavorable results of the alternatives. This decision criterion does not consider the results of all six reconstruction methods but only the most unfavorable result for each branching topology. The calculation procedure is simply searching for the lowest method-specific effective value for each topology. The best value for this criterion is 100; the worst value is 0.

The order of the preferred branching topologies according to both decision criteria is A $\gg$ G $\gg$ B $\gg$ all other topologies. Branching topology A is at the same time the only alternative for which all phylogenetic reconstructions based on distance matrix methods show no branch of length 0 and for which all programs based on

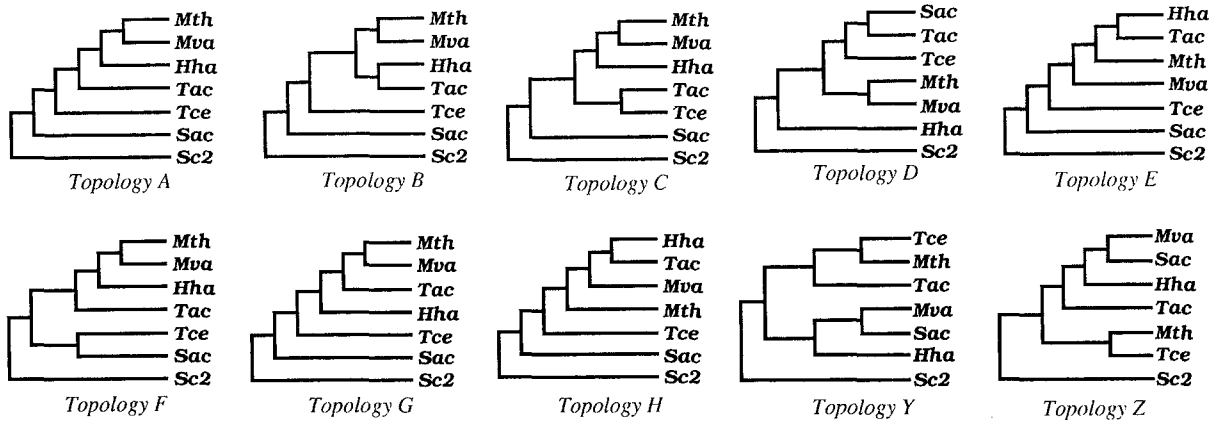


Fig. 4. Relevant branching topologies. Topologies A and B turned out to be the only optimal results when applying the spectrum of phylogenetic reconstruction programs described in Methods to the aligned sequences of the archaeal RNAP component B (or B′ and B″) or their corresponding genes. Topologies Y and Z showed the most unfavorable (worst) scores with the DNA parsimony and protein parsimony method, respectively. Topologies C and D yielded the optimal results in some of the phylogenetic reconstructions with the archaeal RNAP components A′ and A″ (our unpublished data). Assuming that closely related species can substitute for each other, topologies E to H represent phylogenetic trees described in the literature: Topology E represents the branching pattern found by Woese and Olsen (1986) and by Burggraf et al. (1991) when comparing 16S rRNA sequences; topology F represents the branching pattern found by Garrett et al. (1991) with 23S rRNA sequences (not including *T. acidophilum*); topology G represents the branching pattern of *S. acidocaldarius, T. acidophilum, M. vannielii,* and *H. halobium* in an analysis done by Cammarano et al. (1992) with sequences of elongation factors EF-2; topology H represents the branching pattern found with DNA-rRNA cross-hybridizations (Klenk et al. 1986). The *abbreviations* used are the same as in Figs. 1 and 3.

Table 1. Decision matrix[a]

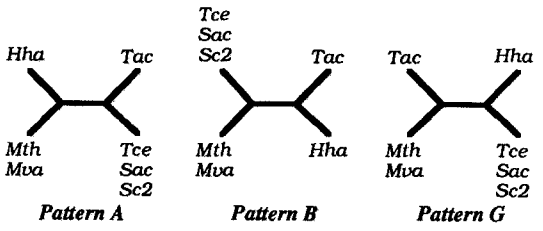
	A	B	C	D	E	F	G	H	Y	Z
Amino acid sequences										
[1.1] PAM 250 matrix	**100**	**94**	85	14	94	71	98	**94**	0	0
[1.2] Structure and genetic matrix	**100**	**100**	89	19	93	65	99	**93**	0	0
[1.3] Genetic code matrix	**100**	**98**	97	31	80	64	98	80	0	0
[1.4] Unitary matrix	**100**	**99**	96	22	88	56	98	88	0	0
[1.5] Problem specific matrix	**100**	**97**	94	22	87	63	97	87	0	0
[1] **Distance matrix method**	**100**	**98**	92	22	88	64	98	88	0	0
[2] **Maximum parsimony method**	**100**	**98**	78	62	**94**	62	**98**	**85**	3	0
[3] **Maximum likelihood method**	**92**	**100**	**79**	50	**95**	62	**88**	**85**	0	3
Nucleotide sequences										
[4] **Distance matrix method**	**100**	92	94	28	81	73	92	81	0	0
[5] **Maximum parsimony method**	**100**	80	72	57	**82**	61	**85**	69	0	2
[6] **Maximum likelihood method**	**100**	**85**	**82**	**66**	**77**	76	**90**	65	0	2
Arithmetical mean criterion	**99**	92	83	48	86	66	92	79	1	1
Minimax criterion	**92**	80	72	21	77	61	85	65	0	0
Confidence index	**67**	51	33	4	29	25	48	23	0	0

[a] In the case of the distance matrix method, values printed in boldface indicate positive branch lengths for all branches of the corresponding topology and plain figures indicate at least one branch of length 0. In the cases of maximum parsimony and maximum likelihood methods values printed in boldface show that the score for the topology is not significantly worse than the score for the best topology (statistical testing was done as proposed by Templeton 1983) and effective values printed with plain numbers indicate that the score for the topology is significantly worse than the score for the best topology. The best value determined for each of the reconstruction methods (lines 1.1–6), and the three highest values for each of the decision criteria as well as for the confidence index, are printed enlarged. The confidence index is discussed in the section about bootstrap and jackknife confidence.

maximum parsimony methods and maximum likelihood method yield only results not significantly worse than the best score (Table 1).

Other Phylogenetic Reconstruction Methods

All phylogenetic reconstructions included in the decision matrix used the sequence alignment shown in Fig. 1 or the corresponding alignment of nucleic acid sequences. Distance matrix trees calculated from pairs of sequences aligned with PIRALIGN (George et al. 1986) or from a multiple sequence alignment done with CLUSTAL (Higgins and Sharp 1989) confirmed topology A as the best branching pattern. Only reconstruction programs defining an optimality criterion as the objective function for evaluating an optimal branching



	Pattern A	Pattern B	Pattern G
Standard parsimony:	**207 (5.5)**	145 (0.5)	164 (0.0)
Evolutionary parsimony:	**81 (3.0)**	39 (1.0)	35 (2.0)

Fig. 5. Analysis of quartets. All six possible quartets containing one representative only of each of the four groups were evaluated and added up for each of the two methods. The elements of the table show the total counts (positions) favoring the corresponding branching pattern and in *brackets* how often each of the three branching patterns is the favored one. The *abbreviations* are the same as in Figs. 2 and 4. The best scores are printed in boldface.

topology and for comparing alternative topologies to one another could be used for the decision matrix. The neighbor-joining technique of Saitou and Nei (1987) does not use such an optimality criterion and therefore allows only the inference of a preferred tree but no comparison of alternative topologies to one another. The preferred trees inferred from this method with the NEIGHBOR program (Felsenstein 1991) confirmed in all cases topology A as the best branching topology. The compatibility method for DNA sequences (Le Quesne 1969) defines an optimality criterion for evaluating an optimal topology and allows comparing alternative topologies to one another. The results inferred from this method once more confirmed topology A as the optimal branching topology but were not included in the decision matrix because the range of the scores obtained for the best and the most unfavorable topology was too small for the calculation of reliable effective values. Unequal weighting of differently conserved positions within the alignment or weighting transversions twice as high as transitions in the DNA parsimony method (both with the appropriate option in PAUP, Swofford 1989) did not change the optimal topology determined without weighting. The results of programs assuming an evolutionary clock will be discussed separately.

Phylogenetic Invariants

The three most-preferred branching topologies (A, B, and G) were also compared in an analysis of quartets by the standard parsimony method and the evolutionary parsimony method (Lake 1987) using the option offered by PAUP (Swofford 1989). The nucleic acid sequences of the seven taxa included in the analysis were distributed into four groups joined in these most-preferred branching topologies: Group 1 containing *S. cerevisiae, S. acidocaldarius,* and *T. celer;* group 2 containing *M. vannielii* and *M. thermoautotrophicum,* and groups 3 and 4 containing only *T. acidophilum* and *H. halobium,* respectively. The patterns A, B, and G drawn in the top line of Fig. 5 show the three possible distributions of the four groups into quartets, each of which represents one of the three most-probable branching topologies (A, B, and G). The standard parsimony method considers the four nucleotides (A, C, G, and T) as independent character states whereas the evolutionary parsimony counts branching patterns supporting and contradicting the phylogenetic invariants (Lake 1987). Pattern A, representing branching topology A, is in both cases the favored pattern and thus clearly supports the result found by analysis of the decision matrix.

Phylogenetic Tree of the Archaea

Figure 6 depicts the phylogenetic tree reconstructed with the DNA maximum likelihood method. The tree shows the most-preferred branching topology A and remarkably short lineages leading to *S. acidocaldarius* and *T. celer.* These lineages (measured from the root of the Archaea given by the outgroup) are the shortest also in the trees obtained by the other phylogenetic reconstruction methods considered above. Therefore the phylogenetic tree shown in Fig. 6 is the best representation both of the branching order and the lengths of the branches describing the evolution of RNAP subunit B (B′ and B″) of the Archaea.

Assumption of a Molecular Evolutionary Clock

All phylogenetic reconstructions described in the previous sections were calculated without the assumption of an evolutionary clock. Both the distance matrix method and the maximum likelihood method (using the DNAMLK program) were also applied under the constraint that branch lengths must be consistent with a molecular clock. Reconstructions of phylogenetic trees from distance matrices (using the KITSCH program) determined either from DNA sequences (using the second nucleotides of the codons only) or from amino acid sequences according to the methods described above

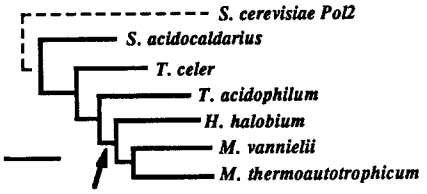


Fig. 6. DNA maximum likelihood tree. The tree is reconstructed with the program DNAML (Felsenstein 1991) from the second nucleotides in the codons only. The *arrow* indicates the phylogenetic location of the splitting event separating the two halves (B′/B″) of the largest RNAP component, which in methanogens and halophiles replace component B of the thermophilic sulfur archaea, and the second-largest component both of *Eucarya* and *Bacteria*, respectively. The *bar* represents 100 expected nucleotide substitutions. The *broken lineage* indicates the outgroup.

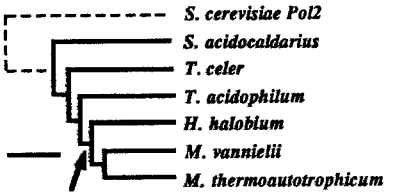


Fig. 7. DNA maximum likelihood tree assuming of an evolutionary clock. The tree is reconstructed with the program DNAMLK (Felsenstein 1991) from the second nucleotides in the codons only. The *arrow* indicates the location of the splitting event which separated the two halves (B′/B″) of RNAP subunit B. The *bar* represents 100 expected nucleotide substitutions. The *broken lineage* indicates the outgroup.

yielded in all cases branching orders which differed from topology A only in the location of the *T. celer* lineage. Without the constraints of a molecular evolutionary clock the *T. celer* lineage is located between the lineages of *S. acidocaldarius* and *T. acidophilum*. Assuming an evolutionary clock *T. celer* was found with equal probability at one of three locations: (1) Between the lineages of *T. acidophilum* and *H. halobium*, (2) between the lineages of *H. halobium* and the common stem of the two methanogens, and (3) together with the lineage of *M. vannielii*. The constraints of the assumed evolutionary clock thus move the remarkably short lineage of *T. celer* away from its position next to the root of the Archaea up in the tree toward or even into the *M. vannielii* lineage.

The best molecular clocks are those depending on nucleotide positions least affected by regional or genomic evolutionary changes in GC content (Zuckerkandl 1987). According to a proposal of E. Zuckerkandl the best clocks should be obtained with sufficiently large sets of second codon positions. The fraction comprising the nucleotides found at the second codon positions of our alignment represents one of these suitable data sets. Phylogenetic reconstructions with the maximum likelihood method for DNA sequences yield the same optimal branching topology either with (Fig. 7) or without (Fig. 6) the assumption of an evolutionary clock. This allows a reasonably legitimate statistical test for the assumption of a molecular clock (Felsenstein 1991). With the program DNAML (not assuming a clock) the logarithm of likelihood (−6,518.6) has been calculated by taking all 11 branch lengths into consideration. With the program DNAMLK (assuming a clock) the logarithm of likelihood (−6,541.6) has been calculated considering only six branching times (in effect six branch lengths). The likelihood ratio test (Felsenstein 1991), which is in fact a χ^2 test, is performed by comparing the double of the difference between the estimated logarithms of likelihood from DNAML and DNAMLK (χ^2*) with a χ^2 distribution with 5 degrees of freedom[6] (Felsenstein 1991). Assuming a significance level of $\alpha = 0.001$ (99.9% certainty of the tested hypothesis) and 5 degrees of freedom, χ^2 (0.999; 5)[7] is 20.41. The double of the difference between the two logarithms of likelihood (χ^2*) is 46. The assumption of an evolutionary clock can thus be rejected since the logarithm of likelihood is significantly increased ($\chi^2* > \chi^2$) by allowing all 11 branch lengths to be estimated instead of just six branching times. To summarize: This analysis demonstrates clearly that the RNAP B subunit genes do not appear to be diverging at a constant rate or with a constant clock speed.

Bootstrap and Jackknife Confidence of the Branching Topologies

The statistical confidence of the existence of sets of species was tested with bootstrap and jackknife resampling methods (Efron 1982) by the DNA and protein parsimony, the DNA maximum likelihood, and the DNA and protein distance matrix programs as described in the PHYLIP package (Felsenstein 1991). The programs were used with 200 replications each. Bootstrap values calculated with different reconstruction programs can be spread over a wide range (Table 2). Some examples: (*Hha, Tce*) from 0 to 15, (*Mva, Mth, Tac*) from 1 to 22, and (*Mth, Hha*) from 13 to 29. The calculation of average bootstrap values from values found with different phylogenetic reconstruction methods is not a commonly used method but the average values (given in the last column) alleviate the potential artifacts of any single reconstruction method. The first four sets of

[6] Reconstructions of phylogenetic trees from seven species not assuming an evolutionary clock possess 5 additional degrees of freedom when compared with reconstructions under the constraint of an evolutionary clock. Each degree of freedom represents the possibility to calculate the length of a branch independent from the lengths of other branches

[7] (0.999; 5) means (1-α; five degrees of freedom)

species, which together constitute the branching topology A, show by far the highest average bootstrap values and their method-specific bootstrap values are in any case higher than all values for all other sets of species. The geometrical mean of the average bootstrap values from the four sets of species constituting any of the branching topologies can be considered as an index of support reflecting the confidence of the branching topology. These values are shown for the 10 branching topologies A to Z in the line termed "confidence index" in Table 1. The rank of the branching topologies according to this criterion ($A \gg B > G \gg$ all other topologies) supports once more the result of the decision matrix. Jackknife values were always in the same range as bootstrap values when calculated with the same reconstruction program (data not shown). All sets of species with an average bootstrap value of at least 3% were represented in Table 2. It can therefore be concluded that any one of the 935 possible branching topologies (seven species can be organized in 945 unrooted branching topologies) not compared in the decision matrix (Table 1) is supported less by the confidence index than is the most preferred branching topology A.

Discussion

The Combined Phylogenetic Reconstruction Method

The strategy used for searching the most-favored phylogenetic tree and for the analysis of its advantage compared with the other probable trees was a combination of different phylogenetic reconstruction methods and methods used in the theory of statistical decisions and with statistical resampling tests. The four steps for the analysis procedure are: (1) Determination of the relevant branching topologies, (2) construction of a decision matrix for the determination of the preferred topologies, (3) confirmation of the favored branching topology in an analysis of quartets including the "evolutionary parsimony" method (Lake 1987), and (4) determination of the statistical confidence of the branching topologies with resampling methods.

It is not unusual to find different branching topologies when analyzing the phylogeny of a marker molecule with different phylogenetic reconstruction methods or when analyzing different phylogenetic marker molecules from the same set of species. For the comparison of several branching topologies in a decision matrix the method-specific scores have to be transformed into effective values. The relevant topologies for this transformation are all method-specific optimal topologies and at least some of the least-preferred topologies found by the same methods. Since not all of the applied reconstruction programs allow the determination of a least-preferred topology, we used only the least parsimonious topologies found by DNA and protein maximum parsimony methods. They proved to be also the least-preferred topologies when compared with other reconstruction methods. Topologies determined in the analysis of other marker molecules can be included in the comparative analysis without disturbing the selection of the most-favored phylogenetic tree.

A decision matrix can be used for the identification of the most-preferred branching topology within the group of conceivable candidates for the "true" phylogenetic branching topology. The most-preferred branching topology should give a satisfying description of the evolution as viewed by various phylogenetic reconstruction methods. Studies trying to determine the efficiency of different phylogenetic reconstruction methods in obtaining the correct tree lead to contradictory results (Saitou 1988; Hasegawa et al. 1991). In selecting the methods used in the decision matrix we neither assumed a superiority of one of the reconstruction methods over the others nor an inequality of results obtained from DNA or amino acid sequences. The distance matrix method as well as the maximum parsimony method and the maximum likelihood method were assumed to be suitable and equivalent methods for the reconstruction of phylogenetic trees from both DNA and amino acid sequences. The two criteria which were applied for the determination of the preferred effective values for all six methods allow not only the determination of the preferred branching topology but also an estimation of the confidence of this decision. The latter is important since consistency among different methods is a poor guide to statistical significance (Felsenstein 1991). Example: Let one branching topology (I) be the optimal solution for all applied reconstruction programs. Let another branching topology (II) be insignificantly worse than topology I for all applied reconstruction programs. The simple ratio between the numbers of favored topologies is all for topology I against nothing for topology II. The ratio of the overall effective values might be 100 for topology I against 99 for topology II, indicating that topology I is not confidently superior to topology II. Therefore the estimation of the confidence of the preferred tree is an essential part of the analysis.

The branching topologies with the best values for the two decision criteria were compared in an analysis of quartets including the method of phylogenetic invariants (Lake 1987). This third step in the combined phylogenetic reconstruction method should confirm the preferred branching topology of the decision matrix or help to discriminate between topologies with insignificant differences in their preferred effective values. At the same time Lake's method of phylogenetic invariants offers the possibility to test for homoplasies—i.e., systematic errors which tend to join diverged sequences as sister groups in unrooted phylogenetic trees. The "evolutionary parsimony method" (Lake 1987) has been

Table 2. Statistical testing of sets of species[a]

Sets of species	Topology										DNA sequences			Amino acid sequences		Average bootstrap value
	A	B	C	D	E	F	G	H	Y	Z	[MP]	[ML]	[DM]	[MP]	[DM]	
(*Mva, Mth, Hha, Tac, Tce*)	X	X	X		X		X	X			85	81	91	98	99	90
(*Mva, Mth, Hha, Tac*)	X	X			X	X	X	X			85	69	62	82	84	76
(*Mva, Mth, Hha*)	X		X			X					56	48	51	43	49	49
(*Mva, Mth*)	X	X	X	X		X	X				51	60	56	56	73	59
(*Hha, Tac*)		X			X			X			5	3	4	29	40	16
(*Mva, Mth, Tac*)							X				4	17	20	22	1	13
(*Mth, Hha, Tac*)					X						19	4	1	8	1	6
(*Tac, Tce*)			X								3	4	11	2	4	5
(*Tac, Sac*)				X							5	7	4	2	0	3
(*Mva, Hha, Tac*)								X			4	1	5	1	3	3
(*Tce, Sac*)						X					1	5	2	0	2	2
(*Tac, Tce, Sac*)				X							2	6	2	0	0	2
(*Mva, Mth, Tac, Tce, Sac*)				X							1	2	2	0	0	1
(*Mva, Hha, Tac, Sac*)										X	2	1	0	0	0	1
(*Mva, Sac*)									X	X	0	1	0	0	0	0
(*Mth, Tce*)									X	X	0	0	0	0	0	0
(*Mth, Tac, Tce*)									X		0	0	0	0	0	0
(*Mva, Hha, Sac*)									X	X	0	0	0	0	0	0
(*Mth, Hha*)											29	17	19	23	13	20
(*Mva, Tac*)											11	19	19	5	5	12
(*Hha, Tac, Tce*)											3	13	3	11	9	8
(*Hha, Tce*)											1	15	8	0	3	5
(*Mva, Mth, Hha, Tce*)											4	5	10	6	0	5
(*Mva, Mth, Hha, Tac, Sac*)											12	10	3	3	0	5
(*Mva, Hha, Tac, Tce*)											0	3	0	10	10	4
(*Mva, Hha*)											0	2	8	1	4	3

[a] Bootstrap values calculated with maximum parsimony [MP], maximum likelihood [ML], and distance matrix [DM] methods were percent values. The first column shows the sets of species to which the bootstrap values in the same line belong. In the second column the sets of species constituting the 10 branching topologies (A and Z) are marked. Each branching topology comprising seven species can be described by four sets of species. The first four sets of species constitute branching topology A; the next 14 sets of species (between the dotted lines) constitute branching topologies B to Z. The last eight sets of species comprise the most probable sets of species not considered in topologies A to Z. The abbreviations used for the species are the same as in Fig. 4.

suggested to serve this purpose by being a rate-independent technique for the analysis of nucleic acid sequences. In our analysis the result of the evolutionary parsimony method clearly supports topology A and contradicts the influence of homoplasies.

The last steps of the analysis were the estimation of the confidence of the joining of species in sets within the branching topologies (Table 2) and the determination of confidence indices for entire branching topologies (Tables 1 and 2). The frequency with which a set of species appeared in replicate trees was interpreted as an index of support for the common stem separating this set of species from all other species. The calculation of average bootstrap values from values determined with different phylogenetic reconstruction methods avoids potential preferences of any single reconstruction method. The rank of the branching topologies according to the confidence index (Table 1) clearly shows once more topology A to be the optimal branching order. Since all sets of species with average bootstrap values of at least 3% were analyzed (Table 2) it can be concluded that no other branching topology is more favored by the confidence index than topology A.

The higher phylogenetic conservation (Fig. 2) and the lower variability of the nucleotide composition (Fig. 3) of the DNA fraction comprising the nucleotides found at the second codon positions led to the exclusion of the more variable codon positions 1 and 3 from phylogenetic reconstructions. The observation that the second position requires a significantly lower number of steps for the reconstruction of the most parsimonious phylogenetic tree than the first and third positions is in line with the detection of different fixation rates for different codon positions in a number of other genes (Kimura 1983). The observation that the GC content for the nucleotides found at the second codon positions shows less variation than for those found at the first and the third codon positions (Fig. 3) is in accordance with the

observation that the GC content changes with the highest rate in the third positions and with the lowest rate in the second positions (Bernardi and Bernardi 1986).

Branching Topology of the Archaea

This analysis shows topology A to be the most-preferred branching topology for the archaeal domain. The deepest bifurcation within the Archaea divides *Sulfolobus* (the only included representative of the Crenarchaeota) from the five Euryarchaeota. The only branching topology (D) which does not show this fundamental bifurcation within the Archaea was found to be the least-preferred of the "real" topologies compared by the decision matrix (not considering topologies Y and Z, which were only included for standardization). A common stem for *M. vannielii* and *M. thermoautotrophicum* is characteristic not only for the preferred branching topology (A) but also for the second (G)- and third (B)-best solutions of the decision matrix. Branching topology E shows the best values within the group of topologies without a common stem for the *Methanococcales* and *Methanobacteriales* but is significantly less preferred than topology A. Topology A locates the event that replaced the RNAP subunit B of the thermophilic sulfur archaea (*S. acidocaldarius, T. celer,* and *T. acidophilum*) by components B′and B″ of all methanogens and halophiles in the branch separating the thermophilic sulfur archaea from the two methanogens and *H. halobium.* Not only this gene split but also the appearance of methanogenesis and the fading of sulfur-dependent modes of life are most simply explained by topology A.

The sets of species included in studies comparing different marker molecules overlap in most cases only partially. Nevertheless, for the comprehension of the evolution of species it is necessary to compare such results for as many molecular markers as possible. We compared our result with some of the branching patterns shown in former studies (topologies E to H in Fig. 4 and Table 1), assuming that closely related species can substitute for each other. According to the results of the decision matrix all of these branching topologies are less preferred than topology A. Branching topology A is not only the most-preferred topology for the large RNAP subunits but also a possible branching topology for most of the marker molecules used before (16S rRNA: Burggraf et al. 1991; 23S rRNA: Woese et al. 1991; EF-2: Cammarano et al. 1992).

Our claim of significance for the superiority of topology A over the other branching topologies mainly depends on the preferred effective values calculated in the decision matrix. Since this is the first application of a decision matrix in a combined phylogenetic reconstruction method it is not absolutely clear whether this method provides a rigorous test for the differentiation between the phylogenetic qualities of different branching topologies. Further studies on the features of this method should follow.

None of the subsets of topology A received more than 90% support of the bootstrap results in Table 2. According to the rule that 95% bootstrap frequency is necessary for statistically significant trees, this result does not support our claim of significance of topology A. Nevertheless, Table 2 allows at least the conclusion that no other branching topology is more supported by bootstrap frequencies than topology A.

In much the same way as the splitting event of the B subunit, the presence of shared deletions-insertions in the aligned sequences can be used to support or refute branching topologies. Some of the deletions-insertions in the alignment (Fig. 1) are too complicated to assign to a specific topology (positions 328–335, 601–617, and 975/976); others support the preferred topology (positions 662/663). Nevertheless, the deletions-insertions at positions 248 and 256–259 clearly not support topology A. They support topologies which join *H. halobium* either with *M. thermoautotrophicum* or with *M. vannielii;* both of them cannot be considered as promising candidates for the overall preferred topology (Tables 1 and 2).

Component B of RNAP as a Phylogenetic Marker Molecule

A molecule has to fulfill certain criteria to be a useful chronometer for evolution: (1) It has to occur in all compared organisms; (2) its rate of change has to be slow enough to compare all species; (3) its size should be large enough to guarantee high confidence in the phylogenetic results. Subunit B of RNAP fulfills all these features: (1) All archaea contain a multicomponent RNAP and subunit B (either complete or split) is present in each of these enzymes; (2) subunit B is one of the important components of the transcription apparatus and therefore highly constrained to evolve with a reasonably slow rate; (3) the 1,092 positions of the alignment used for the phylogenetic reconstructions represent one of the longest sequence data sets used for the investigation of archaeal phylogenies [23S rRNAs: 1,417 positions (Garrett et al. 1991); EF-2: 840 positions (Cammarano et al. 1992); 16S rRNAs: 830 positions (Woese and Olsen 1986); EF-1α: 409 positions (Creti et al. 1991)].

Sequences of RNAP components have one important disadvantage when compared with sequences of rRNAs—the most-used molecular chronometers: They cannot be sequenced directly and therefore rapidly. This drawback is compensated for by three advantages: (1) Amino acid sequences can be aligned more easily than nucleic acid sequences; (2) the amino acid sequences can be used for phylogenetic comparisons in addition to

the nucleic acid sequences, allowing a more extensive spectrum of reconstruction methods; (3) at the nucleotide level the different features of the three codon positions with respect to nucleotide composition and evolutionary rate allow a restriction of phylogenetic reconstructions on the fraction which minimizes the danger of branching-order artifacts.

Branching topologies derived from sequence analysis have to be considered as hypotheses, which have to be tested and either strengthened or rejected on the basis of other kinds of data (Jensen 1985). The subunit pattern of archaeal RNAPs provides such a different kind of data. A suitable description of the phylogeny of the Archaea should localize the splitting event of the RNAP component B at a single position within the phylogenetic tree. The favored branching topology A in this study clearly pinpoints this characteristic event of the archaeal evolution.

Acknowledgments. We thank P. Palm for contributing his computer programs and the sequences of rpoB1 and rpoB2 from *M. vannielii.*

References

Adachi J, Hasegawa M (1992) MOLPHY: programs for molecular phylogenetics, I. PROTML: maximum likelihood inference of protein phylogeny. In: Komazawa T, Nakamura T, Takagi H, Tamura Y-H, Tanabe K, Ohsumi N (eds) Computer science monographs no. 27. The Institute of Statistical Mathematics, Minato-ku, Tokyo, Japan.

Berghoefer B, Kroeckel L, Koertner C, Truss M, Schallenberg J, Klein A (1988) Relatedness of archaebacterial RNA polymerase core subunits to their eubacterial and eukaryotic equivalents. Nucleic Acids Res 16:8113–8128

Bernardi G, Bernardi G (1986) Compositional constraints and genome evolution. J Mol Evol 24:1–11

Burggraf S, Stetter KO, Rouviere P, Woese CR (1991) *Methanopyrus kandleri:* an archaeal methanogen unrelated to all other known methanogens. System Appl Microbiol 14:346–351

Cammarano P, Palm P, Creti R, Ceccarelli E, Sanangelantoni AM, Tiboni O (1992) Early evolutionary relationships among known life firms inferred from elongation factor EF-2/EF-G sequences: phylogenetic coherence and structure of the archaeal domain. J Mol Evol 34:396–405

Creti R, Citarella F, Tiboni O, Sanangelantoni A, Palm P, Cammarano P (1991) Nucleotide sequence of a DNA region comprising the gene for elongation factor 1α (EF-1α) from the ultrathermophilic archaeote *Pyrococcus woesei:* phylogenetic implications. J Mol Evol 33:332–342

Dayhoff MO (1978) A model of evolutionary change in proteins. Matrices for detecting distant relationships. In: Dayhoff MO (ed) Atlas of protein sequence and structure, vol 5, suppl 3. Natl Biomed Res Found, Washington, DC, pp 345–358

Efron B (1982) The jackknife, the bootstrap, and other resampling plans. CBMS-NSF regional conference series in applied mathematics, monograph 38. Society of Industrial and Applied Mathematics, Philadelphia

Felsenstein J (1991). PHYLIP users manual V3.4, University of Washington, Seattle

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

Fitch WM, Margoliash E (1967) Construction of phylogenetic trees. Science 15:279–284

Garrett RA, Dalgaard J, Larsen N, Kjems J, Mankin AS (1991) Archaeal rRNA operons. Trends Biochem Sci 16:22–26

George DG, Barker WC, Hunt LT (1986) The protein identification resource. Nucleic Acids Res 14:11–15

Gropp F, Reiter W-D, Sentenac A, Zillig W, Schnabel R, Thomm M, Stetter KO (1986) Homologies of components of DNA-dependent RNA polymerases of archaebacteria, eukaryotes and eubacteria. System Appl Microbiol 7:95–101

Hasegawa M, Kishino H, Saitou N (1991) On the maximum likelihood method in molecular phylogenies. J Mol Evol 32:443–445

Higgins DG, Sharp PM (1989) Fast and sensitive multiple sequence alignments on a microcomputer. CAIBIOS 5:151–153

Iwabe N, Kuma K-i, Kishino H, Hasegawa M, Miyata T (1991) Evolution of RNA polymerases and branching patterns of the three major groups of archaeabacteria. J Mol Evol 32:70–78

Jensen RA (1985) Biochemical pathways in procaryotes can be traced backwards through evolutionary time. Mol Biol Evol 2:87–120

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism, vol 3. Academic Press, New York, pp 21–132

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England

Klenk H-P, Haas B, Schwass V, Zillig W (1986) Hybridization homology: a new parameter for the analysis of phylogenetic relations, demonstrated with the urkingdom of the archaebacteria. J Mol Evol 24:167–173

Klenk H-P, Schwass V, Zillig W (1992a) Nucleotide sequence of the genes encoding the three largest subunits of the DNA-dependent RNA polymerase from the archaeum *Thermococcus celer.* Nucleic Acids Res 20:4659

Klenk H-P, Renner O, Schwass V, Zillig W (1992b) Nucleotide sequence of the genes encoding the subunits H, B, A′ and A″ of the DNA-dependent RNA polymerase and the initiator tRNA from *Thermoplasma acidophilum.* Nucleic Acids Res 20:5226

Lake JA (1987) Rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. Mol Biol Evol 4:167–191

Lake JA (1991) Tracing origins with molecular sequences: metazoan and eukaryotic beginnings. Trends Biochem Sci 16:46–50

Leffers H, Gropp F, Lottspeich F, Zillig W, Garrett RA (1989) Sequence, organization, transcription and evolution of RNA polymerase subunit genes from the archaebacterial extreme halophiles *Halobacterium halobium* and *Halococcus morrhue.* J Mol Biol 206:1–17

Le Quesne WJ (1969) A method of selection of characters in numerical taxonomy. Syst Zool 18:201–205

Niehans J (1948) Zur Preisbildung bei ungewissen Erwartungen. Schweizerische Zeitschrift für Volkswirtschaft und Statistik 84:433–456

Pühler G, Leffers H, Gropp F, Palm P, Klenk H-P, Lottspeich F, Garrett RA, Zillig W (1989) Archaebacterial DNA-dependent RNA polymerases testify to the evolution of eukaryotic nuclear genome. Proc Natl Acad Sci USA 86:4569–4573

Pühler G, Lottspeich F, Zillig W (1989b) Organization and nucleotide sequence of the genes encoding the large subunits A, B and C of the DNA-dependent RNA polymerase of the archaebacterium *Sulfolobus acidocaldarius.* Nucleic Acids Res 17:4517–4534

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

Saitou N (1988) Property and efficiency of the maximum likelihood method for molecular phylogeny. J Mol Evol 27:261–273

Savage LJ (1951) The theory of statistical decisions. Am Statist Assoc 46:55–67

Schnabel R, Thomm M, Gerardy-Schahn R, Zillig W, Stetter KO,

Huet J (1983) Structural homology between different archaebacterial DNA-dependent RNA polymerases analyzed by immunological comparison of their components. EMBO J 2:751–755

Sidow A, Wilson AC (1990) Compositional statistics: an improvement of evolutionary parsimony and its application to deep branches in the tree of life. J Mol Evol 31:51–68

Sweetser D, Nonet M, Young RA (1987) Procaryotic and eukaryotic RNA polymerases have homologous core subunits. Proc Natl Acad Sci USA 84:1192–1196

Swofford DL (1989) PAUP 3.0 user's manual. Illinois Natural History Survey, Champaign, IL

Templeton AR (1983) Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. Evolution 37:221–244

Tu J, Prangishvilli D, Huber H, Wildgruber G, Zillig W, Stetter KO (1982) Taxonomic relations between archaebacteria including 6 novel genera examined by cross hybridization of DNAs and 16S rRNAs. J Mol Evol 18:109–114

Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 74:5088–5090

Woese CR, Olsen GJ (1986) Archaebacterial phylogeny: perspectives on the Urkingdoms. System Appl Microbiol 7:161–177

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

Woese CR, Kandler O, Wheelis M (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA 87:4576–4579

Woese CR, Achenbach L, Rouviere P, Mandelco L (1991) Archaeal phylogeny: reexamination of the phylogenetic position of *Archaeoglobus fulgidus* in light of certain composition-induced artifacts. System Appl Microbiol 14:364–371

Zillig W, Klenk H-P, Palm P, Pühler G, Gropp F, Garrett RA, Leffers H (1989) The phylogenetic relations of DNA-dependent RNA polymerases of archaebacteria, eukaryotes, and eubacteria. Can J Microbiol 35:73–80

Zuckerkandl E (1987) On the molecular evolutionary clock. J Mol Evol 26:34–46