Experimental Phylogeny of Neutrally Evolving DNA Sequences Generated by a Bifurcate Series of Nested Polymerase Chain Reactions

Gerdine F. O. Sanson, Silvia Y. Kawashita, Adriana Brunstein, and Marcelo R. S. Briones

Departamento de Microbiologia, Imunologia e Parasitologia, Escola Paulista de Medicina, Universidade Federal de São Paulo, São Paulo, Brazil

A known phylogeny was generated using a four-step serial bifurcate PCR method. The ancestor sequence (SSU rDNA) evolved in vitro for 280 nested PCR cycles, and the resulting 15 ancestor and 16 terminal sequences (2,238 bp each) were determined. Parsimony, distance, and maximum likelihood analysis of the terminal sequences reconstructed the topology of the real phylogeny and branch lengths accurately. Divergence dates and ancestor sequences were estimated with very small error, particularly at the base of the phylogeny, mostly due to insertion and deletion changes. The substitution patterns along the known phylogeny are not described by reversible models, and accordingly, the probability substitution matrix, based on the observed substitutions from ancestor to terminal nodes along the known phylogeny, was calculated. This approach is an extension of previous studies using bacteriophage serial propagation, because here mutations were allowed to occur neutrally rather than by addition of a mutagenic agent, which produced biased mutational changes. These results provide for the first time biochemical experimental support for phylogenies, divergence date estimates, and an irreversible substitution model based on neutrally evolving DNA sequences. The substitution preferences observed here (A to G and T to C) are consistent with the high G+C content of the *Thermus aquaticus* genome. This suggests, at least in part, that the method here described, which explores the high *Taq* DNA polymerase error rate, simulates the evolution of a DNA segment in a thermophilic organism. These organisms include the bacterial rod *T. aquaticus* and several Archaea, and thus, the method and data set described here may well contribute new insights about the genome evolution of these organisms.

Introduction

Experimental phylogenetics is a convincing means of understanding basic processes of nucleotide change in phylogenies (Hillis, Mable, and Moritz 1996). Hillis and collaborators evolved the bacteriophage T7 by sequential propagation and generated a known phylogeny which provided for the first time experimental support for phylogeny inference methods (Hillis et al. 1992; Bull et al. 1993). Gene phylogenies can be inferred from sequence data using several algorithms under three basic optimality criteria, namely, parsimony (Fitch 1977), pairwise distances (Saitou and Nei 1987), and maximum likelihood (Felsenstein 1981). Among these, the use of explicit models of sequence evolution, or maximum likelihood, which is computationally very intensive, has been increasing recently because of improvements in computer hardware speed and software optimization (Olsen et al. 1994; Strimmer and von Haeseler 1996; Korber et al. 2000). Gene phylogenies may represent species phylogeny if the substitutions in a particular gene represent orthologous steps (Li and Graur 1991). Divergence dates of genes and species can also be estimated from phylogenetic distances (Rambaut and Bromham 1998; Yoder and Yang 2000). These estimates are based on the concept of a molecular clock (Zuckerkandl and Pauling 1962), either global or local, which can be tested for a set of sequences, models, and trees, using relative rate tests (Sarich and Wilson 1973), triplets rate test (Tajima 1993), and the likelihood ratio test (Felsenstein 1988).

In the maximum likelihood method for phylogeny inference, the explicit model of nucleotide substitution (the state transition matrix) is of primary importance (Felsenstein 1981; Posada and Crandall 1998). This varies from the single parameter Jukes and Cantor model to the general time reversible model (Jukes and Cantor 1969; Rodriguez et al. 1990). However, these models assume reversible matrices; in other words, they assume that the probability of the forward change over time (e.g., A to G) is equal to the probability of the reverse event (G to A). Other models, based on irreversible matrices, were proposed, but require the position of the phylogeny root to be inferred by maximizing the likelihood (Yang and Roberts 1995; Galtier and Gouy 1998; Schadt, Sinsheimer, and Lange 1998; Galtier, Tourasse, and Gouy 1999).

Neutral substitutions are very interesting for phylogenetic purposes, although their use for divergence date estimates and phylogeny inference has not yet been tested explicitly by experimental phylogenetics. In the experimental phylogeny of bacteriophage T7, the accelerated rate of change was induced by the presence of a mutagenic agent which changed not only the tempo, but also the mode of evolution, because the mutagenic agent likely biased substitutions from G to A and C to T (Bull et al. 1993). Also, in the T7 sequential propagation, lineages that replicate faster or more effectively tend to be overrepresented compared with slower replicating lineages, and therefore the system is not neutral (Hillis et al. 1992). Neutral substitutions, with proper corrections for superimposed mutations and stochastic variations, likely reflect the elapsed time, and therefore history, since the divergence of two sequences (Li and Graur 1991). Substitutions in sites with selective pressure

Key words: molecular evolution, experimental phylogenetics, SSU rRNA gene, maximum likelihood.

Address for correspondence and reprints: Marcelo R. S. Briones, Departamento de Microbiologia, Imunologia e Parasitologia, Escola Paulista de Medicina, Universidade Federal de São Paulo, Rua Botucatu, 862, 3° andar. CEP 04023–062, São Paulo, S.P. Brazil.
E-mail: marcelo@ecb.epm.br.

Mol. Biol. Evol. 19(2):170–178. 2002

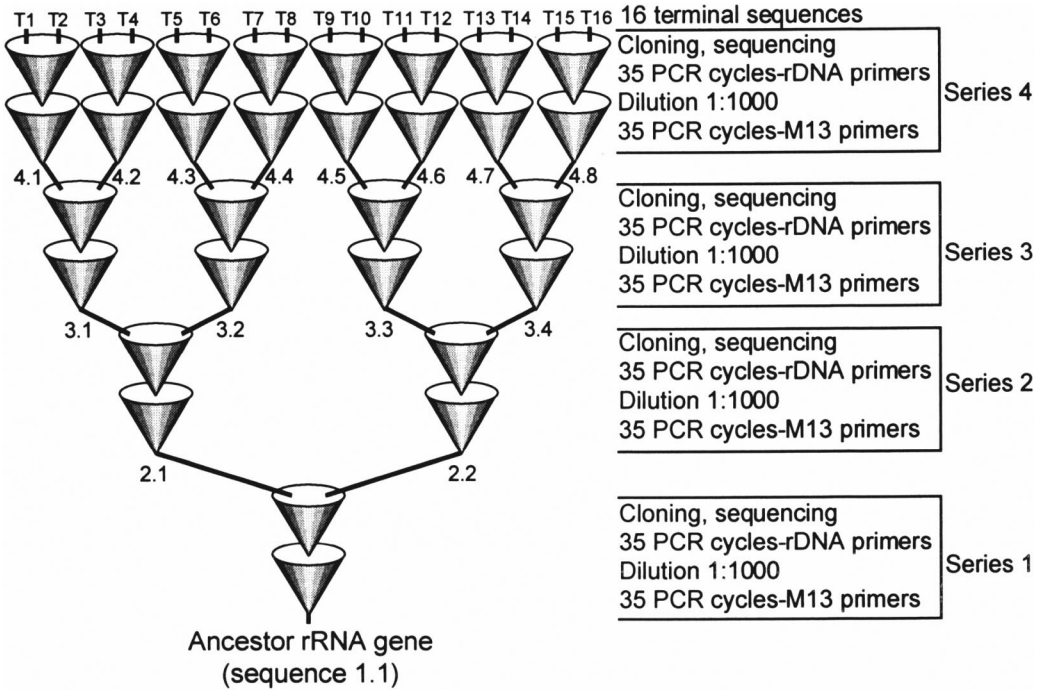

FIG. 1.—Evolution of DNA sequences by a series of bifurcate PCRs. An ancestor SSU rDNA cloned in pBluescript was used as template for series 1 of 70 nested PCR cycles with M13 primers. After the initial 35 cycles, reaction products were diluted 1:1,000 and used as templates for the subsequent 35 cycles, with rDNA primers RIBA and RIBB. After 70 cycles amplicons were cloned, and two clones were picked randomly and used as templates for the next series of nested PCR cycles. Lineages are propagated at random, and therefore the evolution is neutral and behaves as a stochastic process. Tree nodes T1 to T16 indicate terminal sequences, and 1.1 to 4.8, internal ancestors.

mostly reflect the selection regime and thus may cause convergent substitutions, under negative selection, or anomalously long phylogeny branches, under positive selection. Accordingly, divergence dates could be underestimated or overestimated, respectively.

Here we present a method, based on serial PCR, that extends previous studies (Hillis et al. 1992; Bull et al. 1993) by generation of a data set of neutrally evolving sequences. Analysis of this data set provided experimental support for maximum likelihood, with modelfit analysis, for finding the correct topology, reconstruction of ancestor sequences, and divergence date estimates. This data set was also used to calculate a probability transition matrix which describes an irreversible dynamics, based on *Taq* DNA polymerase error rate, and should contribute to further research on coalescence theory (Kingman 1982), phylogeny inference methods, and evolution of thermophilic organisms (Galtier, Tourasse, and Gouy 1999).

Materials and Methods

Gene Amplification and Sequencing

The ancestor sequence (*Trypanosoma cruzi* SSU rDNA, GenBank AF288660) was used as template in a 35 cycle PCR using primers RIBA (5′-CCGAATTCGTCGACAACCTGGTTGATCCTGCCA GT-3′) and RIBB (5′-CCCGGGATCCAAGCTTGATCCTTC TGCAGGTTCACCTAC-3′) which enables the amplification of the complete SSU rDNA sequence. The amplification started with amplification of 0.35 μg of cloned ancestor sequence (0.102 pmol, 6.14×10^{10} molecules) in a 100-μl solution containing 5 units of *Taq* DNA pol (GIBCO), 7 mM $MgCl_2$, 40 nmol each deoxynucleotide triphosphates (dNTP), 20 mM tris(hydroxymethyl)aminomethane (Tris)–HCl pH 8.4, 50 mM KCl, and 10 pmol of each primer. After 35 cycles amplicons were purified from gel (GFX, Amersham-Pharmacia) serially diluted 1:1,000, and 4 μl of the diluted solution were used as template for an additional 35 cycles in 100-μl solutions containing 5 units of *Taq* DNA pol (GIBCO), 7 mM $MgCl_2$, 40 nmol each dNTP, 20 mM Tris–HCl p H8.4, 50 mM KCl, and 10 pmol of each primer (fig. 1). Cycling conditions were 94°C for 1 min, 41°C for 1 min, and 72°C for 2 min, and final extension 72°C for 7 min. After each 70 nested cycles the amplicons were cloned into pBluescript, and two randomly selected clones were picked to be the ancestor of another 70 cycles nested for a total of four rounds of 70 nested cycles. This procedure was repeated four times, resulting in four rounds of 70 nested PCR cycles. Amplification of cloned products was done using M13 forward and reverse primers. The nested reaction has 280 cycles, and all 16 terminal sequences coalesce to ancestor sequence 1.1 (fig. 2). The 70 cycles from each round were divided into two 35 nested cycles, because at 35 cycles the reaction is close to its linear stage. All 31 sequences derived from the process (16 terminal nodes plus 15 ancestors), along with six additional clones of the first 70 cycles, were determined completely, in a total of 37 complete SSU rDNA sequences (total of 82,806 assembled bases). Sequencing was done using BigDye Terminators (Applied Biosystems) in an ABI377/96 auto-

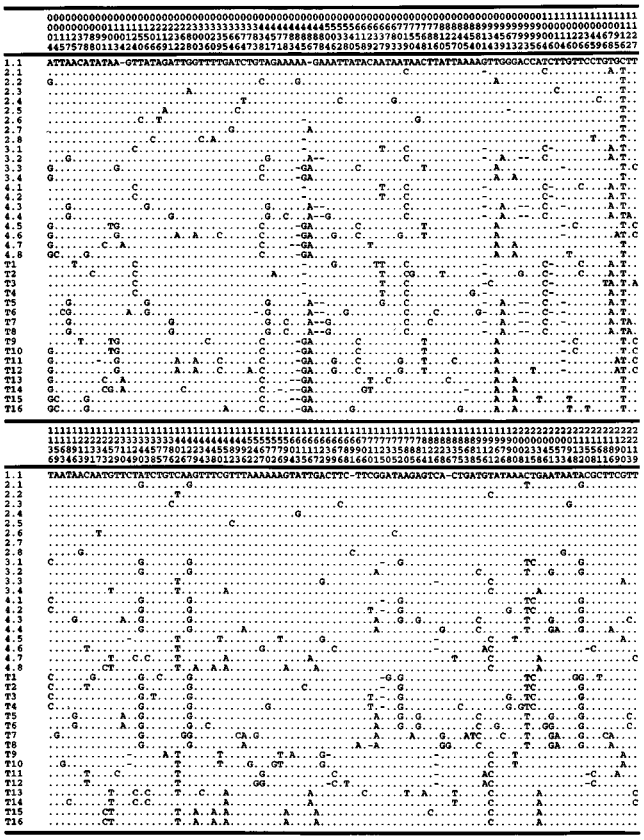

FIG. 2.—Polymorphic sites of sequences generated by serial PCR neutral evolution. Sequences 1.1 to 4.8 represent the internal ancestors, and T1 to T16, the terminal sequences. Sequences 2.3 to 2.8 were not used in subsequent propagation, except to estimate the *Taq* DNA polymerase error rate at 70 cycles. Numbers above sequences indicate the position number in the alignment (total number of positions considered is 2,238). Dots indicate residues that are identical to sequence 1.1.

mated sequencer, by primer walking using internal primers of the 18S rRNA gene. Sequences were assembled using PHRED+PHRAP+CONSED (Gordon, Abajian, and Green 1998) from ABI chromatograms, available upon request. Quality of assembled sequences ranged from phred scores 30 to 40 as estimated by CONSED (Gordon, Abajian, and Green 1998). Sequences presented here have been deposited in GenBank under accession numbers AF288660 (ancestor 1.1) and AF359461 to AF359496 (from 2.1 to T16). Supplementary data, such as alignments and phylogenies, are available on the World Wide Web site (http://compbio.epm.br/ievol) of one of us.

Phylogenetic Analysis

Terminal sequences were aligned by eye using SEAVIEW sequence editor (Galtier, Gouy, and Gautier 1996). Trees were constructed using PAUP 4.0b6 (Swofford 1998) and TREE-PUZZLE (Strimmer and von Haeseler 1996), and modelfit tests were performed using the hierarchical likelihood ratio test implemented in MODELTEST 3.04 (Posada and Crandall 1998). Maximum likelihood trees were constructed using the model selected by MODELTEST (TVMef [Posada and Crandall 1998] with equal base frequencies; the rate matrix A–C = 0.4397, A–G = 13.2362, A–T = 4.9778, C–G = 0.2123, C–T = 13.2362, and G–T = 1.0; proportion of invariable sites = 0; and equal rates for all sites) with heuristic tree-bisection–reconnection (TBR) search. Parsimony (Fitch 1977) trees were built using accelerated transformation (ACCTRAN) and TBR searching with collapse option. Distance trees were constructed using Neighbor-Joining (Saitou and Nei 1987) with a maximum likelihood distance matrix using the model selected by MODELTEST. The standard errors of branch lengths were estimated using PAUP 4.0b6 (Swofford 1998). The number of molecules after 280 cycles was calculated by quantitation of templates and PCR products by absorbance at 260 nm, applying corresponding corrections for dilutions, and conversion to number of molecules using Avogadro's number. Divergence dates with low and high confidence intervals and associated evolutionary rates were estimated using the quartet analysis implemented in QDATE version 1.11 (Bromham et al. 1998; Rambaut and Bromham 1998). The number of substitutions of each rate category was directly quantitated from polymorphic sites along the real phylogeny.

Model Construction

The instantaneous rate matrix (Q-matrix) was built from the observed number of changes of each of 16 categories, in the real topology. Each element Q_{ij} was calculated by dividing the observed number of changes from nucleotide i to nucleotide j by the total number of changes. The diagonal elements Q_{ii} were calculated by using the relation $Q_{ii} = -\Sigma_{j \neq i} Q_{ij}$. The Q-matrix was then written as in table 1.

The substitution probability matrix (*P*-matrix) was calculated from the relation $P(t) = e^{Qt}$ (Swofford et al. 1996). In order to find the elements of the *P*-matrix we had to decompose the Q-matrix into its eigenvalues and eigenvectors. The eigenvalues were obtained by calculating the λ values (v_n) that satisfied the equation $\det(Q - \lambda I) = 0$ (where I is the identity matrix), namely, $v_1 = -0.5187$, $v_2 = -0.4218$, $v_3 = -0.0595$, and $v_4 = -5.7958 \times 10^{-6}$. For each λ value the respective eigenvector ($\mathbf{V}_\lambda$) was determined as that satisfying the equation $(Q - \lambda I)\mathbf{V}_\lambda = 0$. These vectors are: $\mathbf{V}_1 = (-0.6060, -0.1701, 0.1669, 0.7589)^T$; $\mathbf{V}_2 = (0.5025, -0.2510, -0.2801, 0.7785)^T$; $\mathbf{V}_3 = (0.5243, -0.5155, 0.6207, -0.2722)^T$; and $\mathbf{V}_4 = (-0.4999, -0.5001, -0.4998, -0.5000)^T$. The $P_{ij}(t)$ elements in the *P*-matrix were calculated using standard transformation procedures of the Q-matrix, using the previously calculated eigenvalues and eigenvectors. All mathematical calculations were confirmed using MATHEMATICA software (Wolfram Research).

Results and Discussion

Here we wanted to simulate neutral evolution and therefore we present a model which explores the high error rate of *Taq* DNA polymerase (Saiki et al. 1988) (fig. 1). A known SSU rRNA gene sequence was used as the ancestor to start the process of sequential PCR

Table 1
Observed Substitutions Along the True Phylogeny (fig. 3*A*)

Position	Substitution	Position	Substitution	Position	Substitution	Position	Substitution
4	A > G	632	A > T	1219	C > T	1725	DelG
14	T > C	637[a]	A > T, A > G	1231	A > G	1730	A > G
15	T > C	679	T > C	1243	T > C	1752	A > G
17	A > G	703	A > G	1252	G > T	1780	A > T
25	A > T	719	T > C	1279	T > C	1785	G > A
37	C > T	750	A > G	1310	T > A	1786	A > G
78	A > G	768	C > T	1314	DelC	1814	G > A
88	T > C	781	T > C	1329	T > C	1821	T > C
90	DelA	786	T > C	1340	A > G	1826	DelC
91	T > C	810	A > T	1343	T > C	1828	A > G
101[a]	A > G, A > T	825	T > C	1358	C > T	1836	InsG
103	A > G	827	T > C	1375	T > C	1837	C > T
104	InsA	840	A > G	1377	G > A	1855	T > C
112	G > A	845	DelA	1402	C > T	1863	G > A
124	T > C	854	A > G	1416	A > G	1888	A > T
156	A > G	880	A > G	1427	A > G	1915	T > C
221	A > G	911	DelG	1429	G > A	1916	G > A
232	T > A	934	T > C	1434	T > C	1921	T > C
262	T > C	943	T > A	1443	T > C	1962	A > G
300	G > A	959	G > A	1448	T > A	1976	T > C
306	T > C	973	G > A	1450	C > T	1998	A > G
339	T > C	992[a]	DelA, A > G	1451	G > A	2000	A > T
355	G > A	993	DelC	1493	T > C	2008	A > G
377	T > A	995	C > T	1496	T > A	2021	C > T
383	DelG	996	A > T	1522	A > G	2035	T > C
438	T > C	1004	T > C	1542	A > G	2038	G > A
441	A > G	1006	DelC	1567	A > G	2046	A > G
457	G > A	1014	DelT	1570	A > G	2051	A > G
471	DelA	1016	G > T	1572	A > G	2053	T > A
478	A > C	1020	T > C	1576	A > T	2118	T > G
483	DelA	1026	T > C	1599	G > A	2132	A > G
484[a]	A > G, DelA	1036	C > T	1604	T > A	2150	DelC
485[a]	DelA, A > G	1066	G > T	1615	T > C	2158	G > C
486	InsA	1078	T > A	1616	T > A	2161	C > T
487	DelG	1095	G > A	1617	G > A	2181	T > C
488	DelA	1116	C > T	1622	A > G	2186	T > A
504	A > G	1122	T > A	1639	DelC	2199	C > A
506	A > G	1127[a]	T > C, T > A	1669	T > C	2200	G > A
538	T > G	1136	T > C	1688	C > T	2213	T > C
540	A > G	1159	A > G	1696	T > A	2219	T > C
615	T > C	1163	A > G	1706	T > C		
618	A > G	1184	T > C	1710[a]	C > T, DelC		
629	C > T	1196	A > G	1711	G > A		

NOTE.—Ins = insertion; Del = deletion.
[a] Multiple substitutions.

evolution (fig. 1). Every 70 cycles the PCR products were cloned, and two clones were completely sequenced and used as templates (ancestors) for the next round of 70 cycles (fig. 1). Accordingly, the ancestor 1.1 originated 16 terminal sequences or terminal nodes, T1 to T16, after 280 PCR nested cycles (generations) (fig. 1). The full-length sequences of the 16 terminal nodes and the 15 internal nodes (2.1 to 4.8, fig. 3*A*), or ancestors, were determined. The alignment of all 37 sequences used here revealed 196 polymorphic sites, including gaps, and the alignment of the 16 terminal sequences had 169 polymorphic sites (fig. 2). After PCR evolution of 280 generations (fig. 1), the total number of molecules, 9.92×10^{11}, was estimated from PCR product quantitation. Sequences 2.3 to 2.8 were used only to calculate the mutation rate of *Taq* DNA polymerase after 70 PCR cycles, as described later.

The real phylogeny obtained (fig. 3*A*) has a series of 15 dichotomies from the initial ancestor to the 16 terminal sequences, and substitutions that occurred along the phylogeny involved 7.6% of the total number of positions (table 1). We observed that the number of substitutions per time interval (along a branch length) varies significantly. This might be due to stochastic effects once lineages are sampled and propagated randomly. The 16 terminal sequences (T1 to T16) were used to reconstruct the real phylogeny by maximum likelihood, parsimony, and neighbor-joining (Fitch 1977; Felsenstein 1981; Saitou and Nei 1987). Topologies obtained by the three methods were identical and found the real topology. In fig. 3*B* we show the reconstructed maximum likelihood phylogeny which has a topology identical to the true tree (fig. 3*A*). This phylogeny has ln(Likelihood) = −4259.7384 and was inferred using a

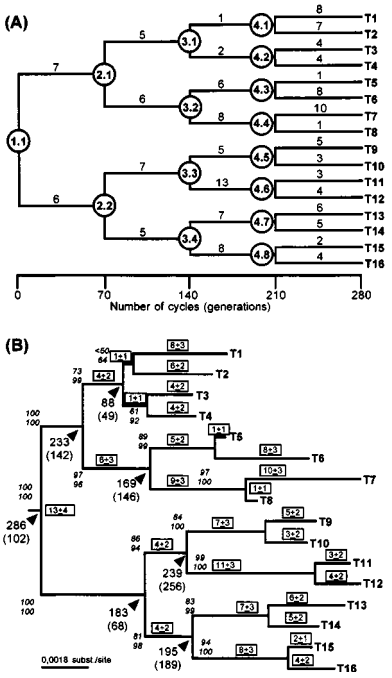

FIG. 3.—Comparison of real phylogeny with inferred maximum likelihood phylogeny (Felsenstein 1981; Posada and Crandall 1998; Swofford 1998). The serial PCR in vitro evolution resulted in the topology depicted *(A)* with varying branch lengths whose ancestors (1.1 to 4.8, circled) and terminal sequences (T1 to T16) were sequenced in full length. Scale bar indicates the number of cycles between tree internodes and nodes. The inferred phylogeny *(B)* has a topology identical to the real tree *(A)* and 9 out of 30 branch lengths were estimated correctly. Boxed numbers indicate branch lengths (number of substitutions), numbers in italics represent the percentage of a given cluster in 100 bootstrap replicates, with reestimation of parameters at each bootstrap replicate (top), and without reestimation at each replicate (bottom). Numbers below arrows indicate the estimated divergence (cycles ago), with the low–high confidence interval range (in parenthesis) as calculated by maximum likelihood quartet analysis (Rambaut and Bromham 1998). Numbers of substitutions, with corresponding standard errors, in the inferred tree *(B)* were calculated by multiplying the branch lengths (in substitutions per site) by the total number of positions (2,238 bp).

submodel of the General Time Reversible Model (TVMef) with equal rates for all sites, with parameter determination by hierarchical likelihood ratio tests (Rodriguez et al. 1990; Posada and Crandall 1998). The best parsimony tree found had 82 informative sites, 157 steps, rescaled consistency index of 0.962, homoplasy index of 0.038, and ln(Likelihood) = −4259.7384, but was not statistically different from the six other parsimony trees with up to 160 steps, as tested by the Kishino–Hasegawa topology test (Kishino and Hasegawa 1989). The maximum likelihood tree differs from the real tree by 30% (9 out of 30) with respect to branch lengths, particularly in clusters where the number of substitutions where higher (fig. 3). However, the differences in branch lengths between the known phylogeny (fig. 3*A*) and the inferred phylogeny (fig. 3*B*) are within the range of the maximum likelihood standard error for branch lengths, as estimated in the inferred tree (fig. 3*B*), and therefore, these differences were expected by the inference method. In addition to MODELTEST, optimization of model parameters was also done with simultaneous estimation of tree searching. For this, an initial topology was inferred by neighbor-joining with simultaneous estimation of nucleotide frequencies and the rate matrix (R-matrix). This initial topology was then rearranged by TBR with simultaneous estimation of the gamma distribution rate parameter and the proportion of invariant sites, and reestimation of nucleotide frequencies and the R-matrix. This approach is as close as possible, in the scope of the present study, to escape from point estimates of model parameters, as done by MODELTEST, and to letting parameters behave freely during tree searching. The tree obtained by this heuristic full optimization (ln[Likelihood] = −4254.74116) found that the real topology (fig. 3*A*) and the parameters were very similar to those obtained by MODELTEST, except in the proportion of invariant sites, which in the full optimization was 0.43. At least in the case of the sequences presented here, MODELTEST gave results similar to those of the heuristic full optimization. A complete full optimization would have to be done using either an exhaustive search of all trees (more than 2×10^{14} trees for 16 taxa) or an exact algorithm, such as branch-and-bound, with simultaneous parameter optimization, and this would take an impractical computational time.

To verify the evolutionary rate of the real phylogeny (*Taq* DNA polymerase error) we used eight sequences, 2.1 to 2.8 (fig. 2), of the first 70 nested cycles. Among 2,238 positions, we observed 44 polymorphic sites, being 2 insertions, 4 deletions, and 48 base substitutions. The average change from the ancestor 1.1 to sequences 2.1 to 2.8 (fig. 2) is six misincorporations (excluding indels), which gives 0.0027 errors per position and an evolutionary rate of 0.38×10^{-4} substitutions per site per generation. This is higher than other estimates for *Taq* DNA polymerase (0.27×10^{-4}) (Bracho, Moya, and Barrio 1998) and might be because of the concentration of $MgCl_2$ (7 mM) used in our study, in order to accelerate the substitution rate. The substitutions at terminal sequences are evenly distributed along the SSU rRNA sequence length (table 1) which demonstrates that we generated a neutral evolution simulation. Therefore, selection is not responsible for the branch length variation observed in the real tree (fig. 3*A*).

We also tested if divergence times could be estimated from inferred phylogenies. The 16 terminal sequences were analyzed by a maximum likelihood quartet method (Rambaut and Bromham 1998), using the same substitution model and parameters used to infer the maximum likelihood phylogeny (fig. 3*B*). All terminal pairs diverged 70 cycles earlier and were clustered into quartets to infer the divergence of internal nodes. Divergence times of ancestors 1.1, 2.1, 2.2, 3.2, and 3.4 were estimated correctly, and ancestors 3.1 and 3.3 had dates outside the 95% confidence interval. The inferred tree, however, passes the likelihood ratio test for molecular clocks (Felsenstein 1988) as the difference between the likelihoods of the molecular clock constrained tree and the unconstrained tree is not statistically significant at 95% confidence level ($2 \times \Delta\text{LnL} = 21.2818$, and the

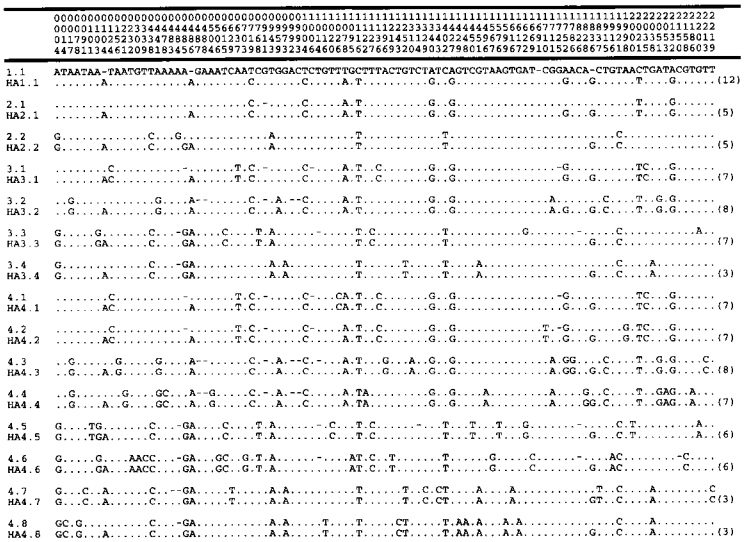

FIG. 4.—Comparison between ancestor sequences in the real phylogeny (fig. 3*A*) and maximum likelihood reconstruction of ancestral states, HA. Only polymorphic positions are shown. Dots indicate residues identical to sequence 1.1. Numbers above the alignment indicate the position numbers. Numbers in parenthesis indicate the differences between the HA and their corresponding real ancestor sequences.

Chi-square critical value for 14 degrees of freedom at $P < 0.05$ is 23.685). The maximum likelihood quartet analysis also estimated the evolutionary rate between 0.24×10^{-4} and 0.42×10^{-4} substitutions per site per generation.

Ancestor sequences 1.1 to 4.8 were also predicted from the maximum likelihood tree (hypothetical ancestors, HA) and compared with ancestors from the real phylogeny (fig. 4). Ancestors of the real phylogeny were reconstructed with accuracy from 99.46% to 99.87% (3 to 12 differences) by maximum likelihood. Most of the inaccurately assigned ancestor states were in regions with insertions and deletions and in two positions with two substitutions at the same site, positions 101 and 637 (fig. 4 and table 1). The reconstruction with the most errors (12) was ancestor 1.1, and the most accurate reconstructions were from ancestors 3.4, 4.7, and 4.8 (fig. 4). This suggests that as we move deeper in the tree, ancestor sequence reconstructions might be more sensitive to insertion and deletion events and multiple substitutions at the same site. Insertions and deletions occurred much more frequently after runs of homopolymeric regions, as observed elsewhere, and might be caused by the slippage of *Taq* DNA polymerase (Bracho, Moya, and Barrio 1998).

Table 2
Rate-Matrix Selected by Modelfit Analysis (reversible) Compared with the Rate-Matrix Observed from Real Data (in parenthesis)

	To			
From	A	C	G	T
A	—	0.4397	13.2362	4.9778
	—	(0.3335)	(14.6650)	(3.3330)
	−0.3595	*0.0065*	*0.2876*	*0.0654*
C	0.4397	—	0.2123	13.2362
	(0.3335)	—	(0.0000)	(4.9995)
	0.0065	*−0.1046*	*0.0000*	*0.0980*
G	13.2362	0.2123	—	1.0000
	(6.3325)	(0.3332)	—	(1.0000)
	0.1242	*0.0065*	*−0.1503*	*0.0196*
T	4.9778	13.2362	1.0000	—
	(4.6660)	(14.3294)	(0.6666)	—
	0.0915	*0.2810*	*0.0131*	*−0.3856*

NOTE.—Rates are relative (G > T) fixed to 1. The instantaneous rate matrix (Q-matrix), whose values are in italics, indicate the observed rates of substitutions expressed as the observed number of substitutions of each type divided by the total numbers of substitutions.

The substitution model selected by the hierarchical likelihood ratio test was compared with the actual changes (table 2). It shows that the changes in the real tree follow an irreversible model, particularly when A > G-G > A, A > T-T > A, and C > T-T > C changes are compared. Base frequencies in the data set (16 terminal sequences) are f(A) = 0.24938; f(C) = 0.22548; f(G) = 0.26574; and f(T) = 0.25941; among constant sites they are (0.24374, 0.23225, 0.27171, 0.25230) and among variable sites (0.32895, 0.12972, 0.18143, 0.35990). This suggests that the direction of change is not biased toward more abundant residues in the nucleotide pool.

To depict the probability of substitutions along the real phylogeny, the elements of the instantaneous rate matrix, the Q-matrix (table 2), were used to assemble the following equations:

$$dP(A)/dt = -0.3595P(A) + 0.0065P(C) + 0.1242P(G) + 0.0915P(T) \quad (1)$$

$$dP(C)/dt = 0.0065P(A) - 0.1046P(C) + 0.0065P(G) + 0.2810P(T) \quad (2)$$

$$dP(G)/dt = 0.2876P(A) + 0.0000P(C) - 0.1503P(G) + 0.0131P(T) \quad (3)$$

$$dP(T)/dt = 0.0654P(A) + 0.0980P(C) + 0.0196P(G) - 0.3856P(T) \quad (4)$$

Table 3
Substitution Probability Matrix Elements of the *P*-matrix, Derived from Observed Substitutions Along the Real Phylogeny (fig. 3A)

$P_{ij}(t)$	j = A	j = C	j = G	j = T
i = A	$(0.4361e^{-\alpha t} + 0.2991e^{-\beta t} + 0.1199e^{-\gamma t} + 0.1449)$	$(0.2211e^{-\alpha t} - 0.2296e^{-\beta t} - 0.4108e^{-\gamma t} + 0.4192)$	$(-0.3288e^{-\alpha t} - 0.3294e^{-\beta t} + 0.3683e^{-\gamma t} + 0.2898)$	$(-0.3284e^{-\alpha t} + 0.2599e^{-\beta t} - 0.0773e^{-\gamma t} + 0.1459)$
i = C	$(0.1224e^{-\alpha t} - 0.1494e^{-\beta t} - 0.1179e^{-\gamma t} + 0.1449)$	$(0.0621e^{-\alpha t} + 0.1147e^{-\beta t} + 0.4039e^{-\gamma t} + 0.4193)$	$(-0.0923e^{-\alpha t} + 0.1645e^{-\beta t} - 0.3622e^{-\gamma t} + 0.2899)$	$(-0.0922e^{-\alpha t} - 0.1298e^{-\beta t} + 0.0760e^{-\gamma t} + 0.1459)$
i = G	$(-0.1201e^{-\alpha t} - 0.1667e^{-\beta t} + 0.1419e^{-\gamma t} + 0.1449)$	$(-0.0609e^{-\alpha t} + 0.1279e^{-\beta t} - 0.4863e^{-\gamma t} + 0.4193)$	$(0.0905e^{-\alpha t} + 0.1836e^{-\beta t} + 0.4360e^{-\gamma t} + 0.2899)$	$(0.0905e^{-\alpha t} - 0.1448e^{-\beta t} - 0.0915e^{-\gamma t} + 0.1459)$
i = T	$(-0.5461e^{-\alpha t} + 0.4635e^{-\beta t} - 0.0623e^{-\gamma t} + 0.1449)$	$(-0.2770e^{-\alpha t} - 0.3557e^{-\beta t} + 0.2133e^{-\gamma t} + 0.4193)$	$(0.4117e^{-\alpha t} - 0.5104e^{-\beta t} - 0.1912e^{-\gamma t} + 0.2899)$	$(0.4113e^{-\alpha t} + 0.4026e^{-\beta t} + 0.0401e^{-\gamma t} + 0.1459)$

NOTE.—Given $\alpha = 0.5187$, $\beta = 0.4218$, and $\gamma = 0.0595$.

where *P*(A), *P*(C), *P*(G), and *P*(T) are the probabilities of observing the corresponding nucleotides at a given site after a time interval dt.

To construct the substitution probability matrix, the *P*-matrix (table 3; table 2, in italics), the Q-matrix was decomposed into its eigenvalues and eigenvectors and used in calculations as described in *Materials and Methods*. The resulting *P*-matrix (table 3) is a stochastic Markov model (Swofford et al. 1996) which is also irreversible, as are other models proposed (Yang and Roberts 1995; Galtier and Gouy 1998; Schadt, Sinsheimer, and Lange 1998; Galtier, Tourasse, and Gouy 1999). Another variation in the models is regarding their homogeneity along the phylogeny (Galtier and Gouy 1998; Galtier, Tourasse, and Gouy 1999). To verify whether the model is homogeneous or nonhomogeneous along the known phylogeny (fig. 3*A*), we compared the rate matrices (R-matrix) after 280 cycles with the rate matrix after 70 cycles (based on sequences 2.1 to 2.8). The R-matrix at 70 cycles (not shown) is nearly identical to the R-matrix after 280 generations (table 2) which suggests that the model of sequence evolution is homogeneous along the know phylogeny in fig. 3*A*.

The experimental approach to phylogeny inference by a biochemical assay, as presented here, is advantageous over plain computer simulations, because the introduction of variations in sequences along the phylogeny is driven by errors of an enzymatic reaction, which is a better approximation of the real event. The process of substitution along a phylogeny is not understood in its full extent, and therefore, it is very unlikely that a computer simulation will include all the variations and nuances underlying this process. Consequently, computer simulations will be dependent on the particular parameters and bias that the programmer chooses to include.

The results presented here, using a biochemical assay, show that the phylogeny topology was reconstructed correctly, even with differences between the observed model and the maximum likelihood modelfit selected model (table 2). However, inference of branch lengths, divergence dates, and hypothetical ancestors are more sensitive to stochastic rate variations which should be corrected by using a more accurate model. Also, the variation in branch lengths, meaning variation in evolutionary rates, observed in the true tree (fig. 3*A*) occurred in the absence of selection. Interestingly, the molecular clock is not rejected by the likelihood ratio test of the inferred phylogeny (fig. 3*B*) (Felsenstein 1988), which suggests that even with considerable variation of rates among lineages, at least to the extent shown here, divergence date estimates from phylogenies are accurate (Bromham et al. 1998; Rambaut and Bromham 1998). We employed a strategy analogous to the one used by Hillis et al. (1992). However, in our study the number of molecules at each stage and the exact number of generations (PCR cycles) are known. Because we simulated neutral evolution of an SSU rDNA, changes in the secondary structure of the gene product did not interfere with our in vitro evolutionary process. The SSU rRNA in vivo has a functionally conserved secondary struc-

ture, and most substitutions tend to occur within regions of unpaired loops (Hillis and Dixon 1991).

The serial PCR method described here could be used in studies of evolution of thermophilic organisms. The substitution bias observed here (A > G and T > C, table 1) is consistent with the high G + C content of the *Thermus aquaticus* genome, and might be a consequence of specific properties of *Taq* DNA polymerase. However, the polymerase domains of *Taq* DNA polymerase and the Klenow fragment of *Escherichia coli* DNA polymerase I are nearly identical. In *Taq* DNA polymerase, two of the catalytically critical carboxylate residues on the 3′–5′ exonuclease activity are missing (Kim et al. 1995). The simulation described here might well reflect the neutral evolution of a DNA segment in a thermophilic organism, such as in the bacterial rod *T. aquaticus* and in several Archaea, which implies that the method presented here could be developed to address questions about the genome evolution of these organisms (Woese 1987; Pace 1997).

As a perspective, the serial PCR evolution method and the data set presented here can contribute to future studies on coalescence theory (Kingman 1982) and to divergence date estimate methodology (Rambaut and Bromham 1998; Yoder and Yang 2000), because the number of generations, the phylogeny, and the mutation rate per generation are known. The in vitro generation of a phylogeny with no selection and no migration should be particularly useful for estimating the θ parameter (Kuhner, Yamato, and Felsenstein 1995). Nevertheless, this study provides for the first time biochemical experimental support for phylogeny inference from neutral substitutions using maximum likelihood with modelfit optimization.

Acknowledgments

We thank J. F. Perez, Scientific Director of Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for sequencing equipment made available to us through the ONSA Brazilian sequencing network, and the excellent suggestions from the two anonymous referees who reviewed this manuscript. G.F.O.S. and S.Y.K. received graduate and undergraduate fellowships, respectively, from CNPq (Brazil). This work was supported by grants to M.R.S.B. from FAPESP and CNPq (Brazil), and the International Research Scholars Program of the Howard Hughes Medical Institute (United States).

LITERATURE CITED

BRACHO, M. A., A. MOYA, and E. BARRIO. 1998. Contribution of Taq polymerase-induced errors to the estimation of RNA virus diversity. J. Gen. Virol. **79**:2921–2928.

BROMHAM, L., A. RAMBAUT, R. FORTEY, A. COOPER, and D. PENNY. 1998. Testing the Cambrian explosion hypothesis by using a molecular dating technique. Proc. Natl. Acad. Sci. USA **95**:12386–12389.

BULL, J. J., C. W. CUNNINGHAM, I. J. MOLINEUX, M. R. BADGETT, and D. M. HILLIS. 1993. Experimental molecular evolution of bacteriophage-T7. Evolution **47**:993–1007.

FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**:368–376.

———. 1988. Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. **22**:521–565.

FITCH, W. M. 1977. On the problem of discovering the most parsimonious tree. Am. Nat. **111**:223–257.

GALTIER, N., and M. GOUY. 1998. Inferring pattern and process: maximum likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol. **15**:871–879.

GALTIER, N., M. GOUY, and C. GAUTIER. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput. Appl. Biosci. **12**:543–548.

GALTIER, N., N. TOURASSE, and M. GOUY. 1999. A nonhyperthermophilic common ancestor to extant life forms. Science **283**:220–221.

GORDON, D., C. ABAJIAN, and P. GREEN. 1998. Consed: a graphical tool for sequence finishing. Genome Res. **8**:195–202.

HILLIS, D. M., J. J. BULL, M. E. WHITE, M. R. BADGET, and I. J. MOLINEUX. 1992. Experimental phylogenetics: generation of a known phylogeny. Science **255**:589–592.

HILLIS, D. M., and M. T. DIXON. 1991. Ribosomal DNA: molecular evolution and phylogenetic inference. Q. Rev. Biol. **66**:411–453.

HILLIS, D. M., B. K. MABLE, and C. MORITZ. 1996. Applications of molecular systematics: the state of the field and a look to the future. Pp. 515–543 *in* D. M. HILLIS, C. MORITZ, and B. K. MABLE, eds. Molecular systematics. Sinauer Associates, Sunderland, Mass.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–123 *in* H. N. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

KIM, Y., S. H. EOM, J. WANG, D.-S. LEE, S. W. SUH, and T. A. STEITZ. 1995. Crystal structure of *Thermus aquaticus* DNA polymerase. Nature **376**:612–616.

KINGMAN, J. F. C. 1982. The coalescent. Stochast. Proc. Appl. **13**:235–248.

KISHINO, H., and M. HASEGAWA. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominidea. J. Mol. Evol. **29**:170–179.

KORBER, B., M. MULDOON, J. THEILER, F. GAO, R. GUPTA, A. LAPEDES, B. H. HAHN, S. WOLINKSY, and T. BHATTACHARYA. 2000. Timing the ancestor of the HIV-1 pandemic strains. Science **288**:1789–1796.

KUHNER, M. K., J. YAMATO, and J. FELSENSTEIN. 1995. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics **140**:1421–1430.

LI, W. H., and D. GRAUR. 1991. Fundamentals of molecular evolution. Sinauer Associates, Sunderland, Mass.

OLSEN, G. J., H. MATSUDA, R. HAGSTROM, and R. OVERBEEK. 1994. fastDNAmL: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput. Appl. Biosci. **10**:41–48.

PACE, N. R. 1997. A molecular view of microbial diversity and the biosphere. Science **276**:734–740.

POSADA, D., and K. A. CRANDALL. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics **14**:817–818.

RAMBAUT, A., and L. BROMHAM. 1998. Estimating divergence dates from molecular sequences. Mol. Biol. Evol. **15**:442–448.

RODRIGUEZ, F., J. F. OLIVER, A. MARIN, and J. R. MEDINA. 1990. The general stochastic model of nucleotide substitution. J. Theor. Biol. **142**:485–501.

SAIKI, R. K., D. H. GELFAND, S. STOFFEL, S. J. SCHARF, R. HIGUCHI, G. T. HORN, K. B. MULLIS, and H. A. ERLICH. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science **239**:487–491.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

SARICH, V. M., and A. C. WILSON. 1973. Generation time and genomic evolution in primates. Science **179**:1144–1147.

SCHADT, E. E., J. S. SINSHEIMER, and K. LANGE. 1998. Computational advances in maximum likelihood methods for molecular phylogeny. Genome Res. **8**:222–233.

STRIMMER, K., and A. VON HAESELER. 1996. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Mol. Biol. Evol. **13**:964–969.

SWOFFORD, D. L. 1998. PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4b6. Sinauer Associates, Sunderland, Mass.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, and D. M. HILLIS. 1996. Phylogenetic inference. Pp. 407–514 *in* D. M. HILLIS, C. MORITZ, and B. K. MABLE, eds. Molecular systematics. Sinauer Associates, Sunderland, Mass.

TAJIMA, F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics **135**:599–607.

WOESE, C. R. 1987. Bacterial evolution. Microbiol. Rev. **51**: 221–271.

YANG, Z., and D. ROBERTS. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol. Biol. Evol. **12**:451–458.

YODER, A. D., and Z. YANG. 2000. Estimation of primate speciation dates using local molecular clocks. Mol. Biol. Evol. **17**:1081–1090.

ZUCKERKANDL, E., and L. PAULING. 1962. Molecular disease, evolution and genetic heterogeneity. Pp. 189–225 *in* M. MARSHA and B. PULLMAN, eds. Horizons in biochemistry. Academic Press, New York.

WOLFGANG STEPHAN, reviewing editor

Accepted September 26, 2001