RNA Sequence Evolution With Secondary Structure Constraints: Comparison of Substitution Rate Models Using Maximum-Likelihood Methods

Nicholas J. Savill,[1] David C. Hoyle and Paul G. Higgs

School of Biological Sciences, University of Manchester, Manchester M13 9PT, United Kingdom

Manuscript received July 17, 2000
Accepted for publication September 11, 2000

ABSTRACT

We test models for the evolution of helical regions of RNA sequences, where the base pairing constraint leads to correlated compensatory substitutions occurring on either side of the pair. These models are of three types: 6-state models include only the four Watson-Crick pairs plus GU and UG; 7-state models include a single mismatch state that combines all of the 10 possible mismatches; 16-state models treat all mismatch states separately. We analyzed a set of eubacterial ribosomal RNA sequences with a well-established phylogenetic tree structure. For each model, the maximum-likelihood values of the parameters were obtained. The models were compared using the Akaike information criterion, the likelihood-ratio test, and Cox's test. With a high significance level, models that permit a nonzero rate of double substitutions performed better than those that assume zero double substitution rate. Some models assume symmetry between GC and CG, between AU and UA, and between GU and UG. Models that relaxed this symmetry assumption performed slightly better, but the tests did not all agree on the significance level. The most general time-reversible model significantly outperformed any of the simplifications. We consider the relative merits of all these models for molecular phylogenetics.

THERE are several classes of RNA molecules where sequences are available over a wide range of species and where multiple sequence alignments are well established, *e.g.*, transfer RNA, 5S ribosomal RNA, small and large subunit ribosomal RNA, and ribonuclease P RNA. Secondary structure is strongly conserved over long time periods, indicating that selection is acting to maintain a structure that is essential for the function of these molecules. The helical regions of the molecule are often quite variable in sequence. This shows that the precise sequence of bases within the helical regions is of relatively little importance as long as the positioning of these regions within the secondary structure is correct.

The mode of evolution within the helical regions is via pairs of compensatory neutral mutations; *i.e.*, a mutation on one side of a pair disrupts the structure and is slightly deleterious, but a second mutation of the other side of the pair restores the pairing ability. Compensatory pair changes form the basis of the comparative method for deducing RNA secondary structures (Woese and Pace 1993; Gutell 1996) and have also been the subject of several evolutionary studies (Wheeler and Honeycutt 1988; Rousset *et al.* 1991; Vawter and Brown 1993; Gatesy *et al.* 1994; Kirby *et al.* 1995). For an illustrative example of the degree of variability seen in the helical regions of tRNA see Figure 1 of Higgs (1998).

The mathematical theory of compensatory mutations was first discussed by Kimura (1985), who showed that compensatory changes can occur rapidly if the two sites are closely linked. The theory was developed by Iizuka and Takefu (1996) and was studied specifically in the context of RNA helices by Stephan (1996). These articles treat mutation as irreversible and calculate the time until fixation of the double mutant as a function of the mutation rate, the population size, and the strength of selection against the intermediate (single mutant). Higgs (1998) has considered the same problem in the case of reversible mutations and has calculated frequency distributions for the different possible paired states.

RNA sequences are often used in constructing molecular phylogenies (*e.g.*, Olsen and Woese 1993). If the phylogenies are constructed using the helical regions of RNA molecules then an understanding of the way in which these parts of the sequences evolve is important if reliable estimates of distances are to be obtained. Typically Markov models are used to represent the substitution process in the sequences. With 16 possible pairs that can be formed with four bases a 16-state Markov model, with a 16×16 rate matrix to describe relative rates from one state to another, is needed to study the evolution of the helical regions of RNAs. Models of this type have been proposed by Schöniger and von Haeseler (1994), Muse (1995), and Rzhetsky (1995). However, only the 6 matching pairs AU, GU, GC, UA,

Corresponding author: Paul G. Higgs, School of Biological Sciences, University of Manchester, Oxford Rd., Manchester M13 9PT, United Kingdom. E-mail: paul.higgs@man.ac.uk

[1]*Present address:* Department of Mathematics, Heriot-Watt University, Edinburgh EH14 4AS, United Kingdom.

UG, and CG occur frequently in helices, whereas the other 10, so-called mismatch, pairs occur rarely, presumably because they are deleterious mutations that destabilize the secondary structure. Since the mismatch states are rare, it is reasonable to consider 6-state models that allow only the 6 matching pairs, as was done by Tillier (1994) and Tillier and Collins (1995). A third alternative, rather than ignore mismatches completely, is to group all the 10 mismatch (MM) pairs together into a single state. This gives a 7-state model (Tillier and Collins 1998; Higgs 2000).

There is now a large variety of slightly different models. The principle aim of this article is to compare these alternatives to see which is best able to describe real sequence data. Some of these models involve a relatively small number of parameters and make assumptions about the symmetry of the rate matrix. This allows analytical solution of the rate equations in several cases. More complex models are straightforward to construct. Increasing the complexity of a model will, in general, improve the quality of the fit to the data. However, very large numbers of parameters are sometimes not justified because the extra parameters simply fit noise in the data rather than any underlying trends. We therefore require statistical techniques for this model selection process. Comparison of models of differing complexity has been carried out for models of single nucleotide substitution by Yang *et al.* (1994), and here we carry out a similar comparison for paired-site models. We analyze a set of ribosomal RNA sequences using 18 different models. For each model we obtain the maximum-likelihood solution for the frequency of the states, the rate matrix, and the branch lengths of an example phylogenetic tree. Statistical tests are then used to compare the maximum-likelihood values of the different models.

The most important issues to be considered when comparing models are introduced at this point. First, we might expect that the frequency of GC pairs should be equal to that of CG, that the frequency of GU should be equal to that of UG, and that the frequency of AU should be equal to that of UA. We refer to this as "base pair reversal symmetry." We wish to determine whether models that allow arbitrary base pair frequencies fit the data significantly better than models that assume base pair reversal symmetry. Note that the first-mentioned base in the pair is the one closer to the 5′ end of the molecule. Thus, for example, it is possible to unambiguously distinguish a GC from a CG pair by this rule, and the equivalence of GC and CG pairs does not follow as a trivial point.

Second, we wish to determine how to treat mismatches. The 6- and 7-state models clearly throw away information by ignoring mismatches or by treating them in a simplified way. However, since the mismatch states are rare, it may be that they give very little phylogenetic information in any case, and it may be difficult to estimate the parameters determining the rates of change to and from mismatch states with any accuracy. Hence it is not clear *a priori* whether 16-state models give any advantage.

A third important issue is whether double substitutions should be permitted in the rate matrix. It is frequently observed that pairs of closely related species differ by a pair of compensatory substitutions; *e.g.*, a GC in one species is replaced by an AU in the other. Since mutation rates are very low in real organisms it is unlikely that these two changes occurred in a single organism in a single generation. There were presumably some individuals with single mutant genotypes (in this case probably GU) at some point in time. From the population genetics viewpoint the compensatory change can happen in two ways. The first is by fixation of the slightly deleterious mutation (*i.e.*, GU sequences rise to a high frequency in the population) followed by fixation of the second mutation, which is now slightly advantageous (*i.e.*, AU sequences arise and replace the GU sequences). The second method is by the compensatory substitution mechanism discussed by Kimura (1985), Stephan (1996), and Higgs (1998). In this case, slightly deleterious single mutant sequences are created continually by recurrent mutations, but their frequency is kept very low by selection. If one of these sequences undergoes a second mutation this can create a sequence that is almost neutral with respect to the majority of the population. For example, GC sequences are in the majority, GU sequences are created in very small numbers, and one of these mutates to an AU. The neutral variant can then sometimes replace the original dominant variant due to drift in gene frequencies. In this example the consensus sequence would change from GC to AU in a single step, and the GU sequences would remain as a minor variant throughout.

In phylogenetic studies, there is usually only one sequence available for each species, and there is no information available on minor sequence variants that might exist in the population. The substitution rates therefore represent changes in the consensus sequence of the population and do not represent rates of mutation in individual copies of a gene. Thus it is perfectly reasonable to allow double substitutions in the rate matrix, even though double mutations in single genes probably almost never occur. Tillier and Collins (1998) and Higgs (2000) have used models that allow double substitutions and find that the observed values of double substitution rates appear to be high. In this article we use likelihood methods to determine if models that permit double substitutions fit the data significantly better than models that disallow double substitutions.

MATERIALS AND METHODS

Definition of models: A model is defined by the matrix $\boldsymbol{r}$, where each element, r_{ij}, gives the rate of substitution to state j given that the base pair is currently in state i. The theoretical

TABLE 1

Definitions of models

Model ID no.	Frequency parameters	Rate parameters	Constraints	Free parameters	Reference
6A	6: $\pi_1, \pi_2 \ldots \pi_6$	15: α_{ij}	2	19	
6B	6: $\pi_1, \pi_2 \ldots \pi_6$	3: $\alpha_s, \alpha_d, \beta$	2	7	
6C	3: π_1, π_2, π_3	3: $\alpha_s, \alpha_d, \beta$	2	4	Tillier (1994)
6D	3: π_1, π_2, π_3	2: α_s, β	2	3	Tillier (1994)
7A	7: $\pi_1, \pi_2 \ldots \pi_7$	21: α_{ij}	2	26	Higgs (2000)
7B	4: $\pi_1, \pi_2, \pi_3, \pi_7$	21: α_{ij}	2	23	
7C	7: $\pi_1, \pi_2 \ldots \pi_7$	10: α_{ij}	2	15	
7D	7: $\pi_1, \pi_2 \ldots \pi_7$	4: $\alpha_s, \alpha_d, \beta, \gamma$	2	9	Tillier and Collins (1998)
7E	7: $\pi_1, \pi_2 \ldots \pi_7$	2: α_s, γ	2	7	Tillier and Collins (1998)
7F	4: $\pi_1, \pi_2, \pi_3, \pi_7$	4: $\alpha_s, \alpha_d, \beta, \gamma$	2	6	
16A	10: $\pi_1 \ldots \pi_{16}$	5: $\alpha_s, \alpha_d, \beta, \gamma, \varepsilon$	2	19	
16B	16: $\pi_1, \pi_2 \ldots \pi_{16}$	1: μ	2	15	Schöniger and von Haeseler (1994)
16C	7: $\pi_1 \ldots \pi_6, \pi_m$	5: $\alpha_s, \alpha_d, \beta, \gamma, \varepsilon$	2	10	
16D	4: $\pi_A, \pi_C, \pi_G, \pi_U$	4: $\alpha, \beta, \lambda, \phi$	2	6	
16E	4: $\pi_A, \pi_C, \pi_G, \pi_U$	3: α, β, λ	2	5	Muse (1995) modified HKY
16F	4: $\pi_A, \pi_C, \pi_G, \pi_U$	3: α, β, λ	2	5	Muse (1995) GU model
16G	0	3: α, β, γ	1	2	Rzhetsky (1995)
16H	0	2: μ, λ	1	1	Muse (1995)

treatment of rate matrices has been developed in the context of single site models (see, for example, Li and Gu 1996; Li 1997; Waddell and Steel 1997). The rate equations used here are equivalent to those used for single site models, with the exception that the number of states is 6, 7, or 16, instead of 4. The probability $P_{ij}(t)$ that a base pair is in state j at time t given that its ancestor was in state i at time zero satisfies

$$\frac{dP_{ij}}{dt} = \sum_k P_{ik} r_{kj}. \tag{1}$$

The diagonal elements of the matrix must satisfy

$$r_{ii} = -\sum_{j \neq i} r_{ij} \tag{2}$$

to conserve probability, and this constraint is included in the definition of the models. At large times $P_{ij}(t)$ tends to π_j, the equilibrium frequency of state j, irrespective of the initial state i. In some models the equilibrium frequencies are parameters of the model; hence when fitting data we need to apply the constraint

$$\sum_i \pi_i = 1. \tag{3}$$

In other models, the frequencies are defined as functions of other parameters, in which case constraint (3) applies automatically. When comparing models it is useful to have a common time scale. We choose the time scale so that an average of one substitution event per base pair happens in 1 time unit, hence the constraint

$$\sum_i \pi_i \sum_{j \neq i} r_{ij} = 1. \tag{4}$$

This constraint can be imposed by multiplication of all elements of the matrix by a constant factor. In addition, all the models considered here are time reversible; *i.e.*, they satisfy

$$\pi_i r_{ij} = \pi_j r_{ji}. \tag{5}$$

Table 1 shows the models tested and summarizes the parameters involved. The number of free parameters in a model is the number of frequency parameters plus the number of rate parameters minus the number of constraints. The constraints are Equation 4, for all models, and Equation 3, where appropriate. The models are assigned identification codes A, B, C, etc. in order of decreasing numbers of free parameters. In all the models, states 1–6 refer to the principal paired states in the following order: AU, GU, GC, UA, UG, CG. In 7-state models, state 7 is MM. In 16-state models, states 7–16 refer to the 10 possible mismatch states in alphabetical order.

A general reversible model is the most general matrix of a given number of states that satisfies Equation 5 (Li and Gu 1996; Waddell and Steel 1997). The most general reversible 7-state model, labeled 7A, has 26 free parameters and is shown in Figure 1. This model was used by Higgs (2000), but has not previously been used with maximum-likelihood methods. Since this model has many parameters we wish to consider whether simpler models will fit the data equally well. One natural simplification to make is to impose base pair reversal symmetry on model 7A by setting $\pi_4 = \pi_1$, $\pi_5 = \pi_2$, and $\pi_6 = \pi_3$. This gives model 7B. Another possible simplification is to set all the α parameters corresponding to double substitutions to zero, giving model 7C. Changes to and from the MM state are treated as single substitutions and are not set to zero in 7C. Tillier and Collins (1998) have also defined a 7-state model, here called 7D, and shown in Figure 1. This has 7 frequency parameters and allows double substitutions. The many independent α_{ij} parameters in 7A are simplified to just 4: α_s controls the single substitution rate, α_d controls the double substitution rate, β controls the double transversion rate, and γ controls substitutions to and from the mismatch state. Following the usual convention, we define a transition as a substitution from one purine to another, or one pyrimidine to another, and a transversion as a substitution from a purine to a pyrimidine or vice versa.

The three models 7B, 7C, and 7D are *nested* in model 7A; *i.e.*, the simpler models are special cases of the more general model obtained by setting some parameters equal or some parameters to zero. Further simplifications of these models are possible. If the double substitutions are set to zero in 7D, we obtain 7E. If base pair reversal symmetry is imposed on 7D, we obtain 7F. The relationship between all these models

Model 7A

		1	2	3	4	5	6	7
		AU	*GU*	*GC*	*UA*	*UG*	*CG*	*MM*
1	*AU*	*	$\pi_2\alpha_{12}$	$\pi_3\alpha_{13}$	$\pi_4\alpha_{14}$	$\pi_5\alpha_{15}$	$\pi_6\alpha_{16}$	$\pi_7\alpha_{17}$
2	*GU*	$\pi_1\alpha_{12}$	*	$\pi_3\alpha_{23}$	$\pi_4\alpha_{24}$	$\pi_5\alpha_{25}$	$\pi_6\alpha_{26}$	$\pi_7\alpha_{27}$
3	*GC*	$\pi_1\alpha_{13}$	$\pi_2\alpha_{23}$	*	$\pi_4\alpha_{34}$	$\pi_5\alpha_{35}$	$\pi_6\alpha_{36}$	$\pi_7\alpha_{37}$
4	*UA*	$\pi_1\alpha_{14}$	$\pi_2\alpha_{24}$	$\pi_3\alpha_{34}$	*	$\pi_5\alpha_{45}$	$\pi_6\alpha_{46}$	$\pi_7\alpha_{47}$
5	*UG*	$\pi_1\alpha_{15}$	$\pi_2\alpha_{25}$	$\pi_3\alpha_{35}$	$\pi_4\alpha_{45}$	*	$\pi_6\alpha_{56}$	$\pi_7\alpha_{57}$
6	*CG*	$\pi_1\alpha_{16}$	$\pi_2\alpha_{26}$	$\pi_3\alpha_{36}$	$\pi_4\alpha_{46}$	$\pi_5\alpha_{56}$	*	$\pi_7\alpha_{67}$
7	*MM*	$\pi_1\alpha_{17}$	$\pi_2\alpha_{27}$	$\pi_3\alpha_{37}$	$\pi_4\alpha_{47}$	$\pi_5\alpha_{57}$	$\pi_6\alpha_{67}$	*

Model 7D

		1	2	3	4	5	6	7
		AU	*GU*	*GC*	*UA*	*UG*	*CG*	*MM*
1	*AU*	*	$\pi_2\alpha_s$	$\pi_3\alpha_d$	$\pi_4\beta$	$\pi_5\beta$	$\pi_6\beta$	$\pi_7\gamma$
2	*GU*	$\pi_1\alpha_s$	*	$\pi_3\alpha_s$	$\pi_4\beta$	$\pi_5\beta$	$\pi_6\beta$	$\pi_7\gamma$
3	*GC*	$\pi_1\alpha_d$	$\pi_2\alpha_s$	*	$\pi_4\beta$	$\pi_5\beta$	$\pi_6\beta$	$\pi_7\gamma$
4	*UA*	$\pi_1\beta$	$\pi_2\beta$	$\pi_3\beta$	*	$\pi_5\alpha_s$	$\pi_6\alpha_d$	$\pi_7\gamma$
5	*UG*	$\pi_1\beta$	$\pi_2\beta$	$\pi_3\beta$	$\pi_4\alpha_s$	*	$\pi_6\alpha_s$	$\pi_7\gamma$
6	*CG*	$\pi_1\beta$	$\pi_2\beta$	$\pi_3\beta$	$\pi_4\alpha_d$	$\pi_5\alpha_s$	*	$\pi_7\gamma$
7	*MM*	$\pi_1\gamma$	$\pi_2\gamma$	$\pi_3\gamma$	$\pi_4\gamma$	$\pi_5\gamma$	$\pi_6\gamma$	*

FIGURE 1.—Definition of the rate matrix for models 7A and 7D.

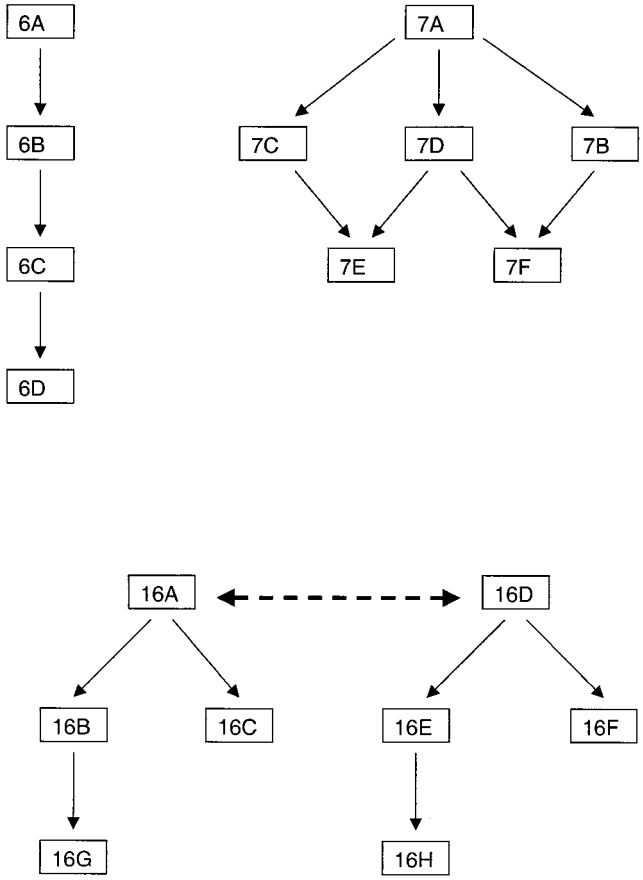

FIGURE 2.—Relationships between the three groups of models. Each of the solid arrows indicates that the model at the head of the arrow is nested within the model at the tail. Statistical tests are made for each pair of models related in this way. The dashed arrow indicates that a statistical test is made between the models that does not require the models to be nested.

is shown in Figure 2, where an arrow indicates that the model at the head of the arrow is nested in the model at the tail.

The 6-state models are similar to the 7-state models, except that they lack the MM state. Model 6A is the general reversible 6-state model. Models 6B and 6C are obtained by eliminating the MM state from models 7D and 7F, respectively. Model 6D is obtained by setting double transitions to zero in 6C. These 6-state models form a simple nested series, as shown in Figure 2. Models 6C and 6D were originally proposed by TILLIER (1994).

In principle we could define a general reversible 16-state model with 134 free parameters; however, we do not believe such a complex model would be practical, and we have not attempted this. To facilitate comparison between the 6- and 7-state models and the 16-state models we have introduced models 16A and 16C, which are similar in spirit to model 7D. The full matrix for 16A is shown in Figure 3. There are 16 frequency parameters for the 16 states. The rate parameters for the 6 principal states are the same as those in 7D. Rates of single substitutions to and from mismatch states are controlled by a parameter γ, and rates of single substitutions between mismatch states are controlled by a parameter ε. Model 16C further simplifies the treatment of mismatches by setting the frequencies of all 10 mismatches to a single parameter π_m. Models 16A and 16C are the only 16-state models that allow a nonzero rate of double substitutions.

In the model proposed by SCHÖNIGER and VON HAESELER (1994), the rates are defined as $r_{ij} = \pi_j$ if states i and j differ by a single substitution and zero otherwise. To apply Equation 4 we introduce an extra factor μ, so that $r_{ij} = \mu\pi_j$, and then scale μ to satisfy (4). This model, termed 16B, is identical to that of SCHÖNIGER and VON HAESELER (1994), except for the timescale. Since the timescale does not affect the maximum-likelihood value, the statistical tests on 16B also apply to the model as originally defined.

MUSE (1995) proposed three models. The simplest, 16H, has only one free parameter after scaling the time. The second model, 16E, was termed the "modified Hasegawa-Kishino-Yano (HKY) model" as it has several features in common with the model of HASEGAWA *et al.* (1985) for single site evolution. It distinguishes between transition and transversion rates and allows the frequencies of the four bases to differ. The third model, 16F, is similar to 16E, but differs in its treatment of GU and UG pairs. In 16E, GU and UG pairs behave exactly as mismatches, whereas in 16F they behave exactly as Watson-Crick pairs. In natural RNA sequences, GU and UG frequencies are considerably lower than the Watson-Crick states, but considerably greater than the mismatch states. We have therefore introduced a model 16D by adding an extra parameter ϕ that enables the GU and UG pairs to have intermediate frequencies. The full rate matrix for 16D is given in Figure 4. The equilibrium frequency π_{XY} of a base pair XY is related to the frequencies of the two bases π_X and π_Y by $\pi_{XY} = \kappa\pi_X\pi_Y\lambda^2$ if X and Y form a Watson-Crick pair; $\pi_{XY} = \kappa\pi_X\pi_Y\phi^2$ if X and Y are GU or UG; $\pi_{XY} = \kappa\pi_X\pi_Y$ if X and Y form a mismatch. The constant κ is determined by

$$1/\kappa = 2(\lambda^2 - 1)(\pi_A\pi_U + \pi_G\pi_C) + 2(\phi^2 - 1)\pi_G\pi_U + 1. \quad (6)$$

This model reduces to 16E if $\phi = 1$ and to 16F if $\phi = \lambda$. In addition, model 16E reduces to model 16H if all the four

	AU	*GU*	*GC*	*UA*	*UG*	*CG*	*AA*	*AC*	*AG*	*CA*	*CC*	*CU*	*GA*	*GG*	*UC*	*UU*
AU	*	$\alpha_s\pi_2$	$\alpha_d\pi_3$	$\beta\pi_4$	$\beta\pi_5$	$\beta\pi_6$	$\gamma\pi_7$	$\gamma\pi_8$	$\gamma\pi_9$	0	0	$\gamma\pi_{12}$	0	0	0	$\gamma\pi_{16}$
GU	$\alpha_s\pi_1$	*	$\alpha_s\pi_3$	$\beta\pi_4$	$\beta\pi_5$	$\beta\pi_6$	0	0	0	0	0	$\gamma\pi_{12}$	$\gamma\pi_{13}$	$\gamma\pi_{14}$	0	$\gamma\pi_{16}$
GC	$\alpha_d\pi_1$	$\alpha_s\pi_2$	*	$\beta\pi_4$	$\beta\pi_5$	$\beta\pi_6$	0	$\gamma\pi_8$	0	0	$\gamma\pi_{11}$	0	$\gamma\pi_{13}$	$\gamma\pi_{14}$	$\gamma\pi_{15}$	0
UA	$\beta\pi_1$	$\beta\pi_2$	$\beta\pi_3$	*	$\alpha_s\pi_5$	$\alpha_d\pi_6$	$x\pi_7$	0	0	$\gamma\pi_{10}$	0	0	$\gamma\pi_{13}$	0	$\gamma\pi_{15}$	$\gamma\pi_{16}$
UG	$\beta\pi_1$	$\beta\pi_2$	$\beta\pi_3$	$\alpha_s\pi_4$	*	$\alpha_s\pi_6$	0	0	$\gamma\pi_9$	0	0	0	0	$\gamma\pi_{14}$	$\gamma\pi_{15}$	$\gamma\pi_{16}$
CG	$\beta\pi_1$	$\beta\pi_2$	$\beta\pi_3$	$\alpha_d\pi_4$	$\alpha_s\pi_5$	*	0	0	$\gamma\pi_9$	$\gamma\pi_{10}$	$\gamma\pi_{11}$	$\gamma\pi_{12}$	0	$\gamma\pi_{14}$	0	0
AA	$\gamma\pi_1$	0	0	$\gamma\pi_4$	0	0	*	$\varepsilon\pi_8$	$\varepsilon\pi_9$	$\varepsilon\pi_{10}$	0	0	$\varepsilon\pi_{13}$	0	0	0
AC	$\gamma\pi_1$	0	$\gamma\pi_3$	0	0	0	$\varepsilon\pi_7$	*	$\varepsilon\pi_9$	0	$\varepsilon\pi_{11}$	0	0	0	$\varepsilon\pi_{15}$	0
AG	$\gamma\pi_1$	0	0	0	$\gamma\pi_5$	$\gamma\pi_6$	$\varepsilon\pi_7$	$\varepsilon\pi_8$	*	0	0	0	0	$\varepsilon\pi_{14}$	0	0
CA	0	0	0	$\gamma\pi_4$	0	$\gamma\pi_6$	$\varepsilon\pi_7$	0	0	*	$\varepsilon\pi_{11}$	$\varepsilon\pi_{12}$	$\varepsilon\pi_{13}$	0	0	0
CC	0	0	$\gamma\pi_3$	0	0	$\gamma\pi_6$	0	$\varepsilon\pi_8$	0	$\varepsilon\pi_{10}$	*	$\varepsilon\pi_{12}$	0	0	$\varepsilon\pi_{15}$	0
CU	$\gamma\pi_1$	$\gamma\pi_2$	0	0	0	$\gamma\pi_6$	0	0	0	$\varepsilon\pi_{10}$	$\varepsilon\pi_{11}$	*	0	0	0	$\varepsilon\pi_{16}$
GA	0	$\gamma\pi_2$	$\gamma\pi_3$	$\gamma\pi_4$	0	0	$\varepsilon\pi_7$	0	0	$\varepsilon\pi_{10}$	0	0	*	$\varepsilon\pi_{14}$	0	0
GG	0	$\gamma\pi_2$	$\gamma\pi_3$	0	$\gamma\pi_5$	$\gamma\pi_6$	0	0	$\varepsilon\pi_9$	0	0	0	$\varepsilon\pi_{13}$	*	0	0
UC	0	0	$\gamma\pi_3$	$\gamma\pi_4$	$\gamma\pi_5$	0	0	$\varepsilon\pi_8$	0	0	$\varepsilon\pi_{11}$	0	0	0	*	$\varepsilon\pi_{16}$
UU	$\gamma\pi_1$	$\gamma\pi_2$	0	$\gamma\pi_4$	$\gamma\pi_5$	0	0	0	0	0	0	$\varepsilon\pi_{12}$	0	0	$\varepsilon\pi_{15}$	*

FIGURE 3.—Definition of the rate matrix for model 16A.

bases have equal frequency and there is no difference between transitions and transversions.

RZHETSKY (1995) also proposed a 16-state model, labeled 16G. This model is a simplification of 16B, which is obtained by setting the frequencies of the four Watson-Crick states to be equal, the frequencies of GU and UG to be equal, and the frequencies of the mismatch states to be equal. The relationship between the 16-state models is shown in Figure 2.

The maximum-likelihood calculation: A set of eubacterial small subunit ribosomal RNA sequences was obtained from the rRNA database (VAN DE PEER *et al.* 1998). The object is not to test the tree topology but to test the evolutionary models; therefore we chose five species for which the rRNA phylogeny is well established. *Bacillus subtilis* is a member of the gram-positive bacteria and is an outgroup to the remaining four proteobacteria. *Rhodomicrobium vannielii* and *Sphingomonas capsulata* are examples of the alpha proteobacteria subdivision, while *Escherichia coli* and *Pseudomonas aeruginosa* are examples of the gamma proteobacteria subdivision. The tree is shown in Figure 5 in unrooted form. This tree is the one given by both the National Center for Biotechnology Information taxonomy browser (LEIPE and SOUSSOV 1995) and the Ribosomal Database Project (MAIDAK *et al.* 1999). In addition we confirmed the tree topology using the dnaml and dnapars programs in the Phylip package (FELSENSTEIN 1995).

The sequence alignments and the positions of the conserved secondary structures given by VAN DE PEER *et al.* (1998) were taken to be correct. This analysis uses only the paired regions of the sequence and ignores the loop regions. We have previously analyzed the frequencies of base pairs in a set of over 400 sequences from the same database that includes a representative from each genus of eubacteria (HIGGS 2000). In general it is found that GU and UG pairs have low frequencies, as would be expected if they are slightly deleterious. However, a small fraction of the paired sites have a GU or UG pair in a majority of sequences. This suggests that these pairs are positively selected when they occur at particular points in the structure (see also ROUSSET *et al.* 1991; GAUTHERET *et al.* 1995). Such sites violate the assumptions of all the models discussed in this article, which treat GU and UG pairs as slightly deleterious alternatives to Watson-Crick pairs. For this reason, paired sites at which the observed GU frequency or UG frequency in the full set of sequences was >50% were excluded from the analysis carried out in this article. The analysis was carried out on 296 pairs (*i.e.*, 592 sites) that satisfied these criteria. The effect of inclusion *vs.* exclusion of these pairs is discussed in more

	AU	*GU*	*GC*	*UA*	*UG*	*CG*	*AA*	*AC*	*AG*	*CA*	*CC*	*CU*	*GA*	*GG*	*UC*	*UU*
AU	*	$\alpha\pi_G\phi/\lambda$	0	0	0	0	$\beta\pi_A/\lambda$	$\alpha\pi_C/\lambda$	$\beta\pi_G/\lambda$	0	0	$\beta\pi_C/\lambda$	0	0	0	$\beta\pi_U/\lambda$
GU	$\alpha\pi_A\lambda/\varphi$	*	$\alpha\pi_C\lambda/\phi$	0	0	0	0	0	0	0	0	$\beta\pi_C/\phi$	$\beta\pi_A/\phi$	$\beta\pi_G/\phi$	0	$\beta\pi_U/\phi$
GC	0	$\alpha\pi_U\phi/\lambda$	*	0	0	0	0	$\alpha\pi_A/\lambda$	0	0	$\beta\pi_C/\lambda$	0	$\beta\pi_A/\lambda$	$\beta\pi_G/\lambda$	$\beta\pi_U/\lambda$	0
UA	0	0	0	*	$\alpha\pi_G\phi/\lambda$	0	$\beta\pi_A/\lambda$	0	0	$\alpha\pi_C/\lambda$	0	0	$\beta\pi_G/\lambda$	0	$\beta\pi_C/\lambda$	$\beta\pi_U/\lambda$
UG	0	0	0	$\alpha\pi_A\lambda/\varphi$	*	$\alpha\pi_C\lambda/\phi$	0	0	$\beta\pi_A/\phi$	0	0	0	0	$\beta\pi_G/\phi$	$\beta\pi_C/\phi$	$\beta\pi_U/\phi$
CG	0	0	0	0	$\alpha\pi_U\phi/\lambda$	*	0	0	$\beta\pi_A/\lambda$	$\alpha\pi_A/\lambda$	$\beta\pi_C/\lambda$	$\beta\pi_U/\lambda$	0	$\beta\pi_G/\lambda$	0	0
AA	$\beta\pi_U\lambda$	0	0	$\beta\pi_U\lambda$	0	0	*	$\beta\pi_C$	$\alpha\pi_G$	$\beta\pi_C$	0	0	$\alpha\pi_G$	0	0	0
AC	$\alpha\pi_U\lambda$	0	$\alpha\pi_G\lambda$	0	0	0	$\beta\pi_A$	*	$\beta\pi_G$	0	$\beta\pi_C$	0	0	0	$\beta\pi_C$	0
AG	$\beta\pi_U\lambda$	0	0	0	$\beta\pi_U\phi$	$\beta\pi_C\lambda$	$\alpha\pi_A$	$\beta\pi_C$	*	0	0	0	0	$\alpha\pi_G$	0	0
CA	0	0	0	$\alpha\pi_U\lambda$	0	$\alpha\pi_G\lambda$	$\beta\pi_A$	0	0	*	$\beta\pi_C$	$\beta\pi_U$	$\beta\pi_G$	0	0	0
CC	0	0	$\beta\pi_G\lambda$	0	0	$\beta\pi_G\lambda$	0	$\beta\pi_A$	0	$\beta\pi_A$	*	$\alpha\pi_U$	0	0	$\alpha\pi_U$	0
CU	$\beta\pi_A\lambda$	$\beta\pi_G\phi$	0	0	0	$\beta\pi_G\lambda$	0	0	0	$\beta\pi_A$	$\alpha\pi_C$	*	0	0	0	$\alpha\pi_U$
GA	0	$\beta\pi_U\phi$	$\beta\pi_C\lambda$	$\beta\pi_U\lambda$	0	0	$\alpha\pi_A$	0	0	$\beta\pi_C$	0	0	*	$\alpha\pi_G$	0	0
GG	0	$\beta\pi_U\phi$	$\beta\pi_C\lambda$	0	$\beta\pi_U\phi$	$\beta\pi_C\lambda$	0	0	$\alpha\pi_A$	0	0	0	$\alpha\pi_A$	*	0	0
UC	0	0	$\beta\pi_G\lambda$	$\beta\pi_A\lambda$	$\beta\pi_G\phi$	0	0	$\beta\pi_A$	0	0	$\alpha\pi_C$	0	0	0	*	$\alpha\pi_U$
UU	$\beta\pi_A\lambda$	$\beta\pi_G\phi$	0	$\beta\pi_A\lambda$	$\beta\pi_G\phi$	0	0	0	0	0	0	$\alpha\pi_C$	0	0	$\alpha\pi_C$	*

FIGURE 4.—Definition of the rate matrix for model 16D.

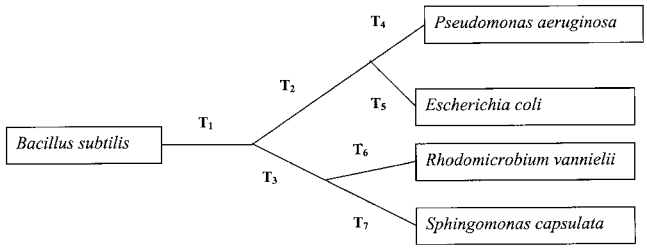

FIGURE 5.—The phylogenetic tree used for testing the models.

detail by HIGGS (2000), using a different method of sequence analysis. Elimination of these unusual pairs does not favor any one model over any other and therefore does not influence the model selection criteria in this article.

When testing the models we assumed that the topology of the tree was fixed, but not the branch lengths. For each model we obtained the parameter set that maximized the likelihood of observation of the given sequences on the given tree. Methods of calculating the likelihood are discussed by FELSENSTEIN (1981), SWOFFORD *et al.* (1996), and LI (1997). In our case, the adjustable parameters consist of the parameters defining the model (as defined in the previous section and in Table 1) and the seven branch lengths in the unrooted tree (as shown in Figure 5).

When using the six-state models, a decision had to be made as to what to do with the mismatch states that do actually occur in the real sequences. Positions where there are many mismatches do not count as paired sites in the consensus secondary structure in the database, and therefore these sites are not considered. However, there remain sporadic mismatches occurring rather randomly throughout the sequence alignment in positions where almost all the other sequences are properly paired. We did not wish to discard the complete column of data from the sequence alignment, simply because a single sequence in the set had a mismatch at that position. Therefore, any mismatch states that occurred were treated as being the same as the Watson-Crick pair that has the same 5′ base; *i.e.*, AA, AC, and AG are treated as AU, while GA and GG are treated as GC, etc. In this way the mismatch states are distributed roughly equally between the four main states.

The maximum-likelihood values of the parameters for each model were calculated as follows. An initial estimate of the state frequencies was obtained by measuring the average frequencies in the data. The initial rate matrix was estimated by calculating the frequency of changes of states between pairs of sequences. The individual rates were estimated by calculating the number of differences of each type between sequence pairs and normalizing using Equation 4. The rates were then averaged over all pairs of sequences. The initial times were estimated by finding the maximum-likelihood divergence time for pairs of sequences given the initial estimation of the rate matrix.

Once initial values of the parameters had been estimated an iterative algorithm was used to calculate the maximum likelihood of the data. The algorithm was iterated until the maximum likelihood converged (typically 1000–2000 iterations but this depends on the model being studied). At each iteration a small random change was made to a single rate parameter and a single frequency parameter. $P_{ij}(t)$ was solved by numerical integration of Equation 1 and the maximum-likelihood values of the times were calculated by a hill climbing method in time-space. If the new value of the maximum likelihood was greater than the best value found so far, the new parameter values were retained. Otherwise, the changes made to the rates and frequencies were discarded. Checks were made that the algorithm converged to the same optimal parameter values from different starting points. For some of the models we also checked the results from our optimization algorithm against results obtained using a simple simulated annealing algorithm (KIRKPATRICK *et al.* 1983; KIRKPATRICK 1984) to locate the maximum-likelihood values of times and model parameters. No significant differences were found between the results of our hill climbing algorithm and the simulated annealing algorithm, indicating that the hill climbing algorithm was indeed adequate for the task. For each set of rates, the optimum times were calculated to the nearest 0.001 unit for the 6- and 7-state models, and to the nearest 0.01 unit for the 16-state models, since these required greater computer time to calculate.

Statistical tests: Having estimated the maximum-likelihood parameter set for each of the models, we have a likelihood value L for each model, which is the likelihood of the given sequences on the given tree assuming the optimized values of the parameters. In general a larger L indicates a better fit of the model to the data; however, in choosing between models it is not sufficient to select the one with the highest L. There is a tendency for models with more parameters to give higher L values. In fact, in cases where models are nested, the one with the larger number of parameters will always give the higher L. However, the use of additional parameters is sometimes not justified statistically.

A criterion often used to compare models is the Akaike information criterion (AIC; LINHART and ZUCCHINI 1986), defined as AIC = $-\ln L$ + number of free parameters (or sometimes as twice this). Theory suggests that the model with the lowest AIC is to be preferred. The AIC thus penalizes models with too many parameters.

The likelihood-ratio test (LRT) makes a direct comparison between two models H_0 and H_1, where H_0 is the simpler model nested within the more general model H_1. If L_0 is the likelihood of the data according to H_0, and L_1 is the likelihood of the data according to H_1, then the LRT proceeds by calculating the logarithm of the likelihood ratio: $\delta = \ln(L_1/L_0)$. Even if H_0 is perfectly valid δ will still be positive, since H_1 has a larger number of parameters with which to fit the data. Theory shows (LINHART and ZUCCHINI 1986) that if the simpler model is true then 2δ will be distributed according to a χ^2 distribution with the number of degrees of freedom equal to the difference in the number of parameters between the two models. As a significance test we can calculate the probability P that 2δ from the χ^2 distribution will be greater than the observed value of 2δ. A small P indicates that the observed result is unlikely to occur by chance if H_0 is an adequate model, and hence that H_1 is a significantly better fit to the data. A large P indicates that introduction of the extra parameters in H_1 does not significantly improve the fit given by H_0.

The proofs of the AIC and LRT rely on the asymptotic assumption, *i.e.*, that there is a very large amount of data. For the case of phylogenetic methods this means that the sequences should be extremely long. GOLDMAN (1993) has cast doubt on the validity of these asymptotic tests for analysis of biological sequences of realistic length and has proposed that Cox's test should be used instead. This test (COX 1962) works by calculating δ for two models as in the LRT. Although the distribution of δ cannot be calculated analytically if the asymptotic results do not apply, it can still be simulated numerically. After fitting the real data and calculating the maximum-likelihood parameters according to both models, a large number of sets (typically 100) of simulated sequences are generated using the model H_0. Each of the simulated sets is then refitted using both models, and a histogram of δ values is obtained for the simulated sets. The significance probability P is the

TABLE 2

Maximum-likelihood and AIC values

Model ID no.	$-\ln L$	AIC
6A	1154.95	1180.95
6B	1180.76	1194.76
6C	1185.00	1196.00
6D	1201.56	1211.56
7A	1200.41	1233.41
7B	1205.27	1235.27
7C	1222.51	1244.51
7D	1228.87	1244.87
7E	1243.36	1257.36
7F	1233.07	1246.07
NP	911.60	NA
16A	1246.24	1272.24
16B	1274.12	1296.12
16C	1261.03	1278.03
16D	1274.60	1287.60
16E	1291.04	1303.04
16F	1308.42	1320.42
16G	1850.62	1859.10
16H	1365.52	1373.52

NA, not applicable.

fraction of the simulated sets that have δ higher than the observed value for the real data. This test is to be preferred over the other two since it involves no assumptions on the distribution of δ; however, it requires very much greater computer time. In cases where the tree topology is not known, it is usual to estimate the maximum-likelihood tree topology for both the real data and each set of simulated data when doing the Cox test. We did not do this. When fitting the simulated data the tree topology was kept fixed while the branch lengths and rate parameters were varied, in the same way as was done for fitting the real data. As long as the same procedure is used for fitting the real and the simulated data, the statistical test is valid.

The adequacy of the most general models cannot be tested by comparison to any of the other models. However, they can be compared to a nonparametric (NP) model (Goldman 1993). A position along the aligned sequences will exhibit a combination C of states called a pattern. This pattern occurs with a certain frequency in the aligned sequences. The maximum-likelihood solution for the NP model takes the form

$$L = \prod_C (N_C/N)^{N_C}, \tag{7}$$

where N_C is the number of occurrences of the pattern C in the aligned sequences. We carried out the Cox test for model 7A *vs.* the seven-state NP model.

RESULTS

The values of the log-likelihood and the AIC statistical test as well as the optimal phylogenetic tree times for each model are shown in Table 2. Since the lowest AIC is to be preferred, we see that the general reversible 6-state model 6A is the best of the 6-state models, the general reversible 7-state model is the best of the 7-state models, and model 16A is the best of the 16-state models. Note that AIC values cannot be compared between the three groups of models because the states are different, and the sequence data are treated in a different way. The more possible states there are, the smaller the likelihood of change between any two individual states. Hence models with more states have lower likelihoods and higher AICs, but this does not tell us anything about the relative merits of the three groups of models.

Table 3 shows the outcomes of the statistical tests between the model pairs. Tests are carried out for pairs of models linked by arrows in Figure 2. The number of degrees of freedom (d.f.) for each test is also given. The significance values for the LRT are P values obtained from a χ^2 table. For all but two of the Cox tests, 100 replicates were performed. In many cases there were no simulated δ values greater than the real value, and we quote this as a significance of $P < 1/100$. For cases where there were n simulated values higher than the real value out of m replicates we have quoted $P < (n + 1)/m$. Where the level of significance P was small and the outcome of the Cox test critical in affecting our final choice of model, we have performed a larger number of replicates to reduce the expected error in the estimated level of significance obtained from the Cox test. This was the case for the comparison of 16A and 7A, where we performed 270 replicates, and for the comparison of 7B and 7A, where we performed 200 replicates.

We discuss the 6- and 7-state models together, since the conclusions are very similar. The question of whether base pair reversal symmetry is valid is addressed by the comparison 7B-7A in Table 3. The more general model gives a significantly better fit with $P < 0.025$ according to the LRT and a marginally significant better fit with $P < 0.055$ according to the Cox test. The distributions are shown in Figure 6. Once again the more general model has the lower AIC and consequently overall we consider 7A to be a better model than 7B. The same question is also addressed by the comparisons 7F-7D and 6C-6B. In both these cases the more general model has the lower AIC, suggesting that we should not make the assumption of base pair symmetry. However, the two tests for 7F-7D and 6C-6B give $P = 4$ or 5%, which is only marginally significant. Higgs (2000) has also observed that GC frequency is considerably higher than CG frequency in several large datasets of RNA sequences, but it is not clear what the cause of this could be. Thus, taken together, the results suggest that there may be some effect present that breaks the symmetry between these apparently equivalent states, but in absence of a theory as to why this should happen, and in view of the borderline statistical significance, it is not possible to rule out that the apparent loss of base pair reversal symmetry is a result of chance alone.

The comparison 7C-7A tests whether double substitution rates may be set to zero. The answer is clearly no: 7A is a better fit than 7C with very significant P values according to both pairwise tests. Also, the AIC is lower for 7A than 7C; hence all three tests are in agreement.

TABLE 3

Likelihood-ratio and Cox's tests

Model comparison					
H_0	H_1	d.f.	$\delta = \ln L_1 - \ln L_0$	P(LRT)	P(Cox)
6B	6A	12	25.81	<0.001	<1/100
6C	6B	3	4.24	<0.04	<5/100
6D	6C	1	16.56	<0.001	<1/100
7B	7A	3	4.83	<0.025	<11/200
7C	7A	11	22.10	<0.001	<1/100
7D	7A	17	28.46	<0.001	<1/100
7E	7C	6	20.85	<0.001	<1/100
7E	7D	2	14.49	<0.001	<1/100
7F	7B	17	27.82	<0.001	<1/100
7F	7D	3	4.20	<0.042	<4/100
7A	NP	NA	288.81	NA	<1/100
16B	16A	4	27.88	<0.001	—
16C	16A	9	14.80	<0.007	—
16G	16B	13	576.51	<0.001	—
16E	16D	1	16.44	<0.001	—
16F	16D	1	33.82	<0.001	—
16H	16E	4	58.78	<0.001	—
16D	16A	NA	28.36	NA	<1/100
16A	16D	NA	−28.36	NA	<100/100
16A	7A	NA	45.82	NA	<21/270
16A	7D	NA	17.37	NA	<28/100
16A	6A	NA	92.29	NA	<45/100
16A	6B	NA	65.48	NA	<59/100

NA, not applicable.

The histogram of δ values from Cox's test is shown in Figure 6, in comparison to the $\frac{1}{2}\chi^2$ distribution (expected according to the LRT) and to the real value in the data (denoted by an arrow). It can be seen that the real value is completely outside the range of the distribution, hence P is very much <1% and it would require many more than 100 replicates to estimate a true P value. The question of zero *vs.* nonzero rates of double substitutions is also addressed by the comparison of 7E-7D, in which the parameters α_d and β are set to zero, and 6D-6C in which α_d is set to zero (note that β cannot be zero in the 6-state model, otherwise the states are divided into two inaccessible subsets of three). In both these two comparisons the model with the nonzero rates gives a much better fit (very low P values) and also gives a lower AIC.

Model 7D is of interest because it is the most complex of the models for which an analytical solution is available (Tillier and Collins 1998). Model 7D has 17 fewer parameters than 7A, which gives it a large advantage on the AIC test. Nevertheless, the general model has a much lower AIC. It can also be seen (Figure 6 and Table 3) that 7A gives a highly significant improvement over 7D.

Even though the general model 7A outperforms the alternatives, comparison with the nonparametric model in Table 3 indicates that it is still not an adequate description of the data. A likely reason for this is that we have assumed equal rates of substitution at each site. Relaxing this assumption may give better models but also may increase the number of parameters to fit. Comparison with the NP model is a very stringent test, and it seems unlikely that any reasonably tractable model would ever pass the test when applied to real sequence data. This test is rather unhelpful since it tends to reject models without proposing any better alternative.

One point that can be seen for all the results with the 6- and 7-state models is that the Cox test with numerical simulation of the δ distribution always gives very similar results to the much simpler LRT. The conclusion reached on significance is the same in every case, and the simulated distributions differ rather little (if at all) from the $\frac{1}{2}\chi^2$ distributions assumed by the LRT. Therefore it would seem that the LRT is an appropriate test for analysis of these sequences despite the original doubts that there may not be sufficient data to be in the asymptotic regime. It was also found that the Cox test on the 16-state models was extremely slow (note that every comparison requires 200 runs of the maximum-likelihood program). For these two reasons we did not perform a Cox test for every pair of 16-state models and decided to rely on the results of the LRT. The Cox test was only performed for the 16D-16A and 16A-16D comparisons, for which the LRT is not valid because the models are not nested.

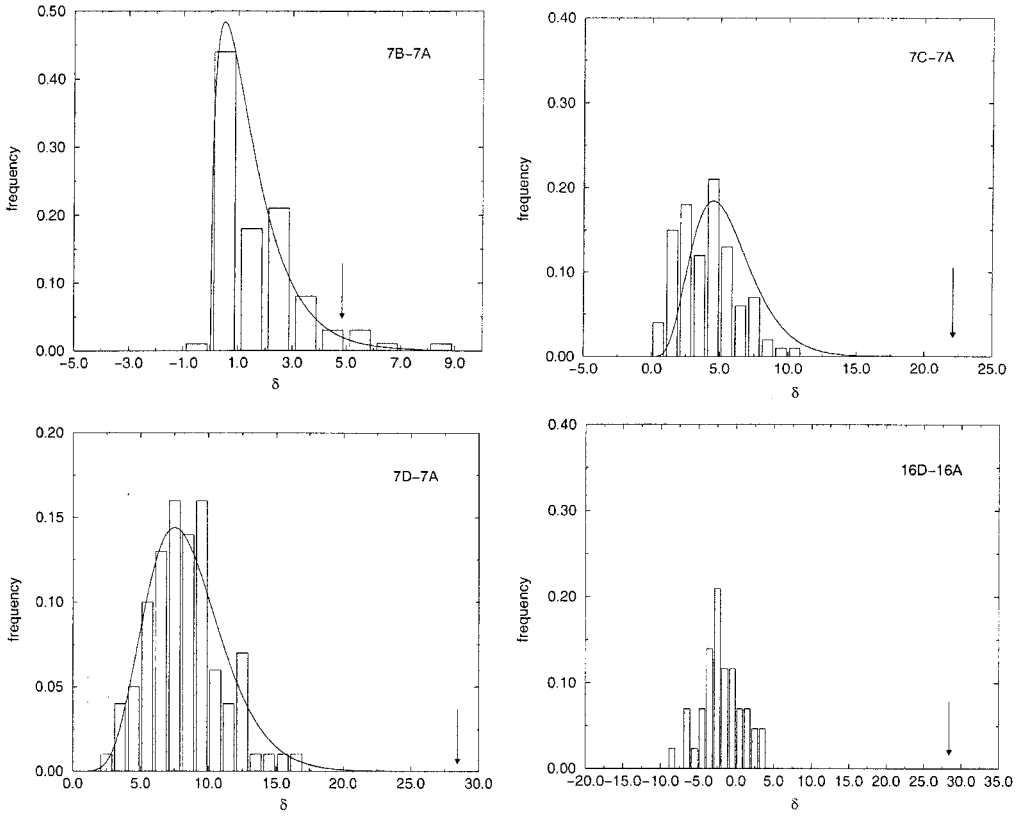

FIGURE 6.—Distributions of δ for four of the pairwise comparisons between models. The histograms indicate distributions simulated by the Cox test. The continuous distributions in the first three graphs are $\frac{1}{2}\chi^2$ distributions. Observed values of δ are indicated by arrows.

The two best 16-state models according to the AIC are 16A and 16C. These are the two models that have nonzero rates of double substitution. The model of SCHÖNIGER and VON HAESELER (1994), 16B, is nested in 16A (by setting α_d and β to zero and $\alpha_s = \gamma = \varepsilon$). Model 16A is significantly better than 16B by the LRT and the AIC. Model 16C also performs less well than 16A, although it is still better than any of the other 16-state models. In model 16C the frequencies of all the mismatch states are set equal to one another. This result shows that there is a significant improvement in the likelihood if all the mismatch frequencies are allowed to vary independently.

Of the two models 16E and 16F proposed by MUSE (1995), 16E is the better. We proposed model 16D as a generalization of both of these two to allow the LRT to be performed. The comparisons 16E-16D and 16F-16D by the LRT indicate that 16D is a significant improvement over both the others. The maximum-likelihood values of the parameters in 16D were $\lambda = 13.41$ and $\phi = 2.93$. The fact that ϕ is closer to 1 than to λ indicates that the GU and UG pairs are closer in behavior to the mismatch pairs than to Watson-Crick pairs according to this model.

The AIC suggests that all three models 16D, 16E, and 16F perform less well than 16A and 16C. This cannot be checked by the LRT because the models are not nested. We therefore performed a Cox test between 16D and 16A [in Table 3, with the result that 16A is significantly better than 16D ($P < 0.01$)]. The simulated distribution of δ values is shown in Figure 6, from which it can be seen that the majority of simulated values are negative; *i.e.*, if 16D were the true model, then 16A would usually not fit the data better than 16D. In contrast, δ is positive for the real data, indicating that it is better explained by 16A. This test can also be performed in the reverse direction by using 16A to simulate the data. In this case δ is negative for the real data, and none of the simulated data give a δ as low as this (*i.e.*, $P < 100/100$). Thus if 16A were the true model then it would be highly unlikely that the difference in likelihood between the two models would be as high as it is. From this point of view we can conclude that 16A is not an entirely adequate description of the real data. Nevertheless it is still better than any of the alternatives considered.

Through use of the Cox test we can compare the performance of models with differing numbers of states. We have performed Cox tests between the best models in each class, *i.e.*, between 16A and 7A, and between 16A and 6A. We have also compared 16A with 7D and 6B, since these two models have exactly the same form of the matrix for the principal states as 16A. The tests give *P* values between 7 and 59%; *i.e.*, there is no evidence for rejecting 16A. The 7% value for the 16A-7A comparison is the smallest of these four values, which gives some degree of support to choosing 7A as the single best model. It is unfortunate that these tests can-

TABLE 4

Branch lengths

Model ID no.	T_1	T_2	T_3	T_4	T_5	T_6	T_7
6A	0.306	0.043	0.071	0.160	0.162	0.143	0.126
6B	0.296	0.050	0.079	0.162	0.153	0.144	0.132
6C	0.295	0.049	0.081	0.163	0.151	0.145	0.130
6D	0.597	0.120	0.172	0.380	0.351	0.337	0.277
7A	0.338	0.050	0.084	0.195	0.150	0.143	0.144
7B	0.335	0.050	0.086	0.193	0.144	0.146	0.140
7C	0.887	0.141	0.244	0.520	0.409	0.437	0.390
7D	0.330	0.053	0.090	0.194	0.132	0.145	0.149
7E	0.673	0.109	0.201	0.366	0.290	0.300	0.303
7F	0.329	0.052	0.093	0.195	0.130	0.147	0.147
16A	0.60	0.02	0.17	0.37	0.24	0.29	0.32
16B	0.98	0.16	0.32	0.67	0.43	0.49	0.49
16C	0.35	0.06	0.10	0.21	0.14	0.15	0.16
16D	0.85	0.12	0.27	0.49	0.39	0.41	0.33
16E	0.83	0.12	0.27	0.48	0.37	0.38	0.30
16F	1.06	0.15	0.36	0.64	0.47	0.52	0.49
16G	0.44	0.09	0.12	0.25	0.26	0.25	0.19
16H	0.78	0.04	0.15	0.41	0.32	0.33	0.31

not be performed in the reverse direction. If 7-state models are used to generate the data, then the data will contain MM states that cannot be treated properly in 16-state models or in 6-state models. If 6-state models are used to generate the data, there will be no mismatch states of any kind, hence there will be no point in fitting the simulated data to 7-state and 16-state models. Thus we cannot reject 16A in favor of any of the other models, but, equally, we cannot reject the others in favor of 16A. These tests are therefore not very helpful in deciding how many states to use.

When choosing a rate model for molecular phylogenetics, it is not just the likelihood value that needs to be taken into account, but also the speed of calculation. Sixteen-state models require considerably more time for likelihood calculations; therefore, since we have been unable to demonstrate a clear advantage for any of the 16-state models, we propose not to use them in our future phylogenetic studies. Our preference is for model 7A, from the results of the above tests, and model 7D, since this is the best of the models that is analytically solvable. The analytical solution again allows savings in computer time in ML methods and easy calculations of pairwise distances for use in distance matrix methods.

The ML values of the seven branch lengths of the tree are given in Table 4. Since the rates are normalized so that there is one substitution event per unit time on average, these branch lengths are the mean number of substitution events per base pair on each branch. Here an "event" is either a single or double substitution. The sets of branch lengths are almost identical for the first three 6-state models, but for 6D they are all much larger. This is because 6D disallows double substitutions; hence wherever there have been compensatory changes this counts as two changes in state. In the other models a double substitution usually occurs as a single step. The same effect shows up in the 7-state models, where the branch lengths are apparently much larger in models 7C and 7E where double substitution rates are zero. The times are also larger for 7-state models and corresponding 6-state models (*e.g.*, 7A and 6A) because the 7-state models count changes to and from mismatch states that are not counted in 6-state models. In terms of relative branch lengths, however, there is not much difference between the 6- and 7-state models. It is the relative values of the lengths that are most important, because the absolute values only have a meaning if a molecular clock calibration is used to assign times to the different branch points on the tree (which we have not attempted with these data). For the 16-state models, there is considerable difference in the branch lengths according to the different models. This is another reason why we prefer the 6- and 7-state models to the 16-state models.

DISCUSSION

The work in this article was begun as a statistical support for the study of RNA helix evolution by Higgs (2000), which analyzes large sequence datasets of tRNAs, rRNAs, and ribonuclease P RNAs using model 7A. The emphasis in the previous article was on understanding the mechanism of evolution and the effects of thermodynamics on sequence evolution, whereas this article emphasizes model selection. Since 7A has been shown to be one of the best models for the small set of

TABLE 5

Maximum-likelihood values of the parameters in model 7A

Base pair	Frequency	Mutability
GC	0.362	0.49
CG	0.298	0.63
AU	0.113	1.34
UA	0.180	1.03
GU	0.015	4.13
UG	0.019	7.07
MM	0.013	8.00

		Elements of the rate matrix						
	Base pair	1	2	3	4	5	6	7
		AU	GU	GC	UA	UG	CG	MM
1	AU	—	0.10	0.54	0.35	0.09	0.13	0.13
2	GU	0.73	—	2.0	0.00	0.90	0.38	0.12
3	GC	0.17	0.08	—	0.01	0.07	0.12	0.04
4	UA	0.22	0.00	0.03	—	0.20	0.46	0.12
5	UG	0.53	0.72	1.27	1.87	—	1.10	1.58
6	CG	0.05	0.02	0.15	0.28	0.07	—	0.06
7	MM	1.18	0.14	1.03	1.74	2.44	1.47	—

sequences used here, we compare these results to those in the previous article. The complete maximum-likelihood rate matrix for 7A is also shown in Table 5, together with the maximum-likelihood values of the frequencies and the mutabilities. The mutability of a state is the net rate of substitution from that state to all other states (*i.e.*, it is the negative of the element on the diagonal of the rate matrix). A mutability of 1.0 indicates that the state changes at the average rate for the whole sequence.

It can be seen that GC and CG pairs have low mutabilities and high frequencies, that AU and UA pairs have moderate frequencies and mutabilities, and that GU, UG, and MM pairs have low frequencies and high mutabilities. The order of the base frequencies is the same as that of the thermodynamic stability of stacking interactions. This shows that sequences are selected to increase the thermodynamic stability of the secondary structure. It can also be seen that rates of double transitions are large; *e.g.*, AU to GC and UA to CG are actually higher than the rates of the single transitions AU to GU and UA to UG. This shows that the single-step compensatory mutation mechanism is occurring frequently in these sequences. The five sequences used here are a subset of the rRNA-1 set used by Higgs (2000). The values of the parameters are close to those obtained for the rRNA-1 set but not identical because of the method of fitting the model to the data. The ML method used here has the advantage of allowing rigorous statistical tests, but is limited to a small number of sequences because of the computer time required. We consider it an open question which of these methods gives a more accurate estimation of the true rate parameters, since the method given in Higgs (2000) efficiently uses information from very many sequences, even though it is less rigorous than the ML method. We emphasize that the results obtained in this article support the conclusions given in the previous article regarding the relative rates of different types of substitution and the influence of thermodynamics on the substitution process.

The ranking order of the models given by the tests used here applies only to the particular set of sequences used and should therefore be treated with some caution. However, we believe that very similar selective effects are occurring in helical regions of many types of RNA, as was discussed fully by Higgs (2000). Therefore we expect that the models that perform best in this analysis will also generally be the best models for describing other RNAs with conserved structure. This will be tested in further work.

Although it is clear from the statistical analysis that general models such as 7A give significantly higher likelihoods than analytically tractable models like 7D, the analysis does not say why this is. However, it is not difficult to see why there should be a lack of symmetry in the rate matrix. First, real molecules may be subject to mutational bias, such that new bases do not arise with equal frequency. Second, there are many different selective effects. As discussed above, we believe that double substitutions are occurring via a single-step compensatory mechanism. The rate at which this occurs is strongly influenced by the fitness of the intermediate state. Double transitions between Watson-Crick pairs occur via GU and UG intermediates, whereas double transversions have to pass via true mismatches like GG or CC. True mismatches have a much greater destabilizing effect on

the helix and therefore presumably a much lower fitness. This accounts for the slow rate of double transversions with respect to double transitions. The thermodynamics of stacking interactions is far from symmetric (see Higgs 2000 and references therein); *i.e.*, not all mismatches are equivalent. Also, interactions occur between stacked pairs within a helix, not just between the two bases in a pair. Therefore, we might expect to see further correlations in substitutions involving more than just two sites. In view of all these effects we do not find it surprising that the rate matrix does not have a simple symmetry.

We became aware of two other models for RNA helix evolution after the analysis of the models in this article was almost complete. M. Schöniger and A. von Haeseler (personal communication) have considered another 16-state model that reduces to the HKY model for single sites (Hasegawa *et al.* 1985) in a limiting case. Apparently this model gives very similar results to their previous model (here called 16B). Also the new model does not allow double substitutions; therefore we would not expect it to do very well in relation to the models considered here. Second, there has been an interesting analytical treatment of the problem by Otsuka *et al.* (1997, 1999). This begins with all 16 states, but the mismatch states and the GU and UG states are treated as hidden variables, and the final model is presented in terms of only the 4 Watson-Crick states. The statistical tests used here would not be applicable to compare this 4-state model with all the other models with 6 or more states; therefore we did not consider it further. On first sight, however, it seems strange to us to remove GU and UG pairs from the model, since these seem to be an essential part of the problem from both a thermodynamic and evolutionary point of view. We also note that Knudsen and Hein (1999) have also used a general reversible 6-state model in their method for RNA secondary structure prediction.

In summary, the results presented here compare a large number of different models to describe RNA sequence evolution. The fact that general reversible models perform best shows that the real sequences are not well described by the simple symmetric assumptions used in some of the previously proposed models. The statistical tests show that we are justified in using large numbers of rate parameters in the model. The helical regions of RNAs are potentially very informative in phylogenetic work. Having a matrix that accurately models the substitution process in these regions should allow reliable phylogenetic inferences to be drawn. There are a large number of potential applications of these rate matrices in constructing phylogenies from RNA sequences.

This work was supported by the UK Biotechnology and Biological Sciences Research Council.

LITERATURE CITED

Cox, D. R., 1962 Further results on tests of families of alternate hypotheses. J. R. Stat. Soc. B **24:** 406–424.

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

Felsenstein, J., 1995 Phylip (Phylogeny Inference Package), version 3.5c.

Gatesy, J., C. Hayashi, R. DeSalle and E. Vrba, 1994 Rate limits for pairing and compensatory change: the mitochondrial ribosomal DNA of antelopes. Evolution **48:** 188–196.

Gautheret, D., D. Konings and R. R. Gutell, 1995 GU base pairing motifs in ribosomal RNA. RNA **1:** 807–814.

Goldman, N., 1993 Statistical tests of models of DNA substitution. J. Mol. Evol. **36:** 182–198.

Gutell, R. R., 1996 Comparative sequence analysis and the structure of 16S and 23S RNA, pp. 15–27 in *Ribosomal RNA: Structure, Evolution, Processing, and Function in Protein Biosynthesis*, edited by R. A. Zimmermann and A. E. Dahlberg. CRC Press, Boca Raton, FL.

Hasegawa, M., H. Kishino and T. Yano, 1985 Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22:** 160–174.

Higgs, P. G., 1998 Compensatory neutral mutations and the evolution of RNA. Genetica **102/103:** 91–101.

Higgs, P. G., 2000 RNA secondary structure: physical and computational aspects. Q. Rev. Biophys. **30**(3).

Iizuka, M., and M. Takefu, 1996 Average time until fixation of mutants with compensatory fitness interaction. Genes Genet. Syst. **71:** 167–173.

Kimura, M., 1985 The role of compensatory neutral mutations in molecular evolution. J. Genet. **64:** 7–19.

Kirby, D. A., S. V. Muse and W. Stephan, 1995 Maintenance of pre-mRNA secondary structure by epistatic selection. Proc. Natl. Acad. Sci. USA **92:** 9047–9051.

Kirkpatrick, S., 1984 Optimization by simulated annealing: quantitative studies. J. Stat. Phys. **34:** 975–986.

Kirkpatrick, S., C. D. Gelatt and M. P. Vecchi, 1983 Optimization by simulated annealing. Science **220:** 671–680.

Knudsen, B., and J. Hein, 1999 RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. Bioinformatics **15:** 446–454.

Leipe, D., and V. Soussov, 1995 NCBI taxonomy browser. http://www.ncbi.nlm.nih.gov/Taxonomy.

Li, W. H., 1997 *Molecular Evolution.* Sinauer Associates, Sunderland, MA.

Li, W. H., and X. Gu, 1996 Estimating evolutionary distances between DNA sequences. Methods Enzymol. **266:** 449–459.

Linhart, H., and W. Zucchini, 1986 *Model Selection.* John Wiley & Sons, New York.

Maidak, B. L., J. R. Cole, C. T. Parker, Jr., G. M. Garrity, N. Larsen *et al.*, 1999 A new version of the RDP (ribosomal database project). Nucleic Acids Res. **27:** 171–173 (http://www.cme.msu.edu/RDP/).

Muse, S., 1995 Evolutionary analyses of DNA sequences subject to constraints on secondary structure. Genetics **139:** 1429–1439.

Olsen, G. J., and C. R. Woese, 1993 Ribosomal RNA: a key to phylogeny. FASEB J. **7:** 113–123.

Otsuka, J., T. Nakano and G. Terai, 1997 A theoretical study of nucleotide changes under a definite functional constraint of forming stable base pairs in the stem regions of ribosomal RNAs. J. Theor. Biol. **184:** 171–186.

Otsuka, J., G. Terai and T. Nakano, 1999 Phylogeny of organisms investigated by the base pair changes in the stem regions of small and large ribosomal subunit RNAs. J. Mol. Evol. **48:** 218–235.

Rousset, F., M. Pelandakis and M. Solignac, 1991 Evolution of compensatory substitutions through GU intermediate state in Drosophila rRNA. Proc. Natl. Acad. Sci. USA **88:** 10032–10036.

Rzhetsky, A., 1995 Estimating substitution rates in ribosomal RNA genes. Genetics **141:** 771–783.

Schöniger, M., and A. von Haeseler, 1994 A stochastic model for the evolution of autocorrelated DNA sequences. Mol. Phylogenet. Evol. **3:** 240–247.

Stephan, W., 1996 The rate of compensatory evolution. Genetics **144:** 419–426.

Swofford, D. L., G. J. Olsen, P. J. Waddell and D. M. Hillis, 1996

Phylogenetic inference, pp. 407–514 in *Molecular Systematics*, Ed. 2, edited by D. M. Hillis. Sinauer Associates, Sunderland, MA.

Tillier, E. R. M., 1994 Maximum likelihood with multi-parameter models of substitution. J. Mol. Evol. **39:** 409–417.

Tillier, E. R. M., and R. A. Collins, 1995 Neighbour joining and maximum likelihood with RNA sequences: addressing the interdependence of sites. Mol. Biol. Evol. **12:** 7–15.

Tillier, E. R. M., and R. A. Collins, 1998 High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. Genetics **148:** 1993–2002.

van de Peer, Y., A. Caers, P. De Rijk and R. De Wachter, 1998 Database on the structure of small ribosomal subunit RNA. Nucleic Acids Res. **26:** 179–182 (http://rrna.uia.ac.be/ssu).

Vawter, L., and W. M. Brown, 1993 Rates and patterns of base change in the small subunit ribosomal RNA gene. Genetics **134:** 597–608.

Waddell, P. J., and M. A. Steel, 1997 General time-reversible distances with unequal rates across sites. Mol. Phylogenet. Evol. **8:** 398–414.

Wheeler, W. C., and R. J. Honeycutt, 1988 Paired sequence difference in ribosomal RNAs: evolutionary and phylogenetic implications. Mol. Biol. Evol. **5:** 90–96.

Woese, C. R., and N. C. Pace, 1993 Probing RNA structure, function and history by comparative analysis, pp. 91–117 in *The RNA World*, edited by R. F. Gesteland and J. F. Atkins. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Yang, Z. H., N. Goldman and A. Friday, 1994 Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. Mol. Biol. Evol. **11:** 316–324.

Communicating editor: S. Yokoyama