Copyright © 1995 by the Genetics Society of America

Evolutionary Analyses of DNA Sequences Subject to Constraints on Secondary Structure

Spencer V. Muse

Institute of Molecular Evolutionary Genetics, Department of Biology, The Pennsylvania State University, University Park, Pennsylvania 16802-5301

Manuscript received August 24, 1994
Accepted for publication November 16, 1994

ABSTRACT

Evolutionary models appropriate for analyzing nucleotide sequences that are subject to constraints on secondary structure are developed. The models consider the evolution of pairs of nucleotides, and they incorporate the effects of base-pairing constraints on nucleotide substitution rates by introducing a new parameter to extensions of standard models of sequence evolution. To illustrate some potential uses of the models, a likelihood-ratio test is constructed for the null hypothesis that two (prespecified) regions of DNA evolve independently of each other. The sampling properties of the test are explored via simulation. The test is then incorporated into a heuristic method for identifying the location of unknown stems. The test and related procedures are applied to data from ribonuclease P RNA sequences of bacteria.

EVOLUTIONARY analyses of DNA sequences typically assume, either explicitly or implicitly, that neighboring nucleotides have evolved independently of one another. To understand the evolutionary dynamics of a molecular sequence constrained by secondary structure, it is necessary to devise statistical methods that relax the assumption of independence among sites. In this work I present evolutionary models of nucleotide substitution that incorporate the effects of constraints on secondary structure. Among the appropriate uses of these models are phylogeny construction and distance estimation for sequences with known structural features. As an example, I demonstrate the utility of the models by constructing a likelihood-ratio test of the null hypothesis that two segments of a DNA sequence evolve independently of one another *vs.* the alternative hypothesis that they are constrained to form a stem structure.

The issue of secondary structure has been studied for a variety of molecules: ribosomal RNAs (FOX and WOESE 1975), transfer RNAs (SPRINZL *et al.* 1987), group I introns (CECH *et al.* 1983), small nuclear RNAs (SILICIANO *et al.* 1987) and precursor messenger RNAs (STEPHAN and KIRBY 1993). Both experimental and analytical methods have been used. The phylogenetic-comparative approach of NORMAN PACE and his colleagues with ribonuclease P RNA (PACE *et al.* 1989; BROWN and PACE 1991) is especially persuasive, and results have been reinforced with experimental findings (HAAS *et al.* 1991). All of the phylogenetic analyses rest on some fairly *ad hoc* guidelines. For example, a hypothesized stem is considered "proven" if two or more cases of compensatory changes are observed (BROWN and PACE 1991). Despite its usefulness in determining secondary structures, the phylogenetic-comparative approach is limited in the extent to which it can be used in evolutionary analyses such as phylogeny construction and distance estimation, because it does not use an explicitly-defined evolutionary model.

The fundamental idea used in the phylogenetic-comparative studies serves as the starting point of this work: constraints on secondary structure should leave evidence in the form of an altered pattern of nucleotide substitution in stem regions. Compensatory changes should be more frequent than expected by chance, and substitutions that discourage base pairing should be less frequent than expected under independence. One of the difficulties with current methodologies for identifying stems is that the degree of sequence divergence is ignored. For example, the rule of two compensatory changes mentioned above is applied universally, regardless of the amount of change the data have undergone. STEPHAN and KIRBY (1993) pointed out the need to account for levels of sequence divergence when making claims about the maintenance of secondary structures. The challenge is to formulate a framework that allows regions with substitution patterns different from the surrounding regions to be identified. It is desirable to quantify the statistical significance of these deviations, accounting for observed levels of divergence. Such a framework is presented below, and it is shown that it provides useful tools both for confirming the existence of hypothesized stems and for identifying stems when no *a priori* knowledge as to their location is available.

STATISTICAL FRAMEWORK

Modified Jukes-Cantor model: As stated above, there have been only a few evolutionary models presented that account for nonindependence of nucleotide sites.

Corresponding author: Spencer V. Muse, Institute of Molecular Evolutionary Genetics, Department of Biology, 208 Mueller Lab, The Pennsylvania State University, University Park, PA 16802-5301.
E-mail: svm1@psuvm.psu.edu

CHURCHILL (1989) developed a Hidden Markov Chain approach in which adjacent sites were more likely to have the same substitution rates than were sites located far apart. MUSE and GAUT (1994) took into consideration the dependencies between nucleotides within codons and presented a model suitable for phylogenetic estimation and testing using coding regions. An approach similar to the latter will be used here.

The procedure begins by extending a simple model of nucleotide substitution (a 4×4 transition matrix) to a model for independently evolving pairs of nucleotides (a 16×16 transition matrix). By introducing a new parameter λ, it is possible to describe a new model that favors the maintenance of stem structures between two regions of the sequence. I will call this a "stem" model. The independent-sites model is a special case of the stem model, formed when $\lambda = 1$. The basic steps are to derive the instantaneous rate matrix for the case of independent evolution of pairs of nucleotides and to modify this rate matrix using λ to account for pairing.

For simplicity, consider the null hypothesis that k sequences of n nucleotides have evolved according to the model of JUKES and CANTOR (1969) (the JC model). Other models will be treated later. The transition probabilities for this model may be expressed as

$$P_{i,j}(t) = \begin{cases} \frac{1}{4}(1 + 3e^{-\mu t}) & i = j \\ \frac{1}{4}(1 - e^{-\mu t}) & i \neq j. \end{cases} \quad (1)$$

$P_{i,j}(t)$ denotes the probability that a nucleotide initially in state i ($i = A, C, G, T$) is in state j after t units of time have passed. Using this parameterization, the rate of substitution per nucleotide per unit time is $3\mu/4$. These expressions will now be used to find the 16×16 transition matrix, $\mathbf{P}^0(t)$, for pairs of nucleotides evolving according to the JC model. (For the case of stems, the pair of nucleotides are not adjacent to each other but are instead nucleotides constrained to form a complementary pair.) If one assumes that each of the two sites in a pair evolves independently according to a JC model, $P^0_{i,j}(t)$ is simply the product of two Jukes-Cantor probabilities. Note that in this context i and j are each one of the 16 nucleotide pairs. Each of the 256 entries of $\mathbf{P}^0(t)$ takes one of three forms:

$$P^0_{i,j}(t) = \begin{cases} \frac{1}{16}(1 + 3e^{-\mu t})^2 \\ \quad \text{No differences } (e.g., i = AC, j = AC), \\ \frac{1}{16}(1 + 3e^{-\mu t})(1 - e^{-\mu t}) \\ \quad \text{1 difference } (e.g., i = AC, j = AG), \\ \frac{1}{16}(1 - e^{-\mu t})^2 \\ \quad \text{2 differences } (e.g., i = AC, j = TG). \end{cases} \quad (2)$$

An alternate description of this evolutionary model will be useful. The matrix $\mathbf{P}^0(t)$ may be obtained from an instantaneous rate matrix, $\mathbf{A}^0$, using the relationship

$$\mathbf{P}^0(t) = e^{\mathbf{A}^0 t}. \quad (3)$$

The instantaneous matrix that produces $\mathbf{P}^0(t)$ is simply the matrix derivative of $\mathbf{P}^0(t)$ with respect to t, evaluated at $t = 0$ (KARLIN and TAYLOR 1981). The entries of $\mathbf{A}^0$ are as follows:

$$A^0_{ij} = \begin{cases} -\frac{3\mu}{2} & \text{No differences,} \\ \frac{\mu}{4} & \text{1 difference,} \\ 0 & \text{2 differences} \end{cases} \quad (4)$$

and the matrix takes the following form:

$$\mathbf{A}^0 = \begin{array}{c} \\ AT \\ TA \\ CG \\ GC \\ AA \\ AC \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccccc} AT & TA & CG & GC & AA & AC & \cdots \end{array} \\ \begin{pmatrix} -3\mu/2 & 0 & 0 & 0 & \mu/4 & \mu/4 & \cdots \\ 0 & -3\mu/2 & 0 & 0 & \mu/4 & 0 & \cdots \\ 0 & 0 & -3\mu/2 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & -3\mu/2 & 0 & \mu/4 & \cdots \\ \mu/4 & \mu/4 & 0 & 0 & -3\mu/2 & \mu/4 & \cdots \\ \mu/4 & 0 & 0 & \mu/4 & \mu/4 & -3\mu/2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix} \end{array}$$

This matrix can now be modified to account for constraints on secondary structure. There is no obvious choice as to what changes should be made. Fortunately, there are some intuitive notions that suggest certain types of alterations. If a stem structure is favored, the relative probability of a change from an unpaired state to a paired state should be greater than the corresponding probability when sites are independent. Similarly, changes from a paired state to an unpaired state should occur with lower frequencies than those predicted by an independent-sites model. To capture these features in the model, one can introduce a "pairing parameter", λ, and modify $\mathbf{A}^0$ to form a new matrix, $\mathbf{A}$. Instantaneous rates for changes from an unpaired state to a paired state are multiplied by λ (relative to the independent model in $\mathbf{A}^0$); those for changes from a paired state to an unpaired state are multiplied by $1/\lambda$. For the moment, only Watson-Crick (AT or CG) pairs are considered. It is possible to incorporate GT pairings by making minor changes in $\mathbf{A}$. This is discussed in a later section. These definitions create an evolutionary model that favors the formation of paired states whenever $\lambda > 1$:

$$\mathbf{A} = \begin{array}{c} \\ AT \\ TA \\ CG \\ GC \\ AA \\ AC \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccccc} AT & TA & CG & GC & AA & AC & \cdots \end{array} \\ \begin{pmatrix} -3\mu/2\lambda & 0 & 0 & 0 & \mu/4\lambda & \mu/4\lambda & \cdots \\ 0 & -3\mu/2\lambda & 0 & 0 & \mu/4\lambda & 0 & \cdots \\ 0 & 0 & -3\mu/2\lambda & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & -3\mu/2\lambda & 0 & \mu/4\lambda & \cdots \\ \mu\lambda/4 & \mu\lambda/4 & 0 & 0 & -\mu(1+\lambda/2) & \mu/4 & \cdots \\ \mu\lambda/4 & 0 & 0 & \mu\lambda/4 & \mu/4 & -\mu(1+\lambda/2) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix} \end{array}$$

The off-diagonal elements can be summarized as follows:

$$A_{ij} = \begin{cases} \mu\lambda/4 & 1 \text{ difference, pairing gained} \\ & (e.g., i = AC, j = AT), \\ \mu/4 & 1 \text{ difference, pairing unchanged} \\ & (e.g., i = AC, j = AG), \\ \mu/4\lambda & 1 \text{ difference, pairing lost} \\ & (e.g., i = AT, j = AC), \\ 0 & 2 \text{ differences} \\ & (e.g., i = AC, j = TG;\ i = AT, j = GC). \end{cases} \quad (5)$$

By design this model reduces to the independent-sites Jukes-Cantor model when $\lambda = 1$. An analytical form for the corresponding transition matrix, $\mathbf{P}(t)$, could not be found. Transition probabilities must instead be computed numerically, using the fact that

$$\mathbf{P}(t) = e^{\mathbf{A}t} = I + \mathbf{A}t + (\mathbf{A}t)^2/2! + (\mathbf{A}t)^3/3! + \cdots \quad (6)$$

A few important analytical results can be obtained, the first being the equilibrium distribution of the process. It is simple to verify that the asymptotic frequencies of paired states, π_p, and of unpaired states, π_u, are

$$\pi_p = \lambda^2/(12 + 4\lambda^2), \quad \pi_u = 1/(12 + 4\lambda^2). \quad (7)$$

To verify this fact, notice that $\pi\mathbf{A} = 0$, where π is the 1×16 vector of asymptotic frequencies. These quantities are useful for likelihood calculations where sums are evaluated across all possible ancestral states, each term weighted by the appropriate equilibrium frequency.

Having found the asymptotic distribution, it is now straightforward to show that the process is reversible. A sufficient condition for reversibility is that

$$\pi_i A_{ij} = \pi_j A_{ji}, \quad i \neq j. \quad (8)$$

Reversibility is desirable because it reduces the computational burden of likelihood calculations. FELSENSTEIN (1981) showed that the "pulley principle" is applicable when a reversible model is used. This effectively removes one branch from an evolutionary tree by limiting likelihood evaluations to unrooted trees.

Finally, the expected number of substitutions per site may be obtained using the fact that

$$E(S) = -\frac{t}{2}\sum_{i=1}^{16} \pi_i A_{ii}. \quad (9)$$

The factor of $^1/_2$ is used to put the expectation on a per nucleotide basis rather than a per nucleotide-pair basis. Some algebra reveals that

$$E(S) = {}^3/_8(\lambda + 1)\mu t. \quad (10)$$

This expression can be used for computing distance measures between pairs of sequences. Note that when $\lambda = 1$ the expectation for the JC model, $E(S) = {}^3/_4\mu t$, is obtained. (Recall that a somewhat nonstandard parameterization of the JC model is being used.) Equation 10 is consistent with observations that single-stranded regions in sequences with structural constraints evolve more slowly than do stem regions (GUTELL *et al.* 1985; VAWTER and BROWN 1993). It may not be necessary to hypothesize additional constraints on the single-stranded regions to explain the discrepancy in evolutionary rates. [However, VAWTER and BROWN (1993) also noted heterogeneous rates among different classes of single-stranded regions, an observation that suggests that the use of one substitution rate for all single-stranded regions may be inappropriate.]

TESTING FOR CONSTRAINTS ON SECONDARY STRUCTURE

FELSENSTEIN (1981) provided the general framework for likelihood analyses of DNA sequences in an evolutionary context. A simple three-sequence example is sufficient to demonstrate the important concepts. Suppose one has homologous DNA sequences of length l from three species, A, B and C. One of the three possible trees is shown in Figure 1A. Let n_j^i be the nucleotide present at site j in sequence i. If sites are assumed to have evolved independently, the likelihood function for site j is

$$L_j = \sum_{n_j^{0'}} \pi_{n_j^{0'}} P_{n_j^{0'},n_j^C}(t_C) \sum_{n_j^0} P_{n_j^{0'},n_j^0}(t_O) P_{n_j^0,n_j^A}(t_A) P_{n_j^0,n_j^B}(t_B), \quad (11)$$

where t_i is the branchlength leading to sequence i. The units of measurement depend on the properties of $\mathbf{P}(t)$. Application of the pulley principle allows the tree to be treated as unrooted when the substitution process is reversible, in which case the likelihood becomes

$$L_j = \sum_{n_j^0} \pi_{n_j^0} P_{n_j^0,n_j^A}(t_A) P_{n_j^0,n_j^B}(t_B) P_{n_j^0,n_j^C}(t_C), \quad (12)$$

with labeling as in Figure 1B. The likelihood over all l sites is simply the product of individual site likelihoods,

$$L = \prod_j L_j. \quad (13)$$

The parameters involved in the transition probabilities, $P_{i,j}(t)$, may be estimated using maximum likelihood. Numerical methods are typically necessary to maximize the likelihood function, and details depend on the form of $\mathbf{P}(t)$.

Likelihood-ratio tests can be constructed using this framework. Several authors have implemented such tests for a variety of null hypotheses (RITLAND and CLEGG 1987; NAVIDI *et al.* 1991; MUSE and WEIR 1992;

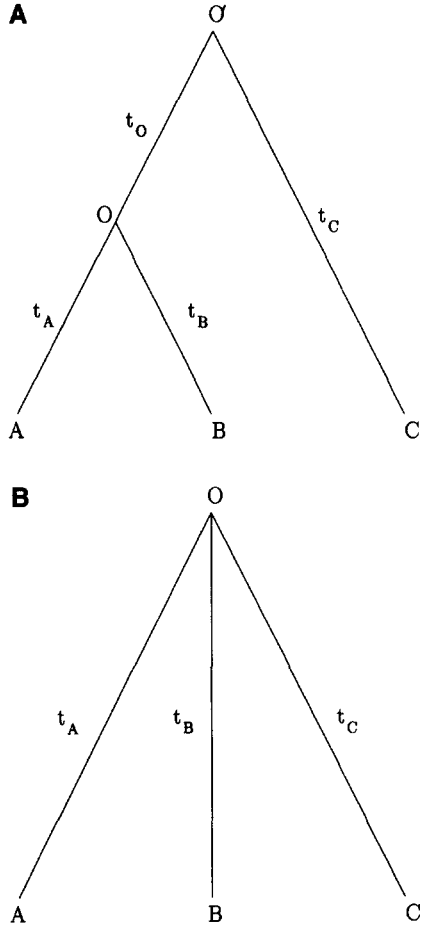

FIGURE 1.—Rooted and unrooted trees for three sequences.

LEARN *et al.* 1993; MUSE and GAUT 1994; GAUT and WEIR 1994). If several homologous DNA sequences are available, an interesting question to consider is whether two regions have been under selective constraint to form a stem structure; in other words, is $\lambda = 1$? A test of this hypothesis can be performed using the models described above. The data must first be divided into two subsets. The first set, $\mathcal{S}_p$, consists of the sites that form the hypothesized stem. If the stem is l_p base pairs in length, there will be $2l_p$ sites in $\mathcal{S}_p$. This set will be represented as l_p pairs of nucleotides. The second set, $\mathcal{S}_u$, contains the remaining l_u unpaired nucleotide sites. As an example consider the sequence

AGCTT CAAGC TTATG GGTTG CCGAA.

If we wish to test the hypothesis that positions 6–10 pair with positions 20–16, the two sets are

$\mathcal{S}_u$ = {AGCTTTTATGCCGAA} and

$\mathcal{S}_p$ = {CG, GG, AT, AT, CG}.

Each of the homologous sequences would contribute in this manner to the two sets.

Although neither assumption is necessary, the present discussion will be restricted to the case where the location of the hypothesized stem is the same for each sequence and where the sequences have a known phylogeny. A combined likelihood function for l_u unpaired sites and a stem region of l_p base pairs can be written as

$$L_T = L_u L_p, \tag{14}$$

where L_u is based on the JC model and uses the sites from $\mathcal{S}_u$, and L_p is based on the stem model and uses the data from $\mathcal{S}_p$. Note that the summation across ancestral nucleotides in Equation 12 is replaced by summation over ancestral pairs when computing L_p. Three key assumptions are that the substitution parameter, μ, is assumed to be the same for both the paired and unpaired regions of the sequence, that the pairing parameter, λ, is the same for all parts of the tree, and that base-paired sites within stems evolve independently of other pairs. The first assumption is consistent with the idea that differences in substitution rates and pattern between paired and unpaired regions are due solely to constraints on base pairing, not to other constraints on the paired region. The second assumption is one of convenience and could be dropped if a large number of sites were included in the stem region. The set $\mathcal{S}_p$ is expected to be small and probably will not provide enough information to estimate separate pairing parameters for each branch. Finally, the assumption of independence is necessary to compute the likelihood function. This assumption is almost certainly violated, because purifying selection is probably more intense once a single mispairing is introduced into a stem. Maximum-likelihood estimates (MLEs) of a μt (substitution rates and time are naturally confounded parameters) for each branch and of a single λ are found by numerical maximization of the likelihood function.

The null hypothesis of no constraints on the potential stem region is equivalent to the case $\lambda = 1$. A likelihood-ratio test of this hypothesis is easily constructed. The numerator of the likelihood ratio is obtained by computing MLEs of the μt under the (independent pairs) JC model. For this maximization the nucleotide sites from both $\mathcal{S}_p$ and $\mathcal{S}_u$ are combined into a single set of unpaired sites, and estimation of the pairing parameter is unnecessary. The denominator is found by maximizing the likelihood function from Equations 12–14 using the partitioned data. This step includes finding the MLE for λ. The standard likelihood-ratio statistic, $-2 \ln (L_0/L_A)$, has a distribution that is asymptotically chi-square with one degree of freedom. Large likelihood ratios suggest that the nucleotides are not evolving in an independent fashion. It is necessary to check the estimate of λ to see if pairing is favored ($\lambda > 1$) or avoided ($\lambda < 1$). If the alternative hypothesis of interest

is H_A: $\lambda > 1$, it is appropriate to halve the P value obtained from the likelihood-ratio test.

It should be clear that the idealized situation just described will rarely arise in practice. As usual, there will often be violations of model assumptions. Of more concern is the fact that it is most likely for the two regions of interest to be selected by prior analysis to find regions of high complementarity. This is implicitly a multiple-test situation. Not only are the two selected regions being tested, but all other possible pairs of regions are effectively tested during the screening process. The interpretation of p values becomes problematic, and this difficulty is not easily circumvented. In theory, an appropriate p value that accounts for multiple tests could be calculated using a parametric bootstrap approach. One would simulate a set of sequences using the MLEs of substitution rates obtained from the null model. All possible pairs of regions of the length used in the test would then be subjected to the likelihood-ratio test just described, and the maximum value recorded. The process would then be repeated many times, providing an estimate of the distribution of the supremum of the test statistics. This distribution, rather than the chi-square distribution, would provide the test's p value. Unfortunately, current computing technology does not make this a practical option. Simulation of many replicate sets of data, applying single tests (or perhaps a moderate number of tests) and tabulating the distribution of the maximum test statistic might provide a reasonable approximation to the full-blown parametric bootstrap procedure. It would be necessary to investigate the number of tests required to obtain an adequate approximation before making strong claims from such a method.

The phylogenetic-comparative method suffers from the multiple-test problem to the same degree as does the likelihood-ratio test. It is likely that the hypotheses one might wish to test using the likelihood-ratio test will be suggested by prior exploration using comparative sequence analysis. The major advantage offered by the likelihood-ratio test is a more rigorous criterion for claiming support for potential stems. The rule of two compensatory changes is replaced by a rejection rule that considers the observed level of sequence divergence. The likelihood-ratio test also makes more efficient use of the data. Compensatory changes are not the only evidence that should be left by maintained stem structures. There is also evidence in the types of changes that do not occur. It is a curious fact that the comparative method is not able to identify a stem that is absolutely conserved. To identify such a feature, one must compare the probability of not seeing any changes under a null model of independence to the corresponding probability under a model that favors pairing. The likelihood methodology incorporates all information regarding the observed pattern of substitution rather than using only information about covariation.

EXTENSIONS

Modified Hasegawa, Kishino and Yano model: There is no reason to limit these procedures to the JC model of evolution. Empirical results are known for a number of models of nucleotide evolution (KIMURA 1980; FELSENSTEIN 1981; TAMURA 1992; TAMURA and NEI 1993). A quite general model (the HKY model) was presented by HASEGAWA *et al.* (1985). It allows for both unequal base frequencies and separate transition and transversion rates. In a sense, it is a combination of the models of KIMURA (1980) and FELSENSTEIN (1981). The formulation proceeds as before, and the resulting instantaneous matrix has the following form:

$$A = \begin{array}{c} \\ AT \\ TA \\ CG \\ GC \\ AA \\ AC \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccccc} AT & TA & CG & GC & AA & AC & \cdots \end{array} \\ \begin{pmatrix} * & 0 & 0 & 0 & \beta\pi_A\lambda & \alpha\pi_C/\lambda & \cdots \\ 0 & * & 0 & 0 & \beta\pi_A/\lambda & 0 & \cdots \\ 0 & 0 & * & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & * & 0 & \mu/4\lambda & \cdots \\ \beta\pi_T\lambda & \beta\pi_T\lambda & 0 & 0 & * & \beta\pi_C & \cdots \\ \alpha\pi_T\lambda & 0 & 0 & \alpha\pi_G\lambda & \beta\pi_A\lambda & * & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix} \end{array}$$

The diagonal elements are defined to make the elements of each row sum to 0, and they are omitted for readability. Once this model is formed, likelihood-ratio tests may be conducted as in the JC case. Each branch now has two substitution parameters to maximize, rather than one. The asymptotic frequencies are also different:

$$\pi_u = \kappa\pi_i\pi_j \quad \pi_p = \kappa\pi_i\pi_j\lambda^2, \tag{15}$$

where π_i and π_j are the base frequencies associated with the two nucleotides in the pair and

$$\kappa = 1/[1 + 2(\pi_A\pi_T + \pi_C\pi_G)(\lambda^2 - 1)]. \tag{16}$$

For completeness it can also be shown that the expected number of substitutions per (nucleotide) site over t units of time is

$$\begin{aligned} E(S) = 2t\kappa\{&\lambda\beta(\pi_A\pi_T + \pi_C\pi_G) + \beta(\pi_A\pi_C + \pi_G\pi_T) \\ &+ \alpha(\pi_A\pi_G[\pi_A + \pi_G] + \pi_C\pi_T[\pi_C + \pi_T]) \\ &+ \lambda\alpha(\pi_A\pi_T[\pi_C + \pi_G] + \pi_C\pi_G[\pi_A + \pi_T])\}. \end{aligned}$$

Allowing GT pairings: When large numbers of species are examined, GT (GU in RNA) pairings are frequently observed in some species at sites where compensatory changes seem to have occurred (ROUSSET *et al.* 1991). It has often been suggested that GT acts as a (slightly deleterious) intermediate state between different canonical pairings and that most, if not all, compen-

satory changes occur via this pathway (ROUSSET *et al.* 1991). A simple modification of the **A** matrix defined in the previous section accommodates much of the nature of the GT-intermediate hypothesis:

$$A_{ij} = \begin{cases} \alpha\pi_t\lambda & \text{transition difference,} \\ & \text{pairing grained } (e.g.,\ i = AC, j = AT), \\ \beta\pi_t\lambda & \text{transversion difference,} \\ & \text{pairing gained } (e.g.,\ i = CT, j = AT), \\ \alpha\pi_t & \text{transition difference,} \\ & \text{pairing unchanged } (e.g.,\ i = AT, j = GT), \\ \beta\pi_t & \text{transversion difference,} \\ & \text{pairing unchanged } (e.g.,\ i = AC, j = AG), \\ \alpha\pi_t/\lambda & \text{transition difference,} \\ & \text{pairing lost } (e.g.,\ i = AT, j = AC), \\ \beta\pi_t/\lambda & \text{transversion difference,} \\ & \text{pairing lost } (e.g.,\ i = AT, j = AA), \\ 0 & \text{2 differences} \\ & (e.g.,\ i = AC, j = GG;\ i = AT, j = GC). \end{cases} \quad (17)$$

In the above definitions, π_t refers to the frequency of the "target" nucleotide. For instance, when considering a change from AT to AC, the target nucleotide is C. The model described by this matrix lacks some features of the GT-intermediate model. Most noticeably, GT pairings are given the same fitness as are canonical pairings. Nonetheless, this adjustment should incorporate most of the desired behaviors. (A worthwhile exercise would be to formulate a similar model that allows a test of the GT-intermediate hypothesis. Such a model would treat GT as a slightly deleterious state and would presumably require the introduction of at least one new parameter.) As has been shown, it is useful to have expressions for the asymptotic frequencies to facilitate likelihood calculations. Equilibrium frequencies for the model of Equation 17 retain the same form as the previous models. The frequencies of paired and unpaired states are still given by Equation 15, with a new value of κ:

$$\kappa = 1/[1 + 2(\pi_A\pi_T + \pi_C\pi_G + \pi_G\pi_T) \times (\lambda^2 - 1)]. \quad (18)$$

SAMPLING PROPERTIES OF THE TEST STATISTICS

Given that stem regions are typically quite short, the adequacy of the asymptotic chi-squared distribution for the likelihood-ratio statistic is questionable. If the small-sample distribution under the null hypothesis is not approximated satisfactorily by a chi-square distribution, it may be necessary to resort to a procedure devised by COX (1961, 1962) and applied to problems in sequence evolution by GOLDMAN (1993). Of course, if P values are estimated using the parametric bootstrap, deviations from the asymptotic distribution are of no consequence. At any rate, this question was studied using Monte Carlo simulations of the test procedure. Two different trees for four sequences were used (see Figure 2). Tree A displays clocklike evolution, whereas Tree B does not. Data were generated according to the (independent sites) HKY model. Parameter values for each branch are shown in the figure. Transition rates (α) are above, and transversion rates (β) are below. These values are similar to those found in the RNase P RNA data examined later in this paper. Nucleotide frequencies for each case were $\pi_A = \pi_C = 0.15$, $\pi_G = \pi_T = 0.35$. Five hundred replicate data sets were generated for each tree. The random sequences were each of length 520. The likelihood-ratio tests based on both the JC and HKY models were computed for each simulated set of data. The alternative hypothesis was that a stem of length 10 was formed between sites 1–10 and 20–11. The results are shown in Table 1. In both cases the tests appear to reject the null hypothesis with the frequency predicted by the chi-squared distribution. Therefore, it should be safe to apply either of the tests to real data with phylogeny and branchlengths similar to the model trees, in the sense that excessive numbers of false positive results are not expected. The results listed are for the upper-tailed test, which rejects only when $\hat{\lambda} > 1$. The rejection levels for the lower-tailed test (and consequently for the two-tailed test as well)

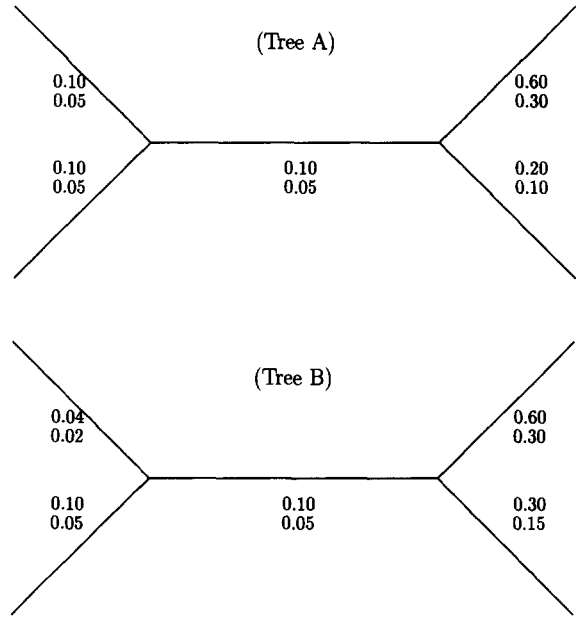

FIGURE 2.—Model trees for simulation studies.

TABLE 1

Null distribution of the likelihood-ratio test statistic

n	Tree	0.025	0.005	0.025	0.005
10	A	0.008	0.000	0.022	0.002
10	B	0.012	0.002	0.026	0.006
50	A	0.002	0.000	0.034	0.008
50	B	0.002	0.000	0.026	0.008

Results of simulation study investigating the rejection probabilities of the likelihood-ratio test when the null hypothesis that $\lambda = 1$ is true. If the asymptotic chi-square distribution is applicable, the test statistic should reject at the 0.025 and 0.005 levels at the nominal frequencies. The length of region tested for base pairing is n. Tree indicates which model tree (see Figure 2) is being used. The table entries indicate the proportion of 500 replicates that reject H_O at the level indicated by the column heading (0.025 or 0.005).

TABLE 2

Power of the likelihood-ratio test

		JC		HKY	
Tree	λ	0.025	0.005	0.025	0.005
A	1.1	0.01	0.00	0.03	0.00
	1.2	0.03	0.02	0.04	0.02
	1.3	0.08	0.02	0.15	0.05
	1.4	0.08	0.02	0.16	0.06
	1.5	0.05	0.02	0.21	0.05
	1.6	0.34	0.14	0.51	0.28
	1.7	0.23	0.12	0.38	0.18
	1.8	0.60	0.36	0.67	0.54
	1.9	0.59	0.39	0.73	0.50
	2.0	0.59	0.33	0.75	0.56
	2.1	0.70	0.50	0.82	0.67
	2.2	0.81	0.63	0.91	0.77
	2.3	0.68	0.36	0.81	0.53
	2.4	0.82	0.58	0.93	0.73
	2.5	0.95	0.86	0.98	0.92
	3.0	0.94	0.87	0.99	0.94
B	1.1	0.03	0.00	0.06	0.02
	1.2	0.07	0.03	0.19	0.05
	1.3	0.12	0.04	0.21	0.10
	1.4	0.08	0.03	0.17	0.06
	1.6	0.10	0.01	0.20	0.06
	1.7	0.34	0.13	0.51	0.28
	1.8	0.69	0.45	0.80	0.65
	1.9	0.39	0.24	0.54	0.31
	2.1	0.67	0.41	0.77	0.60
	2.2	0.69	0.48	0.81	0.64
	2.3	0.86	0.72	0.94	0.83
	2.4	0.69	0.49	0.80	0.63

Results of simulation study investigating the power of the likelihood-ratio test for various values of λ. Table entries indicate the proportion of 100 replicates that reject H_O at the level indicated by the column heading (0.025 or 0.005). The column labeled λ indicates the value of λ used to generate the replicate data sets summarized on that row.

are significantly higher than the levels predicted by the chi-squared distribution (results not shown). Significant test results for $\hat{\lambda} < 1$ should not be trusted.

A second feature to study is the power of the testing procedures. A series of simulations was performed using the HKY stem model to generate random data sets. For each of the trees in Figure 2, 100 replicate data sets were generated using values of λ ranging from 1.0 to 2.5. The rejection probabilities appear in Table 2. Notice that power increases rapidly as λ increases beyond 1.5. As was the case under the null hypothesis, the JC stem test applied to data generated under the HKY model has somewhat reduced power. The values of λ expected in nature are unknown, so it is premature to make judgments as to the power of the tests for real data. In the next section the tests are applied to actual sequence data, resulting in estimates of $\lambda \sim 5$. The tests should have excellent power in such cases. There seems to be little difference between the two model trees. Power seems to be higher for Tree B for small values of λ, but Tree A appears to provide more power for larger values of λ. This is likely due to the overall length of the trees.

APPLICATION TO RNASE P RNA

The enzyme ribonuclease P (RNase P) functions *in vivo* as an RNA-protein complex, but it is unique in the fact that its catalytic center is composed of RNA rather than protein (PACE and SMITH 1990). A great deal of effort has been spent trying to determine the structure of the RNase P RNA with hopes that such knowledge will enhance the understanding of its function (JAMES *et al.* 1988; PACE *et al.* 1989; BROWN and PACE 1991). This work has provided a model of RNase P secondary structure along with persuasive evidence that the model is mostly correct. However, the analyses follow a somewhat *ad hoc* route. One particularly interesting rule of thumb is that a stem structure is considered "proven" if at least two examples of compensatory changes (covariations of base pairings that maintain stem structure) are found in the species being studied. This criterion was originated by FOX and WOESE (1975), who were studying the secondary structure of ribosomal RNAs. The RNase P sequence data provide an excellent opportunity to explore the utility of the test proposed above and to provide additional support for the secondary structure model of RNase P RNA.

The data and sequence alignment were taken from PACE *et al.* (1989) and consist of sequences from the four Gram-positive bacteria *Bacillus subtilis* (Bsu), *B. stearothermophilus* (Bst), *B. megaterium* (Bme) and *B. brevis* (Bbr). The analysis used 433 aligned sites along with the core secondary structure model given in BROWN and PACE (1991). This model is shown in Figure 3. The phylogeny of the species was assumed to be

TABLE 3

Tests of RNaseP core structure

Stem	Location	$\hat{\lambda}$	LRT
1	4–14, 385–375	4.09	32.03
2	15–21, 328–322	6.62	29.25
3	23–29, 43–37	3.85	20.59
4	52–56, 379–375	∞	47.47[a]
5	58–90, 243–235	4.31	15.55
6	93–96, 110–107	5.57	15.07
7	114–116, 129–127	∞	16.05[a]
8	197–204, 209–202	6.87	36.53
9	248–250, 264–262	∞	15.94[a]
10	268–274, 297–291	4.78	24.71
Model		5.29	224.30

Each of the features labeled in Figure 3 was tested using the likelihood-ratio test. Columns 1 and 2 provide the label numbers from Figure 3 and the actual location of the features in the RNaseP sequence. The final two columns show the maximum-likelihood estimate of λ and the value of the likelihood-ratio test statistic.

[a] If the two regions being tested show only paired states at all sites in all sequences, the MLE of λ is ∞. The value of the test statistic shown in the table is a lower bound for the true likelihood-ratio statistic. Numbering of sites follows BROWN and PACE 1991.

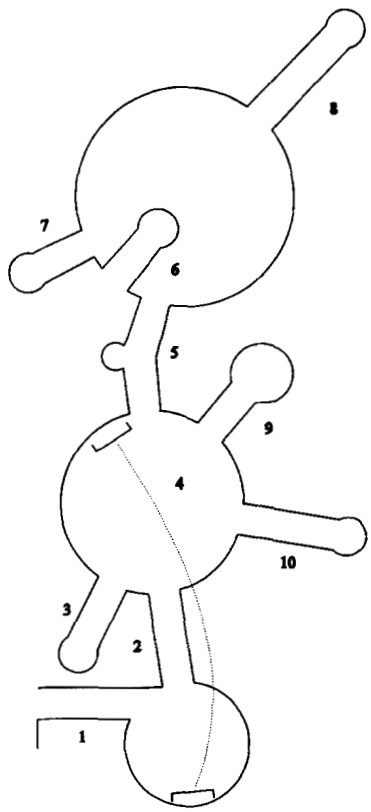

FIGURE 3.—Core secondary structure model for RNase P. From BROWN and PACE (1991).

that given in BROWN and PACE (1991): {Bbr, [Bme, (Bst, Bsu)]}. GT pairings were not considered in the analysis.

The first question posed was whether or not the first stem structure hypothesized by the Brown and Pace model (sites 4–14 paired with sites 385–375) has been maintained by selective constraint. The likelihood-ratio test described above was applied, using the exact boundaries of the proposed structure. The value of the chi-square test statistic was 32.0 ($P \ll 0.001$), and the estimate of λ was 4.1. Therefore, the conclusion is that the nucleotides in these two regions have not evolved independently but have instead evolved in tandem to maintain a stem structure. Each of the other stems from the core structure were also tested, and these results are shown in Table 3. A test of the entire core model was also performed. As mentioned earlier, there is clearly a difficulty with proper interpretation of these P values, because previous analysis of the data suggested the appropriate tests. Some help is provided by the simulations from the previous section. The largest chi-square statistic from any of those tests when the null hypothesis was correct was 9.24. The smallest value from the RNaseP data was 15.07. When data were simulated with $\lambda > 1$, an individual chi-square statistic above 15 was not observed until λ was 1.6, and the average value of the chi-square statistic did not exceed 15 until λ was 2.5. This gives some measure of credibility to the claim that there is statistically significant support for the stems listed in Table 3. Even with the problems interpreting P values, the level of rigor is considerably higher than inferences drawn from the phylogenetic-comparative approach.

Perhaps a more interesting problem than testing for the significance of a stem proposed by other analyses of the same data is the search for stems when their locations are unknown. The likelihood-ratio test was incorporated into a procedure designed to perform this task. Ideally, one would want to perform the test on all possible stem structures, allowing for arbitrary loop sizes and stem lengths. However, it takes only a moment to realize that there are far too many possibilities to perform such an analysis, even with relatively small data sets. Some restrictions must be placed on the search. To obtain the results shown in Figure 4, the search was restricted to stems of length 10 bp and a maximum loop size of 20 nucleotides. It is important to note that such a restriction does not prohibit the procedure from finding stems of other lengths. A signal is expected whenever the correct nucleotides are paired together, even if some nonpairing sites are included in the hypothesized stem region. The explored region consists of 200 nucleotides at the beginning of the RNaseP sequence. Features 3, 6 and 7 from Figure 3 are included in this region. In Figure 4 the vertical and horizontal axes represent the starting positions of two 10-base regions that are tested for pairing constraint. A point is

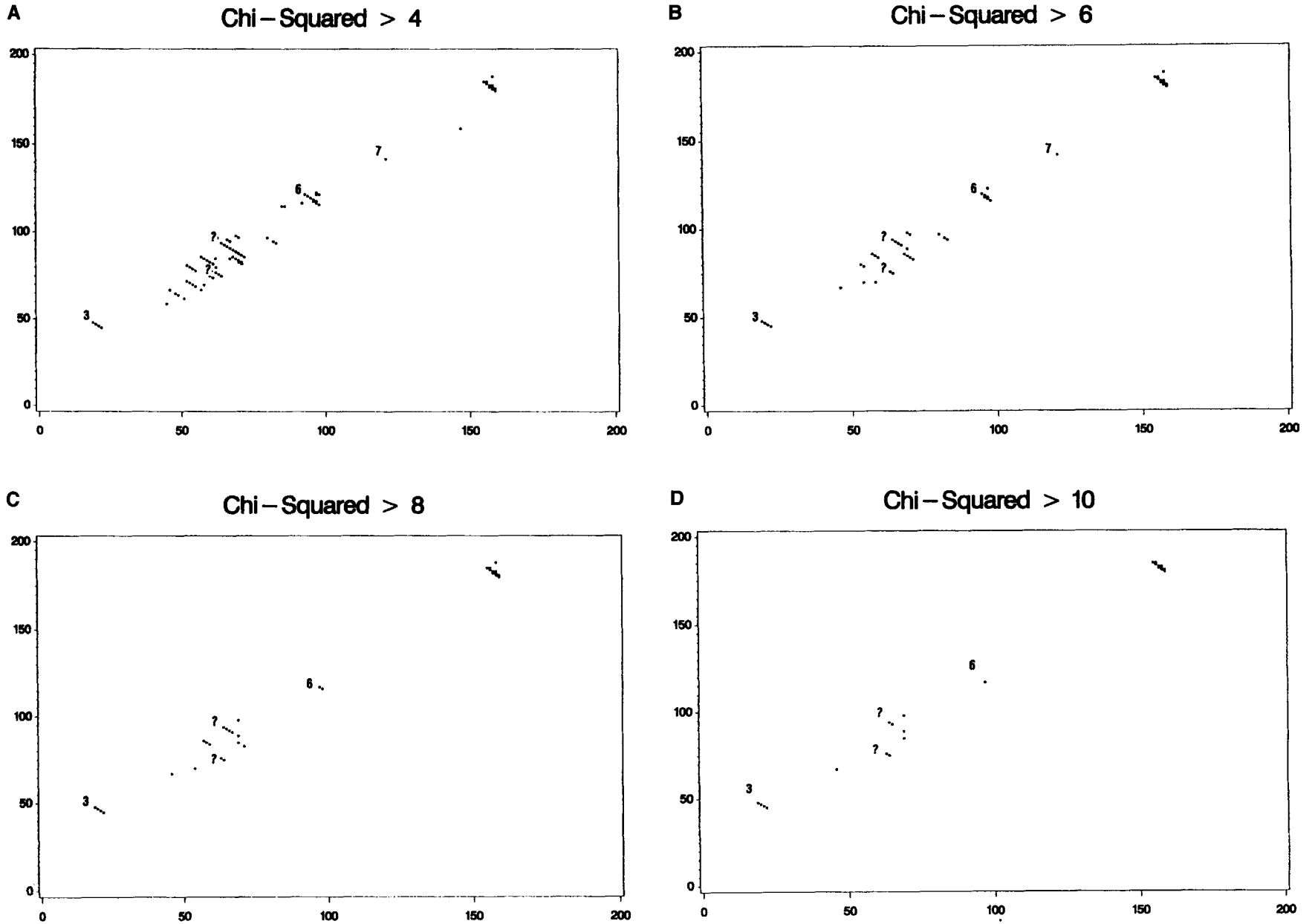

FIGURE 4.—Application to RNaseP sequence data. A point is plotted at (x, y) if the test pairing nucleotides $x - x + 9$ with nucleotides $y + 9 - y$ is greater than the critical value for the panel. For example, a point is plotted at (40, 70) in Figure 4A if the test pairing nucleotides 40–49 with nucleotides 79–70 results in a test statistic above 4.0. Numbers (3, 6 and 7) indicate structures from the RNaseP core structure (see Table 2). ?'s indicate a potential pairings found neither in the core structure nor in the proposed Bacillus secondary structure (BROWN and PACE 1991).

plotted if the likelihood-ratio test statistic is greater than a given cutoff value. For example, in Figure 4A a point would be plotted at (41, 55) if the test pairing sites 41–50 with sites 64–55 resulted in a chi-square value above 4. Not surprisingly, Figure 4A contains many points because of the large number (3591) of tests performed. As the significance level is lowered, though, the number of significant test results quickly falls. In Figure 4D virtually all the significant tests are for previously identified paired regions. Features 3 and 6 each provide strong signals, as do several features that are present in Bacillus RNaseP RNA but not in the core structure (BROWN and PACE 1991). Feature 7 is supported only with chi-square values above 6. Deletions in that region reduce the power of the testing procedure. The low number of false positives is encouraging. Only two regions (63–72, 85–76 and 64–73, 94–83) indicate unsupported pairings with chi-square values above 10. Curiously, the 5′ regions of these pairs include one domain (62–66) of the pseudoknot (feature 4). These results serve as confirmation for both the core model and the search procedure.

DISCUSSION

New evolutionary models appropriate for regions that form stem structures have been developed. Using these models, a test for detecting constraints on secondary structure has been formulated. In turn, this test has been included in a procedure for identifying the locations of unknown stem structures. The procedures were applied to ribonuclease P RNA sequence data from bacteria with positive results: "known" stems were found to be under selective constraint, and evidence for correctness of the core structure was provided.

The evolutionary models are also appropriate for computing evolutionary distances and performing phylogenetic analyses for sequences with pairing constraints. In fact, it might be the case that the most useful applications of these models are in the area of phylogeny reconstruction. Although violations of the independently and identically distributed assumption have long been acknowledged when using molecules such as ribosomal RNAs for phylogenetic inference, little has been done to account for the violations. The models presented here are immediately useful for maximum-likelihood estimation of evolutionary trees with the caveat that at least some parts of the secondary structure are known. The models allow maximum-likelihood estimation of evolutionary distances that account for correlated changes at paired sites and also provide correct variance estimates via the curvature of the likelihood surface near the maximum. Estimates of evolutionary distances can be found by replacing the parameters in Equation 10 or 15 with their MLEs. By the invariance property of maximum-likelihood estimators, this estimate is also a MLE.

With the tremendous increase in computing power over the last few years, it is now time to move from models of DNA sequence evolution that assume that sites evolve independently to more complex and realistic ones. The models presented in this work are a step in that direction, as were the models in CHURCHILL (1989) and MUSE and GAUT (1994). Such models should provide more reliable evolutionary analyses, because they can account for aspects of DNA sequence evolution that are known to occur but are not accommodated by older simpler models. More complex models also offer, for the first time, rigorous tests of some biological hypotheses. One of the fundamental goals of the study of molecular evolution is to obtain an understanding of the mode in which molecular sequences evolve. The evolutionary process may be very complicated, and it is only by developing more realistic mathematical models that we can develop methods for testing the adequacy of simpler models.

I thank ANDREW CLARK for pointing out this problem and for useful discussion during the course of the study. This work has been supported in part by National Institutes of Health grants GM-16250 to S.V.M. and GM-45876 to ANDREW G. CLARK. Computing support was provided by the Pennsylvania State University Center for Computational Biology.

LITERATURE CITED

BROWN, J. W., and N. R. PACE, 1991 Structure and evolution of ribonuclease P RNA. Biochimie **73:** 689–697.

CECH, T. R., N. K. TANNER, I. TINOCO, JR., B. R. WEIR, M. ZUCKER *et al.* 1983 Secondary structure of the *Tetrahymena* ribosomal RNA intervening sequence: structural homology with fungal mitochondrial intervening sequences. Proc. Natl. Acad. Sci. USA **80:** 3903–3907.

CHURCHILL, G. A., 1989 Stochastic models for heterogeneous DNA sequences. Bull. Math. Biol. **51:** 79–84.

COX, D. R., 1961 Tests of separate families of hypotheses, pp. 105–123 in *Proceedings of the 4th Berkeley Symposium,* Vol. 1. University of California Press, Los Angeles.

COX, D. R., 1962 Further results on tests of separate families of hypotheses. J. R. Statist. Soc. B **24:** 406–424.

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

FOX, G. E., and C. R. WOESE, 1975 5S RNA secondary structure. Nature **256:** 505–507.

GAUT, B. S., and B. S. WEIR, 1994 Detecting substitution-rate heterogeneity among regions of a nucleotide sequence. Mol. Biol. Evol. **11:** 620–629.

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

GUTELL, R. R., B. WEISER, C. R. WOESE and H. F. NOLLER, 1985 Comparative anatomy of 16s-like ribosomal RNA. Prog. Nucleic Acid Res. Mol. Biol. **32:** 155–216.

HAAS, E. S., D. P. MORSE, J. W. BROWN, F. J. SCHMIDT and N. R. PACE, 1991 Long-Range Structure in Ribonuclease P RNA. Science **254:** 853–856.

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.

JAMES, B. D., G. J. OLSEN, J. LIU and N. R. PACE, 1988 The secondary structure of ribonuclease P RNA, the catalytic element of a ribonucleoprotein enzyme. Cell **52:** 19–26.

JUKES, T. H., and C. R. CANTOR, 1969 Evolution of protein molecules, pp. 21–132 in *Mammalian Protein Metabolism,* Vol. 3, edited by H. N. MUNRO. Academic Press, New York.

KARLIN, S., and H. M. TAYLOR, 1981 *A Second Course in Stochastic Processes*. Academic Press, New York.

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

LEARN, G. H., J. S. SHORE, G. R. FURNIER, G. ZURAWSKI and M. T. CLEGG, 1992 Constraints on the evolution of plastid introns: The group II intron in the gene encoding tRNA-val (UAC). Mol. Biol. Evol. **9:** 856–871.

MUSE, S. V., and B. S. GAUT, 1994 A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates with application to the chloroplast genome. Mol. Biol. Evol. **11:** 715–724.

MUSE, S. V., and B. S. WEIR, 1992 Testing for equality of evolutionary rates. Genetics **132:** 269–276.

NAVIDI, W. C., G. A. CHURCHILL and A. VON HAESLER, 1991 Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. Mol. Biol. Evol. **8:** 128–143.

PACE, N. R., and D. SMITH, 1990 Ribonuclease P: function and variation. J. Biol. Chem. **265:** 3587–3590.

PACE, N. R., D. K. SMITH, G. J. OLSEN and B. D. JAMES, 1989 Phylogenetic comparative analysis and the secondary structure of ribonuclease P RNA—a review. Gene **82:** 65–75.

RITLAND, K., and M. T. CLEGG, 1987 Evolutionary analysis of plant DNA sequences. Am. Nat. **130:** S74–S100.

ROUSSET, F., M. PÉLANDAKIS and M. SOLIGNAC, 1991 Evolution of compensatory substitutions through G·U intermediate state in *Drosophila* rRNA. Proc. Natl. Acad. Sci. USA **88:** 10032–10036.

SILICIANO, P. G., M. H. JONES and C. GUTHRIE, 1987 *Saccharomyces cereviseae* has a U1-like small nuclear RNA with unexpected properties. Science **237:** 1484–1487.

SPRINZL, M., T. HARTMANN, F. MEISSNER, J. MOLL and T. VORDERWULBECKE, 1987 Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res. (Suppl.) **15:** r53–r188.

STEPHAN, W., and D. A. KIRBY, 1993 RNA folding in Drosophila shows a distance effect for compensatory fitness interactions. Genetics **135:** 97–103.

TAMURA, K., 1992 Estimation of the number of nucleotide substitutions when there are strong transition and G + C content biases. Mol. Biol. Evol. **9:** 678–687.

TAMURA, K., and M. NEI, 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. **10:** 512–526.

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

Communicating editor: G. B. GOLDING