Empirical Models for Substitution in Ribosomal RNA

Andrew D. Smith, Thomas W. H. Lui, and Elisabeth R. M. Tillier

Department of Medical Biophysics, University of Toronto, and Ontario Cancer Institute, University Health Network, Toronto, Ontario, Canada

Empirical models of substitution are often used in protein sequence analysis because the large alphabet of amino acids requires that many parameters be estimated in all but the simplest parametric models. When information about structure is used in the analysis of substitutions in structured RNA, a similar situation occurs. The number of parameters necessary to adequately describe the substitution process increases in order to model the substitution of paired bases.

We have developed a method to obtain substitution rate matrices empirically from RNA alignments that include structural information in the form of base pairs. Our data consisted of alignments from the European Ribosomal RNA Database of Bacterial and Eukaryotic Small Subunit and Large Subunit Ribosomal RNA (Wuyts et al. 2001. Nucleic Acids Res. **29**:175–177; Wuyts et al. 2002. Nucleic Acids Res. **30**:183–185). Using secondary structural information, we converted each sequence in the alignments into a sequence over a 20-symbol code: one symbol for each of the four individual bases, and one symbol for each of the 16 ordered pairs. Substitutions in the coded sequences are defined in the natural way, as observed changes between two sequences at any particular site. For given ranges (windows) of sequence divergence, we obtained substitution frequency matrices for the coded sequences. Using a technique originally developed for modeling amino acid substitutions (Veerassamy, Smith, and Tillier. 2003. J. Comput. Biol. **10**:997–1010), we were able to estimate the actual evolutionary distance for each window. The actual evolutionary distances were used to derive instantaneous rate matrices, and from these we selected a universal rate matrix.

The universal rate matrices were incorporated into the Phylip Software package (Felsenstein 2002. http://evolution.genetics.washington.edu/phylip.html), and we analyzed the ribosomal RNA alignments using both distance and maximum likelihood methods. The empirical substitution models performed well on simulated data, and produced reasonable evolutionary trees for 16S ribosomal RNA sequences from sequenced Bacterial genomes.

Empirical models have the advantage of being easily implemented, and the fact that the code consists of 20 symbols makes the models easily incorporated into existing programs for protein sequence analysis. In addition, the models are useful for simulating the evolution of RNA sequence and structure simultaneously.

Introduction

The conserved nature of ribosomal RNA structure, facilitated by compensating substitutions of paired bases, means that the evolution of bases in structurally related positions may be highly dependent. Ribosomal RNA (rRNA) sequences thus violate a key assumption of many phylogeny methods: that different sites in a sequence have independent evolutionary rates (Tillier and Collins 1995). Because many rRNA structures are well known, we are now in a position to incorporate structural information into the analysis of rRNA. Indeed, several methods and models have already done so (reviewed in Savill, Hoyle, and Higgs 2001), recognizing the likely evolutionary dependence between sites.

Probabilistic methods of phylogenetic analysis, such as Bayesian phylogeny (Yang and Rannala 1997), maximum likelihood (Felsenstein 1981), and modified distance measures (Kimura 1981), make use of substitution models that define probabilities for the transition from one base to another. When base substitution is adequately modeled using a low number of states (or under restricting assumptions), parametric models are feasible. Simultaneously modeling the evolution of sequence and structure has been done with several parametric models, but this approach has been shown to be accurate only when large numbers of parameters are used (Savill, Hoyle, and Higgs 2001). In the common case where a reasonable description of the substitution process requires a large number of parameters (e.g., for models of amino acid substitution), empirical models are often used. For empirical models, the relative substitution probabilities are derived from databases of alignments; the PAM model of amino acid substitution is a well-known example (Dayhoff, Schwartz, and Orcutt 1978). The estimation of empirical substitution rates in rRNA has been attempted previously (Tillier and Collins 1998; Higgs 2000).

Databases of rRNA sequences are now sufficiently large (with over 50,000 sequences represented) to form the basis of an empirical model of substitution for rRNA. Here we present an empirical model of substitution for both individual and paired bases in structurally annotated rRNA. Our model is based on sequence alignments from RNA databases and conserved reference secondary structures. The method we used to obtain substitution matrices from rRNA data is an extension of a procedure we recently applied to derive models of amino acid substitution (Veerassamy, Smith, and Tillier 2003). We accomplish this by converting the rRNA sequences into sequences over a 20-symbol code. Although the 20-symbol code reflects the number of individual bases and ordered pairs of bases, the size of the alphabet is fortuitous: the resulting rRNA models have 20×20 substitution rate matrices, and can be used in many applications designed to analyze the evolution of proteins using 20×20 amino acid substitution rate matrices. Such programs were modified to allow the analysis of rRNA sequences with the same methods.

Key words: rRNA evolution, empirical substitution model, RNA structure, simulation.

E-mail: e.tillier@utoronto.ca.

Mol. Biol. Evol. 21(3):419–427. 2004
DOI: 10.1093/molbev/msh029
Advance Access publication December 5, 2003
Molecular Biology and Evolution, vol. 21 no. 3

Methods

Data

Sequence alignments were obtained from the European Databases on Small and Large Subunit Ribosomal RNAs (Wuyts et al. 2001; 2002). For each alignment, a reliable reference sequence, with a known secondary structure, was selected from the comparative RNA Web site (CRW) database (Cannone et al. 2002). The reference structures we used were *Escherichia coli* (J01695) for both the Large Subunit (LSU) and Small Subunit (SSU) Bacterial alignments, *Mit. Zea mays* (X00794) for the SSU Mitochondrial alignment, *Mit. Xenopus laevis* (M10217) for the LSU Mitochondrial alignment, *Saccharomyces cerevisiae* (U53879) for both SSU and LSU Eukaryotic alignments, and *Methanococcus jannaschii* (U67517) for the SSU Archaeal alignment (because of insufficient data, we did not apply our method to the LSU Archaeal alignment). We used the secondary structure as a reference to identify paired bases in all sequences in the alignments. This assumes that the ribosomal RNA structures are similar for species within an alignment.

Sequence Coding

We converted each rRNA sequence into a sequence over a 20-symbol alphabet in order to treat single-stranded and double-stranded bases uniformly. The code is given in figure 1a, and an example conversion is shown in figure 1b. The code uses four symbols to represent the single-stranded bases: the unpaired bases or bases paired with a gap (which are also single stranded) and 16 symbols to represent ordered pairs of bases, as given by the reference secondary structure. Bases that were not A, C, G, U, or – were coded as unknowns (X), as were pairs having the form (–,–), and unknowns were ignored in subsequent analyses. The symbol for each pair appears exactly once in the converted sequence, and the symbols are ordered according to the position of the 3′ base in a pair.

Deriving the Empirical Model

Our model was constructed using an approach similar to that of Veerassamy, Smith, and Tillier (2003), designed to obtain an amino acid substitution model from the BLOCKS database (Henikoff and Henikoff 1992). The model uses empirical data in the form of matrices that count substitutions between pairs of coded sequences (from the alignment) having a distance falling within a particular range (window). We were able to obtain count matrices for the sliding windows, each containing data corresponding to a different degree of sequence divergence. As described in more detail below, the count matrices are converted into mutability matrices. Each mutability matrix is a function of two unknowns: a universal rate matrix, and an actual evolutionary distance. The approach first estimates the actual evolutionary distance using a differential equation. The estimated actual evolutionary distance is then used to derive an instantaneous rate matrix for each window. From the set of instantaneous rate matrices, we select one as our estimate of the universal rate matrix.

Substitution Frequency Matrices and Observed Distances

From the alignments, each pair of converted sequences was compared, and their observed sequence divergence (frequency of substitution) was recorded. Those sites containing X were disregarded. Sequence pairs were disregarded altogether when there were fewer than 200 comparable sites between them. This was done to prevent the artificial situation of observing very high divergence between two sequences purely because very few sites were comparable.

A series of overlapping windows representing the full range of sequence divergence were defined. The first window included all sequence pairs with divergence less than or equal to the fixed window size w. The ith window included all pairwise sequence comparisons with a relative divergence between i and $w + i$. For each alignment, each type of substitution was counted for all of the sequence pairs in each window. In this manner, a count matrix was obtained for each window, and these were converted into frequency matrices. Because direction of substitution is not known, each substitution was counted for both directions, and the diagonal entries in the matrix were doubled. Using this method, consecutive windows are expected to contain a nearly identical set of pairs, and windows corresponding to sufficiently high divergence are expected to be empty. The sliding-window strategy allowed us to ensure that we had observations from many different levels of sequence divergence. By manipulating the window size we were able to control the amount of data in each window. The sizes for the sliding windows were selected according to the quality of matrices produced for the sizes (see below).

Estimating Actual Evolutionary Distances

Each count matrix was converted to a frequency matrix, denoted by F. From the frequency matrices, we calculated the vector π of observed frequencies for each code symbol

$$\pi_i = \frac{\sum_{j=1}^{20} F_{ij}}{\sum_{i'=1}^{20} \sum_{j=1}^{20} F_{i'j}}, \quad (1)$$

for $1 \leq i \leq 20$. For each row of the frequency matrix, we divide the entries in that row by the row sum, resulting in mutability matrix M:

$$M_{ij} = \frac{F_{ij}}{\sum_{j'=1}^{20} F_{ij'}}, \quad (2)$$

where $1 \leq i, j \leq 20$. Mutability matrices were only computed for windows where the frequency matrices had nonzero row sums in order to prevent division by zero in equation 2. Mutability matrices describe the frequency of any symbol being substituted by any other (including itself). Because each element of a mutability matrix falls between 0 and 1, and the sum of the elements in each row is equal to 1, mutability matrices are stochastic matrices. The mutability matrices derived from the frequency

(a) RNA Code:

A ⇒ A	AA ⇒ C	CA ⇒ H	GA ⇒ M	UA ⇒ T
C ⇒ R	AC ⇒ Q	CC ⇒ I	GC ⇒ F	UC ⇒ W
G ⇒ N	AG ⇒ E	CG ⇒ L	GG ⇒ P	UG ⇒ Y
U ⇒ D	AU ⇒ G	CU ⇒ K	GU ⇒ S	UU ⇒ V

(b) Example Recoding:

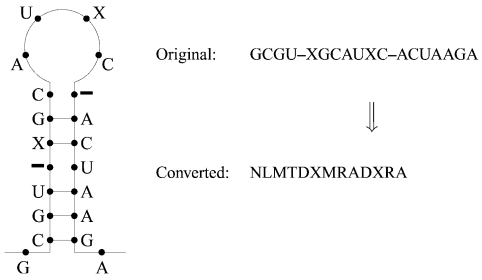

FIG. 1.—rRNA sequence conversion. *a*. The conversion code for the rRNA single-stranded bases and double-stranded base pairs to the 20-symbol code. *b*. An example of how a sequence and its structure are denoted with the new code.

matrices are reversible and thus fulfill the detailed balance equation:

$$\pi_i M_{ij} = \pi_j M_{ji}. \tag{3}$$

The average substitution frequency can be calculated from the mutability matrices and the frequency of the code symbols using the following formula:

$$D(M) = 1 - \sum_{i=1}^{20} \pi_i M_{ii}. \tag{4}$$

The mutability matrices describe the observed frequencies of substitution (observed distance) expected after an unknown actual evolutionary distance P. Therefore, in addition to the definition given above, a mutability matrix can also be considered a function of P. To reflect the dependence of a mutability matrix on the (unknown) evolutionary distance of the sequences from which M was derived, we let $M(P)$ denote the mutability matrix M as a function of evolutionary distance P. The value of P is unknown, because sequences ancestral to those in the alignment are unknown, and because multiple substitutions may have actually occurred at any site. Because actual evolutionary distance is additive, taking the square (square root) of a mutability matrix will double (halve) the actual evolutionary distance. In general, for any mutability matrix M, and any number n,

$$M(P)^n = M(nP), \tag{5}$$

which expresses a special form of the Chapman-Kolmogorov equation for Markov chains. We used this property to derive the instantaneous rate matrix from the mutability matrices. Because this required taking fractional powers and logarithms of the mutability matrices, the mutability matrices were used only if their eigenvalues were all non-negative.

We estimated the derivatives of the observed distances with respect to actual distance numerically, using the five-point formula (Burden and Faires 1985),

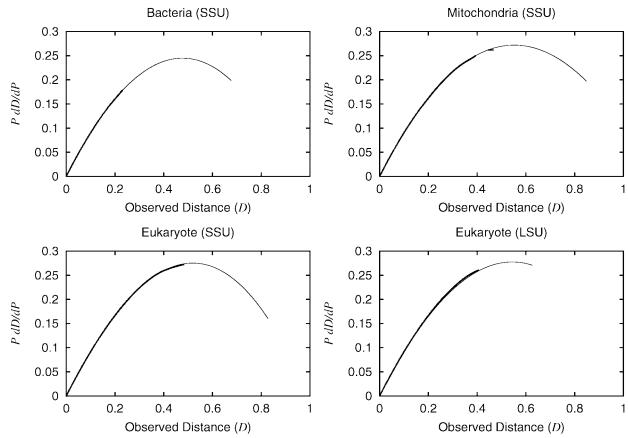

FIG. 2.—Graph of PdD/dP vs D. These graphs indicate the quality of fit, and the range of valid data compared with the range of observable data.

$$\frac{dD(M(P))}{dP} = \frac{1}{12(\lambda)P}\left(D(M((1-2\lambda)P)) - 8D(M((1-\lambda)P)) + 8D(M((1+\lambda)P)) - D(M((1+2\lambda)P))\right) \tag{6}$$

with $\lambda = 0.01$:

$$\frac{dD(M(P))}{dP} = \frac{1}{12(0.01)P}\left(D(M(0.98P)) - 8D(M(0.99P)) + 8D(M(1.01P)) - D(M(1.02P))\right). \tag{7}$$

We used the additivity of mutability matrices and the memoryless property of Markov chains (eq. 5) to obtain the following equivalent formula:

$$P\frac{dD(M(P))}{dP} = \frac{1}{0.12}\left(D(M(P)^{0.98}) - 8D(M(P)^{0.99}) + 8D(M(P)^{1.01}) - D(M(P)^{1.02})\right). \tag{8}$$

The expression on the right can be estimated by a polynomial in $D(M(P))$. For notational convenience, let $D = D(M(P))$. We found that a cubic polynomial with four unknown coefficients was sufficient to characterize the observed data (fig. 2), yielding the differential equation:

$$P\frac{dD}{dP} \approx a_3 D^3 + a_2 D^2 + a_1 D + a_0, \tag{9}$$

We also know the initial conditions because, at sufficiently low distance there are no overlapping substitutions and the number of actual substitutions is equal to the number of observed substitutions. Therefore,

$$\lim_{P\to 0} D(M(P)) = 0 \quad \text{and} \quad \lim_{P\to 0} \frac{dD(M(P))}{dP} = 1, \tag{10}$$

allowing us to conclude that $a_0 = 0$ and $a_1 = 1$. The differential simplifies to the quadratic

$$\frac{P}{D}\frac{dD}{dP} \approx c_2 D^2 + c_1 D + 1, \tag{11}$$

with only two unknown coefficients, c_2 and c_1. The differential is separable and easily solved. The solution

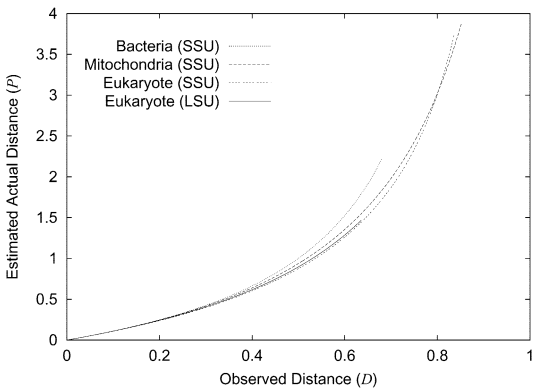

FIG. 3.—Corrected distances. The corrected evolutionary distance is shown as a function of the observed evolutionary distance for 3 SSU rRNA data sets and 1 LSU rRNA data set.

defines a multiple-hit correction formula from which the estimated actual P distances can be estimated from observed D distances

$$P = \text{corrected}(D) \tag{12}$$

The correction formula depends on the data set, and several examples are shown in figure 3.

Instantaneous and Universal Rate Matrices

Once the actual evolutionary distances had been estimated, we had sufficient information to derive instantaneous rate matrices for each window. The following formula was used to obtain the instantaneous rate matrix:

$$A = \ln(M)/P. \tag{13}$$

Equation 13 can only be used when the logarithm is well defined, which requires that all eigenvalues of M be nonnegative. A valid instantaneous rate matrix is required to have both a corresponding valid mutability matrix, and also not have any negative values off the main diagonal.

This scheme can provide up to $n(1 - w) + 1$ instantaneous rate matrices, but many fewer were found to be valid. Both the negative eigenvalues in mutability matrices and the negative off-diagonal rates in instantaneous rate matrices have been found to be associated with insufficient observations in the count matrix and/or high divergence.

From the set of valid instantaneous rate matrices, one was selected as the universal rate matrix. The selection criteria were based on how well an instantaneous matrix could predict the observed mutability matrices, given the estimated actual evolutionary distance. For each valid instantaneous rate matrix A, we evaluated the sum of relative residuals. This is calculated as the difference in the norm (largest singular value) between the exponential of the estimated log of the matrix and the original matrix:

$$\sum_i \frac{\|\exp(AP_i) - M_i\|}{\|M_i\|}, \tag{14}$$

where the sum is over all valid mutability matrices. The instantaneous matrix minimizing this quantity was selected as the universal rate matrix. Examples of rate matrices are shown in figure 4.

Simulations Experiments

Experiments were conducted to assess how well the matrices could recover phylogenetic trees obtained from simulated data. Our Bacterial 16S rRNA model was incorporated into PSeq-Gen (Grassly, Adachi, and Rambaut 1997), a program designed to generate protein sequences from a given tree and an amino-acid substitution model such as PAM (Dayhoff, Schwartz, and Orcutt 1978). Because the matrices that describe our model have the same dimensions as amino acid substitution matrices, programs using amino acid substitution matrices can easily be modified to implement our empirical model instead. We modified the PSeq-Gen program to incorporate the SSU bacterial empirical rRNA substitution model.

We simulated sequences using a 20-species tree (see fig. 5), and the resulting sequences were analyzed with our own programs, and programs from the PHYLIP package (Felsenstein 2002). We used both the dnaml program and the combination of the dnadist and neighbor programs from PHYLIP on sequences obtained by converting the simulated sequences (over the 20-symbol code) back into

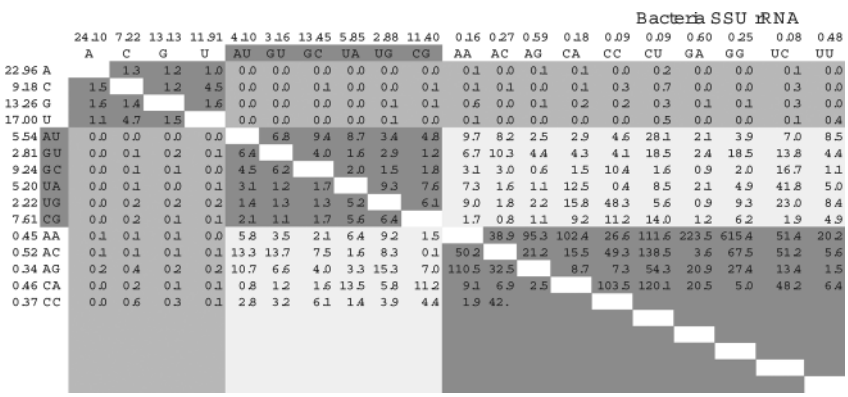

FIG. 4.—Rate matrices for SSU rRNA. The entries of instantaneous rate matrices derived for the bacteria and eukaryote data sets are divided by the frequency of the base/ base pair for that column. The resulting matrices are symmetrical. The Bacterial matrix is shown on the upper right and the Eukaryotic matrix in the lower left. The order of the rows/columns has been changed to show them in order of single-stranded bases, Watson-Crick and GU base pairs, and non-Watson-Crick pairings.

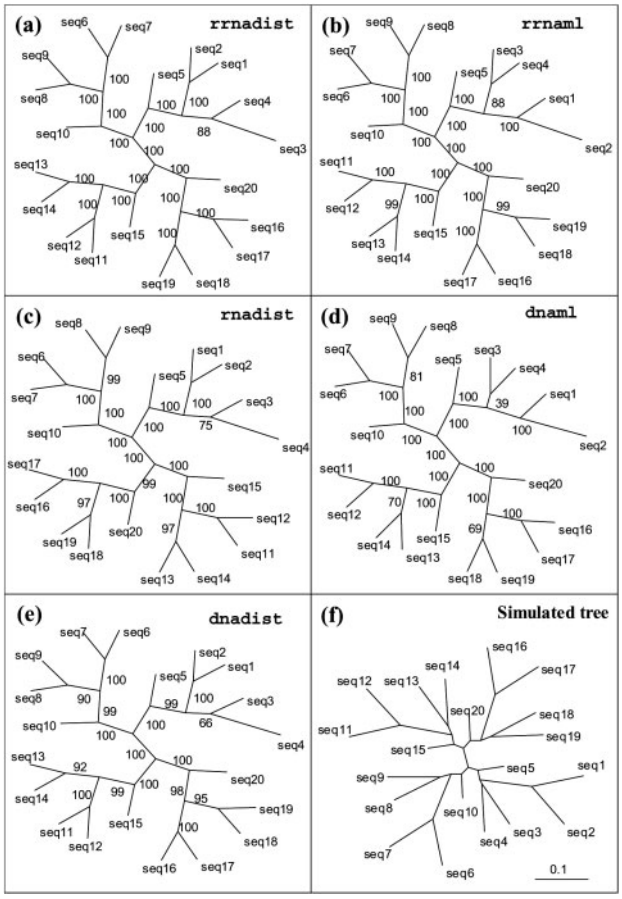

FIG. 5.—Simulation results. The numbers at the nodes indicate how often the correct branching was found out of 100 simulations using five different analysis methods. The simulated tree topology is shown in panel *f*.

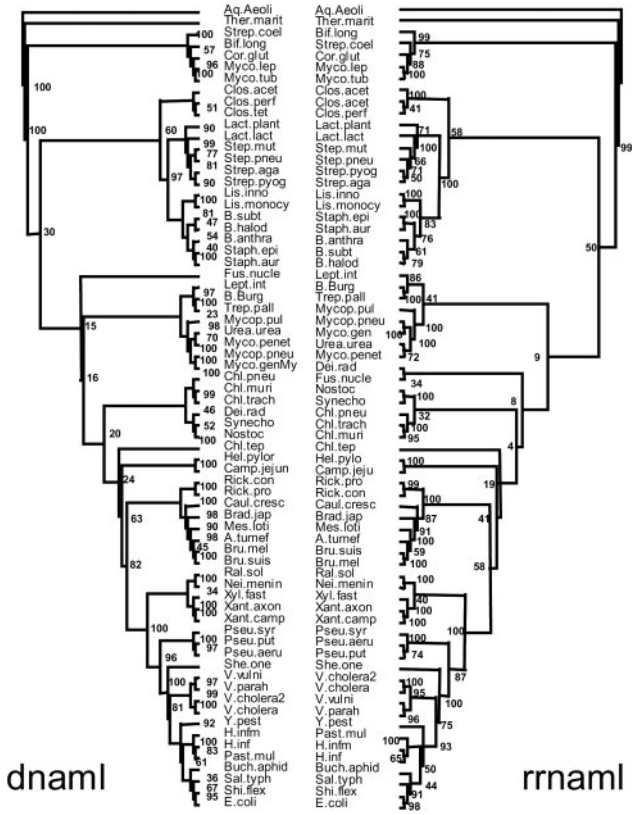

FIG. 6.—Bacterial SSU rRNA trees. Results of 100 bootstrap samples for the bacterial 16S rRNA sequences using maximum likelihood methods. dnaml was run with a discrete approximation of gamma-distribution with a gamma parameter of 1 and four categories of sites. rrnaml was run with the same gamma parameter and only two categories. The values at the nodes indicate the number of bootstrap samples with the corresponding branching. Although the trees are almost identical, the bootstrap values for the dnaml tree are generally higher.

the four bases of RNA (using our program called 20to4). We also used the programs rrnadist and rrnaml, which are modified versions of the PHYLIP protdist and protml that incorporate our 16S rRNA matrices. Finally, our own distance program rnadist, which implements the parametric model OTRNA from Tillier and Collins (1998), was also used.

Example Data Set: Bacterial Phylogeny

We selected 16S rRNA sequences from the set of completely sequenced genomes for which there was a representative alignment in the European Ribosomal RNA Database (a total of 70 full-length sequences). We used *Escherichia coli* as a reference sequence and the secondary structure obtained from the RNA database (Cannone et al. 2002), and we converted the sequences to the 20-letter code with our program 4to20. We analyzed this data set using the same programs listed above for the simulated data. We allowed for gamma-distributed rates, with a gamma parameter of 1, for all methods; for the maximum likelihood methods, we used two categories in the Hidden Markov Model (HMM). All trees were bootstrapped 100 times. The trees obtained using the maximum likelihood methods are shown in figure 6.

Results

The rRNA Matrices

We applied our method to several alignments from the European RNA databases (Wuyts et al. 2001; 2002) attempting to obtain matrices specific to Bacterial, Eukaryotic, Archeal, and Mitochondrial ribosomal RNAs, for both the Small Subunit and the Large Subunit. Some properties of these alignments are given in table 1.

The alignments were of varying sizes. The number of observations that can be extracted from an alignment is proportional to the number of pairs of sequences and the number of comparable sites in each pair. The number of sequences in the alignments ranged from 156 for Eukaryote LSU, to 12,107 for Bacterial SSU. The alignment lengths ranged from 1,000 to 2,500 bases, but the number of comparable sites varied considerably, even within an alignment. The most important factor affecting the number of observations for any alignment is the number of sequences in that alignment. Only four alignments (Bacterial, Mitochondrial, Eukaryote SSU, and Eukaryote LSU) produced a universal rate matrix. For the remainder, the alignments were too small (too few observations). The Eukaryote LSU alignment is an

Table 1
Properties of Alignments Used to Derive rRNA Matrices

	Bacteria		Mitochondria		Eukaryote		Archaea
Database	SSU	LSU	SSU	LSU	SSU	LSU	SSU
Number of sequences	12,107	399	1,039	659	6,590	156	590
Alignment length (bp)	1,064	2,035	1,504	1,258	1,303	2,488	1,013
Window size	0.10	0.10	0.10	0.10	0.20	0.20	0.10
Number of windows							
All windows	828	1049	1430	1071	1302	2117	483
Valid windows	144	0	19	0	532	530	0
Maximum divergence							
All windows	0.680	0.416	0.852	0.753	0.836	0.636	0.378
Valid windows	0.162	NA	0.203	NA	0.351	0.412	NA
Mean relative residual							
All windows	0.184	NA	0.109	NA	0.167	0.103	NA
Valid windows	0.073	NA	0.048	NA	0.048	0.050	NA

NOTE.—The reference structures were: *Escherichia coli* (J01695) for Bacteria, *Mit. Zea mays* for Mitochondria (SSU) and *Mit. Xenopus laevis* for Mitochondria (LSU), *Saccharomyces cerevisiae* (U53879) for Eukaryotes, and *Methanococcus jannaschii* (U67517) for Archaea.

exception to this trend. Having only 156 sequences, this alignment still produced many valid instantaneous rate matrices from which to select a universal rate matrix.

Window sizes were selected to produce the greatest number of valid matrices. For the Bacterial, Mitochondrial, and Archaeal alignments, a window size of 10% of the alignment length was found to produce very good results, and deviation from this size did not produce significantly more valid matrices. For Eukaryote (SSU and LSU) alignments, increasing the window size resulted in improvements up to 20%. Because we wish to cover the largest range of actual divergences, the window size was selected to balance the sometimes conflicting requirements of producing instantaneous rate matrices corresponding to large divergence, and also to keep the residuals low.

Table 1 shows the number of windows in each alignment from which observations were made. The total number of windows is the number of count matrices with non-zero row sums. All of these were used in our analysis when testing the universal rate matrices according to predictive ability. Not all of these could be used when fitting the curves needed to estimate the actual distances. Those windows for which the mutability matrix had negative eigenvalues could not be used in Equation 8. We also required that the numerically estimated derivatives from Equation 8 be non-negative. The mutability matrices corresponding to all windows fitting this criterion were used to fit the curve given by equation 11. Figure 2 gives the curves for Bacterial SSU, Mitochondrial SSU, and Eukaryote SSU and LSU, where the data points used to estimate the curve are plotted, along with the resulting curve. The fitted curve is shown over the range of divergence for non-empty windows. Figure 3 shows the resulting distance correction for the different datasets.

The valid windows are those that meet the requirements for fitting the curve; they also produced a valid instantaneous rate matrix with no negative values off the diagonal. Table 1 shows that the Eukaryote alignments produced many valid windows, whereas the mitochondrial alignment produced very few.

The most important statistic in table 1 is the average relative residual of the observed mutability matrices with the mutability matrices predicted using the universal rate matrix. The relative residual measures how well the selected universal rate matrix can predict the observed data (the mutability matrices). We calculated residuals predicting mutability matrices from all non-empty windows; we also predicted mutability matrices from those windows with a valid instantaneous rate matrix. When predicting mutability matrices for those windows producing valid instantaneous matrices, we observed that the universal rate matrix had residuals less than 0.1. The residual for predicting Bacterial matrices was based on 144 comparisons. For the Mitochondrial SSU matrix, the residual with respect to valid matrices was 0.048, but it was averaged over only 19 matrices, each corresponding to a window close to the one used to obtain the universal rate matrix. The Eukaryotic alignments produced a large number of valid matrices.

Simulations

Because of the unusual nature of our parametric substitution model for recoded RNA sequences, and because we would be using protein-analysis programs for the analysis of rRNA data, we first established the feasibility of our approach by means of simulated sequences. The Bacterial SSU rRNA matrix was incorporated into PSeq-Gen (Grassly, Adachi, and Rambaut 1997) and used to generate 100 simulated data sets which we then analyzed with several methods; the results are given in figure 5. The values on the branches are the number of simulations in which the correct tree (given in figure 5f) was found. The topology and branch lengths of the tree in figure 5f were chosen to reflect a range for the degree of difficulty in correctly identifying the branching by standard methods. The rRNA methods (rrnaml [fig. 5b] and rrnadist [fig. 5a] using our Bacterial SSU rRNA matrix with default parameters) did well in recovering the trees. This is not surprising because these methods applied the

same model that was employed to generate the sequences. Nevertheless, this result means that such models are useful for the simulation of rRNA evolution and for phylogenetic analysis of such sequences. The simulated sequences were also analyzed with the distances calculated according to the OTRNA model (figure 5c), which is a parametric model for RNA evolution (Tillier and Collins 1998). Although rnadist does not do as well as the rrnadist (fig. 5a), it does perform better than the standard DNA methods dnadist (fig. 5e) and dnaml (fig. 5d) on the same data. The DNA methods were applied after reverse coding the sequences back to the four-base RNA code, so that the secondary structure information was lost. The observed reduction in the methods' ability to obtain the correct tree shows that the consideration of structural information is important for obtaining the correct phylogeny.

Bacterial Phylogeny

Because the rrnadist and rrnaml programs were successful in simulations, we were confident in applying the empirical models to actual rRNA data. We chose the sequenced Bacterial genomes because of our interest in these genomes, and because their true phylogeny is unknown. The resulting rrnaml tree is compared to the dnaml tree in figure 6. The trees are remarkably similar, but with generally lower bootstrap values for the rrnaml tree. The distance trees obtained with dnadist, rnadist, and rrnadist are also not significantly different from one another (data not shown).

Discussion

Obtaining the matrices that describe our model required making several methodological assumptions, as well as various assumptions and approximations about the data sets we used.

The assumptions of our method for deriving the empirical matrix have largely been addressed by Veerassamy, Smith, and Tillier (2003). Some minor methodological changes were made to apply the method to the coded rRNA sequences. Unlike the amino acid substitution frequency matrices from the BLOCKS database (Henikoff and Henikoff 1992), the rRNA substitution frequency matrices often had negative eigenvalues for higher sequence divergence levels and thus could not be used for estimating a rate matrix. Negative eigenvalues in mutability matrices are often associated with an insufficient number of observations in the corresponding count matrix. Instead of using the clustering approach of Blosum (Henikoff and Henikoff 1992), we used a sliding-window strategy. The purpose of the sliding window was to obtain a large number of count matrices, corresponding to many different levels of sequence divergence, and to ensure that each count matrix contained a high number of counts.

Negative eigenvalues are also observed when the frequency matrix is too far from identity (Devauchelle et al. 2002). Biologically speaking, this is the situation of too many overlapping substitutions to determine a set of positive substitution rates that produced the observed data. This phenomenon has been observed before and accounts for the difficulty in obtaining distance estimates from rRNA (Hoyle and Higgs 2003). Negative eigenvalues occur earlier (i.e., at lower levels of observed divergence) in the Bacterial data set than in the Eukaryotic data set. A likely cause is higher rates of simultaneous compensatory base substitutions in Bacteria, and stronger conservation of secondary structure in Bacteria.

Certain assumptions underlying our model are independent of our methods. The most evident assumption is that the sequence data and the alignments used are correct. The huge alignments from the rRNA databases represent solutions to an extremely difficult global multiple sequence alignment problem, and they cannot realistically be expected to be optimal with respect to any measure of alignment quality. The nature of our empirical model requires that very large multiple alignments be used, so we must assume that the alignment given is sufficiently accurate to form a basis of inference.

We also analyzed alignments obtained from the RDPII database (Maidak et al. 2001) and obtained similar results (data not shown). We chose to focus on the European rRNA database, so that we could refine our methods for a specific database. The European rRNA database was more complete in terms of 23S sequences. Also, in the RDPII database, the crucial distinction between deleted nucleotides and unknown (unsequenced) nucleotides was not always clear.

Other assumptions had to do with the correctness of the reference structures, and how well conserved is the secondary structure in any of the alignments used. The structures we used were obtained from the CRW database (Cannone et al. 2002). The structures were largely obtained by comparative sequence analysis (Gutell, Lee, and Cannone 2002), which also assumes conserved structures throughout sequences. Although the structure of rRNA is quite conserved, it is not immutable. Particularly the Eukaryote sequences have shown changes in structure (Wuyts, Van de Peer, and Wachter 2001). These variations are reflected in the universal rate matrix we derived for them, which showed some significant substitution rates between paired and unpaired bases (see fig. 4). These rates reflect changes in structure or problems with the alignments. The matrices will be refined by considering more reference secondary structures in an alignment. The same methods can easily be applied in the extreme case where each sequence has a known secondary structure.

The idea that the base pairs rather than the bases in the sequences are the independently evolving units in an structured RNA sequence has been used before (Tillier and Collins 1995), and sophisticated parametric models considering up to the 16 base combinations possible in a pairing have been developed. The disadvantages of these models are that they can be slow and difficult to implement, although some such models have been implemented in a Bayesian phylogenetic framework (PHASE) (Jow et al. 2002; Hudelot et al. 2003), and for distance analyses (rrnadist). It has also been shown that RNA models models require many free parameters to be accurate (Savill, Hoyle, and Higgs 2001) because of differences in the rates of substitution between the base pairs. However, the models are usually applied only to

rRNA molecules for phylogenetic analysis. The large size of the rRNA database makes it possible to empirically derive the substitution rates between the base pairs, and including the single-stranded bases in the model allowed us to easily implement the combined analysis of single-stranded and double-stranded regions. This task was made even easier by the fact that 20×20 empirical matrices have commonly been implemented for the analysis of protein sequences.

A recoding approach could be applied to other RNA molecules (tRNAs for example), but different recoding alphabets might also be used on other types of sequences. For example, it is conceivable to derive an empirical model for codon evolution (a 64×64 matrix). The method used to then derive a matrix is independent of the recoding. Our method for deriving the empirical matrices (from Veerassamy, Smith, and Tillier 2003), requires large amounts of data (as does the approach of Muller and Vingron [2000]). Other approaches for deriving matrices could be used in more specific cases with smaller datasets (Arvestad and Bruno 1997; Whelan and Goldman 2001; Devauchelle et al. 2002).

The phylogenetic trees obtained with the new method show reduced levels of statistical support, as expected, because of the reduced number of independently evolving characters considered in the recoded sequences (Tillier and Collins 1995). The new method therefore does not allow better resolution of the tree; rather it gives a more accurate estimate of the (generally low) confidence in the branch estimates. The branching order is not significantly different from the one obtained with the standard DNA approaches, however. This similarity is reassuring in two ways, the first being that the standard DNA methods on this data set show some robustness with respect to the violation of the assumption of independence of sites, and second because it shows that the methods using our empirical models do at least as well as the parametric approaches.

We have shown that empirical matrices can be used for the study of rRNA evolution with both simulations and with an example phylogenetic analysis of Bacterial 16S rRNA sequences. The ability to analyze and simulate rRNA sequence evolution including secondary structure constraints should be very useful in other studies.

Supplementary Material

All programs and matrices developed here are available from the URL www.uhnres.utoronto.ca/tillier/rRNA/rna.html.

Literature Cited

Arvestad, L., and W. J. Bruno. 1997. Estimation of reversible substitution matrices from multiple pairs of sequences. J. Mol. Evol. **45**:696–703.

Burden, R. L., and J. D. Faires. 1985. Numerical analysis, 3rd edition. PWS Publishers, Boston.

Cannone, J. J., S. Subramanian, M. N. Schnare, et al. (14 co-authors). 2002. The Comparative RNA Web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics, **3**:1–2; http://www.rna.icmb.utexas.edu/.

Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. 1978. A model of evolutionary change in proteins. Pp. 345–352 *in* (M. O. Dayhoff, ed.) Atlas of protein sequence and structure, vol. 5. National Biomedical Research Foundation.

Devauchelle, C., A. Grossmann, A. Hnaut, M. Holschneider, M. Monnerot, J. L. Risler, and B. Torrsani. 2002. Rate matrices for analyzing large families of protein sequences. J. Comput. Biol. **8**:381–399.

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

———. 2002. PHYLIP (phylogeny inference package) version 3.6.3. http://evolution.genetics.washington.edu/phylip.html.

Grassly, N. C., J. Adachi, and A. Rambaut. 1997. PSeq-Gen: an application for the Monte Carlo simulation of protein sequence evolution along phylogenetic trees. Comput. Appl. Biosci. **13**:559–560.

Gutell, R. R., J. C. Lee, and J. J. Cannone. 2002. The accuracy of ribosomal RNA comparative structure models. Curr. Opin. Struct. Biol. **12**:301–310.

Henikoff, S., and J. Henikoff. 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA **89**:10915–10919.

Higgs, P. G. 2000. RNA Secondary structure: physical and computational aspects. Q. Rev. Biophys. **33**:199–253.

Hoyle, D. C., and P. G. Higgs. 2003. Factors affecting the errors in the estimation of evolutionary distances between sequences. Mol. Biol. Evol. **20**:1–9.

Hudelot, C., V. Gowri-shankar, H. Jow, M. Rattray, and P. Higgs. 2003. RNA-based phylogenetic methods: application to mammalian mitochondrial RNA sequences. Mol. Phylogenet. Evol. **28**:241–252.

Jow, H., C. Hudelot, M. Rattray, and P. G. Higgs. 2002. Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. Mol. Biol. Evol. **19**: 1591–1601.

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

Maidak, B. L., J. R. Cole, T. G. Lilburn, C. T. Parker, P. R. Saxman, R. J. Farris, G. M. Garrity, G. J. Olsen, T. M. Schmidt, and J. M. Tiedje. 2001. The RDP-II (Ribosomal Database Project). Nucleic Acids Res. **29**:173–174.

Muller, T., and M. Vingron. 2000. Modeling amino acid replacement. J. Comput. Biol. **7**:761–776.

Savill, N. J., D. C. Hoyle, and P. G. Higgs. 2001. RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. Genetics **157**:399–411.

Tillier, E. R. M., and R. A. Collins. 1995. Neighbor-Joining and maximum likelihood with RNA sequences: addressing the inter-dependence of sites. Mol. Biol. Evol. **12**:7–15.

———. 1998. High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. Genetics **148**:1993–2002.

Veerassamy, S., A. Smith, and E. R. M. Tillier. 2003. A transition probability model for amino acid substitutions from BLOCKS. J. Comput. Biol. **10**:997–1010.

Whelan, S., and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach. J. Mol. Evol. **18**: 691–699.

Wuyts, J., P. D. Rijk, Y. Van de Peer, T. Winkelmans, and R. D. Wachter. 2001. The European Large Subunit Ribosomal RNA Database. Nucleic Acids Res. **29**:175–177.

Wuyts, J., Y. Van de Peer, and R. D. Wachter. 2001. Distribution of substitution rates and location of insertion sites in the tertiary structure of ribosomal RNA. Nucleic Acids Res. **29**:5017–5028.

Wuyts, J., Y. Van de Peer, T. Winkelmans, and R. D. Watcher. 2002. The European Database on Small Subunit Ribosomal RNA. Nucleic Acids Res. **30**:183–185.

Yang, Z., and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo Method. Mol. Biol. Evol. **14**:717–724.

Brian Goldman, Associate Editor

Accepted September 24, 2003