JOURNAL OF MOLECULAR EVOLUTION

Toward Assigning Helical Regions in Alignments of Ribosomal RNA and Testing the Appropriateness of Evolutionary Models

Michael Schöniger,[1] Arndt von Haeseler[2]

[1] Theoretische Chemie, Technische Universität München, Lichtenbergstr. 4, D-85747 Garching, Germany
[2] Max-Planck-Institut für Evolutionäre Anthropologie, Inselstr. 22, D-04103 Leipzig, Germany

Received: 4 March 1999 / Accepted: 10 May 1999

Abstract. We suggest a nucleotide substitution model that takes correlation between base-paired nucleotides into account. The model includes the estimation of the transition–transversion ratio and allows inference of the shape parameter of a discrete gamma distribution to include rate heterogeneity. A Cox-test statistic, applied to a diatom ribosomal RNA alignment, shows that the suggested correlation model explains evolution of the stem region better than usual independence models. Moreover, the Cox-test procedure is extended to shed some light upon the problem of assigning helical regions in a secondary structure based alignment. This approach provides an estimate of the percentage of stem positions that do not appear to be correlated.

Key words: Evolutionary models — Correlated sequence sites — Secondary structure — Alignment — Rate heterogeneity — Phylogenetic analysis — Maximum-likelihood inference — Monte Carlo simulation

Introduction

Most methods of inferring phylogenetic relationships from sequence data are based on models of sequence evolution most prominently expressed in maximum-likelihood approaches (Felsenstein 1981) or in distance correction methods (Jukes and Cantor 1969; Kimura 1980; Tavaré 1986; Tamura and Nei 1993; Yang 1994a; Zharkikh 1994). The definition of a model usually requires some assumptions about the evolutionary process. Typically it is assumed that nucleotide sites evolved independently of each other. This assumption is clearly violated for sequences that display a distinct secondary structure, e.g., ribosomal RNA (rRNA) or transfer RNA. Nucleotides in the stem regions of these molecules obviously do not evolve independently of their base-pairing counterparts. Consequently, there have been several attempts to incorporate dependencies induced by base-pairing into a Markov model of sequence evolution (Schöniger and von Haeseler 1994; Tillier 1994; Muse 1995; Rzhetsky 1995; Tillier and Collins 1995, 1998). However, only the model by Rzhetsky (1995) accounts for rate heterogeneity among base-paired nucleotides. In this paper we suggest a slightly modified version of our previous model (Schöniger and von Haeseler 1994) that encompasses the Hasegawa–Kishino–Yano model (Hasegawa et al. 1985) as a special case. The current implementation of the model in the PUZZLE program (Strimmer and von Haeseler 1996) includes the estimation of the transition–transversion parameter and the amount of rate heterogeneity assuming a discrete gamma distribution (Yang 1994b).

Given this elaborate model of sequence evolution of base-paired regions of rRNA, we address the question about the appropriateness of this description. Thus, we apply the approach of Goldman (1993), who employed a test statistic suggested by Cox (1961, 1962) to check the adequacy of stochastic models. This approach was ap-

Correspondence to: Michael Schöniger; *e-mail:* schoeniger@theochem.tu-muenchen.de

plied successfully to various estimation problems in studies of molecular evolution (Goldman and Yang 1994; Yang et al. 1994, 1995; Huelsenbeck and Rannala 1997). Here we test the correlation model as an alternative to independence models like the one introduced by Hasegawa et al. (1985) that allows for arbitrary, stationary base frequencies and any transition–transversion bias. For the sake of illustration we analyze the stem regions of the diatom rRNA alignment by Medlin et al. (1996a).

To apply a correlation model one needs to find the appropriate assignment of helical regions in an rRNA alignment. Helices (stems) may be defined in a relatively straightforward way (Zuker and Stiegler 1981; Hofacker et al. 1994) for *single* rRNA sequences. But the task of deciding which columns of an entire *alignment* belong to a helix remains tedious and very time consuming, since the necessary adjustments must be made manually. In this paper, we demonstrate how the Cox-test methodology can be employed to estimate the percentage of stem positions that do not appear to be correlated.

Methods and Data

Models of Nucleotide Substitution

The substitution process at a given site is modeled as a homogeneous stationary Markov process where the instantaneous rate of change from state (nucleotide) i to j is typically defined by an $n \times n$ rate matrix $\mathbf{R^1}$ (Swofford et al. 1996). For example, the HKY (Hasegawa et al. 1985) matrix has entries

$$R^1_{ij} = \begin{cases} \alpha\pi_j & \text{for transitions: A} \leftrightarrow \text{G, C} \leftrightarrow \text{U} \\ \beta\pi_{\mathrm{j}} & \text{for transversions: A, G} \leftrightarrow \text{C, U} \end{cases} \tag{1}$$

where π_j (j = A, G, C, U) is the equilibrium frequency of nucleotide j. The diagonal elements of the rate matrix are defined by setting row sums equal zero. $\mathbf{R^1}$ describes the evolution of a nucleotide site. If we assume that each site in a sequence evolves according to $\mathbf{R^1}$ and independently of the rest, then this model of sequence evolution is applicable in a maximum-likelihood framework.

To model the evolution of base pairs in stem regions of RNA molecules, the state space is extended to the 16 possible dinucleotides with stationary frequencies π_μ (μ = AA, AG, AC, . . ., UU). The 16 × 16 rate matrix $\mathbf{R^2} = (R^2_{\nu\mu})$ is given by

$$R^2_{\nu\mu} = \begin{cases} \alpha\pi_\mu/\pi_i & \text{if} \quad D(\nu,\mu) = 1 \quad \text{for transitions} \\ \beta\pi_\mu/\pi_{\mathrm{i}} & \text{if} \quad D(\nu,\mu) = 1 \quad \text{for transversions} \\ 0 & \text{if} \quad D(\nu,\mu) = 2 \end{cases} \tag{2}$$

where $D(\nu,\mu)$ is the number of nucleotide differences between doublet ν and doublet μ, and π_i is the stationary marginal distribution of the nucleotide i, which remains unaffected when substituting doublet ν by doublet μ. This definition ensures that the HKY model is a special case of the doublet correlation model (DC). Instantaneous substitution rates of two nucleotides in one doublet are assumed to be zero. However, given the predominant frequencies of the admissible base-paired doublets, it is quite likely that a substitution at a non-base-pairing doublet will lead to a base-paired doublet within a relatively short time interval, representing a so-called compensatory mutation. The doublet frequency parameters π_μ are estimated from the aligned data. Throughout this study we use a symmetrized version of the DC model with parameters $\pi_{ij} = \pi_{ji}$ (e.g., $\pi_{\mathrm{GC}} = \pi_{\mathrm{CG}}$).

Since different positions in a sequence evolve at different rates, a site-dependent relative factor r is introduced and the rate matrix for that site, i.e., $\mathbf{R^1}$ or $\mathbf{R^2}$**,** is multiplied by r to emulate the effect of rapidly or slowly evolving sites. The distribution of relative rates is assumed to follow a gamma distribution,

$$g_a(r) = \frac{a^a r^{a-1}}{e^{ar}\Gamma(a)} \tag{3}$$

with expectation 1 and variance $1/a$ (Uzzell and Corbin 1971; Wakeley 1993). The parameter a specifies the shape of the distribution and thus the amount of rate heterogeneity along the sequence. If a tends to infinity, then rates are homogeneous. For $a \approx 1$ we observe a bell-shaped distribution indicative of weak rate heterogeneity. If $a \ll 1$, then strong rate heterogeneity is obtained, and the corresponding distribution of rates is L-shaped. If an evolutionary model includes rate heterogeneity, a Γ is appended to the abbreviation of the substitution model, e.g., DCΓ represents the doublet correlation model together with a gamma distribution.

Estimation of Model Parameters

Based on the model of sequence evolution M the likelihood $\ell(T \mid M, \mathscr{A})$ of a tree T for a sequence alignment $\mathscr{A}$ can be computed where the model parameters are estimated from the data. A tree $\widehat{T_M}$ is called the maximum-likelihood estimate if

$$\ell(\widehat{T_M} \mid M,\mathscr{A}) = \max_{T \in \tau} \{\ell(T|M,\mathscr{A})\} \tag{4}$$

where τ is the space of all possible trees. Note that $\widehat{T_M}$ represents the tree topology together with branch lengths and parameter estimates of M. We use version 4.0a of the PUZZLE program (Strimmer and von Haeseler 1996; Strimmer 1997) to compute $\widehat{T_M}$. The shape parameter a is estimated assuming the discrete gamma model (Yang 1994b) as implemented in PUZZLE. In all analyses we assumed five rate categories.

Statistical Tests Using Monte Carlo Simulation

Different substitution models can produce different trees and different likelihoods. In a statistical framework a method is required to decide which model fits the data better. If the models are nested, i.e., the simple model is a special case of the more complex one, then the usual χ^2 approximation to the likelihood-ratio statistics may apply (Navidi et al. 1991). Unfortunately, one faces, among other things, a serious sample size problem that casts some doubts on the reliability of this approach (Goldman 1993). Goldman (1993) suggested, based on work by Cox (1961, 1962), a method to decide which one of two models or hypotheses provides a better fit to the data. The test makes use of the log-likelihoods $S_0 = \log\ell(\widehat{T_{M_0}} \mid M_0, \mathscr{A})$ and $S_1 = \log\ell(\widehat{T_{M_1}}|\mathrm{M}_1, \mathscr{A})$ of two competing models M_0 and M_1, namely,

$$\delta_{M_1-M_0} = S_1 - S_0 \tag{5}$$

Since the distribution of statistic (5) is not known, it is estimated by Monte Carlo simulation (Goldman 1993). A large number (1000 in this study) of simulated data sets is generated under the null hypothesis that $\widehat{T_{M_0}}$ represents the evolutionary history of the sequences (cf. Schöniger and von Haeseler 1995; Rambaut and Grassly 1997; unpublished modifications of the programs). For each simulated alignment i = 1, . . ., 1000, the differences δ_i according to Eq. (5) are computed. Note that the computation of δ_i requires the maximum-likelihood estimation of two trees $\widehat{T_{M_0}}$ and $\widehat{T_{M_1}}$, whose topologies are not necessarily

HELIX 1

1	Pelagomonas	cuggu....gccag		cg ua gc gc ug
2	Aulacoseira	cuggu....gccag		cg ua gc gc ug
3	Melosira	cuggu....gccag		cg ua gc gc ug
4	Stephanopyxis	cuggu....gccag	⟶	cg ua gc gc ug
5	Rhizosolenia	cuggu....gccag		cg ua gc gc ug
6	Ditylum	cuggu....gccag	rearrangement	cg ua gc gc ug
7	Fragilaria	cuggu....gccag		cg ua gc gc ug
8	Cymatosira	cuggu....gccag		cg ua gc gc ug
9	Chaetoceros	cuggu....gccag		cg ua gc gc ug

HELIX 4

1	Pelagomonas	cuca.....ugag		cg ua cg au
2	Aulacoseira	Auua.....ugaG		AG ua ug au
3	Melosira	Auua.....ugaG		AG ua ug au
4	Stephanopyxis	AuuU.....UgaG	⟶	AG ua ug UU
5	Rhizosolenia	cuca.....ugag		cg ua cg au
6	Ditylum	cuca.....ugag	rearrangement	cg ua cg au
7	Fragilaria	cuca.....ugag		cg ua cg au
8	Cymatosira	cuca.....ugag		cg ua cg au
9	Chaetoceros	cuca.....ugag		cg ua cg au

Fig. 1. Helices 1 and 4 of the secondary structure-based alignment of diatom small-subunit rRNA. Medlin et al. (1996b) used *uppercase letters* for bases that do not belong to helices; i.e., in their secondary structure models, helix 4 may consist of two, three, or four canonical base pairs.

identical. If $\delta_{M_1-M_0}$ falls below the 95th percentile of the empirical distribution obtained from the δ_i values, then M_0 is not rejected.

However, it is not necessary to compare two models of sequence evolution together with the resulting maximum-likelihood trees. One may also ask how well are the data described by $\widehat{TM}_0$ compared to the unconstrained hypothesis (Navidi et al. 1991; Goldman 1993). The unconstrained hypothesis makes no phylogenetic inferences but uses only the observed frequencies of character patterns ξ in the alignment. If independence is assumed, there are 4^N patterns for N aligned sequences, whereas the dinucleotide model (DC) has 16^N patterns. Let L_ξ be the number of occurrences of pattern ξ in an alignment of length L, then the log-likelihood of the unconstrained hypothesis (UC) is calculated to be

$$S_1 = \sum_\xi L_\xi \log L_\xi - L \log L \qquad (6)$$

(Navidi et al. 1991). We note that for dinucleotide models, the length of the alignment reduces to $L/2$.

Data

Medlin et al. (1996a) published an alignment of 34 small-subunit (SSU) rRNA sequences of diatoms consisting of 2076 nucleotides.[1] To keep the computation time reasonable, we selected 9 of the 34 species, namely, the 4 clade I diatoms *Aulacoseira distans, Melosira varians, Stephanopyxis* cf. *broschii,* and *Rhizosolenia setigera;* the 4 clade II diatoms *Ditylum brightwelli, Fragilaria striatula, Cymatosira belgica,* and *Chaetoceros rostratus;* and *Pelagomonas calceolata* as outgroup. Medlin et al. (1996b) and Chesnick et al. (1997) suggested, for a variety of diatom SSU rRNAs, secondary structure models that were used as guidelines to assign helical regions to the SSU rRNA alignment. As Fig. 1 illustrates, this is sometimes straightforward (Fig. 1, top). Sometimes, however, it is difficult to decide which columns of the alignment constitute the overall helix (Fig. 1, bottom). Although the DC model explicitly allows for intermediates like AG or UU, it is not clear which columns should be included in an overall helix. Based on visual inspection of the available secondary structure information, we produced a stem data set, called STEM960, consisting of 480 doublets (960 paired nucleotides).

Results

Figure 2 displays the estimated maximum-likelihood trees for STEM960 assuming an HKY model (Fig. 2, left) or a DC model (Fig. 2, right). The branching patterns of the trees are slightly different, e.g., the clade II diatoms constitute a monophyletic group in the $\widehat{T}_{\mathrm{DC}}$ tree. Both trees are not fully resolved. The multifurcations are possibly due to the short alignment length. Medlin et al. (1996a) found clade I diatoms as the sister group of clade II diatoms, which is not supported by the trees in Fig. 2. However, the polytomies can be resolved, and the monophyly of the clade I diatoms is confirmed, if a combined

[1] Some positions were excluded due to alignment uncertainties indicated by the authors (Medlin, personal communication).

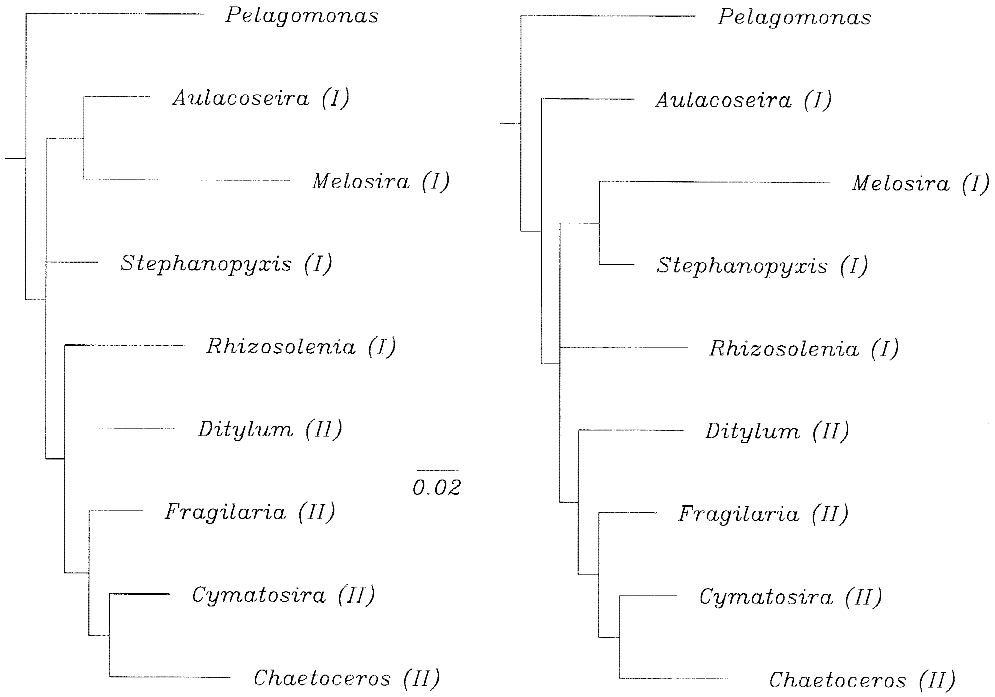

Fig. 2. Trees of the diatom small-subunit rRNA data set STEM960 obtained with PUZZLE using HKY (**left**) and DC (**right**). *Roman numbers* indicate the affiliation of species to clade I and II (Medlin et al. 1996a). Branch lengths are proportional to the number of substitutions per nucleotide site.

data set of stem and loop positions is analyzed with PUZZLE, no matter whether DC or HKY is used for the stem regions (results not shown). Trees $\widehat{T}_{\mathrm{HKY}}$ and $\widehat{T}_{\mathrm{DC}}$ were used as the basis for the computation of the empirical δ distribution.

The score S_0 of $\widehat{T}_{\mathrm{HKY}}$ equals −3661.67 and S_1, the score of $\widehat{T}_{\mathrm{DC}}$, is equal to −3123.89, indicating a substantial increase in the log-likelihood value. This improvement is highly significant. Figure 3 shows the empirical δ distribution if HKY serves as the null model that is tested against the DC model. The mean of this distribution is $\delta = 1.43$, with an empirical standard deviation of $\sigma = 2.22$. Thus the observed value of $\delta = 537.78$ is 242 standard deviations away from $\bar{\delta}$. Therefore, we conclude that DC is a better model for the stem region than HKY.

One may now ask if this improvement is a typical value for sequences that evolved under a DC model. To answer this question, 1000 sets of sequences were generated under $\widehat{T}_{\mathrm{DC}}$ (the right tree in Fig. 2) and analyzed assuming HKY as well as DC. The resulting empirical distribution of the gain in improvement has a mean of 685.61 and a standard deviation of 30.37 (see Fig. 3). The observed improvement $\delta = 537.78$ is −4.87 units of standard deviations away from the average improvement, that is, left from the expected distribution of improvements. In other words, the improvement for the data provided by the DC model is too poor compared to simulated sequences. Thus, there must be some sites in data set STEM960 that do not evolve according to the DC model. One possible explanation is that the data set contains base pairs that are not correlated. This hypothesis is corroborated by the following experiment.

A certain proportion x of the 960 nucleotide positions evolved according to $\widehat{T}_{\mathrm{DC}}$, whereas the remaining part $1 - x$ evolved according to $\widehat{T}_{\mathrm{HKY}}$. One thousand simulated alignments were generated for each $x = 0, 10, 20, \ldots,$ 100% and the averages $\delta_{\mathrm{DC-HKY}}(x)$ for each fraction x were calculated. Figure 4 displays $\delta_{\mathrm{DC-HKY}}(x)$ as a function of x, together with the empirical 95% confidence limits. The graph shows that the average improvement of the DC model increases as the proportion x of sites evolving according to DC increases. If $x = 100\%$, the average improvement equals 685.61.

Since STEM960 led to an improvement of 537.78, Fig. 4 was used to estimate the fraction of sites that evolve according to DC. About 92% of the positions are in accord with the DC model, with an approximate confidence interval of 89–96%. Thus, STEM960 contains about 76 nucleotide positions that do not conform to the DC model. Unfortunately, we have no information about the location of these sites.

Reevaluation of the Secondary Structure-Based rRNA Alignment

The assignment of helical regions in multiple rRNA alignments can be very difficult. The results of our preceding analysis suggest that STEM960 includes assignments of base pairs that we compiled incorrectly. Therefore, we reevaluated the alignment, taking into account the secondary structure data provided in the original data

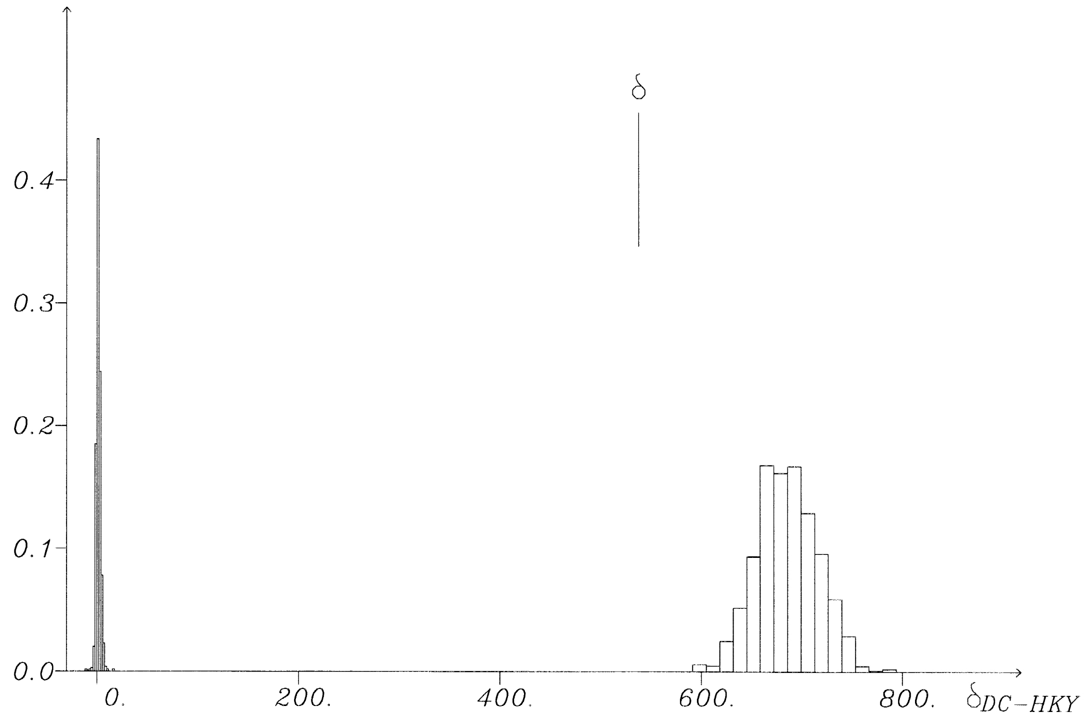

Fig. 3. Distributions of the statistic $\delta_{\text{DC–HKY}}$ for STEM960 (480 doublets). The *left* distribution belongs to the null model $\widehat{T_{\text{HKY}}}$, whereas $\widehat{T_{\text{DC}}}$ serves as the model tree for the *right* distribution. One thousand replicates were simulated using the model trees in Fig. 2. The *line* shows the observed $\delta_{\text{DC–HKY}}$ value of 537.78.

set (Medlin, personal communication). We found three instances in the data set STEM960 where we had ignored a looped-out nucleotide in one strand of the helical region. This led to a wrong assignment of base pairs in the rest of the helix, i.e., we got many nonclassical doublets like UC, UU, and AA.

Some positions of STEM960 contained, in addition to classical base pairs (GC, AU, GU), sporadically columns with a high frequency of nonstandard base pairs (like AA or AC). These columns represent inner loops rather than stems. Although the DC model explicitly includes such intermediates along the evolutionary path as elements of stable helices, it is not very likely to observe their fixation in many species. Therefore we excluded positions from the analysis that showed at least eight (of nine) nonclassical doublets. This reduced the number of columns in helical regions by 18 doublets (36 nucleotides). Together with the reexamination of the frame shifts, the data set STEM924 of length 462 doublets (924 nucleotides) was created and is used in the rest of the paper.

Application of PUZZLE assuming HKY and DC provided maximum-likelihood trees that served as the basis for the computation of the empirical $\delta_{\text{DC–HKY}}$ distribution. The branching patterns of these trees are very similar to those in Fig. 2.

Toward More Complex Models

For STEM924 the Cox test to compare DC and HKY is repeated. The improvement of the score due to the introduction of the DC model compared to HKY is significant (Table 1, test 1). Moreover, the $\delta_{\text{DC–HKY}}$ value of 701.20 falls inside the distribution of improvements if sequences actually evolve according to DC (Fig. 5). It is only $1.00\sigma = 30.96$ away from the mean $\bar{\delta}_{\text{DC–HKY}} = 732.18$. Hence, we have no reason to mistrust the DC model. However, comparison of the log-likelihoods of the DC model and the UC model for 16 dinucleotides shows that DC does not adequately describe the evolution of the data (Table 1, test 2).

If we introduce rate heterogeneity, then the DCΓ model provides again a significant improvement compared to the DC model (Table 1, test 3). However, the DCΓ model is still not appropriate to describe the evolutionary processes that have led to the aligned data. However, if the goodness of fit is measured in units of the empirical standard deviations, DCΓ is, with a $\tilde{\delta} = 4.00$, much closer to the mean of the simulated distribution $\delta_{\text{UC}-M_0}$ than the DC model ($\tilde{\delta} = 10.36$) (cf. Table 1, tests 2 and 4). Thus, one might speculate that the entire sequences or subsets thereof have evolved under a different, yet unresolved model.

Discussion

We have shown that a model which takes correlation into account significantly enhances the description of the evolutionary forces acting upon the stem region of ribosomal RNA. A simple HKY model cannot account for

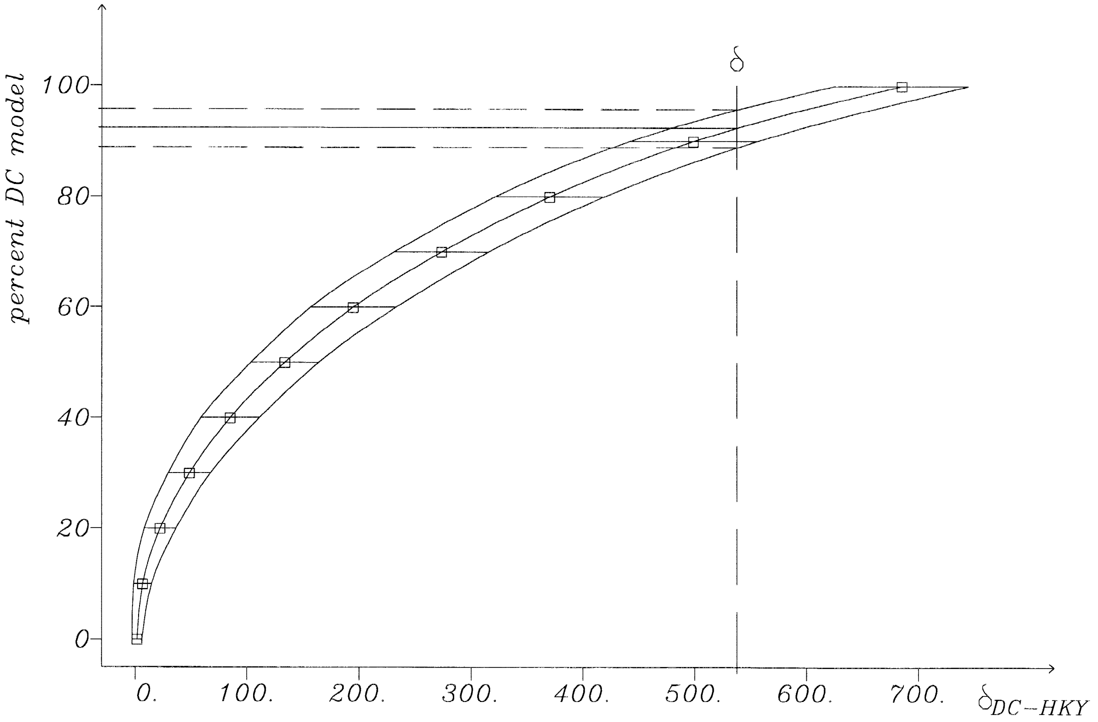

Fig. 4. Averages and approximate 95% confidence intervals (average ± 2 SD) of the $\delta_{\mathrm{DC-HKY}}$ distribution obtained from simulation assuming a mixture model: some of the positions evolve according to DC, and the rest according to HKY. The *vertical line* shows the value of 537.78 for STEM960. The *horizontal lines* display the estimated proportion (*solid*) of positions evolving according DC and its 95% confidence interval (*dashed*).

Table 1. Comparison of different models of sequence evolution for STEM924

Test No.	M_0	M_1	S_0	S_1	δ	$\bar{\delta}$	σ	$\tilde{\delta}$	M_0
1	HKY	DC	−3479.40	−2778.20	701.20	1.54	2.11	331.67	Rejected
2	DC	UC	−2778.20	−1611.64	1166.56	719.32	43.17	10.36	Rejected
3	DC	DCΓ	−2778.20	−2539.47	238.73	0.00	0.97	246.66	Rejected
4	DCΓ	UC	−2539.47	−1611.64	927.83	753.87	43.43	4.00	Rejected

[a] S_0 and S_1 are the log-likelihood of the corresponding hypotheses; δ is the observed difference between S_0 and S_1 for the data; $\tilde{\delta} = (\delta - \bar{\delta})/\sigma$ is the normalized difference, where $\bar{\delta}$ and σ are the empirical mean and standard deviation based on the simulated distribution of δ values.

the complexities inherent in the data. The goodness of fit can be further improved if we allow for rate heterogeneity among the doublets. A DCΓ model describes the data better than a DC model. This observation matches results for models assuming independently evolving sites (Goldman and Yang 1994; Yang et al. 1994, 1995). Although the DCΓ model is the "best" model, it does not suffice to explain the full variability in terms of doublet patterns in the alignment compared to the unconstrained hypothesis. Thus, there is ample space to refine the model suggested here. One should also note that we have only studied one data set. The analysis of further examples will shed more light on the suitability of DCΓ.

While the above-mentioned methodology to compare different models is straightforward, we suggest the use of the empirical distribution of δ-values, if sequences actually evolved under a more complex model, to detect deviations from the complex model. As an example, we showed how this approach may be used to estimate the amount of positions that do not evolve according to a DC model but, rather, to an HKY model. For STEM960 we estimated that we had falsely assigned about 8% of the sites to helical regions, which turned out to coincide fairly well with the number of positions that were modified or excluded after reanalysis of the alignment. This procedure, however, is tedious and time-consuming. It is therefore desirable to develop an automated version of our approach. However, we note the danger of circularity. If the alignment is modified to fit the model, then of course the model will fit better. Thus, independent data are needed to substantiate the claim that the alignment and the model are both better. Moreover, it is certainly worthwhile to investigate the applicability of the presented ideas by analyzing more data sets. In addition, it

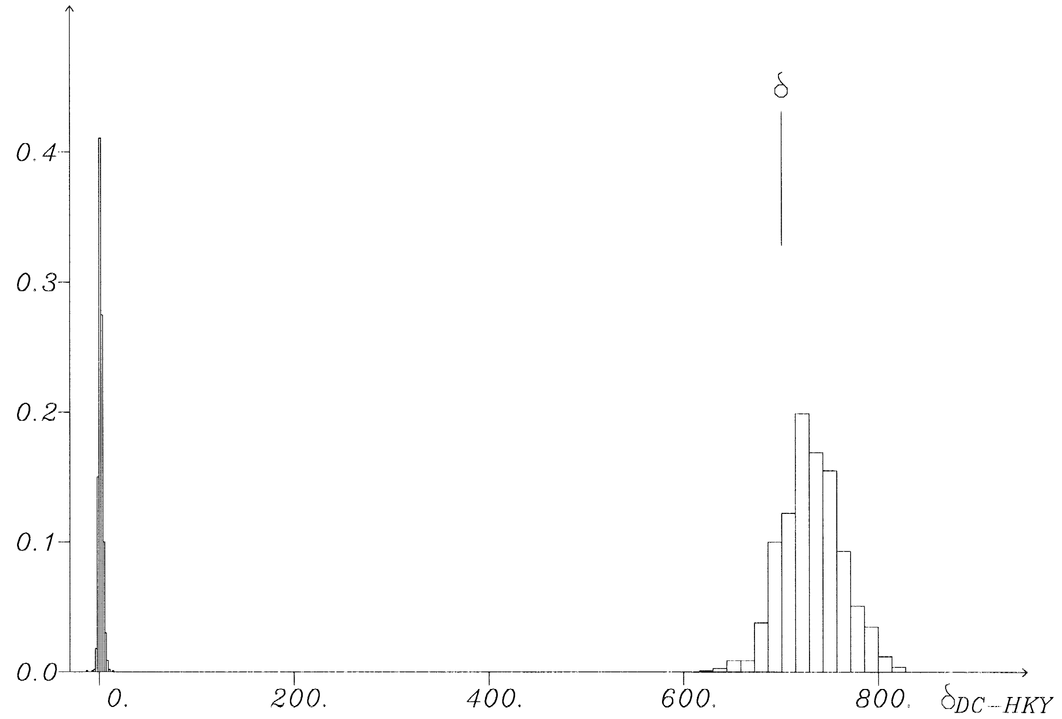

Fig. 5. Distributions of the statistic $\delta_{\mathrm{DC-HKY}}$ for STEM924 (462 doublets). The *left* distribution belongs to the null model $\widehat{T_{\mathrm{HKY}}}$, whereas $\widehat{T_{\mathrm{DC}}}$ serves as the model tree for the right distribution. One thousand replicates were simulated using the model trees obtained by PUZZLE, which are only slightly different from the ones shown in Fig. 2. The *line* shows the observed $\delta_{\mathrm{DC-HKY}}$ value of 701.20 for STEM924.

should be possible to use our test methodology to detect regions of misaligned positions.

Acknowledgments. We are grateful to Linda Medlin for kindly providing us with the aligned diatom data, useful information regarding secondary structure, and helpful discussions. We thank David Pollock for discussions on nestedness of stochastic models, Korbinian Strimmer for the latest versions of PUZZLE, and David Swofford for discussions and valuable suggestions. Financial support by the Deutsche Forschungsgemeinschaft and the Fonds der Chemischen Industrie is gratefully acknowledged.

References

Chesnick JM, Kooistra WHCF, Wellbrock U, Medlin LK (1997) Ribosomal RNA analysis indicates a benthic pennate diatom ancestry for the endosymbionts of the dinoflagellates Peridinium foliaceum and Peridinium balticum (Pyrrhophyta). J Euk Microbiol 44:314–320

Cox DR (1961) Tests of separate families of hypotheses. Proceedings of the 4th Berkeley Symposium (University of California Press) 1:105–123

Cox DR (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

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

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 725–736

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

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P (1994) Fast folding and comparison of RNA secondary structures. Monatsh Chem 125:167–188

Huelsenbeck JP, Rannala B (1997) Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. Science 276:227–232

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

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Medlin LK, Kooistra WHCF, Gersonde G, Wellbrock U (1996a) Evolution of the diatoms (Bacillariophyta). II. Nuclear-encoded small-subunit rRNA sequence comparisons confirm a paraphyletic origin for the centric diatoms. Mol Biol Evol 13:67–75

Medlin LK, Kooistra WHCF, Gersonde G, Wellbrock U (1996b) Evolution of the diatoms (Bacillariophyta). III. Molecular evidence for the origin of the Thalassiosirales. Nova Hedw 112:221–234

Muse SV (1995) Evolutionary analysis of DNA sequences subject to constraints on secondary structure. Genetics 139:1429–1439

Navidi WC, Churchill GA, von Haeseler A (1991) Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. Mol Biol Evol 8:128–143

Rambaut A, Grassly NC (1997) Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci 13:235–238

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

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

Schöniger M, von Haeseler A (1995) Simulating efficiently the evolution of DNA sequences. Comput Appl Biosci 11:111–115

Strimmer K (1997) Maximum likelihood methods in molecular phylogenetics, PhD thesis. Herbert Utz Verlag, München

Strimmer K, von Haeseler A (1996) Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. Mol Biol Evol 13:964–969

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK (eds) Molecular systematics, 2nd ed. Sinauer Associates, Sunderland, MA, pp 407–514

Tamura K, Nei M (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

Tavaré S (1986) Some statistical aspects of the primary structure of nucleotide sequences. In: Miura RM (ed) Lectures on mathematics in the life sciences, Vol 17. American Mathematical Society, Providence, RI, pp 57–86

Tillier ERM (1994) Maximum likelihood with multiparameter models of substitution. J Mol Evol 39:409–417

Tillier ERM, Collins RA (1995) Neighbor joining and maximum likelihood with RNA sequences: Addressing the interdependence of sites. Mol Biol Evol 12:7–15

Tillier ERM, Collins RA (1998) High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. Genetics 148:1993–2002

Uzzell T, Corbin KW (1971) Fitting discrete probability distributions to evolutionary events. Science 172:1089–1096

Wakeley J (1993) Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J Mol Evol 37:613–623

Yang Z (1994a) Estimating the pattern of nucleotide substitution. J Mol Evol 39:105–111

Yang Z (1994b) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J Mol Evol 39:306–314

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

Yang Z, Goldman N, Friday A (1995) Maximum likelihood trees from DNA sequences: A peculiar statistical estimation problem. Syst Biol 44:384–399

Zharkikh A (1994) Estimation of evolutionary distances between nucleotide sequences. J Mol Evol 39:315–329

Zuker M, Stiegler P (1981) Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 9:133–148