6184–6193 *Nucleic Acids Research, 2009, Vol. 37, No. 18* *Published online 15 September 2009*
doi:10.1093/nar/gkp600

Accurate and efficient reconstruction of deep phylogenies from structured RNAs

Roman R. Stocsits[1,*], Harald Letsch[1], Jana Hertel[2,*], Bernhard Misof[3] and Peter F. Stadler[2,4,5,6,7]

[1]Zoologisches Forschungsmuseum Alexander Koenig, Bonn, [2]Bioinformatics Group, Department of Computer Science, Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, [3]UHH Biozentrum Grindel & Zoologisches Museum, Hamburg, [4]Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, [5]RNomics Group, Fraunhofer Institut für Zelltherapie und Immunologie, Deutscher Platz 5e, D-04103 Leipzig, Germany, [6]Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria and [7]Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA

Received March 25, 2009; Revised June 29, 2009; Accepted July 01, 2009

ABSTRACT

Ribosomal RNA (rRNA) genes are probably the most frequently used data source in phylogenetic reconstruction. Individual columns of rRNA alignments are not independent as a consequence of their highly conserved secondary structures. Unless explicitly taken into account, these correlation can distort the phylogenetic signal and/or lead to gross overestimates of tree stability. Maximum likelihood and Bayesian approaches are of course amenable to using RNA-specific substitution models that treat conserved base pairs appropriately, but require accurate secondary structure models as input. So far, however, no accurate and easy-to-use tool has been available for computing structure-aware alignments and consensus structures that can deal with the large rRNAs. The `RNAsalsa` approach is designed to fill this gap. Capitalizing on the improved accuracy of pairwise consensus structures and informed by *a priori* knowledge of group-specific structural constraints, the tool provides both alignments and consensus structures that are of sufficient accuracy for routine phylogenetic analysis based on RNA-specific substitution models. The power of the approach is demonstrated using two rRNA data sets: a mitochondrial rRNA set of 26 Mammalia, and a collection of 28S nuclear rRNAs representative of the five major echinoderm groups.

INTRODUCTION

Ribosomal RNAs (rRNAs) are the most widely used source of phylogenetic information, although protein-coding genes, often derived from Expressed Sequence Tag (EST) sequencing or from sequencing complete mitogenomes, have provided an increasingly large amount of new genomic data. The SSU and LSU rRNA genes have been sequenced for thousands of taxa throughout the metazoan kingdom, providing a much denser taxon coverage than what is available for any particular protein-coding gene. Since sequence conservation varies dramatically between different regions of rRNA genes, these data are informative on a wide range of phylogenetic time scales, ranging from recent to ancient splits (1,2).

This variation in substitution rates, however, is also a major technical obstacle for using rRNA in molecular phylogenetics. The correct assignment of homologous characters, i.e. alignment columns, is the crucial first step in molecular systematics on which all subsequent analyses depend. The high variability of substitution rate along the sequence, combined with similar variations in insertion and deletion rate, makes it impossible in practice to construct unambiguous alignments of the more variable regions by means of standard sequence alignment programs.

The rRNAs, however, are highly structured, with large parts of the molecules exhibiting very strong conservation of their base pairing patterns. Therefore, it is natural to improve alignment accuracy by incorporating secondary structure conservation. Indeed, this approach has been advocated repeatedly in the literature, e.g. (3–7).

*To whom correspondence should be addressed. Tel: +49 341 97 16686; Fax: +49 341 97 16679; Email: jana@bioinf.uni-leipzig.de
Correspondence may also be addressed to Roman R. Stocsits. Tel: +49 228 9122 352; Fax: +49 228 9122 295; Email: roman.stocsits@gmail.com

In practise, however, the application of this idea has remained a hard and tedious task, mostly because of the difficulties in obtaining a correct structural annotation. If good structure annotations were readily available, we could simply employ one of the alignment tools that explicitly incorporate secondary structure information (8–15).

The accuracy of thermodynamic folding algorithms falls sharply as the length of the RNA increases. Secondary structures can be computed with acceptable precision based on experimentally measured thermodynamic parameters only for short fragments with a length $\lesssim$100 nt (16,17). This limitation is in part due to inaccuracies in the 'nearest neighbor model' and its parameters (18,19), in part caused by the kinetics of folding process and tertiary interactions (20). Even more importantly, the RNA and protein components of the ribosome are tightly packed and thus mutually influence their folds (21). The functional rRNA structures, therefore, cannot reasonably be expected to be identical with the structures of isolated rRNAs—which is what the thermodynamic folding algorithms compute.

The latter effect can be captured at least in part by considering patterns of sequence covariation that provide information on evolutionarily conserved base pairs. The `RNAalifold` approach, for instance, has demonstrated that the accuracy of biological structure predictions can be increased to acceptable levels by using the consensus structures of a set of closely related sequences and by explicitly taking information on base pair covariation into account (22,23). Designed for relatively closely related sequences, `RNAalifold` unfortunately requires a sequence alignment as input.

`RNAsalsa` is designed to overcome this limitation by combining the prediction of consensus structures of closely related sequences with prior knowledge that constrains the set of acceptable structures. Consensus structures for groups of related sequences are used to generate high-quality alignments by funneling structure information into the alignment scoring function. Thus, `RNAsalsa` uses both structure information for adjusting and refining the sequence alignment and sequence information contained in the alignment to refine the structure predictions. We designed `RNAsalsa` primarily for phylogenetic applications. In this context, RNA secondary structure is of importance at two levels: first, changes in secondary structures can be useful phylogenetic markers in their own right (24). Clearly, accurate structure predictions are a necessary prerequisite to utilize structural differences in this way. Second, knowledge about conserved secondary structures allows the use of more detailed models of RNA sequence evolution. In this contribution we focus on the latter aspect.

RNA-specific substitution models (25–30) have been introduced to capture the effects of covariation among paired sites of rRNA sequences. Slightly deleterious substitutions at one side of a helix, which would disrupt the structure, are frequently compensated by a second substitution at the pairing site, so that the pairing ability (31) is restored. This leads to a strong correlation of paired positions within rRNA sequences. The corresponding alignment columns, therefore, do not display independent phylogenetic information. Since paired sites are strongly correlated but treated as independent, phylogenetic information is scored twice, leading to unjustified high support for some trees and erroneously low support for alternative trees (32,33). The application of RNA-specific substitution models for phylogenetic analyses requires a structural annotation of the input alignment. It is important that in particular the regions that are treated as paired are aligned in a structurally correct way because the effect of substitutions is evaluated with respect to specific column pairs.

The `RNAsalsa` software, which is implemented in `C`, is specifically optimized for phylogenetic applications. The source code as well as pre-compiled executables for various platforms, together with a detailed manual providing some guidelines for practical use, may be downloaded from http://www.rnasalsa.zfmk.de/ and http://www.bioinf.uni-leipzig.de/Software/RNAsalsa.

MATERIALS AND METHODS

Workflow and algorithms

`RNAsalsa` implements a workflow that makes use of several well-established algorithms for both RNA secondary structure prediction and structure enhanced alignment (Figure 1). It consists of three major stages: (i) the determination of structural constraints for each input sequence that emphasize the common features; (ii) the computation of a detailed structural model for each input sequence; and (iii) the construction of a final multiple sequence alignment together with a consensus structure.

The starting point for `RNAsalsa` is an initial alignment $\mathbb{A}^0$ of a collection $\{x^1, \ldots, x^N\}$ of homologous RNA sequences (produced e.g. simply by `clustalW`) and an *a priori* secondary structure constraint σ for a single sequence x^0 which is contained in the alignment $\mathbb{A}^0$. The sequence x^0 and its structural model are used only to initialize the structure prediction and alignment process.

In the first step, `RNAsalsa` checks the consistency of the initial alignment $\mathbb{A}^0$ and the initial constraint σ^0: for each base pair in σ^0, we evaluate whether the corresponding aligned positions of a sufficient number of sequences in $\mathbb{A}^0$ can also pair. If so, we retain the base pair, otherwise it is removed from the constraint. The resulting 'relaxed' constraint

$$\sigma = \text{filter}(\sigma^0|\mathbb{A}) \quad \mathbf{1}$$

can be seen as base pair-wise filtering of the initial constraint σ^0 that removes pairs from σ^0 that are largely inconsistent with the initial alignment. Projecting the relaxed constraint σ separately onto each aligned sequence (i.e. retaining only the canonical base pairs of σ that can be formed by the input sequence x^i), then produces distinct initial structure constraints for each sequence:

$$\sigma^i = \text{projection}(\sigma, x^i|\mathbb{A}) \quad \mathbf{2}$$

Up to this point, the result heavily depends upon the initial alignment. It may not cover the input sequences uniformly,

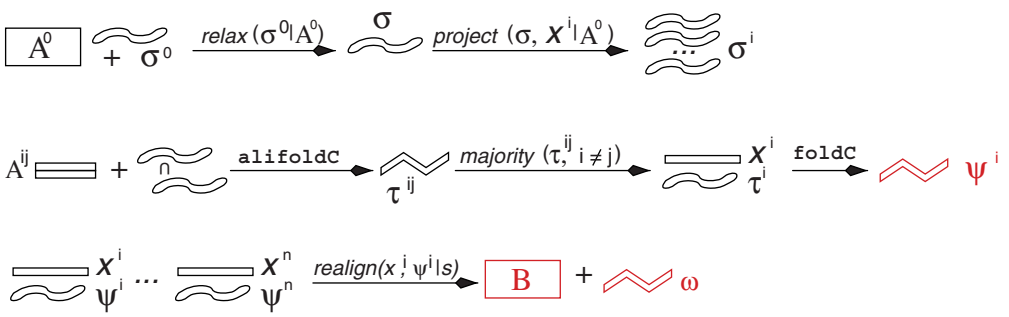

Figure 1. Overview of the RNAsalsa workflow. Starting from an initial sequence alignment $\mathbb{A}^0$ and a structure constraint σ^0, a relaxed consensus constraint and then constraints for each input sequence are produced. Pairwise sequence alignments and the intersections of the two individualized constraints are used to compute a collection of constrained pairwise consensus structures τ_{ij}, from which a refined individualized structure constraint τ^i is extracted by means of a majority voting procedure. Then τ^i is used as constraint in minimum energy folding to obtain the final structural annotation ψ^i for input sequence x^i. The structure-annotated input sequences are now aligned with a structure-aware alignment method, resulting in the final alignment $\mathbb{B}$ and a corresponding consensus structure ω.

in particular it will often be concentrated on the well-conserved (and therefore properly aligned) regions.

In the second step, RNAsalsa utilizes the improved accuracy of the predicted consensus structures. To this end, we construct a collection of pairwise sequence alignments $\mathbb{A}^{ij}$ from the input sequences x^i and x^j. These can be constructed in different ways, either by dynamic programming alignment, or by projecting the corresponding sub-alignment of $\mathbb{A}^0$. For details we refer to the RNAsalsa manual, which is part of the Supplementary Data. For each of the pairwise alignments $\mathbb{A}^{ij}$ we compute the consensus minimum free-energy structures

$$\tau^{ij} = \text{RNAalifold}(\mathbb{A}^{ij}|\sigma^i \cap \sigma^j) \quad \mathbf{3}$$

using the base pairs common to the projected structures σ^i and σ^j, respespectively, as constraint. This step uses the Vienna RNA Package library functions underlying RNAalifold (22) to perform the constrained folding computations.

For each sequence x^i, the collection of structures $\{\tau^{ij}, i \neq j\}$ taken together defines a set of base pairs on x^i that are both thermodynamically plausible and conserved in at least one other sequence of the input set. From this set of pairs, we select a single secondary structure

$$\tau^i = \text{majority}(\{\tau^{ij}, j \neq i\}). \quad \mathbf{4}$$

for sequence x^i using a majority voting procedure. RNAsalsa currently implements a simple greedy procedure that selects the most frequent base pairs first and rejects pairs that would cross previously selected ones to avoid the formation of pseudo-knotted structures. Alternatively, one could also use Nussinov's Maximum Circular Matching algorithm (34) to retrieve a maximum weight sub-set of non-intersecting pairs. The base pairs of τ^i, which by construction typically contain most of the initial constraint-derived pairs σ^i, are now used as a constraint for computing the final secondary structure prediction

$$\psi^i = \text{RNAfold}(x^i|\tau^i), \quad \mathbf{5}$$

for each input sequence x^i. Again, we employ the Vienna RNA Package library here.

The purpose of the entire—rather complex and computationally expensive—procedure is to use as much information as possible in guiding the last step, the computation of the secondary structure models ψ^i for each input sequence. This guiding information is derived from two sources: the initial structural constraint σ and the ensemble of plausible base pairs generated from all pairwise alignments.

In the next step, the sequence–structure pairs (x^i, ψ^i) are realigned. To this end, RNAsalsa uses a hierarchical progressive alignment based on pairwise dynamic programming alignments with affine gap costs (35). The scoring function explicitly incorporates the secondary structure annotation: the (mis)match score $s(x_i, y_j)$ of position i from sequence x with position j from sequence y is defined as follows:

$$s(x_i, y_j) = b_0 s_m(x_i, y_j) + b_1 s_n(x_{\pi(i)}, y_{\pi(j)}) + c s_p(x_i, y_j) \quad \mathbf{6}$$

where $x_{\pi(i)}$ and $y_{\pi(j)}$ denote the pairing partners of x_i and y_j in their respective secondary structures. The coefficient $b_0 = 1$ if both x_i and y_j are paired nucleotides. The coefficient b_1 is set to 1 if x and y share sufficient structural conservation to a certain extent that overcame the precedent filtering steps and if $x_{\pi(i)}$ and $y_{\pi(j)}$ are located either both upstream or both downstream of x_i and y_j, respectively. Otherwise the structural contribution is ignored, $b_1 = 0$. Finally, if one x_i or y_j are unpaired, then $b_0 = b_1 = 0$ and $c = 1$. In regions without structural information we therefore use a pure nucleic acid sequence score s_p, where as in structured regions, the modified scoring functions s_m and s_n are used. For instance, within trusted structural regions **A-G** is scored as a match because both may pair with **U**, while it is not in regions without sufficiently trusted structural information. Default scoring tables are listed in the manual and Supplementary Data. The final result is a global re-alignment $\mathbb{B}$ of the input sequences which integrates all the secondary structure information obtained in the previous steps.

The individual folds ψ^i and the alignment $\mathbb{B}$ are used to derive a consensus structure

$$\omega = \text{consensus}(\{\psi^i\}|\mathbb{B}) \quad \mathbf{7}$$

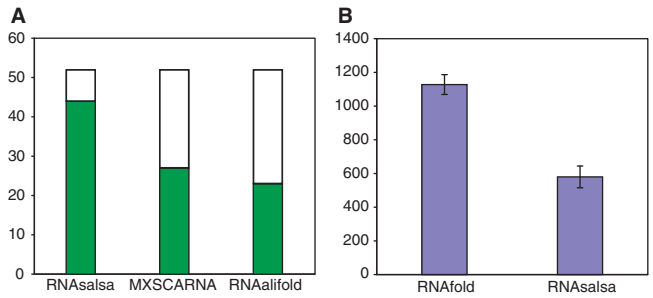

Figure 2. Accuracy of structure prediction. Fraction of correctly predicted helices (green bars) compared with the mammalian 16S rRNA consensus models (36). (**A**) `RNAsalsa` significantly outperforms `MXSCARNA` and `RNAalifold` (default parameter settings; three-sample test for equality of proportions without continuity correction; $\chi^2 = 19.96$, $df = 2$, $P < 0.0001$). Average tree-edit distance (37) (**B**) between predicted individual structures and the mammalian 16S rRNA reference model. `RNAsalsa` predictions conform the consensus model much better (paired sampled t-test; $t = 33.46$, $df = 1$, $P < 0.0001$; $N = 26$).

Since we now have the trusted alignment $\mathbb{B}$, we can again employ a simple voting strategy: we start from the set of all base pairs that appear sufficiently often in superposition of the ψ^i. Again, we use a greedy strategy to avoid conflicting base pairs (note that no conflicts can arise if we consider only base pairs that occur at least $N/2$ times).

Several parameters can be adjusted in the process. In particular, the stringency of the initial filtering of base pairs, Equation (1), and the two majority voting procedures, Equations (4) and (7) can be adjusted by the user to the peculiarities of the data sets. In each case, a threshold for the minimum number of consistent pairs can be specified. Some guidelines for practical use can be found in the `RNAsalsa` manual.

RESULTS

Secondary structure prediction

Performance of `RNAsalsa`'s structure predictions was evaluated in comparison with three other relevant methods: `MXSCARNA` (15) computes pairing probabilities and considers potential stem information in the subsequent alignment process; `RNAfold` (17) produces individual secondary structures of RNA sequences; and `RNAalifold` (22) generates the consensus structure for a given input alignment. We compared `RNAfold` predictions with `RNAsalsa`'s individual predictions ψ^i (Equation 5), while the `MXSCARNA` and `RNAalifold` results are compared with `RNAsalsa`'s final consensus structure ω. As a reference model we employed the mammalian 16S rRNA secondary structure adopted from (36) (Figure 2).

The `RNAsalsa` secondary structure model for the mammalian 16S rRNA sequences is highly congruent to the *Bos taurus* reference model proposed by Burk *et al.* (36), see Supplementary Data for an illustrating graphical representation. In particular, 44 of the 52 helices within the conserved core of the structure are correctly predicted. Much of the remaining discrepancy can be explained by differences in the stringency of rules (Figure 2A). While the literature reference was derived with very conservative rules, we allow more variability among the organisms, and accept stems even if there is no direct evidence from compensatory substitutions. `MXSCARNA` and `RNAalifold` capture only 27 and 23 helices, respectively. In contrast to `RNAsalsa`, they failed to detect in particular long-range interactions.

Furthermore, to demonstrate the capabilities of constraints in thermodynamic folding algorithms, we compared the results of unconstrained foldings by the `RNAfold` software (17) only with those generated by `RNAsalsa`. In both cases, we used the same thermodynamical parameter sets and algorithms as implemented in `RNAfold`. The only difference is the usage of folding constraints in the case of `RNAsalsa`. These structure constraints are automatically generated and optimized for each RNA sequence within the data set.

`RNAsalsa`'s predictions of the individual structures always use an individual structure constraint that was optimized by `RNAsalsa`'s procedure itself up to that point. Therefore, they can match the reference model much better than thermodynamic folds by `RNAfold`, which have been calculated without supporting constraints (Figure 2B). See Supplementary Data for a set of comparisons performed on 16S rRNA.

Structure-aware RNA sequence alignments

The impact of secondary structures on the alignments as well as the overall performance was investigated by comparison with two commonly used sequence alignment methods, the classical `ClustalW` (38) and the more modern `MAFFT` (39) approach, and with the structural alignment method `MXSCARNA`. For this purpose we employed both simulated and real data sets.

Manually curated alignments for rRNAs are a rare commodity. At present, only partial data sets are available. Here, we used the alignments of 26 mammalian mitochondrial rRNAs published in (40,41). We only compared the subset of alignment positions that are marked as 'trusted' in these alignments. For the 16S rRNAs, we find that all four alignment programs reproduce 97.5–99% of the 'trusted' pairwise (mis)matches. The two structural alignment programs, `MXSCARNA` and `RNAsalsa`, have a slightly increased fraction of 'trusted' (mis)matches, Table 1. Very similar results were observed for the 12S mitochondrial rRNAs from the same source.

The relative quality of alignments was assessed following the procedures outlined in (42,43). In particular, we calculated the sum of pairs score (SPS) and the total column score (TC) as implemented in the `baliscore` software (42), as well as the structure conservation index (SCI) (44). Both SPS and TC are similarity metrics for multiple alignments. While SPS evaluates the number of equivalently aligned residue pairs summed over all pairs of sequences, TC counts the number of columns exactly shared by the two alignments. When one of the alignments is used as a benchmark reference, both SPS and TC can be interpreted as measures of alignment quality. The SCI, on the other hand, measures to what extent the aligned sequences form a common secondary structure in such a way that the sequence alignment also represents a

Table 1. Performance of alignment algorithms compared with manually curated mitochondrial 16S rRNAs (40,41)

Method	No. of (mis)matches	'Trusted '	Fraction
Kjer reference	492 456	375 062	0.7616
clustalW	500 357	366 840	0.7330
MAFFT	503 628	371 031	0.7367
MXSCARNA	496 447	370 020	0.7453
RNAsalsa	491 931	365 574	0.7431

We list the total number of (mis)matches in all pairwise sub-alignments of the multiple sequence alignment, and the subset of (mis)matches that coincide with those marked as 'trusted' in the reference alignments.

structural alignment. Originally introduced as a measure of structural conservation given the alignment, it was used in (45) to measure the ability of alignment algorithms to reconstruct conservative consensus folds. Of course, the SCI can only be employed to measure alignment quality when the existence of a common global consensus structure can safely be assumed.

Simulated data were generated as reference alignments using the new software rnasim (46). The tool takes both a structure annotated sequence string and a pre-defined phylogenetic tree as input and simulates the evolution of RNA molecules with fitness constraints derived from thermodynamic stability along the given tree. The tool provides us with the true reference alignment of the leaf sequences at the end of each simulation run. Several sets of examples representing smaller and larger RNAs (tRNAs, mammalian 12S rRNAs, and the *Saccharomyces cerevisiae* 28S rRNA) were used as roots. We used various values between 1 and 100 000 for rnasim's 'branch scaling' option -s, whereas the default was used for all other options. Since the branch scaling parameter has a strong influence on both sequence and structure, we systematically analysed how various alignment algorithms behave as a function of divergence.

Table 2 summarizes the results of some of the RNAsim-derived data sets. For the tRNAs, all programs except clustalW perform similarly well. In case of the 28S rRNA data set, which constitutes a very good example with highly divergent sequences, RNAsalsa performs slightly better than its competitors, albeit on a low level. These examples illustrate a general pattern: most programs, including RNAsalsa, perform with comparable scores on a wide range of test cases.

As a function of the branch length scaling parameter, we observe that all alignment algorithms perform comparably well or poorly in terms of the SPS, TC and SCI values. As expected, when the alignment problem becomes harder, i.e. for large scaling values, the performance metrics decrease uniformly (Supplementary Data). The resulting alignment then becomes more and more divergent between different methods.

We also used the BRAliBase-II set of structural alignments (http://projects.binf.ku.dk/pgardner/bralibase/bralibase2.html) (42), which covers group II introns, 5S rRNA, tRNA, U5 and SRP RNA [following (42), we did not use the SRP RNA alignments]. Again, we observe no major differences in the performance metrics

Table 2. Benchmark results for different alignment programs on rnasim-simulated data sets initialized with tRNAs and 28S rRNAs, respectively

Method[a]	tRNAs			LSU rRNA		
	SPS	TC	SCI	SPS	TC	SCI[b]
RNAsalsa	0.92	0.69	0.92	0.57	0.11	0.18
MAFFT	0.89	0.65	0.80	0.55	0.05	0.11
ClustalW	0.84	0.51	0.38	0.55	0.06	0.13
MXSCARNA	0.94	0.77	0.88	NA	NA	NA[c]

[a]All algorithms were started with default set-ups.
[b]All applied score values approach 1 as the alignments become identical with the reference.
[c]We could not get results using MXSCARNA with LSU rRNA.

(Supplementary Data). As in the case of the simulated data, the alignments become more divergent between methods as the problems become harder.

It appears that they tend to deviate from the reference in different features. RNAsalsa avoids poorly supported (mis)matches and instead tends to introduce more gaps in regions where the alignment becomes ambiguous. While this feature does not positively influence the scores in usual performance metrics, it does have a desirable effect for phylogenetic reconstruction, because gap-rich regions tend to be down-weighted or even excluded in typical phylogenetic applications. In that sense, RNAsalsa apparently reduces conflicting information arising from spurious (mis)matches. The RNAsalsa alignments emphasize the well-supported regions, and—as we have seen in the previous section—lead to quite accurate assignments of the most conserved secondary structure elements.

Exemplary applications in phylogeny reconstruction

In order to demonstrate the usefulness of RNAsalsa in phylogenetic applications, we consider two distinct data sets in detail. For each alignment, we performed phylogenetic analyses using a likelihood-based approach and compared the results with the published analyses of the data set. After individually aligning the LSU and SSU sequences, the alignments were concatenated. Then the program Aliscore (47), a method to identify ambiguously aligned regions in multiple sequence alignments, was used to extract the informative parts of the alignment. Aliscore identifies ambiguously aligned regions in multiple sequence alignments based on comparisons of aligned sequences in a sliding window and a Monte Carlo (MC) approach. The MC re-sampling compares the score of the originally aligned sequences in a given window position with scores of randomly drawn sequences of similar character composition. Maximum likelihood (ML) analysis was conducted using the RAxML software under a GTR + Γ model set up. For further details we refer to the Supplementary Data.

Primates. Tree reconstruction results of the mammalian data are shown in Figures 3 and 4. We focus here on the phylogeny of primates and, in particular, on the exact

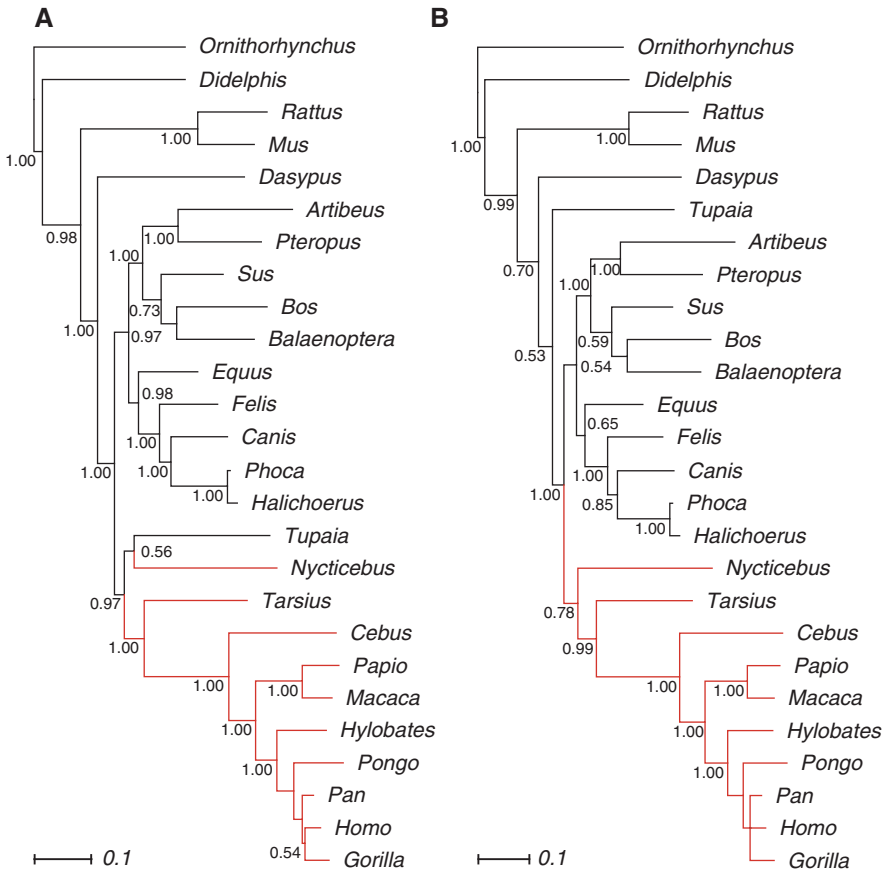

Figure 3. Bayesian tree inferred from the combined mammalian 12S rRNA and 16S rRNA. (**A**) Analysis with *GTR* + Γ model in simple DNA mode. (**B**) Analysis with *GTR* + Γ model in RNA mode for paired positions and DNA mode for loop regions. Numbers indicate Bayesian posterior probabilities. The scale bar denotes the estimated number of substitutions per site.

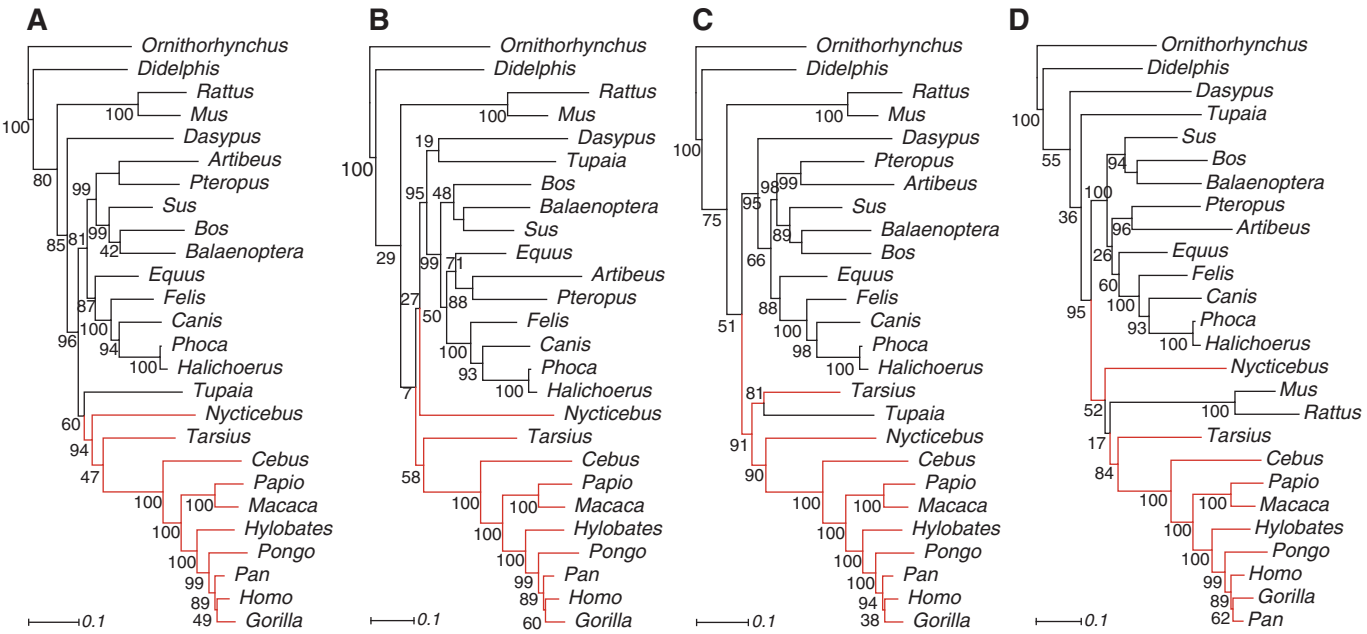

Figure 4. Phylogenies inferred from combined analyses of the mammalian 12S rRNA and 16S rRNA. Sequences are aligned with (**A**) `RNAsalsa`, (**B**) `MAFFT`, (**C**) `ClustalW` and (**D**) `MXSCARNA`. Tree reconstruction is based on ML analyses with *GTR* + Γ model. Numbers indicate Bootstrap support values (1000 replicates). The scale bar denotes the estimated number of substitutions per site.

phylogenetic position of one of the most basal primates, the tarsier. To this end, we re-evaluate a data set specifically compiled for this purpose (48).

Molecular studies on mitochondrial and nuclear DNA data of primates have so far lead to incongruent results. Nuclear DNA data favour haplorrhines ('dry-nosed' primates), i.e. the grouping of anthropoids (human-like apes) and tarsier (49). This hypothesis has gained strong support by the discovery of haplorrhine-specific SINES (50,51). Mitochondrial data, in contrast, mostly support the Prosimian hypothesis that postulates a sister group relationship of *Tarsius* and strepsirrhines ('wet-nosed' primates) (52,53,48).

The ML analysis based on the RNAsalsa alignment shows well-supported monophyletic primates (Figure 4). In contrast, primates do not appear monophyletic in analyses that use other alignments. The MAFFT alignment does not provide any phylogenetic signal to display relationships between anthropoids, the strepsirrhine representative *Nycticebus*, *Tarsius* and all remaining mammalian groups. The ClustalW analysis groups *Tupaia*, a scandentian representative, within primates as sister taxon to *Tarsius*, both forming the sister clade to *Nycticebus* and anthropoids. In the MXSCARNA analysis, primates appear paraphyletic with nested Rodentia.

Within primates, *Tarsius* appears as sister taxon to anthropoids in the RNAsalsa alignments, although with weak bootstrap support. This is also the case for the MAFFT alignment, albeit on the background of largely unresolved mammals. The MXSCARNA alignment leads to well-supported haplorrhines.

Although the placement of the tarsier is only weakly corroborated in the RNAsalsa analysis, these results show that the inclusion of good secondary structure models into the alignment procedure can make a significant difference for phylogeny reconstruction. RNAsalsa performs better than both purely sequence-based alignment approaches and sequence–structure alignments that are based directly on thermodynamic structure predictions.

The non-monophyletic appearance of primates with nested Scandentia and Rodentia in the MAFFT, ClustalW and MXSCARNA analyses must be interpreted as erroneous. A few studies based on mitochondrial genes propose paraphyletic primates with nested Dermoptera (53,54), but this observation has been explained as an effect of base composition bias in the mitochondrial markers (55). Scandentia or even Rodentia never appeared within primates to our knowledge.

An analysis of the whole mitochondrial genome of mammals revealed that heterogeneous substitution rates among different mammalian groups lead to misleading phylogenetic signals in mitochondrial genes (48). Their support for the prosimian hypothesis is thus likely an artefact. RNAsalsa apparently corrects this effect and leads to phylogenies from mitochondrial RNAs that are congruent with the results for nuclear genes.

We compared RNA-specific substitution models with simpler DNA models to determine to what extent they influence topology and/or node support of phylogenetic trees. Unfortunately, RNA substitution models are not implemented in any of the available ML software. They can, however, be used in Bayesian inference software. Here we used the parallel version of MrBayes (version 3.1.2) (56) with the GTR version of the Schoeniger and von Haeseler RNA model as described in (25) to account for character covariance. This model can be used to take interdependence of stem regions in ribosomal sequences into account. It assumes that a base pairing is converted into another one by means of a two-step process via an intermediate state. As this includes no double substitutions, other nucleotide changes are accounted for according to standard DNA models. Character covariation is considered, as the survival rate of a mutation in stem positions depends on the partner nucleotide, i.e. the substitution of the G in a UG pair by an A has a higher survival rate than the substitution by a C. As we understand it, this model currently represents a realistic and applicable approach to reflect nucleotide interdependencies in rRNA sequences. We have, however, opted for MrBayes over the alternative, PHASE (32,33,57), mostly because MrBayes provides a parallel version, which significantly reduces computation time in practice.

Bayesian inference results of the mammalian data set are shown in Figure 3. Application of simple DNA models led to a paraphyletic appearance of primates. *Tupaia* is a sister taxon to the strepsirrhine *Nycticebus*, both forming the first branching clade within the paraphyletic primates. In contrast, the application of mixed RNA/DNA models shows monophyletic primates with at least moderate nodal support. In both analyses, *Tarsius* appears highly supported as the sister taxon to anthropoids, forming monophyletic haplorrhines. Again, the monophyly of primates in the mixed model analysis can be interpreted as a hint that this approach performs better than the application of simple DNA models. These results corroborate the previously proposed superiority of the mixed model approach over simple DNA models (33).

Echinoderms. Our second example tackles the question of inter-class relationships in Echinodermata (Figure 5). This phylum is composed of five extant classes, the Crinoidea (sea lilies), Ophiuroidea (brittle stars), Asteroidea (starfishes), Holothuroidea (sea cucumbers) and Echinoidea (sea urchins). Monophyly in these five classes is well founded. The relationships between the five classes remain subject of ongoing discussion, however.

Several contradicting hypotheses of inter-class phylogeny in Echinodermata have been raised in the past, based on morphological and molecular data. Nevertheless, there is some consensus regarding major aspects of echinoderm phylogeny (58–60). Crinoids are mostly seen as the most basal split within Echinodermata, forming the sister group to the four remaining classes (Eleutherozoa). Furthermore, there is a strong support for a sister group relationship of echinoids and holothurians (Echinozoa). Debates on the phylogenetic position of the stellate forms (starfishes and brittle stars) recently ended up in two competing hypotheses: are the ophiurids alone sister group to Echinozoa (61,62) or do asteroids and ophiuroids form a clade (Asterozoa), which is then the sister taxon to Echinozoa (60)?

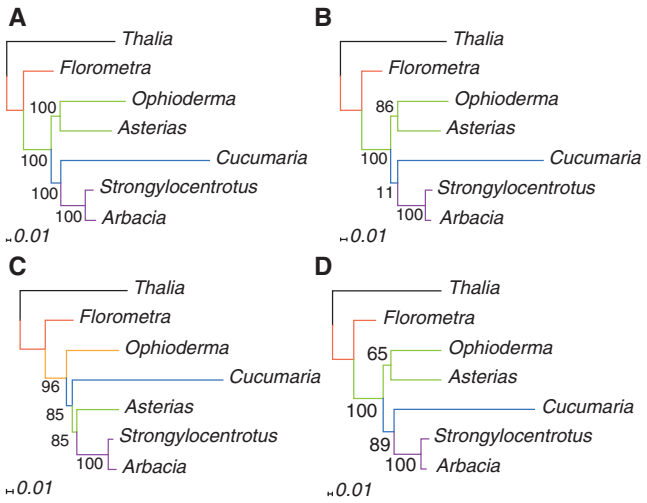

Figure 5. Phylogenies inferred from analyses of the echinoderm 28S rRNA. Sequences are aligned with (**A**) RNAsalsa, (**B**) MAFFT, (**C**) ClustalW and (**D**) MXSCARNA. Tree reconstruction is based on ML analyses with $GTR + \Gamma$ model. Numbers indicate Bootstrap support values (1000 replicates). The scale bar denotes the estimated number of substitutions per site.

Likelihood analyses based on different alignment methods are congruent only in parts of the resulting phylogenies. The sea lily species *Florometra* is the first split within monophyletic Echinodermata, and the two sea urchins *Arbacia* and *Strongylocentrotus* correctly appear monophyletic with highest bootstrap support.

There are, however, striking differences in many other aspects. The RNAsalsa alignment shows monophyletic Echinozoa with *Cucumaria* as sister taxon to the two echinoids. The starfish *Asterias* appears as sister taxon to the brittle star *Ophioderma*. These monophyletic Asterozoa are the sister clade to Echinozoa. All mentioned relationships gain highest bootstrap support. The MAFFT alignment also show monophyletic Asterozoa but with lesser support. Furthermore, there is no phylogenetic signal to resolve relationships between Asterozoa, Echinoidea and Holothuroidea. The ClustalW analysis does not show monophyletic Asterozoa and Echinozoa. Instead, there is a closer relationship between echinoids and *Asterias*. Within Eleutherozoa, the ophiuriod *Ophioderma* is the first split, followed by *Cucumaria* and the *Asterias* + Echinoidea clade.

The results of the MXSCARNA analyses are comparable with those of RNAsalsa. Eleutherozoa, Echinozoa and Asterozoa are monophyletic, the latter ones with lesser support than in the RNAsalsa analyses. Compared with the previous studies on echinoderm phylogeny, the results of the structural alignment methods must be seen as more reasonable. In particular, based on monophyletic Echinozoa their superiority over the two exclusively sequence-based alignment methods is pointed out. Both of those fail to recover monophyletic Echinozoa and the ClustalW alignment erroneously show *Asterias* as sister taxon to the sea urchins.

Overall, we find that structure-aware alignments yield more plausible results than purely sequence-based alignments. RNA-specific substitution models yield better results with the RNAsalsa alignments (which incorporate some prior knowledge on the structure) than structural alignments that are based entirely on unconstrained thermodynamic folding.

DISCUSSION

ML analyses and Bayesian inference both revealed a remarkable influence of rRNA secondary structure consideration on both the sequence alignment and on the subsequent tree reconstruction. This phenomenon is well-known in molecular systematics and has already led to the development of RNA-specific substitution models. The application of these models, however, is confined to a few studies (3,5,32,33,63,64), mostly because of the lack of a convenient way to construct usable secondary structure models and alignments for newly sequenced rRNAs. Structural inference is not only difficult but also very tedious, and hence avoided in the overwhelming majority of published studies.

As a tool for simultaneously computing structure annotation and structure-aware sequence alignments of large RNA molecules, RNAsalsa has been designed specifically to overcome this barrier. While it can also be useful for other tasks, its primary domain of application is phylogenetic inference. Here the relatively large computational cost of the structure prediction (compared with other, less accurate tools) is of little concern, since it is dwarfed by the demands of subsequent ML or Bayesian computations. Extensive tests, and two real-world applications, demonstrate that RNAsalsa can lead to significant improvements in reconstructed phylogenies, positively affecting both tree stability and tree topology. These improvements can be traced back to two sources: alignments that emphasize the unambiguous regions improve the phylogenetic signal; second, more exact automatically generated consensus structures enhance the benefit of RNA-specific substitution models. As our examples show, both types of improvements can offset the problems incurred by unequal substitution rates and long branches.

The modular structure of RNAsalsa lends itself to incorporate further improvements. For example, it is likely beneficial to use a Sankoff-style algorithm such as foldalign (65) or locarna (13) to construct the pairwise alignments $\mathbb{A}^{ij}$ and/or the final alignment $\mathbb{B}$ and its consensus structure ω. The current version of RNAsalsa consistently generates alignments and consensus structures of acceptable quality (compared with the extremely tedious manual curation of such data). It is therefore suitable for routine applications in molecular phylogenetics based on structured RNAs, in particular rRNAs.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Deutsche Forschungsgemeinschaft (under the auspices of SPP-1174 'Deep Metazoan Phylogeny', Projects STA 850/2-1, STA 850/3-1, as well as MI 649/3 and MI 649/6.

Conflict of interest statement. None declared.

REFERENCES

1. Woese,C.R. (1987) Bacterial evolution. *Microbiol. Rev.*, **51**, 221–271.
2. Hillis,D.M. and Dixon,M.T. (1991) Ribosomal DNA: molecular evolution and phylogenetic inference. *Q. Rev. Biol.*, **66**, 411–453.
3. Kjer,K.M. (1995) Use of ribosomal-RNA secondary structure in phylogenetic studies to identify homologous positions – an example of alignment and data presentation from the frogs. *Mol. Phylogenet. Evol.*, **4**, 314–330.
4. Mallatt,J. and Sullivan,J. (1998) 28S and 18S rDNA sequences support the monophyly of lampreys and hagfishes. *Mol. Biol. Evol.*, **15**, 1706–1718.
5. Buckley,T.R., Simon,C., Flook,P.K. and Misof,B. (2000) Secondary structure and conserved motifs of the frequently sequenced domains IV and V of the insect mitochondrial large subunit rRNA gene. *Insect Mol. Biol.*, **9**, 565–580.
6. Misof,B., Niehuis,O., Bischoff,I., Rickert,A., Erpenbeck,D. and Staniczek,A. (2006) A hexapod nuclear SSU rRNA secondary-structure model and catalog of taxon-specific structural variation. *J. Exp. Zool. Part B Mol. Dev. Evol.*, **306B**, 70–88.
7. Mallatt,J. and Winchell,C.J. (2007) Ribosomal RNA genes and deuterostome phylogeny revisited: more cyclostomes, elasmobranchs, reptiles, and a brittle star. *Mol. Phylogenet. Evol.*, **43**, 1005–1022.
8. Hochsmann,M., Voss,B. and Giegerich,R. (2004) Pure multiple RNA secondary structure alignments: a progressive profile approach. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **1**, 53–62.
9. Reeder,J. and Giegerich,R. (2005) Consensus shapes: an alternative to the Sankoff algorithm for RNA consensus structure prediction. *Bioinformatics*, **21**, 3516–3523.
10. Dalli,D., Wilm,A., Mainz,I. and Steger,G. (2006) `STRAL`: progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time. *Bioinformatics*, **22**, 1593–1599.
11. Torarinsson,E., Havgaard,H.J. and Gorodkin,J. (2007) Multiple structural alignment and clustering of RNA sequences. *Bioinformatics*, **23**, 926–932.
12. Lindgreen,S., Gardner,P.P. and Krogh,A. (2007) MASTR: multiple alignment and structure prediction of non-coding RNAs using simulated annealing. *Bioinformatics*, **23**, 3304–3311.
13. Will,S., Missal,K., Hofacker,I.L., Stadler,P.F. and Backofen,R. (2007) Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, **3**, e65.
14. Katoh,K. and Toh,H. (2008) Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. *BMC Bioinformatics*, **9**, 212.
15. Tabei,Y., Kiryu,H., Kin,T. and Asai,K. (2008) A fast structural multiple alignment method for long RNA sequences. *BMC Bioinformatics*, **9**, 33.
16. Zuker,M. and Sankoff,D. (1984) RNA secondary structures and their prediction. *Bull. Math. Biol.*, **46**, 591–621.
17. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.
18. Doshi,K., Cannone,J., Cobaugh,C. and Gutell,R. (2004) Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*, **5**, 105.
19. Layton,D.M. and Bundschuh,R. (2005) A statistical analysis of RNA folding algorithms through thermodynamic parameter perturbation. *Nucleic Acids Res.*, **33**, 519–524.
20. Ding,F., Sharma,S., Chalasani,P., Demidov,V.V., Broude,N.E. and Dokholyan,N.V. (2008) *Ab initio* RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA*, **14**, 1164–1173.
21. Yusupov,M.M., Yusupova,G.Z., Baucom,A., Lieberman,K., Earnest,T.N., Cate,J.H.D. and Noller,H.F. (2001) Crystal structure of the ribosome at 5.5 Å resolution. *Science*, **292**, 883–896.
22. Hofacker,I.L., Fekete,M. and Stadler,P.F. (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.
23. Bernhart,S.H., Hofacker,I.L., Will,S., Gruber,A.R. and Stadler,P.F. (2008) `RNAalifold`: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**, 474.
24. Collins,L.J., Moulton,V. and Penny,D. (2000) Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP. *J. Mol. Evol.*, **51**, 194–204.
25. Schoeniger,M. and von Haeseler,A. (1994) A stochastic model for the evolution of autocorrelated DNA sequences. *Mol. Phylogenet. Evol.*, **3**, 240–247.
26. Rzhetsky,A. (1995) Estimating substitution rates in ribosomal RNA genes. *Genetics*, **141**, 771–783.
27. Tillier,E.R.M. and Collins,R.A. (1995) Neighbor joining and maximum-likelihood with RNA sequences—addressing the interdependence of sites. *Mol. Biol. Evol.*, **12**, 7–15.
28. Stephan,W. (1996) The rate of compensatory evolution. *Genetics*, **144**, 419–426.
29. Tillier,E.R.M. and Collins,R.A. (1998) High apparent rate of simultaneous compensatory basepair substitutions in ribosomal RNA. *Genetics*, **148**, 1993–2002.
30. Parsch,J., Braverman,J.M. and Stephan,W. (2000) Comparative sequence analysis and patterns of covariation in RNA secondary structures. *Genetics*, **154**, 909–921.
31. Savill,N.J., Hoyle,D.C. and Higgs,P.G. (2001) RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. *Genetics*, **157**, 399–411.
32. Jow,H., Hudelot,C., Rattray,M. and Higgs,P.G. (2002) Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. *Mol. Biol. Evol.*, **19**, 1591–1601.
33. Hudelot,C., Gowri-Shankar,V., Jow,H., Rattray,M. and Higgs,P.G. (2003) RNA-based phylogenetic methods: application to mammalian mitochondrial RNA sequences. *Mol. Phylogenet. Evol.*, **28**, 241–252.
34. Nussinov,R., Piecznik,G., Griggs,J.R. and Kleitman,D.J. (1978) Algorithms for loop matching. *SIAM J. Appl. Math.*, **35**, 68–82.
35. Gotoh,O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.
36. Burk,A., Douzery,E.J. and Springer,M.S. (2002) The secondary structure of mammalian mitochondrial 16S rRNA molecules: refinements based on a comparative phylogenetic approach. *J. Mammalian Evol.*, **9**, 225–252.
37. Shapiro,B.A. and Zhang,K. (1990) Comparing multiple RNA secondary structures using tree comparisons. *CABIOS*, **6**, 309–318.
38. Thompson,J.D., Higgs,D.G. and Gibson,T.J. (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties, and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
39. Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.
40. Kjer,K.M. and Honeycutt,R.L. (2007) Site specific rates of mitochondrial genomes and the phylogeny of eutheria. *BMC Evol. Biol.*, **7**, 8.
41. Kjer,K.M., Swigonova,Z., LaPolla,J.S. and Broughton,R.E. (2007) Why weight? *Mol. Phylogenet. Evol.*, **43**, 999–1004.
42. Gardner,P.P., Wilm,A. and Washietl,S. (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res*, **33**, 2433–2439.
43. Wilm,A., Mainz,I. and Steger,G. (2006) An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms Mol. Biol.*, **1**, 19.
44. Washietl,S., Hofacker,I.L. and Stadler,P.F. (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA*, **102**, 2454–2459.
45. Gruber,A.R., Berhart,S.H., Hofacker,I.L. and Washietl,S. (2008) Strategies for measuring evolutionary conservation of RNA secondary structures. *BMC Bioinformatics*, **9**, 122.

46. Guo,S., Wang,L.S. and Kim,J. (2009) Large-scale simulation of RNA macroevolution by an energy-dependent fitness model. *Sys. Biol.*, http://kim.bio.upenn.edu/software/rnasim.shtml.
47. Misof,B. and Misof,K. (2009) A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. *Syst. Biol.*, **58**, 21–34.
48. Schmitz,J., Ohme,M. and Zischler,H. (2002b) The complete mitochondrial sequence of *Tarsius bancanus*: evidence for an extensive nucleotide compositional plasticity of primate mitochondrial DNA. *Mol. Biol. Evol.*, **19**, 544–553.
49. Goodman,M., Porter,C.A., Czelusniak,J., Page,S.L., Schneider,H., Shoshani,J., Gunnell,G. and Groves,C.P. (1998) Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. *Mol. Phylogenet. Evol.*, **9**, 585–598.
50. Zietkiewicz,E., Richer,C. and Labuda,D. (1999) Phylogenetic affinities of tarsier in the context of primate Alu repeats. *Mol. Phylogenet. Evol.*, **11**, 77–83.
51. Schmitz,J., Ohme,M. and Zischler,H. (2001) SINE insertions in cladistic analyses and the phylogenetic affiliations of *Tarsius bancanus* to other primates. *Genetics*, **157**, 777–784.
52. Hayasaka,K., Gojobori,T. and Horai,S. (1988) Molecular phylogeny and evolution of primate mitochondrial DNA. *Mol. Biol. Evol.*, **5**, 626–644.
53. Murphy,W.J., Eizirik,E., Johnson,W.E., Zhang,Y.P., Ryder,O.A. and O'Brien,S.J. (2001) Molecular phylogenetics and the origins of placental mammals. *Nature*, **409**, 614–618.
54. Arnason,U., Adegoke,J.A., Bodin,K., Born,E.W., Esa,Y.B., Gullberg,A., Nilsson,M., Short,R.V., Xu,X. and Janke,A. (2002) Mammalian mitogenomic relationships and the root of the eutherian tree. *Proc. Natl Acad. Sci. USA*, **99**, 8151–8156.
55. Schmitz,J., Ohme,M., Suryobroto,B. and Zischler,H. (2002a) The colugo (*Cynocephalus variegatus*, Dermoptera): the primates' gliding sister? *Mol. Biol. Evol.*, **19**, 2308–2312.
56. Ronquist,F. and Huelsenbeck,J.P. (2003) `MrBayes 3`: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.
57. Gowri-Shankar,V. and Rattray,M. (2006) On the correlation between composition and site-specific evolutionary rate: implications for phylogenetic inference. *Mol. Biol. Evol.*, **23**, 352–364.
58. Littlewood,D.T., Smith,A.B., Clough,K.A. and Emson,R.H. (1997) The interrelationships of the echinoderm classes: morphological and molecular evidence. *Biol. J. Linn. Soc.*, **61**, 409–438.
59. Smith,A.B. (1997) Echinoderm larvae and phylogeny. *Ann. Rev. Ecol. Syst.*, **28**, 219–241.
60. Janies,D. (2001) Phylogenetic relationship of extant echinoderm classes. *Canadian J. Zool.*, **79**, 1232–1250.
61. Smith,A.B. (1988) Fossil evidence for the relationship of extant echinoderm classes and their times of divergence. In Paul,C.R.C. and Smith,A.B. (eds), *Echinoderm Phylogeny and Evolutionary Biology*. Clarendon Press, Oxford.
62. Scouras,A. and Smith,J.M. (2006) The complete mitochondrial genomes of the sea lily *Gymnocrinus richeri* and the feather star *Phanogenia gracilis*: signature nucleotide bias and unique nad4L gene rearrangement within crinoids. *Mol. Phylogenet. Evol.*, **39**, 323–34.
63. Kjer,K.M. (2004) Aligned 18S and insect phylogeny. *Syst. Biol.*, **53**, 506–514.
64. Niehuis,O., Naumann,C.M. and Misof,B. (2006c) Identification of evolutionary conserved structural elements in the mt SSU rRNA of Zygaenoidea (Lepidoptera): a comparative sequence analysis. *Org. Divers. Evol.*, **6**, 17–32.
65. Hull Havgaard,J.H., Lyngso,R., Stormo,G.D. and Gorodkin,J. (2005) Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics*, **21**, 1815–1824.