Accurate and efficient reconstruction of deep phylogenies from structured RNAs

Roman R. Stocsits [e], Harald Letsch [e], Jana Hertel [a],
Bernhard Misof [f], Peter F. Stadler [a,c,b,d]

[a] *Bioinformatics Group, Dept. of Computer Science, Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*

[b] *Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria*

[c] *RNomics Group, Fraunhofer Institut für Zelltherapie und Immunologie, Deutscher Platz 5e, D-04103 Leipzig, Germany*

[d] *Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA*

[e] *Zoologisches Forschungsmuseum Alexander Koenig, Bonn*

[f] *UHH Biozentrum Grindel & Zoologisches Museum, Hamburg, Germany*

Abstract

Ribosomal RNA 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 aligments and consensus structures that can deal with the large ribosomal RNAs. 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 datasets: a mitochondrial rRNA set of 26 Mammalia, and a collection of 28S nuclear rRNAs representative of the five major echinoderm groups.

Key words: alignment, RNA, secondary structure, phylogeny, evolution

1 Introduction

Ribosomal RNAs are the most widely used source of phylogenetic information, although protein-coding genes, often derived from 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 techniques.

Ribosomal RNAs, 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, 4, 5, 6, 7]. 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, 9, 10, 11, 12, 13, 14, 15].

For short RNAs (length $\lessapprox$ 100nt), secondary structures can be computed with satisfactory accuracy based on experimentally measured thermodynamic parameters [16, 17]. In contrast, for large RNAs, such as SSU and LSU ribosomal RNAs, the accuracy of thermodynamic predictions is insufficient. This is in part due to inaccuracies in the “nearest neighbor model” and its parameters [18, 19], and in part because the RNA and protein components of the ribosome are tightly packed and thus mutually influence their folds. The functional rRNA structures, therefore, cannot be expected to be identical with the structures of isolated rRNAs – which is what the thermodynamic folding algorithms compute. The `RNAalifold` approach shows 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 [20, 21]. 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 [22]. Clearly, accurate structure predictions are a necessary pre-requisite to utilize structural differences in this way. Secondly, 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.

The rationale of RNA-specific substitution models [23, 24, 25, 26, 27, 28] is rooted in the effect of covariation in 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, restoring the pairing ability [29]. 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 [30, 31].

RNAsalsa is written in C. The source code and pre-compiled executables for various platforms, as well as 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.

2 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 (*Fig. 1*).

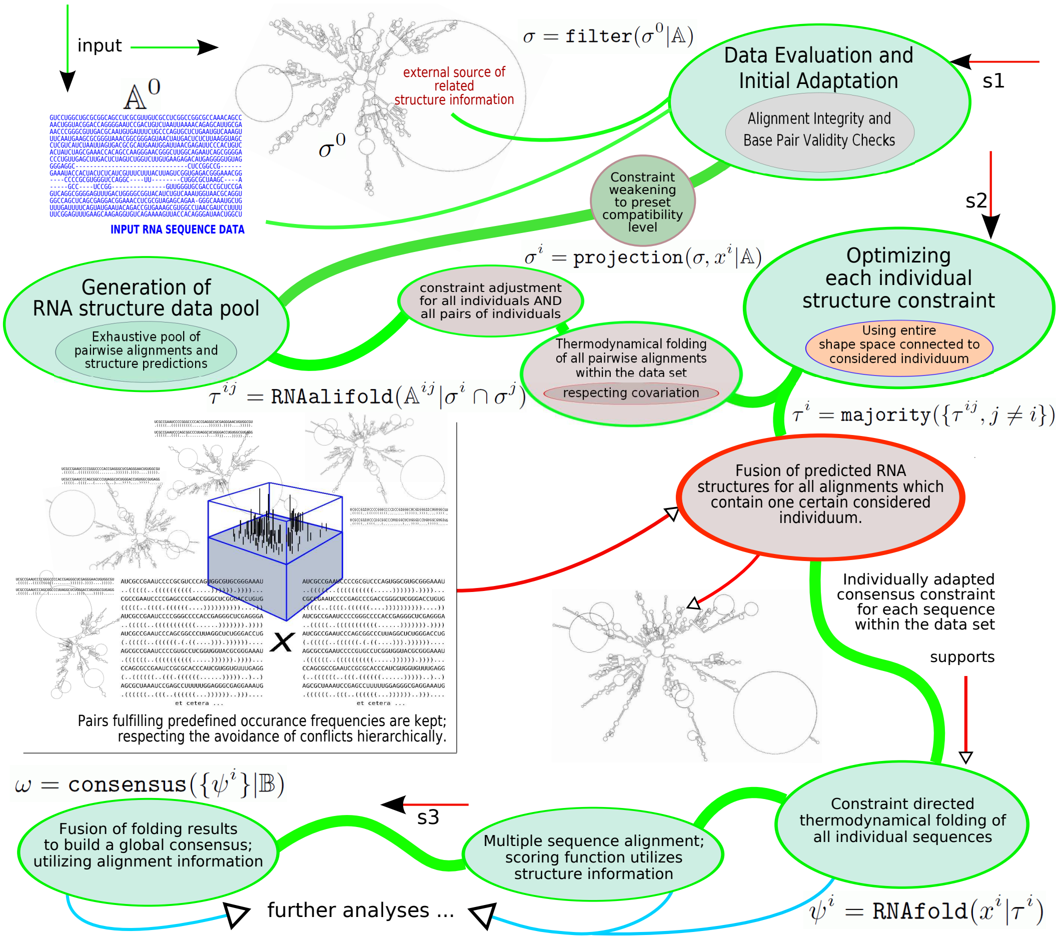
input
$\mathbb{A}^0$
INPUT RNA SEQUENCE DATA
σ^0
external source of related structure information
$\sigma = \texttt{filter}(\sigma^0|\mathbb{A})$
Data Evaluation and Initial Adaptation
Alignment Integrity and Base Pair Validity Checks
s1
s2
Constraint weakening to preset compatibility level
$\sigma^i = \texttt{projection}(\sigma, x^i|\mathbb{A})$
Optimizing each individual structure constraint
Using entire shape space connected to considered individuum
Generation of RNA structure data pool
Exhaustive pool of pairwise alignments and structure predictions
constraint adjustment for all individuals AND all pairs of individuals
Thermodynamical folding of all pairwise alignments within the data set
respecting covariation
$\tau^{ij} = \texttt{RNAalifold}(\mathbb{A}^{ij}|\sigma^i \cap \sigma^j)$
$\tau^i = \texttt{majority}(\{\tau^{ij}, j \neq i\})$
Fusion of predicted RNA structures for all alignments which contain one certain considered individuum.
et cetera ...
et cetera ...
Pairs fulfilling predefined occurance frequencies are kept; respecting the avoidance of conflicts hierarchically.
Individually adapted consensus constraint for each sequence within the data set
supports
$\omega = \texttt{consensus}(\{\psi^i\}|\mathbb{B})$
s3
Fusion of folding results to build a global consensus; utilizing alignment information
Multiple sequence alignment; scoring function utilizes structure information
Constraint directed thermodynamical folding of all individual sequences
further analyses ...
$\psi^i = \texttt{RNAfold}(x^i|\tau^i)$

Fig. 1. The algorithmic concepts throughout the workflow of `RNAsalsa` as a graphical representation. See main text for details and the supplements for an alternative representation.

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* known 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 check 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

$$\sigma = \texttt{filter}(\sigma^0|\mathbb{A}) \tag{1}$$

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 a given input sequence x^i), then produces initial structure constraints separately for each sequence:

$$\sigma^i = \texttt{projection}(\sigma, x^i|\mathbb{A}) \tag{2}$$

Up to this point, the result heavily depends upon the initial alignment. It may not cover the input sequences uniformly, 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. For each of the pairwise alignments, we compute the consensus minimum free energy structures

$$\tau^{ij} = \texttt{RNAalifold}(\mathbb{A}^{ij}|\sigma^i \cap \sigma^j). \tag{3}$$

using the base pairs common to the projected structures σ^i and σ^j, resp., as constraint. This step uses the Vienna RNA Package library functions underlying RNAalifold [20] 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

$$\tau^i = \texttt{majority}(\{\tau^{ij}, j \neq i\}). \tag{4}$$

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 pseudoknotted structures. Alternatively, one could also use Nussinov's Maximum Circular Matching algorithm [32] 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

$$\psi^i = \texttt{RNAfold}(x^i|\tau^i), \tag{5}$$

for each sequence x^i.

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 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 [33]. 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:

$$\begin{aligned} s(x_i, y_j) = {} & b_0 s_m(x_i, y_j) + \\ & b_1 s_n(x_{\pi(i)}, y_{\pi(j)}) + c s_p(x_i, y_j) \end{aligned} \tag{6}$$

where $x_{\pi(i)}$ and $y_{\pi(j)}$ denotes 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, while 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 *supplemental material*. The final result is a global re-alignment $\mathbb{B}$ of the input sequences which respects all 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

$$\omega = \texttt{consensus}(\{\psi^i\}|\mathbb{B}) \tag{7}$$

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, equ.(1), and the two majority voting procedures, equ.(4) and equ.(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.

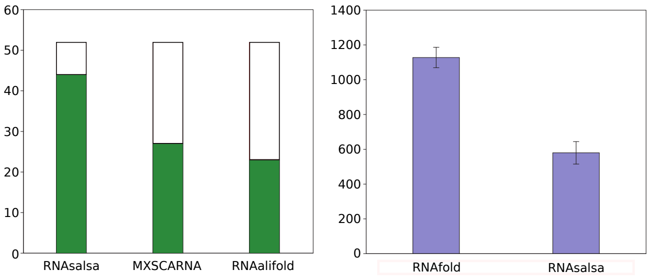

Fig. 2. Accuracy of structure prediction. Fraction of correctly predicted helices (green bars, left) compared to the mammalian 16S rRNA consensus models [34]. `RNAsalsa` significantly outperforms `MXSCARNA` and `RNAalifold` (default parameter settings; 3-sample test for equality of proportions without continuity correction; $\chi^2 = 19.96$, $df = 2$, $p < 0.0001$). On the right: Average tree-edit distance [35] between predicted individual structures and the mammalian 16S rRNA reference model. `RNAsalsa` predictions conform the consensus model much better (paired sampled t-Test; $t = 33.46$, $df = 1$, $p < 0.0001$; $N = 26$.)

3 Results

3.1 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` [20] generates the consensus structure for a given input alignment. We compared `RNAfold` predictions with `RNAsalsa`'s individual predictions ψ^i (equ. 5), while the `MXSCARNA` and `RNAalifold` results are compared with `RNAsalsa`'s final consensus structure ω. See *Fig. 2*.

The `RNAsalsa` secondary structure model for the mammalian 16S rRNA sequences is highly congruent to the *Bos taurus* reference model proposed by [34], see the *supplements* for an illustrating graphical representation. In particular, 44 of the 52 helices within the conserved core of the structure are correctly predicted. The remaining discrepancy is likely not a weakness of `RNAsalsa` but reflects a greater variability of mammalian 16S rRNA structures than present in the data set originally used to construct the reference model, *Fig. 2 (left)*. `MXSCARNA` and `RNAalifold` capture only 27 and 23 helices, resp. In contrast to `RNAsalsa`, they failed to detect in particular long range interactions. Furthermore, `RNAsalsa`'s predictions of the individual structures match the reference model much better than unconstrained thermodynamic folds by `RNAfold`, *Fig. 2 (right)*. Single fold data are provided as *supplements*.

3.2 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` [36] and the more modern `MAFFT` [37] approach, and with the structural alignment method `MXSCARNA`.

As a benchmark system we generated a reference alignment by simulation of tree evolution using `rnasim`. We then compared the reference alignments with the results of each alignment algorithm to estimate alignment quality. The `rnasim` software and an input example representing small tRNAs have been downloaded from `http://kim.bio.upenn.edu/software/rnasim.shtml`. As a second example, we used 28S rRNA from *Saccharomyces cerevisiae* as a root sequence for simulated evolution. Following the procedure of [38, 39] we calculated the *Structure Conservation Index*, *Total Column Score*, and *Sum of Pairs Score* as implemented in the `baliscore` software [38]. The *Sum of Pairs Score* is an accuracy metric for a multiple alignment relative to a reference alignment, based on the number of correctly aligned residue pairs summed over all pairs of sequences. It can equivalently be viewed as a similarity metric between two multiple alignments. This metric is used by the `BaliBase` benchmark. The *Structural Conservation Index* is a measure for alignment quality regarding RNA sequences and emphasizes secondary structure. It values the algorithms ability to reconstruct conservative consensus folds and is the most important measure for RNA alignments because it respects compensatory or consistent sequence variation. The *Structure Conservation Index* (SCI) will be high if the sequences fold together equally well as if folded individually. On the other hand, SCI will be low if no consensus fold can be found [40]. The *Total Column Score* represents the rate of alignment columns that could be reconstructed by the alignment program compared to the reference tree. All programs were used with default settings except for the tree branch length scaling factor in `rnasim` which was set to 100000. `RNAsalsa` always performed best in *SCI* and was second to `Mxscarna` in *TC* and *SPS* with small molecules only. With 28S rRNA `RNAsalsa` performed best w.r.t. all measures. *Table 3.2* summarizes the benchmark results.

3.3 Exemplary applications in phylogeny reconstruction

In order to demonstrate the usefulness of `RNAsalsa` in phylogenetic applications, we consider two distinct datasets 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

Table 1
Benchmark results for different alignment programs for tRNA and 28S rRNA

	tRNAs			LSU rRNA		
Method[1]	SPS	TC	SCI	SPS	TC	SCI[2]
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	n.a.	n.a.	n.a.[3]

[1] All algorithms were started with default setups.
[2] All applied score values approach 1 as the alignments become identical with the reference.
[3] We could not get results using Mxscarna with LSU rRNA.

Aliscore [41], a new method to identify ambiguously aligned regions in multiple sequence alignments, was used to extract the informative parts of the alignment. Maximum likelihood (ML) analysis was performed using a GTR model with gamma distribution (for further details see the *supplements*.

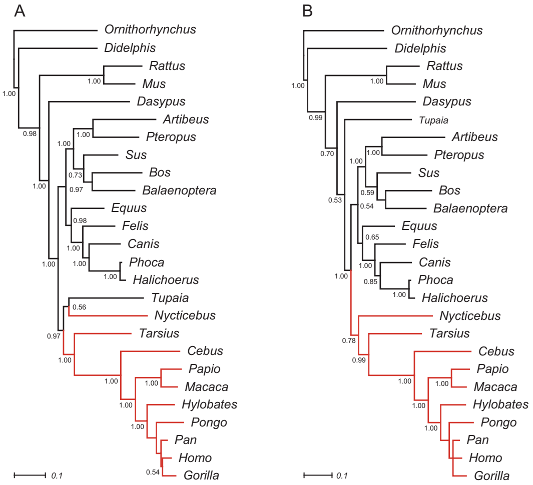

Fig. 3. Bayesian tree inferred from the combined mammalian 12S rRNA and 16S rRNA. (A) Analysis with $GTR + \Gamma$ model in simple DNA mode. (B) Analysis with $GTR + \Gamma$ model in RNA mode for paired positions and DNA mode for loop regions. Numbers indicate Bayesian posterior probabilities. The scale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

Primates

Tree reconstruction results of the mammalian data are shown in *Fig. 3* and *4*. We focus here on the phylogeny of primates and, in particular, on the exact phylogenetic position of one of the most basal primates, the tarsier. To this end, we re-evaluate a dataset specifically compiled for this purpose [42].

Molecular studies on mitochondrial and nuclear DNA data of primates have so far lead to incongruent results. Nuclear DNA data favour haplorrhines, i.e. the grouping of anthropoids and tarsier [43]. This hypothesis has gained strong support by the discovery of haplorrhine-specific SINES [44, 45]. Mitochondrial data, in contrast, mostly support the *prosimian hypothesis* that postulates a sister group relationships of *Tarsius* and strepsirrhines [46, 47, 42].

The Maximum Likelihood analysis based on the `RNAsalsa` alignment shows well supported monophyletic primates, *Fig. 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. `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 resp., must be interpreted as erroneous. A few studies based on mitochondrial genes propose paraphyletic primates with nested Dermoptera [47, 48], but this observation has been explained as an effect of base composition bias in the mitochondrial markers [49]. 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 [42]. 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 Maximum Likelihood software. They can, however, be used in Bayesian inference software. We therefore used MrBayes (version 3.1.2) [50] with a variant of the Schoeniger & von Haeseler model [23] to account for character covariance.

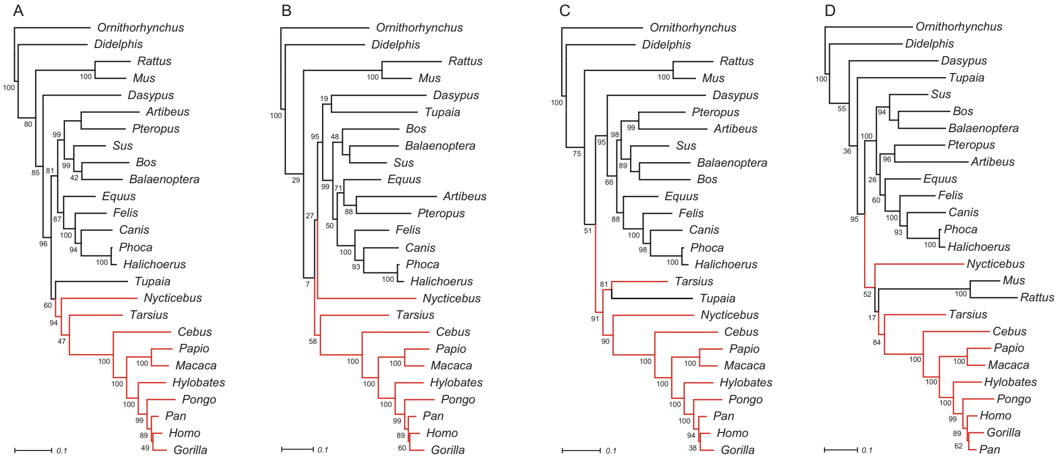

Fig. 4. Phylogenies inferred from combined analyses of the mammalian 12S rRNA and 16S rRNA. Sequences are aligned with (A) RNAsalsa, (B) MAFFT, (C) ClustalW and (D) MXSCARNA. Tree reconstruction is based on Maximum Likelihood analyses with $GTR+\Gamma$ model. Numbers indicate Bootstrap support values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

Bayesian inference results of the mammalian data set are shown in *Fig. 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 [31].

Echinoderms

Our second example tackles the question of inter-class relationships in Echinodermata, *Fig. 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 [51, 52, 53]. Crinoids are mostly seen as the most basal split within Echinodermata, forming the sister group to the four remaining classes (Eleutherozoa). Furthermore, there is 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 [54, 55] or do asteroids and ophiuroids form a clade (Asterozoa), which is then the sister taxon to Echinozoa [53]?

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 to 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 shows *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 which are based entirely on unconstrained thermodynamic folding.

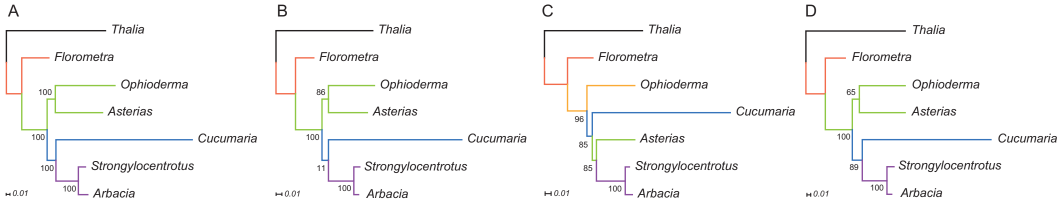

Fig. 5. Phylogenies inferred from analyses of the echinoderm 28S rRNA. Sequences are aligned with (A) `RNAsalsa`, (B) `MAFFT`, (C) `ClustalW` and (D) `MXSCARNA`. Tree reconstruction is based on Maximum Likelihood analyses with $GTR + \Gamma$ model. Numbers indicate Bootstrap support values (1000 replicates).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

4 Discussion

Maximum Likelihood 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, 30, 31, 56, 57], mostly because of the lack of an efficient way to construct good secondary structure models and alignments for newly sequenced rRNAs.

`RNAsalsa` has been designed specifically to overcome this barrier. It is a tool for simultaneously computing high-quality structure annotation and structure-aware sequence alignments of large RNA molecules. 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 to 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: first more accurate alignments 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 incorporating further improvements. For example, it is likely beneficial to use a Sankoff-style algorithm such as `foldalign` [58] 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 to the extremely tedious manual curation of such data). It is therefore suitable for routine applications in molecular phylogenetics based on structured RNAs, in particular ribosomal RNAs.

5 Funding

The work was funded in part by the Deutsche Forschungsgemeinschaft under the auspices of SPP-1174 “Deep Metazoan Phylogeny”, Projects STA 850/2-1 and STA 850/3-1, as well as Projects MI 649/3 and MI 649/6.

References

[1] C. R. Woese. Bacterial evolution. *Microbiological Reviews*, 51:221–271, 1987.

[2] D. M. Hillis and M. T. Dixon. Ribosomal DNA: molecular evolution and phylogenetic inference. *Q. Rev. Biol.*, 66:411–453, 1991.

[3] K M Kjer. 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, 1995.

[4] J. Mallatt and J. Sullivan. 28S and 18S rDNA sequences support the monophyly of lampreys and hagfishes. *Mol. Biol. Evol.*, 15:1706–1718, 1998.

[5] T. R. Buckley, C. Simon, P. K. Flook, and B. Misof. Secondary structure and conserved motifs of the frequently sequenced domains IV and V of the insect mitochondrial large subunit rRNA gene. *Insect Mol. Biol.*, 565-580:9, 2000.

[6] B. Misof, O. Niehuis, I. Bischoff, A. Rickert, D. Erpenbeck, and A. Staniczek. 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, 2006.

[7] J. Mallatt and C. J. Winchell. Ribosomal RNA genes and deuterostome phylogeny revisited: more cyclostomes, elasmobranchs, reptiles, and a brittle star. *Mol. Phylogenet. Evol.*, 43:1005–1022, 2007.

[8] M Hochsmann, B Voss, and R Giegerich. Pure multiple RNA secondary structure alignments: a progressive profile approach. *IEEE/ACM Trans. Comput. Biol. Bioinform*, 1:53–62, 2004.

[9] J Reeder and R. Giegerich. Consensus shapes: an alternative to the Sankoff algorithm for RNA consensus structure prediction. *Bioinformatics*, 21:3516–3523, 2005.

[10] D Dalli, A Wilm, I Mainz, and Steger G. `STRAL`: progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time. *Bioinformatics*, 22:1593–1599, 2006.

[11] E Torarinsson, J H Havgaard, and J. Gorodkin. Multiple structural alignment and clustering of RNA sequences. *Bioinformatics*, 23:926–932, 2007.

[12] S Lindgreen, P P Gardner, and A Krogh. MASTR: multiple alignment and structure prediction of non-coding RNAs using simulated annealing. *Bioinformatics*, 23:3304–3311, 2007.

[13] S. Will, K. Missal, Ivo L. Hofacker, P. F. Stadler, and R. Backofen. Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comp. Biol.*, 3:e65, 2007.

[14] K Katoh and H Toh. Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. *BMC Bioinformatics*, 9:212, 2008.

[15] Y Tabei, H Kiryu, T Kin, and K Asai. A fast structural multiple alignment method for long RNA sequences. *BMC.Bioinformatics*, 9:33, 2008.

[16] Michael Zuker and David Sankoff. RNA secondary structures and their prediction. *Bull. Math. Biol.*, 46:591–621, 1984.

[17] Ivo L Hofacker, Walter Fontana, Peter F Stadler, L Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.

[18] Kishore Doshi, Jamie Cannone, Christian Cobaugh, and Robin Gutell. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*, 5:105, 2004.

[19] D. M. Layton and R. Bundschuh. A statistical analysis of RNA folding algorithms through thermodynamic parameter perturbation. *Nucleic Acids Res.*, 33:519–524, 2005.

[20] Ivo L. Hofacker, Martin Fekete, and Peter F. Stadler. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, 319:1059–1066, 2002.

[21] Stephan H Bernhart, Ivo L Hofacker, Sebastian Will, Andreas R Gruber, and Peter F Stadler. `RNAalifold`: improved consensus structure prediction for rna alignments. *BMC Bioinformatics*, 9:474, 2008.

[22] Lesley J. Collins, Vincent Moulton, and David Penny. Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP. *J. Mol. Evol.*, 51:194–204, 2000.

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

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

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

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

[27] E R M Tillier and R A Collins. High apparent rate of simultaneous compensatory basepair substitutions in ribosomal RNA. *Genetics*, 148:1993–

2002, 1998.

[28] J Parsch, J M Braverman, and W. Stephan. Comparative sequence analysis and patterns of covariation in RNA secondary structures. *Genetics*, 154:909–921, 2000.

[29] Nicholas J. Savill, David C. Hoyle, and Paul G. Higgs. RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. *Genetics*, 157:399–411, 2001.

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

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

[32] R. Nussinov, G. Piecznik, J.R. Griggs, and D.J. Kleitman. Algorithms for loop matching. *SIAM J. Appl. Math.*, 35(1):68–82, 1978.

[33] O. Gotoh. An improved algorithm for matching biological sequences. *J. Mol. Biol.*, 162:705–708, 1982.

[34] A Burk, E J Douzery, and M S Springer. The secondary structure of mammalian mitochondrial 16S rRNA molecules: Refinements based on a comparative phylogenetic approach. *J. Mammalian Evol.*, 9:225–252, 2002.

[35] Bruce A. Shapiro and Khaizhong Zhang. Comparing multiple RNA secondary structures using tree comparisons. *CABIOS*, 6:309–318, 1990.

[36] J. D. Thompson, D. G. Higgs, and T. J. Gibson. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties, and weight matrix choice. *Nucl. Acids Res.*, 22:4673–4680, 1994.

[37] K. Katoh, K. Misawa, K. Kuma, and T. Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res.*, 30:3059–3066, 2002.

[38] P. P. Gardner, A. Wilm, and S. Washietl. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res*, 33:2433–2439, 2005.

[39] A. Wilm, I. Mainz, and G. Steger. An enhanced RNA alignment benchmark for sequence alignment programs. *Alg. Mol. Biol.*, 1:19, 2006.

[40] Andreas R. Gruber, Stephan H. Berhart, Ivo L. Hofacker, and Stefan Washietl. Strategies for measuring evolutionary conservation of RNA secondary structures. *BMC Bioinformatics*, 9:122, 2008.

[41] B. Misof and K. Misof. A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: A more objective means of data exclusion. *Syst. Biol.*, 2008. accepted.

[42] J. Schmitz, M. Ohme, and H. Zischler. The complete mitochondrial sequence of *Tarsius bancanus*: evidence for an extensive nucleotide compositional plasticity of primate mitochondrial DNA. *Mol. Biol. Evol.*,

19:544–553, 2002b.

[43] M. Goodman, C. A. Porter, J. Czelusniak, S. L. Page, H. Schneider, J. Shoshani, G. Gunnell, and C. P. Groves. Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. *Mol. Phylogenet. Evol.*, 9:585–598, 1998.

[44] E. Zietkiewicz, C. Richer, and D. Labuda. Phylogenetic affinities of tarsier in the context of primate Alu repeats. *Mol. Phylogenet. Evol.*, 11:77–83, 1999.

[45] J. Schmitz, M. Ohme, and H. Zischler. SINE insertions in cladistic analyses and the phylogenetic affiliations of *Tarsius bancanus* to other primates. *Genetics*, 157:777–784, 2001.

[46] K. Hayasaka, T. Gojobori, and S. Horai. Molecular phylogeny and evolution of primate mitochondrial DNA. *Mol. Biol. Evol.*, 5:626–644, 1988.

[47] W. J. Murphy, E. Eizirik, W. E. Johnson, Y. P. Zhang, O. A. Ryder, and S. J. O'Brien. Molecular phylogenetics and the origins of placental mammals. *Nature*, 409:614–618, 2001.

[48] U. Arnason, J. A. Adegoke, K. Bodin, E. W. Born, Y. B. Esa, A. Gullberg, M. Nilsson, R. V. Short, X. Xu, and A. Janke. Mammalian mitogenomic relationships and the root of the eutherian tree. *Proc. Natl. Acad. Sci. U.S.A.*, 99:8151–8156, 2002.

[49] J. Schmitz, M. Ohme, B. Suryobroto, and H. Zischler. The colugo (*Cynocephalus variegatus*, Dermoptera): the primates' gliding sister? *Mol. Biol. Evol.*, 19:2308–2312, 2002a.

[50] F. Ronquist and J. P. Huelsenbeck. `MrBayes 3`: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19:1572–1574, 2003.

[51] D. T. Littlewood, A. B. Smith, K. A. Clough, and R. H. Emson. The interrelationships of the echinoderm classes: morphological and molecular evidence. *Biol. J. Linn. Soc.*, 61:409–438, 1997.

[52] A. B. Smith. Echinoderm larvae and phylogeny. *Ann. Rev. Ecol. Syst.*, 28:219–241, 1997.

[53] D. Janies. Phylogenetic relationship of extant echinoderm classes. *Canadian J. Zool.*, 79:1232–1250, 2001.

[54] A. B. Smith. Fossil evidence for the relationship of extant echinoderm classes and their times of divergence. In C. R. C. Paul and A. B. Smith, editors, *Echinoderm Phylogeny and Evolutionary Biology.* Clarendon Press, Oxford, 1988.

[55] A. Scouras and J. M. Smith. 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, 2006.

[56] K. M. Kjer. Aligned 18S and insect phylogeny. *Syst. Biol.*, 53:506–514, 2004.

[57] O. Niehuis, C. M. Naumann, and B. Misof. Identification of evolutionary conserved structural elements in the mt SSU rRNA of Zygaenoidea (Lepidoptera): A comparative sequence analysis. *Organisms Diversity &*

Evolution, 6:17–32, 2006c.

[58] J. H. Hull Havgaard, R. Lyngso, G. D. Stormo, and J. Gorodkin. Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics*, 21:1815–1824, 2005.