Consideration of RNA Secondary Structure Significantly Improves Likelihood-Based Estimates of Phylogeny: Examples from the Bilateria

Maximilian J. Telford,* Michael J. Wise,† and Vivek Gowri-Shankar‡

*Department of Biology, University College London, London, United Kingdom; †School of Biomedical and Chemical Sciences, University of Western Australia, Crawley Perch, Western Australia; and ‡Department of Computer Science, University of Manchester, Manchester, United Kingdom

Sequences from ribosomal RNA (rRNA) genes have made a huge contribution to our current understanding of metazoan phylogeny and indeed the phylogeny of all of life. That said, some parts of this rRNA-based phylogeny remain unresolved. One approach to increase the resolution of these trees would be to use more appropriate models of sequence evolution in phylogenetic analysis.

RNAs transcribed from rRNA genes have a complex secondary structure mediated by base pairing between sometimes distant regions of the rRNA molecule. The pairing between the stem nucleotides has important consequences for their evolution which differs from that of unpaired loop nucleotides. These differences in evolution should ideally be accounted for when using rRNA sequences for phylogeny estimation.

We use a novel permutation approach to demonstrate the significant superiority of models of sequence evolution that allow stem and loop regions to evolve according to separate models and, in common with previous studies, we show that 16-state models that take base pairing of stems into account are significantly better than simpler, 4-state, single-nucleotide models. One of these 16-state models has been applied to the phylogeny of the Bilateria using small subunit rRNA (SSU) and hundreds of large subunit rRNA (LSU) genes.

The current SSU + LSU-based model of bilaterian phylogeny posits two principal clades: the deuterostomes (chordates, echinoderms, and hemichordates) and the protostomes. The protostomes consist of the two major groups Ecdysozoa (arthropods plus various introvertan worms including priapulids and nematodes) and Lophotrochozoa (mollusks, annelids, flatworms, and a number of other phyla) (Aguinaldo et al. 1997). This general scheme has been supported by additional molecular phylogenetic studies (e.g., Haase et al. 2001), although there is also evidence that contradicts this view (e.g., Wolf, Rogozin, and Koonin 2004).

Although this tripartite phylogeny seems on balance to be robust, SSU and LSU phylogenies have not been able to resolve relationships within the ecdysozoan and lophotrochozoan clades. One approach to improving the resolution of the animal tree using SSU and LSU would be to employ more accurate methods of phylogenetic analyses.

An important aspect of rRNA genes that has largely been ignored by zoologists when reconstructing bilaterian phylogenies is that the functional RNAs transcribed from these genes have a complex secondary structure mediated by base pairing between sometimes distant regions of the rRNA molecule. By accounting for this secondary structure in the models of evolution used in phylogeny reconstruction, we might hope to improve the reliability of the bilaterian tree.

RNAs consist of base paired stem and unpaired loop regions. The pairing within the stems involves the Watson-Crick A:U, G:C pairs and the noncanonical G:U pair; other pairings exist, but they are rare enough to be disregarded in the current context. The constraints arising from the need to conserve secondary structure have important effects on the evolution of the stem sequences that differ from those of the loop regions. In essence, because there is selective pressure for the maintenance of rRNA secondary structure potential substitutions affecting one half of a pair of stem nucleotides have a different probability of fixation compared to the equivalent nucleotide in a loop. More specifically, in stem regions, changes from a base paired state to an unpaired state tend to be strongly selected against. Moreover, not all changes between sets of pairing bases are equal. Changes between A:U and G:U and between G:U and G:C each involve a single change and are therefore common. Changing between A:U and G:C or between U:A and C:G can also relatively easily occur involving two changes with the stable transitional state G:U or U:G. Changing from any of A:U, G:U, and G:C to their counterparts U:A, U:G, or C:G on the other hand necessitates either two simultaneous changes which is very unlikely or passage via an unpaired transitional state which is presumably maladaptive. Such changes are correspondingly much rarer.

Stem models have been around for a number of years, and there have been a number of descriptions of the use of such models in phylogeny reconstruction (Schöniger and von Haeseler 1994; Muse 1995; Rzhetsky 1995; Tillier and Collins 1995; Otsuka, Terai, and Nakano 1999; Savill, Hoyle, and Higgs 2001; Jow et al. 2002; Hudelot et al.
Materials and Methods

Data

Bilaterian SSU sequences were downloaded from the European Ribosomal RNA database (http://oberon.fvms.ugent.be:8080/rRNA/) in the dedicated comparative sequence editor (DCSE) format. Our test data set comprised 17 sequences from the deuterostomes (1 xenurbellid, 4 hemichordates, 5 echinoderms, 2 urochordates, 1 cephalochordate, and 4 vertebrates) and 6 out-group sequences (3 ecdysozoans and 3 lophotrochozoans). Sequences from additional taxa were downloaded from GenBank. Our Bilaterian data set contained 72 taxa from the deuterostomes, ecdysozoans, and lophotrochozoans. We used newly developed software Xstem and Ystem to manipulate the data sets and to extract the secondary structure information in DCSE in a format usable by the phylogeny software MrBayes (Huelsenbeck and Ronquist 2001) Tree-Puzzle (Schmidt et al. 2002) and Phase http://bioinf.man.ac.uk/resources/phase/ (see Appendix).

Test Tree

Our initial approach was to evaluate the different models of evolution by comparing their likelihood maximized on one specific test tree. For this model evaluation, a data set of 23 bilaterian SSU sequences was used. Unreliably aligned positions were excluded, leaving 1563 positions in the analysis. A total of 680 sites were paired stem nucleotides and the remaining 883 were loop-bulge nucleotides (alignment available on request). The tree used for all comparisons of models is shown in figure 1. Ideally, one would optimize the topology along with all other parameters when performing the permutation tests, but this is computationally impossible. In spite of this we believe our results to be robust for two reasons. First, our tree is very likely to be approximately correct; it makes excellent sense biologically, and this topology or a very close variant of it is found by all of the models used in an MCMC Bayesian search (results not shown). Second, we have repeated the same permutation tests using a different and probably slightly suboptimal topology (we moved Branchiostoma to be the sister group of the echinoderm-hemichordate-Xenur-bella clade) and got essentially identical results (not shown). We subsequently use the best model identified in this way under a Bayesian framework.

Phylogenetic Models

Five different models of sequence evolution were used:

1. General time reversible (GTR). This condition does not consider the differences between stems and loops. A GTR model with discrete gamma (four categories), proportion of invariant sites (pinv), and base composition was used (written as GTR + I + G4).
2. 2xGTR. Positions were divided into stem and loop positions, and each of these two sets was allowed to evolve as single nucleotides according to a separate model (each of the two models comprising the same parameters as the GTR model.)
3. RNA16A (for stems). All 16 character states are included in this model, and the model permits double substitutions which has been shown to be important by Savill, Hoyle, and Higgs (2001). The rate matrix has 16 frequency parameters for each of the 16 possible
pairs of nucleotides. The rate parameters group together different types of changes: $\alpha$ models the single substitution rate, $\beta$ models the double substitution rate, $\gamma$ models substitutions between paired and mismatched states, and $\epsilon$ models rates between mismatch states (2 parameters are fixed so there are 19 estimated parameters). Please see Savill, Hoyle, and Higgs (2001) for a detailed explanation of this model.

(4) RNA16HKY85 (for stems). Stem-pair evolution was also modeled considering a 16-state model after Schöniger and von Haeseler (1999). This model has 16 frequency parameters for the 16 possible stem pairs and 2 rate parameters. The rate parameters differ significantly from those of RNA16A as in this case we do not consider classes of change between pairs of nucleotides rather considering changes between single nucleotides within the stem pairs (in the case of RNA16HKY85 grouping transitions and transversions). Simultaneous double transitions are disallowed (there are 16 estimated parameters).

(5) RNA16GTR (for stems). This is a more parameter-rich version of the above model; like the GTR model for unpaired nucleotides, it has six rate parameters: one for each single nucleotide change. Simultaneous double transitions are disallowed (20 estimated parameters).

Also included as parameters in models RNA16A, RNA16HKY85, and RNA16GTR were the gamma shape parameter $\gamma$ for the gamma distribution with four discrete rate categories and the proportion of invariant sites.

Permutation Test of 2xGTR Model

We have used a permutation test to compare models rather than the more usual parametric bootstrap. The parametric bootstrap tests the null hypothesis that two nested models (with differing numbers of parameters) are indistinguishable. Parametric bootstrapping uses multiple data sets simulated according to the simpler of the two models under comparison. In contrast, the permutation test approach (Felsenstein 2003), although clearly related to parametric bootstrapping, differs in that (1) the null hypothesis is that the partitioning of characters to different models has no effect and (2) the multiple data sets derive from a randomization of the initial data rather than from simulation. In short, we are asking whether partitioning characters according to our more complex model (e.g., considering stem nucleotides separate from loop nucleotides) is distinguishable, in terms of lnLikelihood, from a random partitioning of these characters. If it is and if the likelihood is higher, then we conclude that the more complex model is preferable to the more simple model.

The first test is whether stem nucleotides and loop nucleotides come from a single population (i.e., they are evolving according to a single GTR model—the simpler model) or whether they come from different populations (and are therefore evolving according to different GTRs—the more complex model). We calculate the lnLikelihood of the more complex model which corresponds to the data set correctly divided into stem and loops (real data set) using separate GTR models for stem and loop. This is compared to the lnLikelihoods calculated under the same dual model using 100 permutated samples in which the same nucleotides are randomly assigned to pseudo stem and loop classes of the same size as the real ones and therefore actually correspond to the simpler model. If the real lnLikelihood is outside the distribution of 95% of the permutated lnLikelihoods, then the stems and loops are inferred to be evolving according to different models (come from different populations) with 95% confidence. An increase in lnLikelihood seen with twin models over a single model implies that we are justified in assigning stems and loops to separate models in our phylogenetic analyses.

To simplify plotting the lnLikelihoods we subtracted the lnLikelihoods of the single GTR from the lnLikelihoods of the double GTRs (2xGTR) for unpermutated and permutated data. The likelihoods were calculated using Phase on the fixed topology shown in figure 1. Branch lengths were also reestimated.

Permutation Tests of Models RNA16A, RNA16HKY85, and RNA16GTR

This permutation test approach was also taken when considering conditions RNA16A, RNA16HKY85, and RNA16GTR. Here, we are asking whether the stem nucleotides evolve according to models in which one side of a pair constrains the evolution of the other side of a pair (i.e., a 16-state, stem-pair model—complex) as opposed to a model in which this is not true and in which there is no correlation (equivalent to a four-state model—simple). For each of the three 16-state models we calculate the lnLikelihood when the model is fitted to the correct stem pairs. This is compared to the lnLikelihoods, calculated under the same model, of 100 permutated samples in which the nucleotides are randomly assigned to the same number of stem pairs; the randomization ensures that there is no correlation between members of pairs. If the lnLikelihood of the correctly paired data set is outside the distribution of 95% of the lnLikelihoods of the permutated data, then we can say with 95% confidence that the stems do not evolve according to the uncorrelated model. An increase in lnLikelihood seen with 16-state models over a 4-state model implies that we would be justified in the use of 16-state models in our phylogenetic analyses.

In order to permutate the data we randomly reordered columns of stem nucleotides while maintaining the set of parentheses indicating the secondary structure. To simplify plotting the lnLikelihoods, we subtracted the lnLikelihood of the GTR from the lnLikelihoods of the 16-state models (models RNA16A, RNA16HKY85, and RNA16GTR) for unpermutated and permutated stem data.

Phylogeny of the Bilatera

The Bayesian framework using the software MCMCPhase was used to estimate the phylogeny of a sample of 72 taxa from within the Bilateria using GTR for the loop regions and RNA16A for the stem regions. Unreliably aligned positions were excluded leaving 1510 positions in
the analysis. A total of 634 sites were paired stem nucleotides and the remaining 876 were loop-bulge nucleotides (alignment available on request). Random starting trees were used, and the parameters of the substitution models were treated as random variables and estimated during the analysis. A total of 1,000,000 initial generations were run which was seen to be more than sufficient for the lnLikelihood to plateau. A total of 1,000,000 additional generations were run with sampling of the tree including branch lengths every 1,000th generation. The analyses were repeated four times with indistinguishable results, supporting the notion that a plateau had been reached. MrBayes ‘sumt’ command was used to generate a consensus tree with branch lengths and Bayesian clad support values from these samples. For comparison, the same analyses were run using a single GTR for both stem and loop positions.

Results and Discussion
Comparing GTR and 2xGTR: Parameter Estimates and Permutation Test

We have used our full test data set (23 species, 680 stem nucleotides, 883 loop nucleotides) and optimized the maximum likelihood estimates of parameters under the single GTR and under the 16-state models RNA16A, RNA16HKY85, and RNA16GTR on our test tree (fig. 1.). The lnLikelihood of the test tree under the 4-state single GTR model is −3551.99, and under the 16-state correlated models RNA16A, RNA16HKY85, and RNA16GTR are −2788.29, −2823.82, and −2820.78, respectively. As before, the use of permutation tests (fig. 3B–D) demonstrate that the lnLikelihoods for the data set with correctly paired nucleotides are far outside the distribution of lnLikelihoods of the permuted data sets. The lnLikelihood is much higher for the correctly paired versus permuted data sets and implies that the sequences are indeed evolving according to the model in which nucleotides changes in stem pairs are correlated. The use of the 16-state, stem-pairing models described is clearly supported by these analyses. The large improvement in lnLikelihood using model RNA16A over models RNA16HKY85 and RNA16GTR seems to confirm its superiority; this is particularly impressive considering that RNA16GTR is actually marginally more parameter rich than RNA16A. We used model RNA16A for our analysis of metazoan phylogeny.

There is very little improvement in lnLikelihood using model RNA16GTR over model RNA16HKY85 (difference = 3.045). As models RNA16HKY85 and RNA16GTR are nested, we are able to compare their lnLikelihoods using a chi-square test with 4 degrees of freedom for the four additional parameters in model RNA16GTR. The chi-square test shows that the improvement in likelihood of model RNA16GTR versus model RNA16HKY85 (using $2 \times$ difference = 6.08 as the test statistic) is nonsignificant.

Superiority of Models that Account for Stem Pairing

We are unaware of a previous use of a permutation approach for studies of stem models. Using this approach we have been able to demonstrate for the first time the superiority of models of nucleotide evolution that divide rRNA sequences into sets of stem and loop nucleotides.

FIG. 1.—Test tree of deuterostomes plus protostomian out-groups used for all permutation test comparisons.
Furthermore, we have used the permutation approach to confirm results (e.g., Schöninger and von Haeseler 1999) showing that 16-state, stem-pair models for stem regions of rRNA genes are appropriate and far superior to 4-state, single-nucleotide models. Models RNA16HKY85 and RNA16GTR were included for completeness because they are included in the popular MrBayes software. Our results show that, although they are almost certainly inferior to Model RNA16A they improve hugely over four-state models and their use should be encouraged.

Phylogeny of the Bilateria

Our Bayesian tree (fig. 4) derived using a combination of 4-state GTR model for loops and 16-state RNA16A model for stems accords well with current ideas of bilaterian phylogeny (Adoutte et al. 2000). We see the three separate major clades of the Deuterostomia, Ecdysozoa, and Lophotrochozoa, although the tree was rooted between Deuterostomia and Protostomia so that the relative topology of these three branches has not been tested. Within the Deuterostomia we recover the expected division between chordates and a clade containing the Ambulacraria (echinoderms plus hemichordates) and the worm Xenoturbella as their sister group. The 99% Bayesian support for

![Graph A](image)

**Fig. 2.**—Comparison of the maximum likelihood estimates of (A) exchangeability parameters (A→C, A→G, etc.) and (B) nucleotide frequencies (fA, fC, fG, fU) for loops and stems analyzed according to separate GTR models for stems and for loops and for combined loops + stems under the single GTR model. Exchangeability parameters cannot be estimated independently of each other, and G→U for all three sets of characters is set to equal 1 as a reference.

![Graph B](image)

**Fig. 3.**—The results of permutation tests of 2xGTR model and models RNA16A, RNA16HKY85, and RNA16GTR. (A) The difference in lnLikelihood between the single GTR and double GTR for 100 permuted data sets have been binned. (B, C, D) The difference in lnLikelihood between the single GTR for stems and the three different 16-state models for 100 permuted data sets have been binned. The difference for the unpermuted data sets are indicated by the arrow and in each case are far outside and greater than the distribution of permuted data, showing that the data do not conform to the simpler model and that the more complex model is superior.
the exclusion of *Xenoturbella* from within the Ambulacraria using this superior model of evolution adds to our confidence in this result (Bourlat et al. 2003), although it must be remembered that Bayesian support generally gives higher values than an equivalent nonparametric bootstrap.

Within the Ecdysozoa there is good support for the basal position of the priapulid and kinorhynch and also nematomorphs. The branching pattern between nematodes, onychophorans, and tardigrades relative to Euarthropoda is not well resolved (low Bayesian support values). Within the Euarthropoda we find two principal branches—the chelicerates plus myriapods on one side and crustaceans plus insects on the other. This result is in keeping with recent studies (Friedrich and Tautz 1995; Boore, Lavrov, and
Brown 1998; Cook et al. 2001), and all these clades receive high support values.

Within the Lophotrochozoa, although certain groups are well resolved (gastrotrichs, platyhelminthes, bivalve mollusks, gastropod mollusks, polyplacophoran mollusks, pogonophorans, nemerteans, and rotifers), for the most part, the relationships between these and other taxa are not well resolved. We do find weak support for a monophyletic assemblage of annelids that includes the sipunculid (now widely considered a derived annelid) and strong support for the groupings of rotifer plus cyclophoran plus enteropcts and gastrotrich plus platyhelminthes, but other branches remain poorly supported. The lack of resolution within the Lophotrochozoa is in common with other analyses that have used SSU (Mallatt and Winchell 2002).

Comparison of the single GTR + RNA16A tree and the GTR tree (not shown) shows a number of differences. Our most significant novel finding using GTR + RNA16A is the observation of two highly supported groups within the Lophotrochozoa; the gastrotrichs plus platyhelminthes in one and all other lophotrochozoan phyla (including the rotifer-cyclophoran clade) in the other. Analyses with a single GTR model group the rotifer-cyclophoran clade with the gastrotrich-platyhelminth clade rather than with the other lophotrochozoans. This difference deserves to be examined further considering the demonstrated superiority of the evolutionary model we have used. The single GTR model also fails to group enteropcts with cyclophora + rotifers and supports a number of unlikely associations—enteropcts with nemerteans, phoronid with annelids, and pogonophorans with gastropod mollusks. In general, the GTR + RNA16A model seems to be more agnostic in its placement of these taxa, resulting in a less well-resolved but more accurate tree. Overall, our method, although certainly not perfect with this data set, seems to have found a more credible tree than the simpler alternative.

Conclusions

It is well established that phylogenetic methods perform better when the model of evolution is appropriate (Felsenstein 1978, Leinert, Kumar, and Albert, 1997; Sullivan and Swofford 1997; Posada and Crandall 2001). We have shown that the simple expedient of considering stem nucleotides independently of loop nucleotides significantly improves the estimated lnLikelihood. Moreover, the 16-state models we have used result in a further major improvement in the fit of the model to the evolution of the pairs of stem nucleotides. Although the consideration of more complex models of sequence evolution inevitably requires additional computer time, with the existence of computationally efficient methods such as the Bayesian MCMC approach used here, the analysis even of relatively large data sets is feasible.

Although clearly desirable in terms of a measurably improved approximation to the true evolution of rRNA sequences, incorporation of secondary structure information is not a panacea. Our consideration of the phylogeny of the Bilateria using SSU sequences shows that much of our tree does not differ from previous analyses using standard four-state models. Additionally, areas of the tree, particularly within the Lophotrochozoa, are still not resolved. rRNA genes are, nevertheless, the most widely sequenced genes across all living organisms—more than 20,000 SSU sequences in the European rRNA database—such a significant improvement in using them for phylogenetic analyses may be hoped to further our understanding of the tree of life. We believe that our tree represents the best-justified SSU-based phylogeny of the Bilateria yet produced.

Acknowledgments

We would like to thank Nick Goldman for advice on permutation tests. The comments of the editor and two anonymous reviewers were also very much appreciated. M.J.T. was supported by the Wellcome trust 060503/Z/00/Z. M.J.W. was supported during the course of this project by a Fellowship at Pembroke College, Cambridge, generously provided by Bristol-Myers Squibb. Xstem (written in python) and Ystem (written in perl) are non–platform specific and are available on request from the authors.

Appendix

Software

The biggest obstacle to using secondary structure information in phylogenetics has been interpreting secondary structure information contained in rRNA sequence databases and translating this into a format that the phylogeny software packages can use. To overcome this we have developed the software applications Xstem and Ystem.

Extracting Secondary Structure Information Using Xstem

Xstem parses the DCSE alignment format file and identifies positions that make up the two halves of each stem pair in each sequence. The file can be output in formats readable by MrBayes, Phase, and in the 20-state character set proposed by Smith, Lui, and Tillier (2004).

Adding Taxa and Altering Alignments Using Ystem

Ystem allows conversion of the DCSE alignment format file and identifies positions that make up the two halves of each stem pair in each sequence. The file can be output in formats readable by MrBayes, Phase, and in the 20-state character set proposed by Smith, Lui, and Tillier (2004).

Literature Cited


