Reconstructing phylogeny from RNA secondary structure via simulated evolution

Will Fischer
Section of Integrative Biology
University of Texas at Austin
Austin, TX 78712 USA
wfischer@uts.cc.utexas.edu

Nicholas Geard
School of Information Technology
and Electrical Engineering
The University of Queensland
St Lucia, Q 4072, Australia
nic@itee.uq.edu.au

August 16, 2002

Abstract

DNA sequences of genes encoding functional RNA molecules (e.g., ribosomal RNAs) are commonly used in phylogenetics (i.e. to infer evolutionary history). Trees derived from ribosomal RNA (rRNA) sequences, however, are inconsistent with other molecular data in investigations of deep branches in the tree of life. Since much of te functional constraints on the gene products (i.e. RNA molecules) relate to three-dimensional structure, rather than their actual sequences, accumulated mutations in the gene sequences may obscure phylogenetic signal over very large evolutionary time-scales. Variation in structure, however, may be suitable for phylogenetic inference even under extreme sequence divergence. To evaluate qualitatively the manner in which structural evolution relates to sequence change, we simulated the evolution of RNA sequences under various constraints on structural change.

Introduction

Phylogenetics

In the last several decades, vast technical advances in molecular biology, computer hardware, and in analytical methods have enabled the rise of molecular phylogenetics. This discipline allows reliable inference of the historical evolutionary relationships of different organisms, and has brought within the realm of possibility the taxonomists' dream of a unified genealogy of the living world.

Phylogenetic methods are well-established for morphological data, but the most widely used and spectacular applications have used data from biological macromolecules, i.e. DNA or protein sequences. Sequences from many different genes have been used to build phylogenies: different genes evolve at various rates, corresponding in large part to varying degrees of functional constraint (i.e.

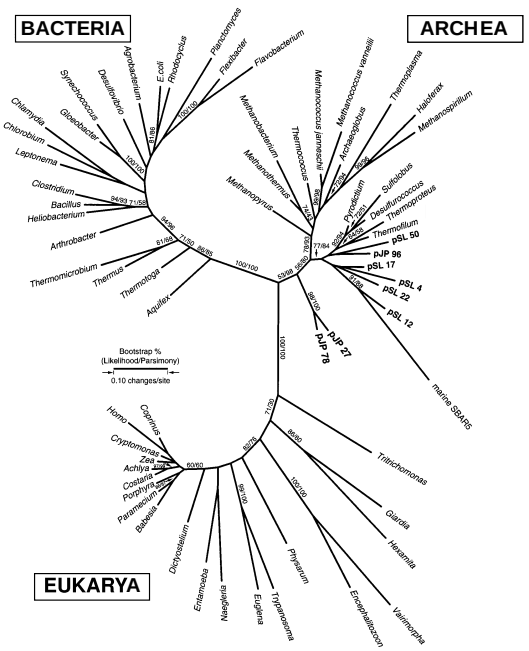

Figure 1: A sequence-based "Universal Tree" inferred from small-subunit ribosomal RNA genes (from Barns et al.[3]), that includes the three "domains."

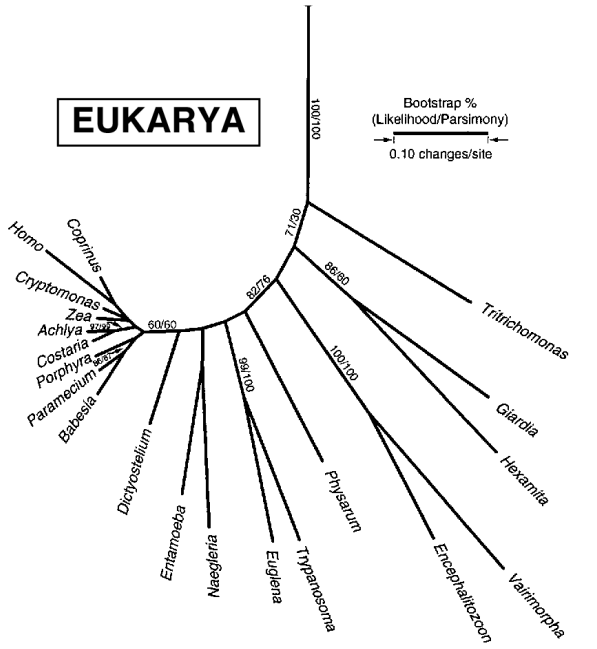

Figure 2: The eukaryotic portion of the tree in Fig. 1. Note the long branch (decorated with isolated long branches) between the "crown group" and the divergence of the three domains.

the strength of natural selection against deleterious mutations), and are therefore useful at different levels of divergence (which correspond very roughly to the time since the separation of the lineages being investigated).

The DNA sequences most commonly used for phylogenetics are the genes that encode ribosomal RNA. There are several reasons for this popularity: these genes (which encode the core of the protein synthesis machinery) are found in all living organisms, include a mix of quickly and slowly evolving domains, and are fairly simple to isolate. Consequently, a huge data set of rRNA sequences, spanning a large proportion of the diversity of life, has been generated by the research community. These sequences are available in public databases (e.g. GenBank), and represent a significant scientific resource.

Unfortunately, although ribosomal data have been of great utility in resolving many relationships, it has proven less useful (and even actively misleading[15,30]) for very distantly related organisms, particularly for the deepest branches in the tree of life[10,20–22,28]. In the most divergent lineages, phylogenetic signal has been nearly obliterated by the large number of nucleotide changes that have occurred[15,20,28]. The canonical small-subunit ribosomal RNA "Tree of Life" (e.g., Fig. 1), shows a similar structure for each of three "domains"[32]: a cluster of not-too-distantly related organisms (the "crown" group) separated from the base by a much longer branch from which a handful of isolated long branches sequentially diverge. Although the crown-group relationships shown on these trees are likely to be correct, the long basal and near-basal branches are probably misplaced.

RNA Structure

Despite the multiple sources of potential error afflicting the sequence-based analysis of rRNA data, the sheer volume of amassed data is a tempting target for analysis, and, hidden within the base sequence is additional information that is inaccessible to standard methods of phylogenetic inference. Ribosomes, like other large cellular RNAs, are mechanical devices[26,27], and their functional activity depends upon their specific three-dimensional structures[6,11,23]; these structures, in turn, are primarily determined by pairings between bases in different parts of the sequence. Since the base sequence determines the set of possible pairings, the three-dimensional structure could in principle be determined from the sequence data alone, and considerable effort has been devoted to this endeavor[7,16–18,33]. In practice, however, structures for large RNAs are generally determined by a combination of X-ray crystallography[2,8,24,31] and comparative sequence analysis[14,19]. A large number of reliable structures have been determined and are available[12,13], but, although there have been some attempts[4,5,9], there is as yet no well-accepted method for determining phylogeny from RNA structure.

Goals of our investigation

The long-range goal of this work is to develop evolutionary models of RNA structure that are generally applicable to phylogenetic inference. We begin this work here by investigating how some established methods for encoding structures perform under a simple scenario with a known phylogeny. We investigated how well a structure-based method of phylogenetic inference performed relative to purely sequence-based methods by simulation, "evolving" RNA sequences with varying constraints on structural change. We then inferred phylogenies for each data set using structure, and also using sequence data, and determined the accuracy of each method relative to the actual (simulated) evolutionary history.

Methods

General Strategy

RNA sequence data sets were simulated for several "generations" as follows:

1. An "ancestral" base sequence was generated;
2. Two descendants were generated:
    - provisional "descendants" were generated from the ancestral sequence by applying a specific number (the mutation parameter) of random mutations.
    - the secondary structure of each mutated sequence was determined

- if the structure differed from that of the immediately ancestral sequence by less than a specified amount (the structural constraint), the provisional descendant was accepted;
- otherwise, a new set of mutations was generated and tested;

3. Each descendant was then used as an ancestor, until a specified number (n) of generations was completed (for n^2 final sequences).

This procedure gave rise to a set of sequences with actual ancestor-descendant relationships that could be accurately represented by a bifurcating"tree" structure (Fig. 3). The terminal sequences were then used as input for various phylogenetic algorithms (see below).

Simulation Details

All simulations were performed using custom software written in C by N.G. RNA secondary structures and distances between structures were determined using computational routines in the Vienna RNA package[16].

For each replicate (i.e. the simulated evolution of a set of sequences), a random RNA sequence, 64 or 96 bases in length, was generated using equal probabilities for each base. Since real evolved sequences are likely to be more thermodynamically stable than random sequences[1], the initial sequences were then mutated for 100-1000 rounds; for each round, a specified number of mutations were applied, and the minimum free energy (MFE) structure was calculated for the resulting sequence. If the new structure had a lower free energy (i.e. was more thermodynamically stable) than the previous one, the mutations were accepted; otherwise, they were rejected and a new set was proposed.

Once a stable starting sequence was obtained, descendants were evolved from it as described above, using for each replicate a single mutation parameter (2, 8, 18, 32, or 50 mutations per tree bifurcation), and structural constraint (i.e., a maximum permitted RNA structure tree-edit distance[16] of 0,2,4,6,8, 10, or ∞). Five rounds of tree bifurcation were performed, generating a final data set of 32 sequences per replicate.

Control data sets were generated having either no structure constraint (allowing a random walk through sequence-structure space) or complete structure constraint (no change in structure allowed, forcing movement through a neutral network[25]). 120 simulations under various conditions (see Results) were performed on an IBM RS6000 SP (with Power4+ "Regatta" processor) and limited to 48 hours maximum run time.

Phylogenetic analysis

The sequence data sets evolved above were analyzed in PAUP*[29] using four different distance metrics and parsimony. To reduce running time, all distance analyses used neighbor-joining, and

parsimony analyses employed a heuristic search strategy. Two nucleotide distances were used (Jukes-Cantor- and P-distance) and two structure distances (HIT and "full" structure distance[16]; these are both string-edit distances on different encodings fo the structure).

Results

The combination of five mutation levels and five maximum constraint levels generated 25 sets of major parameter values. In addition, two minimum structure change levels (0 and 1; i.e. structures were not or were required to change between generations), and two sequence lengths (64 and 96 bases) were used, for a total of 100 initial conditions. Of these 100 simulations, 42 failed to finish within 48 hours: these were conditions with high mutation rates and stringent structure constraints; the longer sequences were less likely to complete within the time limit. In the totally constrained "neutral walk" analyses, similar problems occurred: at mutation rates of 18 or above, mutated sequences with no structural change were not found quickly enough to allow the simulation to complete in the allotted time. Simulations without structural constraints, in contrast, were completed within minutes.

Structural change and phylogenetic performance

The imposition of structural constraints dramatically affected the diversity of structures seen in the evolved data sets (Figures 3 and 4). The constraint levels appeared to be fairly conservative, since most structures within a particular data set were superficially similar (e.g., Figure 4).

For all data sets, we inferred trees using structural distances (e.g. Figure 5), as well as sequence-based methods. Of the two structure methods tested, the "Full" metric performed much better than the "HIT" metric, but, for most of the tested parameter values, neither method performed as well as most of the standard sequence-based methods (Figure 6). Under high mutation rates, however, the performance of the "Full" metric was degraded less than that of the sequence-based metrics; it clearly out-performed Jukes-Cantor distance, and, at the highest mutation level, it performed equally as well or better than p-distances and parsimony. The accuracy of the structure-based distance trees varied both with the mutation rate and with the degree of structural constraint (Figure 8).

Discussion

Expectations

All methods of phylogenetic inference have optimal ranges of data variation. If the variation between taxa (or sequences) is too low, there is an insufficient number of changes to resolve divergences;

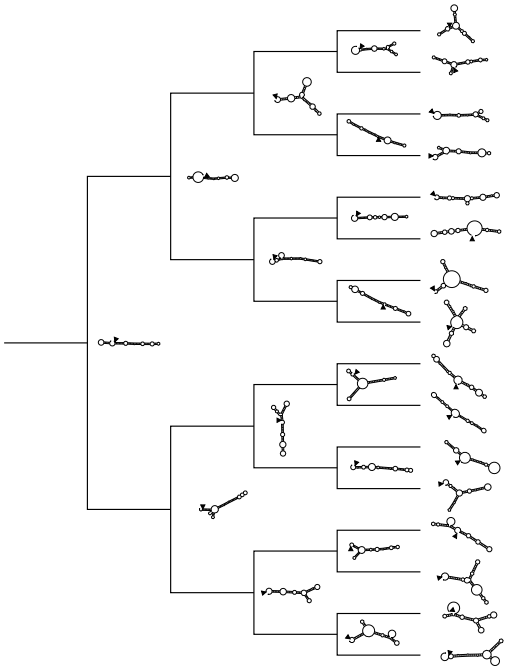

Figure 3: An example of sequence/structure evolution without structural constraints ("random walk"), and a mutation parameter of 50. For clarity, the 3' end of each sequence is marked with a black triangle. Only four of the five generations are shown.

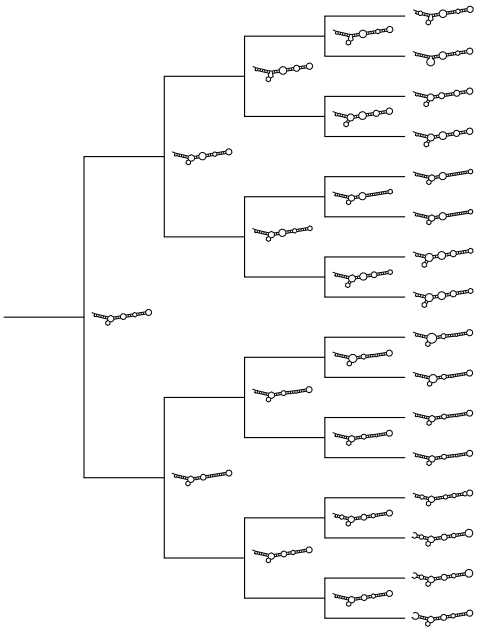

Figure 4: An example of sequence/structure evolution with structural constraints (mutation parameter, 2; constraint parameter, 8).

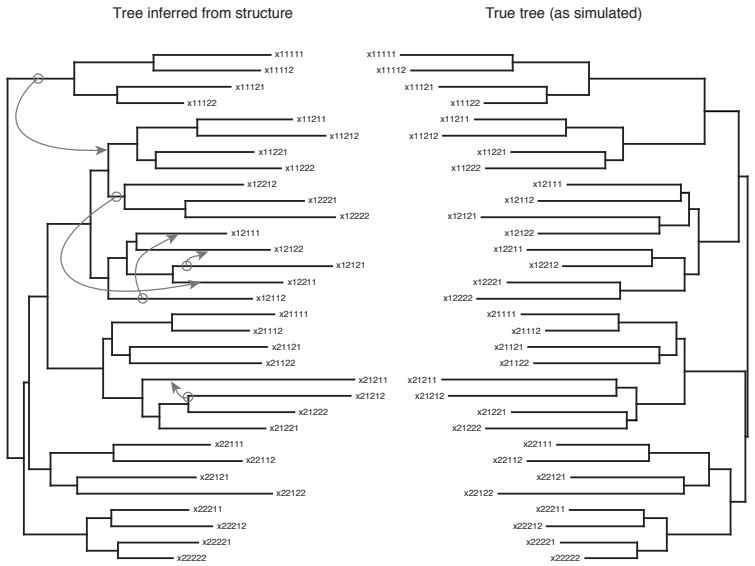

Figure 5: Example of tree inferred from structure, compared with the true tree according to which the evolution of the sequences was simulated. To re-construct the true tree from the inferred tree requires several re-arrangements (arrows).

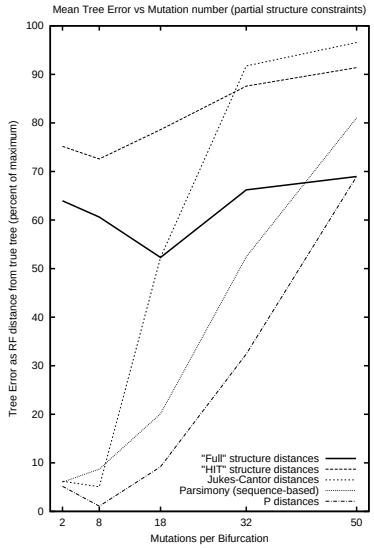

Figure 6: Accuracy of tree reconstruction when RNA structure constrains sequence evolution. For each value of the mutation parameter, the mean tree error[34] (percent of maximum Robinson-Foulds (RF) distancefrom the true tree) of all applicable simulations is shown for each analytical method.

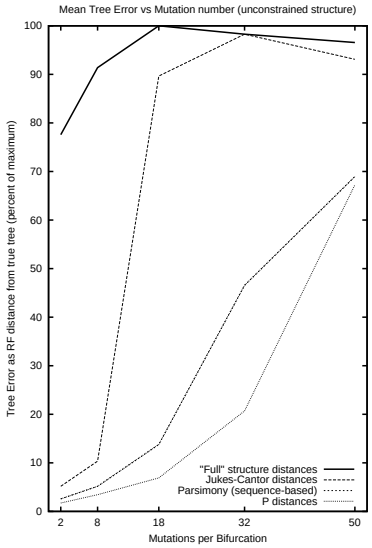

Figure 7: Accuracy of tree reconstruction on sequences evolved in the absence of structural constraints. For each value of the mutation parameter, the mean tree error of all applicable simulations is shown for each method (see Figure 6).

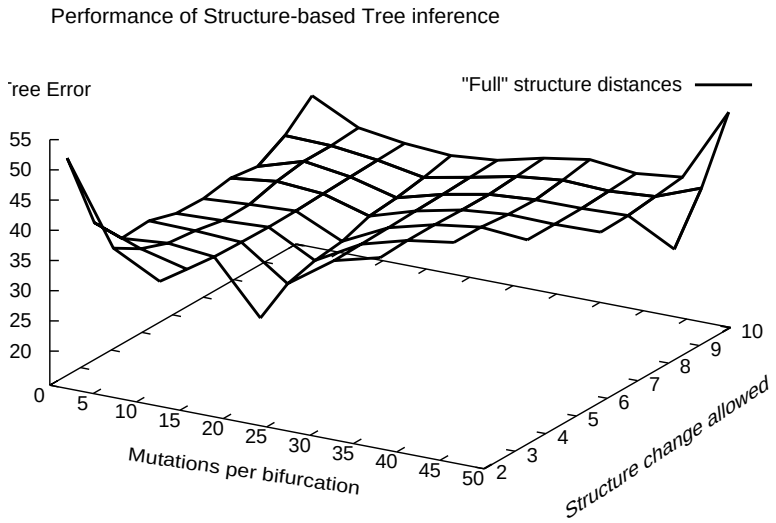

Figure 8: Accuracy of tree reconstruction for "full" distances versus both mutation and constraint parameters. The partial surface is due to incomplete results. In constrast to the error in sequence-based distance metrics, which increases steadily with both mutation and structure change (data not shown), the structure error surface has a pronounced optimum (at $m = 18$ and $constr = 4$.) but is nearly flat otherwise.

in contrast, if the variation is too high, character states that define particular groupings are likely to be overshadowed by random noise, or erased by subsequent changes. Therefore, it is common to see a "U-shaped" curve showing poor performance at the extremes when the performance of a phylogenetic algorithm is plotted against increasing divergence (it is a question of information, signal and noise). The crucial question for any phylogenetic analysis (one which is frustratingly difficult to answer a priori) is whether the data at hand have the appropriate level of divergence for the problem being addressed.

There are three questions of interest for us here: "With regard to mutation rates and structural constraints, what conditions are optimal for structure-based phylogenetic inference?", "What is the relation of that optimum (in position and extent) to the optima for standard methods of sequence-based inference?", and, finally, "Are there conditions under which structure-based methods are likely to out-perform sequence-based methods, and do they apply to any real data?"

We are encouraged that, first, there appears to be an optimum for the structure metric we used, and, second, that at some (limited) parameter values, our structure metric appears to out-perform some of the sequence-based metrics (Figure 6); to answer any of these questions definitively, it will be necessary to sample a broader range of evolutionary parameters, using far more replicates. Future work would benefit from further exploration of methods to encode structure, and of ways to infer phylogeny from structure; it will offer insight into the mechanisms of RNA evolution, and is likely to contribute to our understanding of events in the deep history of life on Earth.

References

[1] L. W. Ancel and W. Fontana. Plasticity, evolvability, and modularity in RNA. *J. Exp Zool.*, 288(3):242–83, Oct. 15 2000.

[2] N. Ban, P. Nissen, J. Hansen, P. B. Moore, and T. A. Steitz. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, 289(5481):905–20, Aug. 11 2000.

[3] S. M. Barns, C. F. Delwiche, J. D. Palmer, and N. R. Pace. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proc. Natl. Acad. Sci., U.S.A.*, 93(17):9188–93, Aug. 20 1996.

[4] B. Billoud, M. A. Guerrucci, M. Masselot, and J. S. Deutsch. Cirripede phylogeny using a novel approach: molecular morphometrics. *Mol. Biol. Evol.*, 17(10):1435–45, Oct. 2000.

[5] G. Caetano-Anollés. Novel strategies to study the role of mutation and nucleic acid structure in evolution. *Plant Cell, Tissue and Organ. Culture*, 67(2):115–132, Nov. 2001.

[6] A. P. Carter, W. M. Clemons, Jr., D. E. Brodersen, R. J. Morgan-Warren, B. T. Wimberly, and V. Ramakrishnan. Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature*, 407:340–348, Sept. 21 2000.

[7] J. H. Chen, S. Y. Le, and J. V. Maizel. Prediction of common secondary structures of RNAs: a genetic algorithm approach. *Nucleic Acids Res.*, 28(4):991–9, Feb. 15 2000.

[8] W. M. Clemons, Jr., J. L. May, B. T. Wimberly, J. P. McCutcheon, M. S. Capel, and V. Ramakrishnan. Structure of a bacterial 30S ribosomal subunit at 5.5 Å resolution. *Nature*, 400 (6747):833–40, Aug. 26 1999.

[9] L. J. Collins, V. Moulton, and D. Penny. Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP. *J. Mol. Evol.*, 51(3):194–204, Sept. 2000.

[10] P. Forterre and H. Philippe. Where is the root of the universal tree of life? *Bioessays*, 21(10): 871–9, Oct. 1999.

[11] J. Frank and R. K. Agrawal. A ratchet-like inter-subunit reorganization of the ribosome during translocation. *Nature*, 406(6793):318–22, July 20 2000.

[12] R. R. Gutell. Collection of small subunit (16S- and 16S-like) ribosomal RNA structures. *Nucl. Acids Res.*, 21(13):3051–4, July 1 1993.

[13] R. R. Gutell, M. W. Gray, and M. N. Schnare. A compilation of large subunit (23S and 23S-like) ribosomal RNA structures: 1993. *Nucl. Acids Res.*, 21(13):3055–74, July 1 1993.

[14] R. R. Gutell, J. C. Lee, and J. J. Cannone. The accuracy of ribosomal RNA comparative structure models. *Curr. Opin. Struct Biol.*, 12(3):301–10, June 2002.

[15] R. P. Hirt, J. M. Logsdon, Jr, B. Healy, M. W. Dorey, W. F. Doolittle, and T. M. Embley. Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. *Proc. Natl. Acad. Sci., U.S.A.*, 96(2):580–5, 1999.

[16] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures: The Vienna RNA package. *Monatschefte für Chemie*, 125:167–188, 1994.

[17] A. B. Jacobson, L. Good, J. Simonetti, and M. Zuker. Some simple computational methods to improve the folding of large RNAs. *Nucleic Acids Res.*, 12(1 Pt 1):45–52, Jan. 11 1984.

[18] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288(5):911–40, May 21 1999.

[19] J. A. Mears, J. J. Cannone, S. M. Stagg, R. R. Gutell, R. K. Agrawal, and S. C. Harvey. Modeling a minimal ribosome based on comparative sequence analysis. *J. Mol. Biol.*, 321(2): 215–34, Aug. 9 2002.

[20] L. Morin. Long branch attraction effects and the status of "basal eukaryotes": phylogeny and structural analysis of the ribosomal RNA gene cluster of the free-living diplomonad *Trepomonas agilis*. *J. Euk. Microbiol.*, 47(2):167–77, Mar.-Apr. 2000.

[21] H. Philippe and A. Germot. Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution. *Mol. Biol. Evol.*, 17(5):830–4, May 2000.

[22] H. Philippe and J. Laurent. How good are deep phylogenetic trees? *Current Opinion in Genetics and Development*, 8:616–623, 1998.

[23] B. T. Porse and R. A. Garrett. Ribosomal mechanics, antibiotics, and GTP hydrolysis. *Cell*, 97(4):423–6, May 14 1999.

[24] F. Schluenzen, A. Tocilj, R. Zarivach, J. Harms, M. Gluehmann, D. Janell, A. Bashan, H. Bartels, I. Agmon, F. Franceschi, and A. Yonath. Structure of functionally activated small ribosomal subunit at 3.3 Å resolution. *Cell*, 102(5):615–23, Sept. 1 2000.

[25] P. Schuster and W. Fontana. Chance and necessity in evolution: lessons from RNA. *Physica D*, 133:427–52, 1999.

[26] A. S. Spirin. Ribosome as a molecular machine. *FEBS Lett.*, 514(1):2–10, 2002.

[27] J. P. Staley and C. Guthrie. Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell*, 92(3):315–26, June 1998.

[28] J. W. Stiller and B. D. Hall. Long-branch attraction and the rDNA model of early eukaryotic evolution. *Mol. Biol. Evol.*, 16(9):1270–9, 1999.

[29] D. L. Swofford. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts., 2000.

[30] C. R. Vossbrinck, J. V. Maddox, S. Friedman, B. A. Debrunner-Vossbrinck, and C. R. Woese. Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes. *Nature*, 326(6111):411–4, Mar. 26-Apr. 1 1987.

[31] B. T. Wimberly, D. E. Brodersen, W. M. Clemons, Jr., R. J. Morgan-Warren, A. P. Carter, C. Vonrhein, T. Hartsch, and V. Ramakrishnan. Structure of the 30S ribosomal subunit. *Nature*, 407:327–339, Sept. 21 2000.

[32] C. R. Woese, O. Kandler, and M. L. Wheelis. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci., U.S.A.*, 87(12):4576–9, June 1990.

[33] M. Zuker. Calculating nucleic acid secondary structure. *Curr. Opin. Struct Biol.*, 10(3):303–10, June 2000.

[34] D. J. Zwickl and D. M. Hillis. Increased taxon sampling greatly reduces phylogenetic error. *Systematic Biology*, 51(4):588–598, 2002.