Assessing the Reliability of RNA Folding Using Statistical Mechanics

Martijn Huynen
Robin Gutell
Danielle Konings

SFI WORKING PAPER: 1996-07-044

SANTA FE INSTITUTE

Assessing the reliability of RNA folding using statistical mechanics

Martijn Huynen[*,†], **Robin Gutell**[‡] **and Danielle Konings**[†,‡]

[*] Theoretical Biology and Biophysics and Center for Nonlinear Studies
Los Alamos National Laboratory, MS-K710
Los Alamos, NM 87545 USA

[†] Santa Fe Institute
1399 Hyde Park Road
Santa Fe, NM 87501 USA

[‡] Molecular, Cellular and Developmental Biology
University of Colorado, Campus Box 347
Boulder, CO 80309 USA

submitted to the Journal of Molecular Biology

Abstract

We have analyzed the base-pair probability distributions of 16S and 16S-like and 23S and 23S-like ribosomal RNAs of Archaea, Bacteria, chloroplasts, mitochondria and Eukarya, as predicted by the partition function approach for RNA folding (McCaskill, 1990). A quantitative analysis of the reliability of RNA folding is done by comparing the base-pairing probability distributions with the structures predicted by comparative sequence analysis (comparative structure). We distinguish two factors that show a relationship to the reliability of RNA minimum free energy structure. The first factor is the dominance of one particular base-pair or the absence of base-pairing for a given base within the base-pairing probability distribution (BPPD): We characterize the BPPD per base, including the probability of not base-pairing, by its Shannon entropy (S). The S value indicates the uncertainty about the base-pairing of a base: low S values result from BPPDs that are strongly dominated by a single base-pair or by the absence of base-pairing. We show that bases with low S values have a relatively high probability that their Minimum Free Energy structure (MFE) corresponds to the comparative structure. The BPPDs of prokaryotes that live at high temperatures (thermophilic Archaea and Bacteria) have, calculated at 37° C, lower S values, than the BPPDs of prokaryotes that live at lower temperatures (mesophilic and psychrophilic Archaea and Bacteria). This reflects an adaptation of the ribosomal RNAs to the environmental temperature.

A second factor that is important to consider with regard to the reliability of MFE folding is a variable degree of applicability of the thermodynamic model of RNA folding for different groups of RNAs. Here we show that among the bases that show low S values, the Archaea and Bacteria have similar, high probabilities (.96 and .94 in 16S and .93 and .91 in 23S, respectively) that the MFE corresponds to the comparative structure. These probabilities are lower in the chloroplasts (16S .91, 23S .79), mitochondria (16S-like .89, 23S-like .69) and Eukarya (18S .81, 28S .86).

Keywords: RNA secondary structure, RNA folding, Boltzmann distribution, base-pairing probability distribution, ribosomal RNA.

Introduction

The higher order structure of RNA is crucial in many of its functions. This is exemplified by its conservation in evolution and by the *in vitro* selection from random sequences of ribozymes for the same function with a similar base-pairing pattern (Ekland *et al.*, 1995). Within RNA structure the secondary structure plays a central role, it covers the dominant energy contributions and provides the major distance constraints for the formation of the tertiary structure. The free energy of RNA secondary structure formation can be approximated by adding up the experimentally determined free energies of its elements (base pairs, hairpin loops, bulges etc.). Predicting RNA secondary structure by finding the structures with the lowest free energies has become a major research tool in experimental and theoretical biology. Here we analyze what affects the reliability of secondary structure prediction by free energy minimization.

We have shown recently that the reliability of secondary structure prediction for ribosomal RNAs (rRNAs) varies between phylogenetic classes (Konings & Gutell, 1995; Fields & Gutell, 1996). The correspondence between minimal-energy folding (MFE) and secondary structure based on comparative sequence analysis (comparative structure) is highest for Archaea followed by (eu)Bacteria. The order of the other three classes, i.e. chloroplasts, mitochondria and Eukarya, varies between the 16S-like RNAs and the 23S-like RNAs. Here we ask the question what underlies this variation in predictability. Is it the current thermodynamic model for calculating RNA secondary structure itself that for various reasons applies less to Eukaryotic rRNAs than to prokaryotic rRNAs ? For example because Eukaryotic rRNAs have more non-standard interactions within the RNA, or because they have more interactions with other molecules ? Or can we distinguish other factors that play a role ? In answering this question we will focus specifically on how the probability of (sub)structures within the Boltzmann distribution of alternative secondary structures is related to their predictability.

Secondary structure prediction by free energy minimization faces the problem that for any sequence there is an exponentially large number of possible structures. Although in thermodynamic equilibrium the structure with the lowest free energy has the highest probability, that probability is very small for long sequences. For example: in thermodynamic equilibrium the probability of the lowest free energy structure within the Boltzmann distribution is, for random sequences of the length of a 16S rRNA (about 1500 nucleotides), generally smaller than 10^{-45}. A more interesting quantity than the probability of a specific large structure is that of small substructures. This approach has been formalized by McCaskill (1990). McCaskill's algorithm focuses on the smallest sub-structure, the base pair. It calculates the comprehensive base-pair probability distribution based on the free energies and resulting probabilities of all structures. In earlier work we have calculated the base-pairing probability distribution of an entire HIV-1 genome (Huynen *et al.*, 1996). We have shown that the RNA secondary structures that are known to be a functional in HIV-1 are relatively "well-defined": i.e. their base-pair probability distribution per base is dominated by a single base-pair interaction or by the absence of base-pairing (e.g. for bases in hairpin-loops). Here we analyze the base-pairing probability distributions of 16S and 16S-like and 23S and 23S-like rRNAs.

We characterize the base-pairing probability distribution per base by its Shannon-entropy (S). Low S values correspond to probability distributions that are dominated by a single or a few probabilities. We then investigate the relationship between the S value of a base and the prediction of its correct base pairing pattern, as given in the comparative structure, by the minimum free energy structure. We discuss the nature of this relationship as revealed by the rRNAs of different phylogenetic groups and with regard to their environmental temperatures.

Methods

analysis of the base-pairing probability matrix

The complete base pairing probability matrix as calculated with the McCaskill algorithm is a symmetric $n \times n$ matrix in which the entry (i, j) is the probability that base i is paired with base j. In practice, we do not use base pairs that occur with a probability of less than 10^{-5} in our analyses. The algorithm is part of the `Vienna RNA Package` (Hofacker *et al.*, 1994; Zuker & Stiegler, 1981) which can be down-loaded with anonymous ftp from ftp.itc.univie.ac.at, directory pub/RNA/ViennaRNA-1.1b. The `Vienna RNA Package` uses free energy parameters from (Freier *et al.*, 1986; Jaeger *et al.*, 1989; He *et al.*, 1991).

We characterize the base-pairing probability distribution per base (i) by its Shannon-entropy (S_i).

$S_i = -\sum_j P_{i,j} \log P_{i,j}$

where $P_{i,j}$ is the probability of base pairing between bases i and j, and $P_{i,j=i}$ is the probability that the base i does not pair with any other base. The lower the S value, the stronger the distribution is dominated by a single or a few base-pairing probabilities. Or, in other words, S values reflect the uncertainty we have about the base-pairing. The average S value of a sequence is the sum of the S values of its bases divided by the sequence length.

In earlier work we characterized the probability distribution per base with its maximum (Huynen *et al.*, 1996), using the term "well-definedness of secondary structure". Although the qualitative results for the analyses presented here for

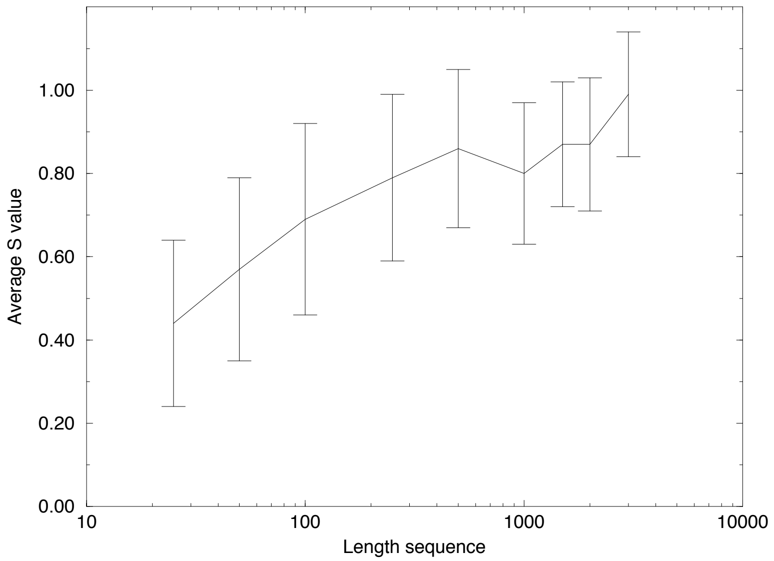

Figure 1: The relation between the average S value per (random)sequence and its length. For sequences of various length classes at least 25,000 nucleotides were folded (250 sequences of length 100, 50 of length 500 etc.). The vertical bars designate one standard deviation. The average S value increases sub-linear with the logarithm of the length of the sequence and starts to saturate at a length of 500.

both measures are the same, the entropy measure shows a higher correlation with the probability that the MFE corresponds to the comparative structure. In a recent paper Zuker and Jacobsen (1995) introduced "well determinedness" of secondary structure, where a structure is well determined if there are no alternative structures within a certain range of free energy. Although the concept is qualitatively very similar to our entropy or well definedness terms, the latter allow for a quantitative much more refined analysis, since they include actual probabilities.

In principle, one should scale the S value with the length of the sequence(N), given that the maximum value of S is the logarithm of N. For long sequences however ($N > 500$) we observed that the average S value per base as a function of sequence length increases less than linear with the logarithm of N, and saturates to a value of about 0.9 (Figure 1), also the distribution of S values per sequence does not change for $N > 500$ (data not shown). Apparently the intrinsic properties of the thermodynamic model of RNA secondary structure prediction result even for very long sequences in only a limited number of bases with which a base has a reasonable chance ($P >= 10^{-5}$) of pairing.

Comparing minimum free energy structures with comparative structures.

Predicting base-pairing patterns by comparative sequence analysis is reviewed in (Gutell *et al.*, 1994). The comparative structures of the RNAs used in this analysis (Gutell *et al.*, 1993; Gutell, 1994) can be found on http://pundit.colorado.edu:8080/. We consider a base predicted correctly in the MFE if it is paired to the same base as in the comparative structure, if it is single-stranded in both the MFE and the comparative structure and if it is single stranded in the MFE and non-canonically base-paired (non A-U, G-C or G-U) in the comparative structure. We also examined other definitions for correctly predicted bases: only considering bases that are base-paired in the comparative structure, or defining that bases that are non-standard base-paired in the comparative structure are always incorrectly predicted in the MFE. We did not observe any significant differences in the relation between the S-value and the probability that a base is predicted correctly in the MFE for the different definitions of correctly predicted bases.

Results

Shannon entropy of base pairing probability distributions

For a set of 16S and 16S-like and 23S and 23S-like rRNA sequences of Archaea, Bacteria, chloroplasts, mitochondria and Eukarya that span the phylogenetic tree and represent the major forms of structural diversity (table 1), we studied how the average S value per sequence relates to the correspondence between the MFE and the comparative structure. Figure 2 shows that there is a negative correlation between the average S value per sequence and the percentage correctly predicted base-pairs in the MFE. Thus, as the uncertainty about the base pairing pattern of the sequence decreases, the MFE of the sequence becomes more reliable. Figure 2 indicates that 16S rRNAs of Archaea have relatively low S values, followed by Bacteria, the chloroplasts, the Eukarya and the mitochondria respectively. Also for the 23S rRNAs the Archaea have the lowest S values followed by the Bacteria, but there is no clear distinction between the other classes. To get a more detailed view of the distribution of S values per class we plotted the distribution of S values of all the sequences per class in cumulative histograms (Figure 3). The Archaea clearly have the most bases with the low S values, followed by the Bacteria. For the other classes the differences are less dramatic, and the order differs for 16S(like) and 23S(like) rRNAs. A number of the Archaea species and two of the Bacterial species in our set are thermophilic or hyper-thermophilic. Since the lowest free energy state within the Boltzmann distribution becomes less dominant with increasing temperature, low S values (calculated at 37° C) might reflect an adaptation to high temperatures. For rRNA sequences that are avail-

Table 1: **Table 1 here**

able over a large range of optimal growth temperatures (Archaea and Bacteria 16S rRNAs and Archaea 23S rRNAs) we analyzed how the average S value per sequences correlates with the optimal growth temperature. A negative correlation between the average S value of the base-pairing probability distribution (BPPD) and the temperature at which the organisms live was indeed observed for all the three groups (Figure 4). In principle, calculations of the BPPD should be done for the temperature at which the secondary structure functions. Although the partition function calculation in the `Vienna RNA Package` allows for calculation of secondary structure at different temperatures, the extrapolation of the free energy parameters is inaccurate for the extremely high temperatures ($> 70°$ C) at which the hyper-thermophiles live. For example, folding the 16S rRNA of *P. occultum* at its optimal growth temperature(105° C), leads to a mostly single stranded structure. As temperature increases the entropic contributions to the free energy become more dominant, effectively melting the secondary structure. The melting of the secondary structure increases the S values, as bases alternate more between single stranded and double stranded states. Folding the same sequence at increasing temperatures does indeed lead to higher S values of the BPPD, and a lower probability of the MFE within the Boltzmann distribution (data not shown). The low S values we observe in the BPPD of the thermophiles (optimal growth temperatures between 45° C and 70° C) and hyper-thermophiles in the Archaea and in the Bacteria can therefore be explained as a result of adaptations to high temperatures. The G+C level of the 16S rRNAs in Archaea is positively correlated with their environmental temperature (Dalgaard & Garrett, 1993). Although high and balanced G and C levels are necessary to get low free energies of secondary structure (Huynen *et al.*, 1992), and prevent melting at high temperatures, they do not necessarily give rise to very well defined structures. On the contrary, random sequences that consist of solely G and C have relatively many alternative structures because they are very "sticky" (Schuster *et al.*, 1994): every nucleotide can in principle pair with 50% of the other nucleotides. Randomizing the sequences but not the base-composition of the hyper-thermophiles resulted in a rise of the average S value for all of the sequences. The average S-value for the hyper-thermophiles rose from 0.45, SD 0.09 to 0.85, SD 0.18, which is no different than that for random sequences in which all nucleotide frequencies are 0.25 (Fig 1). Note that also the 16S rRNA sequences of the mesophiles (optimal growth temperature between 20° C and 45° C) and the psychrophiles (optimal growth temperature below 20° C) have in general S values that are lower than the average for random sequences (Figure 1). In Figure 4 the Archaea have one outlier in the bottom-left: *M. soehngenii*, optimal growth temperature: 37° C, S value: 0.40. The genus Methanotrix has both mesophilic and thermophilic species, the relative low S value value of *M. soehngenii* might be the result of a recent adaptation of the species to the mesophile temperature range which is not yet fully present in the 16S RNA.

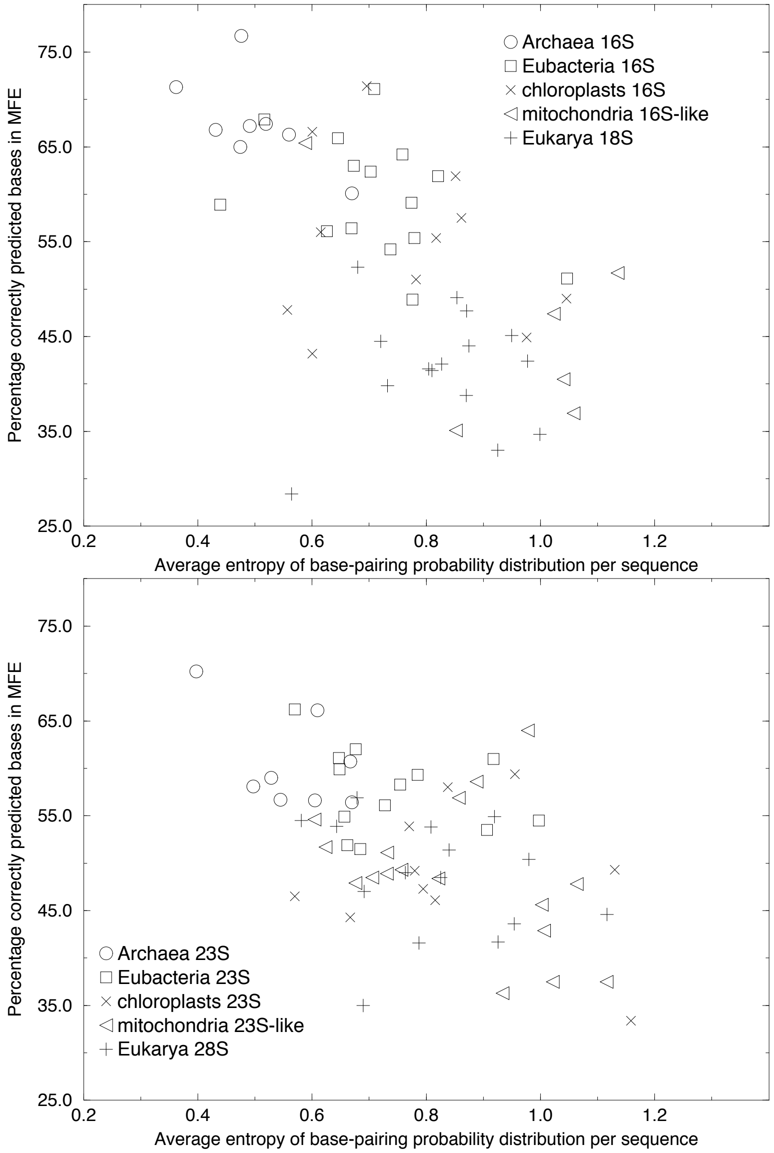

Figure 2: Relation between the average S value and the reliability of the MFE. Shown are the average S value per sequence and the degree to which the MFE of that sequence corresponds to the comparative structure for 16S and 16S-like rRNA (top) and for 23S and 23S-like rRNA (bottom). Both the 16S(like) rRNAs and the 23S(like) rRNAs show a negative correlation between the average S value and the percentage correctly predicted bases. Correlation coefficients over all the sequences are -0.61 and -0.50 for the 16S(like) and the 23S(like) rRNAs respectively.

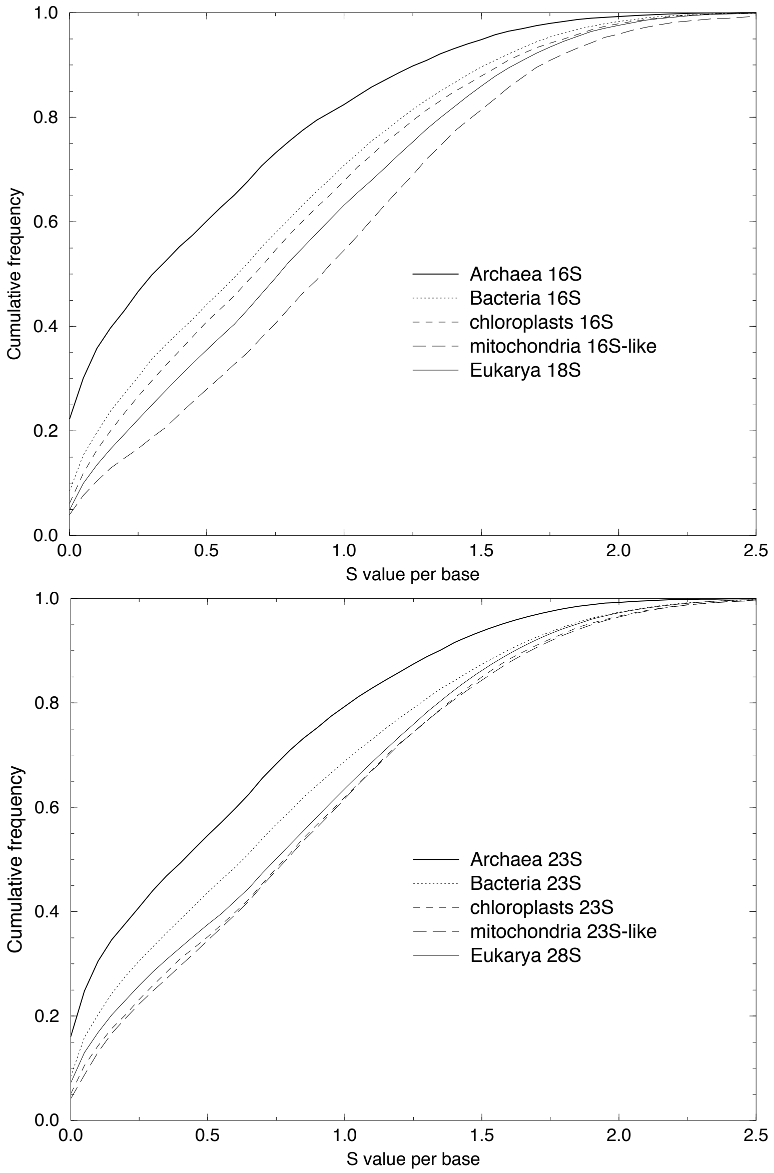

Figure 3: The entropies for the 16S(like) and the 23S(like) ribosomal RNAs per phylogenetic class. The S value per class is presented in a cumulative way. For every class we score which fraction of the bases has an S value between 0 and 0.025, 0.025 and 0.075, 0.075 and 0.125 etc. The fractions are then added. The line shows which fraction of the nucleotides have an S value below the value on the x-axis. For both 16S rRNA(top) and 23S rRNA(bottom) the Archaea have the most bases with low S values, followed by the Bacteria. For 16s rRNA the chloroplasts have a lower S value than Eukarya which in turn have a lower S value than the mitochondria. For the 23S rRNA the S values of the chloroplasts, mitochondria and Eukarya are similar.

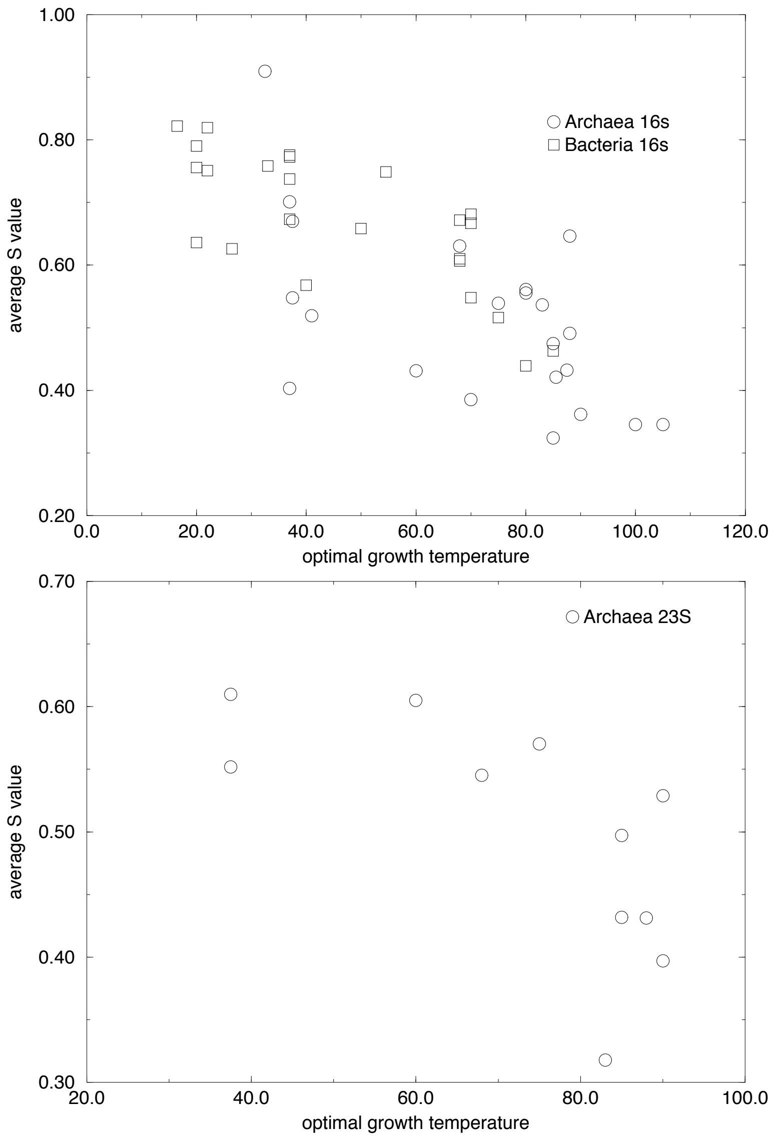

Figure 4: Relation between the optimal growth temperature and the average S value. Shown are the optimal growth temperature and the average S value for 16S rRNA in Archaea and Bacteria (top) and for 23S rRNA in Archaea (bottom). Only sequences with 5 or less unknown nucleotides were used. Data on optimal growth temperatures are from (Holt *et al.*, 1994), unless otherwise noted. When an optimum growth temperature range was specified, the temperature at the center of the range was used. The overall correlation between S values and optimal growth temperatures for 16S rRNA is −.74, for the 23S Archaea it is −.58, for the 16S Bacteria it is −.74. The species with their optimal growth temperatures in degrees Centigrade are:

Archaea 16S: *Acidianus brierleyi* 70, *Acidianus infernus* 88, *Archaeoglobus fulgidus*

83, *Desulfurococcus mobilis* 85, *Metallosphaera sedula* 75, *Methanobacterium formicicum* 41, *Methanobacterium thermoautotrophicum* 68, *Methanococcoides burtonii* 32.5, *Methanococcus vannielli* 37.5, *Methanosphaera stadtmanii* 37, *Methanospirillum hungatei* 37.5, *Methanothermus fervidus* 85.5, *Methanotrix soehngenii* 37, *Pyrococcus furiosus* 100, *Pyrodictium occultum* 105, *Sulfolobus acidocaldarius* 75, *Sulfolobus shibatae* 80, *Sulfolobus solfataricus* 85, *Thermococcus celer* 90, *Thermofilum pendens* 88, *Thermoplasma acidophilum* 60, *Thermoproteus tenax* 90.

Bacteria 16S: *Agrobacterium tumefaciens* 26, *Aquifex pyrophilus* 85 [1], *Arthrobacter globiformis* 26.5, *Bacillus megaterium* 40, *Bacillus psychrophilus* 20, *Bacteroides fragilis* 37, *Borellia burgdorferi* 37, *Carnobacterium alterfunditum* 22 [2], *Carnobacterium funditum* 22 [2], *Desulfurella acetivorans* 54.5, *Escherichia coli* 37, *Fervidobacterium gondwanalandicum* 70, *Flavobacterium salegens* 20, *Frankia alni* 37, *Psychrobacter immobilis* 20, *Renibacterium salmoninarum* 16.5, *Streptococcus thermophilus* 68, *Sulfobacillus thermosulfidooxidans* 50, *Thermoanaerobacter cellulolyticus* 68, *Thermoanaerobacter brockii* 70, *Thermoanaerobacter thermohydrosulfuricus* 68, *Thermoanaerobium lactoethylicum* 70, *Thermotoga maritima* 75, *Thermus thermophilus* 80.

Archaea 23S: *A.fulgidus*, *D.mobilis*, *M.thermoautotrophicum*, *M.hungatei*, *S.acidocaldarius*, *S.solfataricus*, *T.acidophilum*, *T.celer* , *T.pendens*, *T.tenax* .

[1] (Huber *et al.*, 1992) [2] (Franzmann *et al.*, 1991).

Relevance of the thermodynamic model of RNA folding

From the S value per base in a sequence we can analyze whether the correct prediction of a base in the MFE depends on the S value of its base pairing probability distribution. This is a more detailed analysis of the pattern that was presented in Figure 2. Instead of analyzing the correspondence between the MFE and the comparative structure per sequence we analyze the correspondence per S value of the bases. For all bases within one phylogenetic class that have an S value between, say, 0.075 and 0.125 we can count for which fraction of these bases does the MFE correspond to the comparative structure. This is a direct measure of the applicability of the secondary structure model. As we have less uncertainty about the base pairing behavior of a single base we expect that our prediction of the minimum free energy structure of that base will be closer to what is observed experimentally, or in this case, by comparative analysis.

The results (Figure 5) show that for all the classes there is a negative relation between the S value of a base and the chance that it is predicted correctly. This corresponds to the results averaged per sequence in Figure 2. The relation is, however, non-linear. The first part of the slope, between S value 0 and 0.3, is steeper than the rest of the slope. For 16S rRNA we observe that the curve

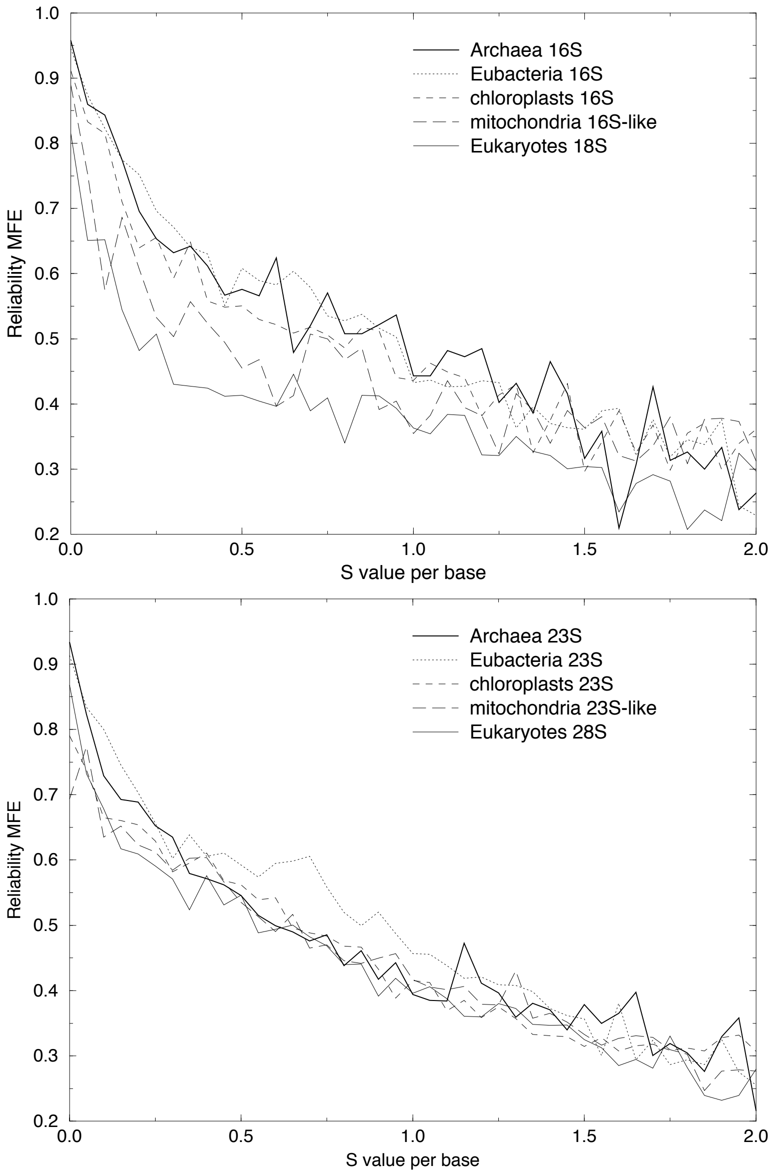

Figure 5: Predictive value of the S value for the reliability of the MFE per base. For all bases that have an S value between 0 and 0.025, 0.025 and 0.075, 0.075 and 0.125 etc., the fraction of bases for which the MFE corresponds to the comparative structure is counted. The figure hence gives a probability that for a base with the given S value the MFE corresponds to the comparative structure. For all the classes we observe a negative relation between the S value of a base and the reliability of the MFE for that base. In other words, as there is, based on the thermodynamic model, less uncertainty of the base-pairing behavior of a base, the correspondence between the MFE and the comparative structure for that base increases. In both the 16S(like) and the 23S(like) rRNAs the prokaryotes score higher than the other phylogenetic classes. For 23S(like) rRNA the differences between the prokaryotes and the other classes are smaller than for the 16S(like) RNAs.

that describes the relation between the S value and the fraction of correctly predicted bases is virtually identical for the Archaea and the Bacteria, but is lower in the chloroplasts and mitochondria, and lowest in the Eukarya. Note that relative to Figure 3, the Eukarya and the mitochondria have switched places: The Eukaryotic 18S rRNA sequences have lower S values than the mitochondrial 16S-like sequences, however the MFE for a base with a low S value is on average more reliable in a mitochondrial sequence than in a Eukaryotic sequence. For the 23S(like) rRNAs we again observe that the curve is highest in Archaea and Bacteria. The differences with the other classes are however smaller than for the 16S(like) rRNA sequences.

Competition between non-canonical and canonical base pairs

type of base pair: in comp. struct.	non-canonical	canonical	single stranded
Archaea	0.325	0.900	0.417
Bacteria	0.416	0.856	0.440
chloroplasts	0.496	0.820	0.474
mitochondria	0.566	0.769	0.556
Eukarya	0.578	0.790	0.572

Probability of being canonically base-paired in the thermodynamic model

Table 2: Average probabilities of canonical base-paring in the thermodynamic model for bases that in the comparative structure form either non-canonical base-pairs, canonical base-pairs (Watson-Crick + G-U) or single stranded bases. The standard deviations are about 0.35 for the non-canonical base-pairs, 0.25 for the canonical base pairs and 0.36 for the single stranded bases, irrespective of the taxonomic class.

The thermodynamic model for secondary structure prediction considers only Watson-Crick and G-U base-pairs. We studied to what extent the non-canonical (excluding G-U) base pairs that have been derived by the comparative analysis compete with canonical base-pairs. In other words, do they tend to occur at positions that would otherwise be paired or single stranded ? We divided the bases in three groups according to their base-pairing in the comparative structures: those that form a non-canonical base-pair (excluding G-U), those that are single stranded and those that form a canonical base-pair (Watson-Crick + G-U). Per group we score the average probability that the bases form a canonical base-pair according to the thermodynamic model. Table 2 shows that the non-canonical base-pairs in the comparative structure occur at positions that have a relatively low probability of forming a canonical base-pair in the thermodynamic model. The probabilities are comparable to the base-pairing probabilities of the positions that are single stranded in the comparative structure. In the Archaea, the base-pairing probability is even lower than for the single stranded bases. Within ribosomal RNAs there appears hence relatively little competition in the thermodynamic model from canonical base-pairs interactions at the positions where we observe non-canonical base-pairs in the comparative structure.

Discussion

We have shown that the local dominance of a single structure within the Boltzmann distribution of alternative secondary structures is strongly correlated with the reliability of the minimum free energy structure. Bases whose base-pairing probability distribution is dominated by a single base pair or by the absence of base pairing are better predicted than the ones that have many alternative states. This pattern is observed in 16S and 16S-like rRNA and 23S and 23S-like rRNA in Archaea, Bacteria, chloroplasts mitochondria and Eukarya. These results are in accordance with the data on 16S rRNA of E.coli by Zuker and Jacobson (1995) . They showed that the parts of the sequence for which a relatively few alternative structures exist within a certain energy range are those that have a high probability of being predicted correctly by minimum free energy structure. Archaea and Bacteria that live at high temperatures have, calculated at 37° C, a more uniquely determined secondary structure than bases of Archaea and Bacteria that live at low temperatures. This appears to reflect an adaptation to their environmental temperature, as the RNA secondary structure for any given sequence becomes less uniquely determined with rising temperature. An interesting observation about hyper-thermophilic Archaea and Bacteria is that their ribosomal RNAs evolve at a relatively low rate (Woese, 1987). An explanation for this is that the thermodynamic constraints imposed on their structure, as reflected in the low S values of their base pair probability distributions, reduce the fraction of neutral mutants. The fact that the S values of the (hyper)thermophiles are significantly lower than those of random sequences and more extreme than the S values of the mesophiles and psychrophiles supports this hypothesis. Environmental temperature is but one factor that affects RNA folding: low PH values and high salt concentrations at which some of the Archaea live stabilize base-pairing. They reduce the repulsion between the negatively charged backbones by protonation of the phosphates (at Ph < 2, (Saenger, 1984)) and by “shielding” the negative charges respectively.

The second factor that affects the reliability of the minimum free energy structure is the applicability of the thermodynamic model for RNA folding itself. The model that is used here for calculating the RNA secondary structure is based on the following principles: the RNA structure is in thermodynamic equilibrium, its free energy is calculated by adding up local contributions, a limited set of experimentally determined parameters of these contributions are included, and interactions other than secondary structure ones are not considered. The interaction with other molecules like proteins (Powers & Noller, 1995) and small nucleolar RNAs (Maxwell & Fournier, 1995), or kinetic effects in the folding of RNAs (Gultyaev *et al.*, 1995) can interfere with the formation of the structure as predicted by the model. By separating two factors that affect reliability of the minimum free energy folding, the first of which is intrinsic to the thermodynamic model itself, our analysis allows for a quantification of effects that interfere with the thermodynamic model.

Non-canonical base-pairs (non Watson-Crick and non G-U) are not part of the thermodynamic model of secondary structure prediction, the assumption is that they are added to the core structure that is formed by the secondary structure *sensu stricto*. We observed that there is indeed relatively little competition at the positions of the non-canonical base-pairs from canonical base-pairing. The effect is strongest in the Archaea and becomes less in respectively Bacteria, chloroplasts, mitochondria and Eukarya. The variation in this effect points to different strengths of selection to prevent canonical base-pair interactions at positions were non-canonical base-pairs occur. Since all the ribosomal RNAs are the product of a selection process on secondary structure, we can, however, from these data not conclude to what extent non-canonical base-pairs interfere with secondary structure formation in random sequences.

We have distinguished two factors to explain the variation that was observed in the reliability of the folding of ribosomal RNAs (Konings & Gutell, 1995; Fields & Gutell, 1996): the dominance within the Boltzmann distribution and the applicability of the thermodynamic folding model. Whereas the higher reliability of the MFEs of Archaea than those of Bacteria is largely due to the better defined secondary structures in Archaea. The lower reliability of the folding in chloroplasts, mitochondria and Eukarya is also due to the fact that the thermodynamic model of secondary structure prediction applies less to these RNAs.

We do fully acknowledge that the base-pairing probability distribution approach represents an entirely different view on RNA structure than does the comparative sequence analysis approach. The base-pairing probability distribution approach is based on statistical mechanics, and does in principle not predict a single structure. The uncertainty it represents in terms of a probability distribution of structures instead of a single structure is assumed to be "real" in thermodynamic equilibrium. Adaptation to this uncertainty by evolving secondary structures that are relatively dominant in their Boltzmann distribution of alternative structures has been shown for tRNAs (Marliere, 1983; Higgs, 1993) and for the functional secondary structures in HIV-1 (Huynen *et al.*, 1996). Here we have shown that this type of adaptation can also compensate the increase of uncertainty that is caused by high environmental temperatures.

Acknowledgments

Characterizing the base-pairing probability distribution by its Shannon-entropy was first suggested by Ivo Hofacker. M.H. thanks Gerhard Hummer for many useful discussions and lessons on thermodynamics. D.K. and R.G. thank the W.M. Keck foundation for its generous support of RNA research on the Boulder Campus. Part of the work was supported by a National Institutes of Health grant (GM-48207). Part of this work was done under the auspices of the U.S. Department of Energy and supported by the Center for Nonlinear Studies at Los Alamos National Laboratory and by the Santa Fe Institute.

References

Dalgaard, J. Z. & Garrett, R. A. (1993). Archaeal hyperthermophile genes. In: *The Biochemistry of Archaea*, (Neuberger, A. & van Deenen, L., eds) pp. 535–536, Amsterdam: Elsevier.

Ekland, E. H., Szostak, J. W., & Bartel, D. P. (1995). Structurally complex and highly active RNA ligases derived from random RNA sequences. *Science,* **269**, 364–370.

Fields, D. S. & Gutell, R. R. (1996). An analysis of large rRNA sequences folded by a thermodynamic method. submitted to *Folding and Design.*

Franzmann, P., Hopfl, P., Weiss, N., & Tindall, B. J. (1991). Psychrotrophic, lactic acid-producing bacteria from anoxic waters in ace lake, Antarctica; *Carnobacterium funditum* sp. nov. and *Carnobacterium alterfunditum* sp. nov. *Archives of Microbiol.* **156**, 255–262.

Freier, S. M., Kierzek, R., Jaeger, J. A., Sugimoto, N., Caruthers, M. H., Neilson, T., & Turner, D. H. (1986). Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci., USA,* **83**, 9373–9377.

Gultyaev, A. P., van Batenburg, F. H., & Pleij, C. W. (1995). The computer simulation of RNA folding pathways using a genetic algorithm. *J. Mol. Biol.* **250**, 37–51.

Gutell, R. R. (1994). Collection of small subunit (16s- and 16s-like)ribosomal RNA structures:1994. *Nucl. Acids Res.* **22**, 3502–3507.

Gutell, R. R., Gray, M. W., & Schnare, M. N. (1993). A compilation of large subunit (23s and 23s-like) ribosomal RNA structures. *Nucl. Acids Res.* **21**, 3055–3074.

Gutell, R. R., Larsen, N., & Woese, C. (1994). Lessons from an evolving rRNA: 16s and 23s rRNA structures from a comparative perspective. *Microbiological Rev.* **58**, 10–26.

He, L., Kierzek, R., SantaLucia, J., Walter, A., & Turner, D. (1991). Nearest-neighbor parameters for GU mismatches. *Biochemistry,* **30**, 11124–11132.

Higgs, P. (1993). RNA secondary structure: a comparison of real and random sequences. *J. de Physique,* **3**, 43–59.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, S., Tacker, M., & Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte f. Chemie,* **125** (2), 167–188.

Holt, J. G., Krieg, N. R., Sneath, P. H., Staley, J. T., & Williams, S. T. (1994). *Bergey's Manual of Determinative Bacteriology.* Baltimore: Williams and Wilkins.

Huber, R., Wilharm, T., Huber, R., Trincone, A., Burggraf, S., Konig, H., Rachel, R., Rockinger, I., Fricke, H., & Stetter, K. (1992). *Aquifex pyrophilus* gen-nov sp-nov represents a novel group of marine hyperthermophilic hydrogen-oxidizing bacteria. *Syst. Appl. Microbiol.* **15**, 340–351.

Huynen, M. A., Konings, D., & Hogeweg, P. (1992). Equal G and C contents in Histone genes indicate selection pressures on mRNA secondary structure. *J. Mol. Evol.* **34**, 280–291.

Huynen, M. A., Perelson, A., Vieira, W., & Stadler, P. (1996). Base-pairing probabilities in a complete HIV-1 genome. *J. Comp. Biol.* **3**, 253–274.

Jaeger, J. A., Turner, D. H., & Zuker, M. (1989). Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci. USA,* **86**, 7706–7710.

Konings, D. & Gutell, R. (1995). A comparison of thermodynamic foldings with comparatively derived structures of 16s and 16s-like rRNAs. *RNA,* **1**, 559–574.

Marliere, P. (1983). Computer building and folding of fictitious transfer-RNA sequences. *Biochimie,* **65**, 267–273.

Maxwell, E. S. & Fournier, M. J. (1995). The small nucleolar RNAs. *Ann. Rev. Biochem.* **64**, 897–934.

McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers,* **29**, 1105–1119.

Powers, T. & Noller, H. F. (1995). Hydroxyl radical footprinting of ribosomal proteins on 16s rRNA. *RNA,* **1**, 194–209.

Saenger, W. (1984). *Principles of Nucleic Acid structure.* New York: Springer-Verlag.

Schuster, P., Fontana, W., Stadler, P. F., & Hofacker, I. (1994). From sequences to shapes and back: a case study in RNA secondary structures. *Proc. Roy. Soc. Lond. B,* **255**, 279–84.

Woese, C. R. (1987). Bacterial evolution. *Microbiological Rev.* **51**, 221–271.

Zuker, M. & Jacobsen, A. B. (1995). "Well-determined" regions in RNA secondary structure prediction: analysis of small subunit ribosomal RNA. *Nucl. Acids Res.* **23**, 2791–2798.

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

Small subunit rRNA (16s and 16s like):
Archaea

Haloarcula marismortui	66.3	0.56
Haloferax volcanii	76.7	0.48
Methanobacterium formicicum	67.4	0.52
Methanococcus vannielli	60.1	0.67
Sulfolobus solfataricus	65.0	0.47
Thermoplasma acidophilum	66.8	0.43
Thermococcus celer	71.3	0.36
Thermoproteus tenax	67.2	0.49

Bacteria

Arthrobacter globiformis	64.2	0.76
Agrobacterium tumefaciens	56.1	0.63
Bacillus subtilis	56.4	0.67
Bacteroides fragilis	63.0	0.67
Borrelia burgdorferi	54.2	0.74
Chlamydia psittaci	61.9	0.82
Desulfovibrio desulfuricans	55.4	0.78
Escherichia coli	59.1	0.77
Frankia alni	48.9	0.78
Mycoplasma gallisepticum	51.1	1.05
Mycoplasma hyopneumoniae	71.1	0.71
Pseudomonas testosteroni	65.9	0.65
Synechococcus sp.6301	62.4	0.70
Thermotoga maritima	67.9	0.52
Thermus thermophilus	58.9	0.44

chloroplasts

Astasia longa	51.0	0.78
Chlamydomonas reinhardtii	44.9	0.98
Chlorella vulgaris	71.4	0.70
Cryptomonas sp.	57.5	0.86
Cyanidium caldarium	61.9	0.85
Euglena gracilis	49.0	1.05
Marchantia polymorpha	55.4	0.82
Nicotiana tabacum	47.8	0.56
Olisthodiscus luteus	66.6	0.60
Palmaria palmata	56.0	0.61

Zea mays	43.2	0.60
mitochondria		
Ascaris suum	36.9	1.06
Aspergillus nidulans	47.4	1.02
Bos taurus	65.4	0.59
Caenorhabditis elegans	40.5	1.04
Saccharomyces cerevisiae	51.7	1.14
Zea mays	35.1	0.85
Eukarya		
Babesia bigemina	34.7	1.00
Cryptococcus neoformans	52.3	0.68
Encephalitozoon cuniculi	33.0	0.93
Giardia ardeae	41.6	0.80
Giardia intestinalis	28.4	0.56
Giardia muris	38.8	0.87
Gracilariopsis sp.	44.5	0.72
Hexamita sp.	39.8	0.73
Homo sapiens	44.0	0.87
Mus musculus	42.1	0.83
Placopecten magellanicus	45.1	0.95
Saccharomyces cerevisiae	42.4	0.98
Tritrichomonas foetus	49.1	0.85
Vairimorpha necatrix	41.4	0.81
Xenopus laevis	47.7	0.87
Large subunit rRNA (23s and 23s like) Archaea		
Halobacterium marismortui	56.4	0.67
Halococcus morrhuae	60.7	0.67
Methanobacterium thermoautotrophicum	56.7	0.55
Methanococcus vannielii	66.1	0.61
Sulfolobus solfatiricus	58.1	0.50
Thermococcus celer	70.2	0.40
Thermoproteus tenax	59.0	0.53
Thermoplasma acidophilum	56.6	0.61
Bacteria		

Bacillus subtilis	58.3	0.75
Borrelia burgdorferi	53.5	0.91
Campylobacter coli	54.9	0.66
Escherichia coli	61.1	0.65
Frankia alni	51.5	0.68
Mycobacterium leprae	56.1	0.73
Pseudomonas aeruginosa	59.3	0.79
Pseudomonas cepacia	54.5	1.00
Rhodobacter sphaeroides	59.9	0.65
Streptomyces ambofaciens	51.9	0.66
Synechococcus sp.6301	61.0	0.92
Thermotoga maritima	66.2	0.57
Thermus thermophilus	62.0	0.68
chloroplasts		
Alnus incana	46.1	0.82
Astasia longa	33.4	1.16
Chlamydomonas eugametos	53.9	0.77
Chlamydomonas reinhardtii	59.4	0.96
Chlorella ellipsoidea	49.2	0.78
Euglena gracilis	49.3	1.13
Marchantia polymorpha	47.3	0.79
Nicotiana tabacum	44.3	0.67
Palmaria palmata	58.0	0.84
Zea mays	46.5	0.57
mitochondria		
Acanthamoeba castellanii	47.8	1.06
Caenorhabditis elegans	37.5	1.02
Chondrus crispus	45.6	1.00
Crossostoma lacustre	64.0	0.98
Dictyostelium discoideum	48.5	0.71
Drosophila melanogaster	42.9	1.00
Gallus gallus	51.1	0.73
Homo sapiens	56.9	0.86
Marchantia polymorpha	48.4	0.82
Paracentrotus lividus	51.7	0.62
Paramecium tetraurelia	48.9	0.73
Prototheca wickerhamii	49.3	0.76

Saccharomyces cerevisiae	37.5	1.12
Schizosaccharomyces pombe	58.6	0.89
Tetrahymena pyriformis	47.9	0.68
Xenopus laevis	54.6	0.60
Zea mays	36.3	0.93
Eukarya		
Arabidopsis thaliana	54.9	0.92
Caenorhabditis elegans	47.0	0.69
Chlorella ellipsoidea	53.8	0.81
Cryptococcus neoformans	51.4	0.84
Giardia intestinalis	35.0	0.69
Giardia muris	44.6	1.12
Herdmania momus	43.6	0.95
Oryza sativa	49.0	0.76
Pneumocystis carinii	53.9	0.64
Physarum polycephalum	41.7	0.93
Phytophthora megasperma	56.9	0.68
Prorocentrum micans	41.6	0.79
Saccharomyces cerevisiae	48.5	0.83
Tetrahymena thermophila	54.5	0.58
Toxoplasma gondii	50.4	0.98

Table1: The species, their percentage correctly predicted bases and their average S values. A base is considered to be predicted correctly in the MFE if it has the same paired to the same base as in the comparative structure, if it is single stranded both in the MFE and in the comparative structure and if it is single stranded in the MFE and non-canonically (non A-U or G-C or G-U) paired in the comparative structure.