Published online 6 June 2017

Nucleic Acids Research, 2017, Vol. 45, No. 14 ***8541–8550***
doi: 10.1093/nar/gkx512

Advanced multi-loop algorithms for RNA secondary structure prediction reveal that the simplest model is best

Max Ward[1,*], Amitava Datta[1], Michael Wise[1,2] and David H. Mathews[3]

[1]Computer Science & Software Engineering, The University of Western Australia, Australia, [2]The Marshall Centre for Infectious Diseases Research and Training, The University of Western Australia, Australia and [3]Department of Biochemistry & Biophysics, Department of Biostatistics & Computational Biology, and Center for RNA Biology, University of Rochester, NY, USA

Received January 16, 2017; Revised May 09, 2017; Editorial Decision May 28, 2017; Accepted May 31, 2017

ABSTRACT

Algorithmic prediction of RNA secondary structure has been an area of active inquiry since the 1970s. Despite many innovations since then, our best techniques are not yet perfect. The workhorses of the RNA secondary structure prediction engine are recursions first described by Zuker and Stiegler in 1981. These have well understood caveats; a notable flaw is the ad-hoc treatment of multi-loops, also called helical-junctions, that persists today. While several advanced models for multi-loops have been proposed, it seems to have been assumed that incorporating them into the recursions would lead to intractability, and so no algorithms for these models exist. Some of these models include the classical model based on Jacobson–Stockmayer polymer theory, and another by Aalberts and Nadagopal that incorporates two-length-scale polymer physics. We have realized practical, tractable algorithms for each of these models. However, after implementing these algorithms, we found that no advanced model was better than the original, ad-hoc model used for multi-loops. While this is unexpected, it supports the praxis of the current model.

INTRODUCTION

Ribonucleic acid (RNA) is an important molecule in biology. We are only now beginning to understand the scope of its role as new functional RNA sequences are discovered (1–3). There has been an explosion of RNA sequence data as technology progressed in recent years (4). A well accepted axiom in functional biology is that molecular structure is tantamount to biological function. This appears to be true for RNA, as its structure is conserved during evolution (5). This should come as no surprise in light of its functional role. For example, eukaryote development has been described as being driven by an 'RNA machine' (6). Other examples include its action as a catalyst (7,8), its role in gene silencing (9), and the RNA world hypothesis (10), which posits that RNA had a fundamental role in the genesis of life. As such, the determination of RNA structure is an important problem.

Unfortunately, the most accurate approaches for determining RNA structure, such as nuclear magnetic resonance and X-ray crystallography, are time consuming, expensive, and require considerable technical expertise, as RNA is more labile than DNA (11). Less involved approaches provide an attractive alternative. These comprise comparative sequence analysis, and *de novo* prediction algorithms. Comparative analysis requires considerable manual effort and a large set of related RNA sequences (5), which is often unavailable. As such, *de novo* computational approaches are an area of intense interest. These methods typically predict the structure of RNA from only the primary nucleotide sequence.

Most *de novo* secondary structure prediction algorithms are derived from work by Zuker and Stiegler (12), who, in 1981, provided a set of recursions defining a dynamic programming algorithm capable of efficiently finding a minimum free energy (MFE) structure. The import of this follows from Anfinsen's thermodynamic hypothesis (13), which posits that biological molecules are likely to be in their MFE state. Because of this, the Zuker and Stiegler algorithm was able to predict RNA secondary structures with reasonable accuracy.

The Zuker and Stiegler algorithm and its derivatives rely on a model of RNA secondary structure folding free energy change to define MFE structures. This model has since been called the *nearest neighbor* model. The nascent form of the model was defined by Tinoco *et al.* (14,15), and by Salser (16). Later it was formalized and expanded by Turner and

*To whom correspondence should be addressed. Tel: +61 8 6488 2238; Email: max.ward-graham@research.uwa.edu.au

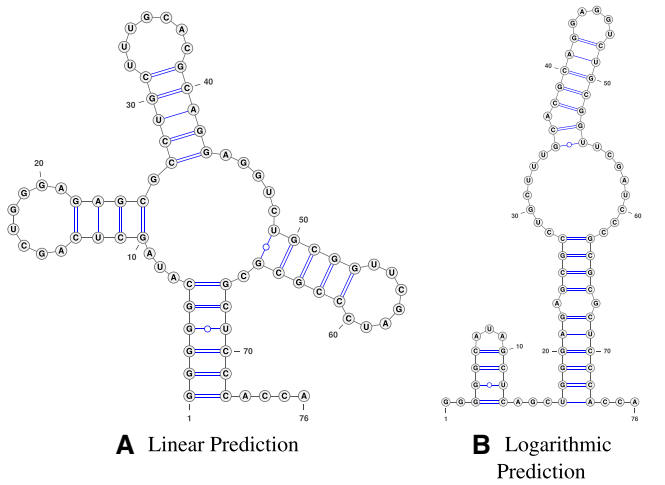

Figure 1. A tRNA secondary structure (Sprinzl ID *RA1661* (43)) that is perfectly predicted by the linear model (PPV = 1, sensitivity = 1), but not by the logarithmic model (PPV = 0, sensitivity = 0). Panel (**A**) shows the prediction from the linear model and panel (**B**) shows the prediction from the logarithmic model. The logarithmic model computes the free energy of the linear prediction to be –30.1 kcal/mol, while its own prediction has a score of –30.2 kcal/mol. The linear model, on the other hand gives these scores of –30.9 kcal/mol and –30.2 kcal/mol respectively.

coworkers (17–22), and is described in full in the Nearest Neighbor Database (NNDB) (23).

The original algorithm described by Zuker and Stiegler (12) gives multi-loops zero folding free energy, effectively ignoring their free energy contribution. Multi-loops are loops from which three or more helices exit, and an example is shown in Figure 1 of a four-way multi-loop. Later, a simple, ad-hoc, linear function of both the number of unpaired nucleotides and the number of branches was used to model multi-loop energy (22,24,25). It was stated that using a more appropriate model, derived from Jacobson–Stockmayer polymer theory (26), would require exponential computation time (27). Other reports have provided only exponential time algorithms that could incorporate the model, which implies that this statement is well accepted (25,28). We report a novel finding: the Jacobson–Stockmayer based model can be incorporated without the exponential time requirement, and we have an efficient, polynomial time algorithm for this model. Using this algorithm, we will compare the effectiveness of the linear model against the Jacobson–Stockmayer based model. This is of practical interest, as the linear function is used by modern software packages for MFE prediction (29–31). The Jacobson–Stockmayer-based model is typically used by the same software packages to evaluate the free energy change of given RNA structures.

We shall refer to the Jacobson–Stockmayer model of multi-loop folding free energy as the *logarithmic* model. This is because it has a logarithmic dependence on the number of unpaired nucleotides in a multi-loop. Let us define the number of unpaired nucleotides in a multi-loop as u, and the number of branches as b, then the logarithmic model is defined, in kcal/mol, by:

$$\Delta G^\circ = \begin{cases} 10.1 - 0.3b - 0.3\,u & \text{if } u \le 6 \\ 10.1 - 0.3b - 0.3 \times 6 + 1.1 \times \ln(u/6) & \text{otherwise} \end{cases} \quad (1)$$

Supplementary Table S1 has reference free energy change values under the logarithmic model for various combinations of b and u. These were computed using the formula exactly as presented here.

The logarithmic term and its coefficient are from Jacobson-Stockmayer polymer theory (26), and appear to have first been suggested by Salser (16). However, the other terms come from the Turner 1999 parameters (17) which were derived by optimizing performance on known structures. For more information about this model and its parametrization we refer the reader to the NNDB (23), and to the derivation of the Turner 1999 parameters (17). For comparison, the linear model parameters we used are taken from work by Mathews *et al.* (32), and are based on linear regression. They are as follows in kcal/mol.

$$\Delta G^\circ = 9.3 - 0.6b + 0\,u \quad (2)$$

Note that these parameters were later published as the basis of different, optimized parameters for the Turner 2004 parameter set (18). However, in 2009, the parameters were reverted to the unoptimized parameters we used (33), and which are currently used in the modern version of RNAstructure (29). The parameters were reverted so that there was no training of the parameters to structure prediction accuracy. This way, a fair comparison against CONTRAfold (34) could be done.

After these multi-loop models were proposed, other more advanced models were proposed. Aalberts and Nandagopal (35) presented a model that applies two-length-scale polymer physics to describe multi-loop free energies in terms of chain entropy. This expands upon the Jacobson–Stockmayer theory by explicitly accounting for the fact that unpaired nucleotides and helix ends have different sizes, and thus contribute differently to the entropy cost of closing the loop. They did not provide a MFE prediction algorithm incorporating their model, but instead re-evaluated the folding free energies of a set of low free energy structures generated by the standard dynamic programming algorithm as had been done previously for the logarithmic model (17). They did find evidence that their model made accurate MFE predictions, and thus was a realistic energy model. We were also able to design a polynomial time MFE prediction algorithm for this model. We here use it to comprehensively test and analyze the performance of the model. The model defines a multi-loop in terms of two different length scales: length-a, and length-b, which represent the typical length between consecutive nucleotides, and the length of crossing a multi-loop branch respectively. The precise values of a and b are defined in angstroms to be $a = 6.2$ and $b = 15$. If we say that the number of length-a segments is N, and the number of length-b segments is M, then the model is defined, in kcal/mol, as:

$$\Delta G^\circ = \frac{59}{36} kT \ln(N^{\frac{6}{5}} a^2 + M^{\frac{6}{5}} b^2) + C \quad (3)$$

Note that k, T and C refer to the Boltzmann constant, the absolute temperature, and a scaling factor respectively. In practice, We fixed the temperature to 310.15 K, as in the Turner rules (17,18) and as done by Aalberts and Nandagopal (35), and C was set to zero as suggested by Aalberts and Nandagopal (35).

Supplementary Table S2 has reference free energy change values under the Aalberts and Nandagopal model for various numbers of branches and unpaired nucleotides. These were computed using the formula exactly as presented here. We explain the correspondence of length-a and length-b segments to the numbers of branches and unpaired nucleotides later, in the section titled *Aalberts and Nandagopal Model*.

In our structure prediction benchmarks of MFE algorithms for the linear, logarithmic, and Aalberts and Nandagopal models, the linear model had the best structure prediction accuracy. This was surprising, and suggests that the currently available packages are already using the best available free energy change model for multi-loop folding, in spite of the origin of the linear model as computationally convenient alternative to the logarithmic model.

MATERIALS AND METHODS

Software

All algorithms were implemented using the C++11 programming language standard. Our algorithms were programmed to avoid isolated base pairs (36), which is common to most popular RNA structure prediction software packages (29,30). The algorithms also include the free energy change contributions of coaxial stacking, dangling ends, terminal mismatches, and end penalties (18,23) fully. The algorithms were implemented *de novo*, but RNAstructure 5.8.1 (29) was used to provide the energy model functions except for multi-loops. A repository containing our code and results can be found at https://github.com/maxhwardg/advanced_multiloops.

Data set

The data set comprised 3948 known RNA primary sequence and structure pairs. These structures are available at (http://rna.urmc.rochester.edu/archiveII.tar.gz), and comprise the 'ArchiveII' data set compiled by the Mathews lab. The data set contains some information about which nucleotides are single stranded for tRNAs. This information was not used while running our algorithms. The full data set was used for the logarithmic model. For the AN model, a reduced data set of 2783 RNA sequences was used. It comprises all RNA sequences from the full data set whose lengths in nucleotides are fewer than or equal to 300 nts. This limit was to ensure that we did not run out of memory while executing the algorithms. The number of RNA in each family for these data sets is included in Supplementary Tables S5 and S6.

Scoring

F-scores were calculated and used for comparison of accuracy. F-score is the harmonic mean of sensitivity and positive predictive value (also called precision, or PPV). Sensitivity is the fraction of known pairs correctly predicted, and positive predictive value is the fraction of predicted pairs that are in the known structure. A summary of prediction statistics including PPV and sensitivity scores for the various algorithms can be found in Supplementary Tables S5 and S6.

Paired t-tests were used to compare the F-scores of different algorithms predicting the same RNA sequences (37). Every set of data used in a paired t-test comparison was checked to see if the assumptions underlying a paired t-test were satisfied. The data was visualized, then skewness and kurtosis measurements were computed to ensure that the data was normally distributed, or at least unimodal and not extremely skewed.

Scoring of base pairs

Most of the known structures used for testing were determined by comparative sequence analysis. At positions where a base pair can have alternative pairing partners, for example one of the two paired nucleotides could alternatively pair with the nucleotide adjacent to its paring partner, there is uncertainty in the true structure. This arises from thermal fluctuations in pairing (38) and also from limitations in the resolution of comparative analysis (39). To ascertain that this did not introduce any problems during our analysis, we reran all of our statistical tests while considering these possible alternative base pairings as correct predictions. No substantial differences were found.

Optimizing parameters

Some of the parameters for the logarithmic model and the Aalberts and Nandagopal (AN) model are derived by optimization. We attempted to re-optimize them because the parameter set we use, the Turner 2004 (18) parameters, differs slightly from the parameter set these models were originally derived for, the Turner 1999 (17,35) parameters. We will provide results for our algorithms using their original parameterization, and our optimized parameters.

Optimization was done by grid search over all parameters near the original parameters for each model. A randomly selected set of 20 tRNAs and 20 5S rRNAs was used as the training set. Performance was judged to be the average F-score of a parameter set when used for prediction on the training set. For the logarithmic model, the initiation, branch, and unpaired costs (originally 10.1, –0.3 and -0.3 respectively in Equation 1) were optimized. Our choice of parameters to be optimized is the same as in the Turner 1999 parameter set (17). For the AN model, the scaling value C (originally set to zero) was optimized. The range of parameters searched, and the best parameters found, are summarized in Table 1. Supplementary Tables S3 and S4 show example free energy changes computed using our optimized parameters.

RESULTS

Logarithmic model

We give a brief description of the algorithm we found for the logarithmic model, then prove its theoretical time and space

Table 1. A summary of parameter optimization for the logarithmic and AN models. For the logarithmic model, we define a to be the initiation cost (10.1 in Equation 1), and b and c to be the branch and unpaired costs respectively (both –0.3 in Equation 1). The step size was chosen to be 0.1 as this is the minimum free energy difference recognized in the RNAstructure energy functions by default. All values are in kcal/mol

	Search Space	Best Parameter Set
Logarithmic Model	$8.1 \leq a \leq 12.1$, $-0.8 \leq b, c \leq 0.5$	$a = 11.0$, $b = -0.8$, $c = -0.5$
AN Model	$-3.0 \leq C \leq 3.0$	$C = -0.5$

requirements. Following these results, we provide the results for the algorithm's accuracy when used for MFE structure prediction. Readers who seek only to understand the practical application of our findings may wish to skip to the accuracy report.

The algorithm we devised for including the logarithmic model is an extension of the typical Zuker and Stiegler formulation, which uses the linear model. Since lucid descriptions of this algorithm already exist (12,28,40,41) we do not include our own in the interest of brevity. We shall define *Paired*(i, j) to be the MFE substructure enclosed by the base pair (i, j). In the case that (i, j) close a multi-loop, another recurrence relation is usually invoked, which contains the optimal internal part of a multi-loop. We modified this recursion as follows. We define the MFE of any internal fragment of a multi-loop between bases i and j inclusive that has at least b branches, and exactly u unpaired nucleotides to be *MultiFragment*(b, u, i, j). Now, the *Paired*(i, j) function can find the MFE multi-loop it could close by calling *MultiFragment*(2, u, $i+1$, $j-1$) for all possible values of u, i.e. all possible numbers of unpaired nucleotides. This works because a multi-loop contains at least three branches, and thus at least two branches not including the closing branch, hence $b = 2$ is sufficient in the recursive call to *MultiFragment*. In addition, upon closing a multi-loop, the number of unpaired nucleotides in that multi-loop is known ahead of time. This allows the logarithmic dependence on the number of nucleotides to be computed in *Paired* when closing a multi-loop. Note that the remaining terms in the model can be computed the same way as in the linear model algorithm.

The definition of the *MultiFragment* recursion has some subtleties. First, let us say that *MultiFragment*(0, 0, i, j) $= 0$ when $i > j$, since an empty fragment is valid if it contains at least zero branches, and exactly zero unpaired nucleotides. Similarly, let *MultiFragment*(b, u, i, j) $= \infty$ when $i > j$ and $b \neq 0$ or $u \neq 0$, since an empty fragment must have exactly zero branches and unpaired nucleotides. With these base cases in mind, the recursive cases become easier to understand. *MultiFragment*(b, u, i, j) either has an unpaired nucleotide at i followed by the rest of the multi-loop fragment, or some branch with its left nucleotide at i followed by the rest of the multi-loop fragment. The recursive definition is:

$$\begin{aligned} &\mathit{MultiFragment}(b, u, i, j) = \\ &\min \begin{cases} \mathit{MultiFragment}(b, u-1, i+1, j) \\ \mathit{Decompose}\ \forall k \ni i < k \leq j \end{cases} \\ &\mathit{Decompose} = \mathit{Paired}(i, k) - \texttt{branch\_cost} \\ &\qquad + \mathit{MultiFragment}(max(0, b-1), u, k+1, j) \end{aligned} \tag{4}$$

Coaxial stacking, dangling ends, terminal mismatches, and end penalties are not included in the description we have given of our algorithm. While their free energy contributions are important, they obfuscate the core idea behind the algorithm as they entail additional cases. These cases have been left out since they follow from the core recursions that we have provided, and can be re-derived.

The complexity analysis of our algorithm is interesting. First, we assume utilization of dynamic programming on the recurrence relations. The time requirement of computing *Paired* remains $O(n^3)$ where n is the number of nucleotides in an RNA. This is because the non-multi-loop cases require only $O(n)$ time (42), trying all possible numbers of unpaired nucleotides requires only $O(n)$ time, and there are only $O(n^2)$ states. The number of states for the *MultiFragment* table is $O(n^3)$ despite having four parameters. This is because $0 \leq b \leq 2$ and thus b contributes only $O(1)$ states. The time requirement for computing a single state for *MultiFragment* is $O(n)$, since k must iterate over all split points. Thus the time requirement overall is $O(n^4)$. The time and space requirement for *MultiFragment* dominate the algorithm, and so the algorithm has $O(n^4)$ and $O(n^3)$ complexities for time and space respectively. It is important to realize that, while our algorithm is fast enough to be used in practice, it has greater time and space requirements compared to the typical algorithm that uses the linear model and requires only $O(n^3)$ time and $O(n^2)$ space (25,42).

Logarithmic model results

The linear and logarithmic models were compared for MFE prediction using a set of RNA sequences with known secondary structures (see *Materials and Methods*). A comparison was made using both the original parameters for the logarithmic model, and our optimized parameters. First we consider the results using the original parameters.

A summary of prediction statistics can be found in Table 2. The linear model was statistically significantly better ($P < 0.05$) for three families of RNA, while the logarithmic model was superior for two. The results for the remaining five families are not statistically significant, but favor the linear model. These results constitute evidence that the linear model is better at predicting secondary structures compared to the logarithmic model when using known parameters sets.

Figure 1 provides an example where the linear model prediction is better than the logarithmic, using a typical tRNA structure. It has a classical, four-way branching multi-loop that is common to many tRNA. The linear model is able to predict this structure perfectly. In contrast, the logarithmic model does not predict a multi-loop at all, as it gives the correct multi-loop a higher energy penalty compared to the linear model. This leads to a poor prediction. The logarithmic model failing to predict any multi-loop is a typical example when the structure prediction is worse than the linear model. For some statistics that show this effect, see

Table 2. A summary of the MFE prediction results comparing the linear model to the logarithmic model. *F*-score values for the predictions are reported. Average scores grouped by RNA family are provided. Statistically significant differences are bold, denoting the better model. This data set comprises 3948 RNAs

	5S rRNA	16S rRNA	23S rRNA	Group I Introns	Group II Introns	RNaseP RNA	SRP RNA	Telomerase RNA	tmRNA	tRNA
Linear average *F*-score	0.599	0.518	0.669	0.484	0.255	**0.529**	0.594	0.465	**0.406**	**0.686**
Logarithmic average *F*-score	**0.618**	0.515	0.667	0.479	0.248	0.502	**0.602**	0.463	0.388	0.652
Paired *t*-test *P*-value (two-tailed)	<0.001	0.518	0.952	0.524	0.120	<0.001	<0.001	0.876	<0.001	<0.001

Supplementary Table S7. In addition, for completeness we provide an example where the logarithmic model makes a better prediction than the linear model in Supplementary Figure S1.

The linear model was also compared to the logarithmic model using our optimized parameters for the logarithmic model. A summary of these results can be found in Table 3. As before, the linear model has more families for which it produced statistically significantly higher *F*-scores ($P < 0.05$). The linear model has two statistically significant advantages, while the optimized logarithmic model is only significantly superior for tRNA in our data set, which were one of the RNA families used to optimize parameters. These results are different to those using the logarithmic model with its original parameters. Specifically, performance on tRNA has increased dramatically. In contrast, performance on SRP RNA has decreased markedly. The average *F*-scores on other families has changed too, but less dramatically.

We compared some of the predictions of the logarithmic model using the original parameters to those obtained using the optimized parameters. This is informative, if qualitative rather than quantitative. The RNA described in Figure 1 represents a typical mistake for the logarithmic model using the original parameters. The same tRNA is perfectly predicted using our optimized parameters. This is true of many tRNA, and for some other RNA in which a multi-loop was overlooked by the original parameters. In contrast, the optimized parameters seem to be overly stabilizing when predicting multi-loops in other RNA. For example, the SRP RNA described in Figure 2, which is well predicted by the original parameters, is predicted with a multi-loop by the optimized parameters. Interestingly, the prediction made by the optimized logarithmic parameters is exactly the same as that made with the AN model. Again, we refer the reader Supplementary Table S7 to see some illustrative statistics on the types of errors the optimized parameters introduce.

Aalberts and Nandagopal model

As with the logarithmic model, we first discuss the algorithm and its complexity, then accuracy results are presented.

The Aalberts and Nandagopal (AN) MFE structure prediction algorithm is similar to that incorporating the logarithmic model. As such, we use the logarithmic recursions as our starting point. The free energy change of a multi-loop in the AN model scales non-linearly with the number of length-*a* and length-*b* segments in the multi-loop (see Equation 3), and cannot be computed like the linear model, or parts of the logarithmic model. When a multi-loop is closed in the AN model, the number of length-*a* and length-*b* segments must be known to correctly score the free energy change of that multi-loop. This is similar to the requirements for logarithmic model algorithm in which the number of unpaired nucleotides must be known when closing a multi-loop. In particular, without considering coaxial stacking, any unpaired nucleotide effectively contributes one length-*a* segment, and a branch one length-*a* and one length-*b* segment. For the sake of brevity, we refer the reader to the original paper (35) for a description of how coaxial stacking can be defined in terms of length-*a* and length-*b* segments. So, for the AN model, we can define *MultiFragment*(N, M, i, j) to be the MFE multi-loop fragment between nucleotides i and j inclusive that has exactly N length-*a* segments, and exactly M length-*b* segments. The modified recurrence relation is:

$$\begin{aligned} &MultiFragment(N,\ M, i,\ j) = \\ &\min \begin{cases} MultiFragment(N-1,\ M, i+1,\ j) \\ Decompose\ \forall k \ni i < k \leq j \end{cases} \\ &Decompose = Paired(i, k) \\ &\qquad + MultiFragment(N-1,\ M-1, k+1,\ j) \end{aligned} \quad (5)$$

Note that in the case of empty segments (for which $i > j$), *MultiFragment*$(N, M, i, j) = \infty$ if $N \neq 0$ or $M \neq 0$, and *MultiFragment*$(N, M, i, j) = 0$ otherwise. Having defined all cases, the optimal multi-loop for some closing pair (i, j) can be found by examining all combinations of N and M for *MultiFragment*$(N, M, i + 1, j - 1)$. Again, we have omitted coaxial stacking, terminal mismatches, end penalties, and dangling ends from our discussion of the algorithm. This is to keep our description concise, and the addition of these terms is straightforward once the core idea behind the algorithm is understood.

The algorithm for AN model requires both more time and space then the linear or logarithmic models. Observe that, under the assumption of dynamic programming, the *MultiFragment* table has $O(n^4)$ cells, since N, $M = O(n)$. Each cell also requires $O(n)$ time to have its value computed since k must iterate over all split points. This leads to a time and space requirement of $O(n^5)$ and $O(n^4)$ respectively. Note that the *Paired* table remains much the same as for the logarithmic model, except that $O(n^2)$ time is required per cell to iterate through values for N and M, which entails a time

Table 3. A summary of the MFE prediction results comparing the linear model to the logarithmic model with optimized parameters. *F*-score values for the predictions are reported. Average scores grouped by RNA family are provided. Statistically significant differences are bold, denoting the better model. This data set comprises 3948 RNAs

	5S rRNA	16S rRNA	23S rRNA	Group I Introns	Group II Introns	RNaseP RNA	SRP RNA	Telomerase RNA	tmRNA	tRNA
Linear average *F*-score	0.599	0.518	0.669	0.484	0.255	0.529	**0.594**	0.465	**0.406**	0.686
Optimized logarithmic average *F*-score	0.595	0.518	0.662	0.474	0.257	0.526	0.556	0.441	0.395	**0.733**
Paired *t*-test *P*-value (two-tailed)	0.361	0.906	0.691	0.107	0.651	0.493	<0.001	0.067	0.002	<0.001

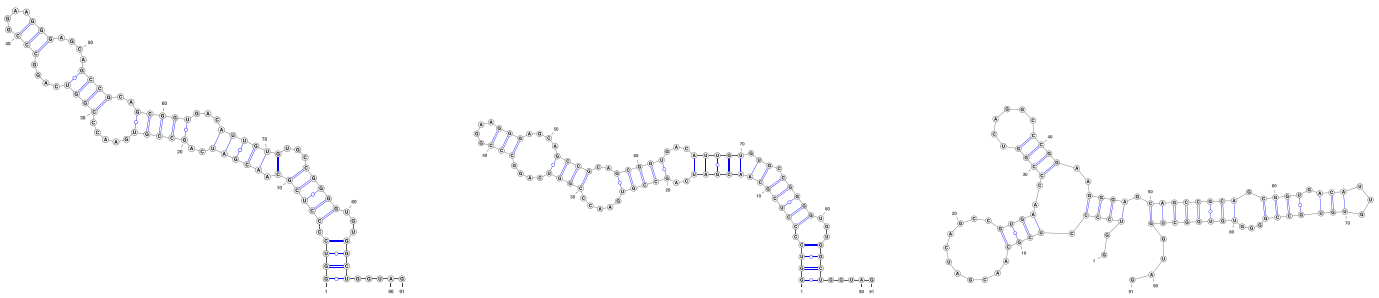

Figure 2. An SRP RNA structure that is almost perfectly predicted by the linear model (PPV = 0.929, sensitivity = 0.963), and poorly predicted by the AN model (PPV = 0, sensitivity = 0). Panel (**A**) is the accepted structure, panel (**B**) is the structure predicted by the linear model and panel (**C**) is the structure predicted by the AN model (35). The AN model gives the linear prediction a score of –32.9 kcal/mol, while its own prediction gets a score of –33.4 kcal/mol. The linear model gives scores of –32.9 kcal/mol and –32.3 kcal/mol respectively.

and space requirement of $O(n^4)$ and $O(n^2)$ respectively, and so the computation of *MultiFragment* dominates.

Aalberts and Nandagopal model results

The linear and AN models were compared for MFE prediction using a subset of the RNA sequences used to test the logarithmic model. This is because the AN algorithm requires a great deal of memory, 874 MBs for an RNA of length 100 nts for example. As such, it was not possible for us to run it for RNA sequences longer than about 300 nts. We first present results for the AN model using its original parameters, then with our optimized parameters.

A summary of results using the original parameters can be found in Table 4. The linear model achieved a higher average *F*-score for every RNA family. Three of these results appeared to be statistically significant ($P < 0.05$). These results provide strong evidence that the linear model is more effective than the AN model for predicting RNA secondary structures.

As with the logarithmic model, we wish to illustrate a common class of prediction error that occurs when using the AN model. Consider the signal recognition particle (SRP) RNA secondary structure found in Figure 2. The true structure contains no multi-loops, and is well predicted by the linear model, which achieves an *F*-score of 0.945, and predicts no spurious multi-loops. The AN model, however, predicts part of the SRP structure to be a small multi-loop, thus poorly predicting the structure. This is an example of a common class of error for the AN model in which a small multi-loop is injected into a structure, or in which small multi-loops are incorrectly favored over larger ones. The reader can see this class of error quantified in Supplementary Table S8. In contrast, Supplementary Figure S2 depicts a case where the AN models predictive propensity toward multi-loops leads to a better prediction.

In addition to comparison against the linear model, we also compared the AN model to the logarithmic model. A summary our results can be found in Table 5. The logarithmic model appeared to yield more accurate predictions for 5S and SRP RNA with statistical significance ($P < 0.05$). The logarithmic model was also better at predicting 16S RNA structure, however, the AN model had an advantage for tRNA. These differences were not statistically significant.

The AN model with optimized parameters was also compared to the linear model. These results are summarized in Table 6. Again, the linear model has statistically significant ($P < 0.05$) higher *F*-scores for more RNA. The AN model appears to be significantly better for no RNA families, while the linear model is superior for four. The results differ somewhat to those obtained with the original parameters. The linear model is closer in performance for RNaseP RNA and tRNA due to small accuracy increases for the AN model after optimization, though it remains statistically significantly better. However, the linear model gains a significant lead on 5S rRNA.

Some qualitative analysis of the predictions of the AN model with optimized parameters versus those obtained using the original parameters was also done. The SRP in Fig-

Table 4. A summary of the MFE prediction results comparing the linear model to the AN model. *F*-score values for the predictions are reported. Average scores grouped by RNA family are provided. Statistically significant differences are bold, denoting the better model. This data set comprises 2783 RNAs

	5S rRNA	16S rRNA	Group I Introns	RNaseP	SRP RNA	tRNA
Linear average *F*-score	0.599	0.630	0.505	**0.511**	**0.597**	**0.686**
AN average *F*-score	0.599	0.611	0.492	0.480	0.580	0.662
Paired *t*-test *P*-value (two-tailed)	0.989	0.485	0.450	0.004	<0.001	<0.001

Table 5. A summary of the MFE prediction results comparing the logarithmic model to the AN model. *F*-score values for the predictions are reported. Average scores grouped by RNA family are provided. Statistically significant differences are bold, denoting the better model. This data set comprises 2783 RNAs

	5S rRNA	16S rRNA	Group I Introns	RNaseP	SRP RNA	tRNA
Logarithmic average *F*-score	**0.618**	0.634	0.492	0.484	**0.604**	0.652
AN average *F*-score	0.599	0.611	0.492	0.480	0.580	0.662
Paired *t*-test *P*-value (two-tailed)	<0.001	0.396	0.990	0.708	<0.001	0.123

Table 6. A summary of the MFE prediction results comparing the linear model to the AN model with optimized parameters. *F*-score values for the predictions are reported. Average scores grouped by RNA family are provided. Statistically significant differences are bold, denoting the better model. This data set comprises 2783 RNAs

	5S rRNA	16S rRNA	Group I Introns	RNaseP	SRP RNA	tRNA
Linear average *F*-score	**0.599**	0.630	0.505	**0.511**	**0.597**	**0.686**
Optimized AN average *F*-score	0.586	0.596	0.480	0.481	0.566	0.673
Paired *t*-test *P*-value (two-tailed)	0.001	0.189	0.218	0.007	<0.001	0.046

ure 2 is mis-predicted by both parameterizations with both making the same prediction. This seems to be the case for many RNA, and few of the RNAs we examined yielded different predictions. However, some notable examples include tRNA for which no multi-loop is predicted using the original parameters, but which are better predicted by the optimized parameters (examples include Sprinzl IDs *RI1180* (43) presented in Supplementary Figure S3). In contrast, and much like the optimized logarithmic parameters, the optimized AN parameters seem to over predict multi-loops in some RNA when compared to the original parameters (see *BX572093* from the SRPDB (44) in Supplementary Figure S4 for an example). This over prediction is quantified in Supplementary Table S8 (Table 7).

The AN model with optimized parameters was also compared to the logarithmic model with optimized parameters. This parallels the comparison we have already described in Table 5. There is a statistically significant ($P < 0.05$) difference for tRNA. This difference favors the logarithmic model. Additionally, the differences between 5S rRNA, and SRP RNA are also significant ($P < 0.05$). These are wins for the logarithmic model, and the AN model respectively. Overall, the logarithmic model has greater statistically significant advantages. The results appear to suggest that the logarithmic model with optimized parameters makes more accurate predictions compared to the AN model with optimized parameters.

DISCUSSION

Our results provide evidence that the linear model leads to better predictions than the logarithmic model using the available thermodynamic parameter sets, and after our parameter optimization method. The linear model seems to have a statistically significant advantage for a majority of RNA families. The linear model also has an advantage for many other families of RNA, although a benchmark of this size cannot demonstrate statistical significance. As such, it appears that the linear model leads to better MFE predictions than the logarithmic model. Anfinsen's thermodynamic hypothesis (13) suggests that this means the linear model could also be a better model of RNA thermodynamics than the logarithmic model.

The results we gathered for the AN model were more pronounced. The AN model never achieved higher performance than the linear model with statistical significance for any RNA family even after parameter optimization. This provides strong evidence that the linear model leads to better MFE predictions compared to the AN model. Again we invoke the thermodynamic hypothesis and suggest that this implies that the linear model might again be the better model of RNA thermodynamics than the AN model.

There are some challenges to our suggestion that these models are poor models of RNA free energy because they are ineffective for MFE prediction. A notable one is that RNA sequences may exist in meta-stable configurations, with several possible 'true' structures (45). This is not captured when using algorithms to find a single MFE structure. However, our data set of sequences with known structures represents a set of structured ncRNA, which are each likely to have a single functional structure *in vivo*.

Another challenge is that our results contradict the findings of Aalberts and Nandagopal (35). They compared their model, the linear model, and the logarithmic model for MFE prediction, much as we have done. They did not use an algorithm to find a true MFE structure under each model,

Table 7. A summary of the MFE prediction results comparing the logarithmic model with optimized parameters to the AN model with optimized parameters. *F*-score values for the predictions are reported. Average scores grouped by RNA family are provided. Statistically significant differences are bold, denoting the better model. This data set comprises 2783 RNAs

	5S rRNA	16S rRNA	Group I Introns	RNaseP	SRP RNA	tRNA
Optimized logarithmic average *F*-score	**0.595**	0.592	0.472	0.498	0.557	**0.733**
Optimized AN average *F*-score	0.586	0.596	0.480	0.481	**0.566**	0.673
Paired *t*-test *P*-value (two-tailed)	0.013	0.747	0.645	0.097	0.027	<0.001

however. Instead, a set of RNA secondary structures was generated by using the standard Zuker and Stiegler style algorithm (using the linear model) to generate all structures within a window of the MFE. Then, a given model was used to re-evaluate the free energy change of each structure. A model's performance was thence judged to be its effectiveness in labeling the most accurate structure as a MFE structure. They found that the models could be ranked by effectiveness in ascending order as linear, logarithmic, then AN. We found that the linear model appears better than both other models. Having said this, we used different parameters for the linear model. The parameters used in the Aalberts and Nandagopal paper were published in 1999 (17), and were updated later (18,32,33) to the parameters used by us. The reasons for our choice of parameters is discussed below. Thus, because different parameters were used for our investigation compared to that of Aalberts and Nandagopal, comparing the corresponding results for the linear model could lead to spurious conclusions. We can, however, consider the logarithmic model compared to the AN model.

Our results comparing the AN model to the logarithmic model partially contradict those of Aalberts and Nandagopal. The RNA families common to both reports are tRNA, SRP RNA and 5S rRNA. For tRNA, Aalberts and Nandagopal found a notable difference in performance in favor of the AN model. In contrast, we found only a minor, not statistically significant difference in favor of the AN model. In addition, we also report a statistically significant advantage for the logarithmic model when predicting 5S rRNA. In contrast, Aalberts and Nandagopal found little difference in performance for 5S rRNA. We were able to produce similar results for SRP RNA, however our results are statically significant, while those reported by Aalberts and Nandagopal are not. To reinforce this, our results using optimized parameters contrast with the findings of Aalberts and Nandagopal even more markedly. For tRNA, there is a large, significant difference favoring the logarithmic model. Further, a significant difference exists for 5S rRNA favoring the logarithmic model, and a significant advantage for SRP RNA appears to exist for the AN model. These findings contradict the originally published results. We propose three explanations for our inability to reproduce Aalberts and Nandagopal's results fully.

First, our method was different; the algorithms we used examine the entire search space of possible structures. In contrast, Aalberts and Nandagopal used a limited set of structures near the MFE. Since the linear model was used to generate the structures in this window, it is likely that the true MFE structures under the models tested were filtered out as not stable enough by the linear model. The second reason is that different sets of RNA sequences were used for testing. Our data set appears to be larger, containing 3948 RNA sequences compared to the 1354 in the set used by Aalberts and Nandagopal. The third and final reason is that different methods were used for comparison. We tested complete MFE prediction algorithms and were able to compare prediction results using *F*-scores. In contrast, Aalberts and Nandagopal used the models to select a best estimate from a set of RNA structures, and thus chose to use the frequency of correct estimates as their accuracy statistic.

It is important to note that the parameters used for the three models were derived differently. For the linear model, the parameters we used came from regression on data from optical melting experiments (32). Thus it should be a reasonable model of free energy, but is grounded in empirical evidence rather than theory. There exist prior sets of parameters for the linear model, but they are entirely optimized for structure prediction accuracy on known RNA sequence and structure pairs (17,18,46), and so we felt it would be disingenuous to use them for comparison. Furthermore, the parameters we chose are those used in the current version of RNA structure (29). For the logarithmic model, some parameters are optimized (17), but the logarithmic term comes from theory (16). The AN model is also largely theoretical, however the *C* term can be used to adjust the model (35).

To ensure that our findings were robust after reparameterization, we re-optimized the relevant parameters in the logarithmic and AN models. The parameter optimization technique we used was quite simplistic. Better results might be realized using regression to derive parameters, as has been done for many of the other parameters in the Turner parameter sets (17,18). In addition, a constraint programming based approach, like that of Andronescu (46), might be used to derive better parameters for these models. A straightforward improvement would be to train on a larger data set. We used only 40 randomly selected RNA from two families. This was due to limitations in computational resources and time. Since our algorithms scale polynomially, it should be feasible to train on a larger set of RNA with more resources. To be sure our findings are valid, deeper investigation of the parameter space of these models is important, and this is a clear direction for future research. Determining why the linear model with current parameters appears to be so effective is a similar open question.

Interestingly, the optimized parameters for both the AN model, and the logarithmic model, improve performance for tRNA, but not for 5S rRNA. This is unexpected, as an equal number of tRNA and 5S rRNA were used for training. Looking at the performance of the parameters on a per RNA basis in the training set suggests that either the perfor-

mance on tRNA could be increased, or the performance on 5S rRNA could be increased, but not both. We hypothesize that this means that the models themselves are insufficient to describe the space of multi-loop free energies completely.

We conclude that, using existing parameters sets, the linear model appears to be superior to the logarithmic and AN models for both structure prediction, and as an energy model. This justifies the persistent use of the linear model in RNA algorithms. Further, it suggests that the logarithmic model should not be used as a standard for judging the free energy of RNA structures. Instead the linear model should be used both to predict and to evaluate RNA secondary structures. Our findings using optimized parameters for these models also supports this claim.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Daniel Aalberts for helpful discussions of the Aalberts and Nandagopal model.

FUNDING

National Institutes for Health [R01GM076485]. Funding for open access charge: INSERM.
Conflict of interest statement. None declared.

REFERENCES

1. Eddy,S.R. (2002) Computational genomics of noncoding RNA genes. *Cell*, **109**, 137–140.
2. Hofacker,I. and Stadler,P.F. (2010) RNAz 2.0: improved noncoding RNA detection. In: *Pacific Symposium on Biocomputing*. Vol. **15**, pp. 69–79.
3. Fu,Y., Xu,Z.Z., Lu,Z.J., Zhao,S. and Mathews,D.H. (2015) Discovery of novel ncRNA sequences in multiple genome alignments on the basis of conserved and stable secondary structures. *PLoS One*, **10**, e0130200.
4. Kodama,Y., Shumway,M. and Leinonen,R. (2012) The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
5. Pace,N.R., Thomas,B.C. and Woese,C.R. (1999) *The RNA World*. 2nd edn. Cold Spring Harbor Laboratory Press, NY.
6. Amaral,P.P., Dinger,M.E., Mercer,T.R. and Mattick,J.S. (2008) The eukaryotic genome as an RNA machine. *Science*, **319**, 1787–1789.
7. Nissen,P., Hansen,J., Ban,N., Moore,P.B. and Steitz,T.A. (2000) The structural basis of ribosome activity in peptide bond synthesis. *Science*, **289**, 920–930.
8. Doudna,J.A. and Cech,T.R. (2002) The chemical repertoire of natural ribozymes. *Nature*, **418**, 222–228.
9. Meister,G. and Tuschl,T. (2004) Mechanisms of gene silencing by double-stranded RNA. *Nature*, **431**, 343–349.
10. Gilbert,W. (1986) Origin of life: the RNA world. *Nature*, **319**, 618.
11. Neidle,S. (2010) *Principles of Nucleic Acid Structure*. Academic Press.
12. Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.
13. Anfinsen,C.B. (1973) Principles that govern the folding of protein chains. *Science*, **181**, 223–230.
14. Tinoco,I., Uhlenbeck,O.C. and Levine,M.D. (1971) Estimation of secondary structure in ribonucleic acids. *Nature*, **230**, 362–367.
15. Tinoco,I., Borer,P.N., Dengler,B., Levine,M.D., Uhlenbeck,O.C., Crothers,D.M. and Gralla,J. (1973) Improved estimation of secondary structure in ribonucleic acids. *Nature*, **246**, 40–41.
16. Salser,W. (1978) Globin mRNA sequences: analysis of base pairing and evolutionary implications. In: *Cold Spring Harbor Symposia on Quantitative Biology* . Cold Spring Harbor Laboratory Press, NY, Vol. **42**, pp. 985–1002.
17. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
18. Mathews,D.H., Disney,M.D., Childs,J.L., Schroeder,S.J., Zuker,M. and Turner,D.H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 7287–7292.
19. Andronescu,M., Condon,A., Turner,D.H. and Mathews,D.H. (2014) The determination of RNA folding nearest neighbor parameters. *Methods Mol. Biol.*, **1097**, 45–70.
20. Xia,T., SantaLucia,J. Jr, Burkard,M.E., Kierzek,R., Schroeder,S.J., Jiao,X., Cox,C. and Turner,D.H. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick base pairs. *Biochemistry*, **37**, 14719–14735.
21. Freier,S.M., Kierzek,R., Jaeger,J.A., Sugimoto,N., Caruthers,M.H., Neilson,T. and Turner,D.H. (1986) Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci. U.S.A.*, **83**, 9373–9377.
22. Jaeger,J.A., Turner,D.H. and Zuker,M. (1989) Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci. U.S.A.*, **86**, 7706–7710.
23. Turner,D.H. and Mathews,D.H. (2010) NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res*, **38**(suppl. 1), D280–D282.
24. Sankoff,D. and Kruskal,J.B. (1983) *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison. In: An Anthology of Algorithms and Concepts for Sequence Comparison*, Addison-Wesley Boston, pp. 265–310.
25. Zuker,M. and Sankoff,D. (1984) RNA secondary structures and their prediction. *Bull. Math. Biol.*, **46**, 591–621.
26. Jacobson,H. and Stockmayer,W.H. (1950) Intramolecular reaction in polycondensations. I. The theory of linear systems. *J. Chem. Phys*, **18**, 1600–1606.
27. Zuker,M., Mathews,D.H. and Turner,D.H. (1999) Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In: *RNA Biochemistry and Biotechnology*. Springer, pp. 11–43.
28. Rivas,E. and Eddy,S.R. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, **285**, 2053–2068.
29. Reuter,J.S. and Mathews,D.H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, **11**, 129.
30. Markham,N. and Zuker,M. (2007) UNAFold: software for nucleic acid folding and hybridization.. *Methods Mol. Biol.*, **453**, 3–31.
31. Lorenz,R., Bernhart,S.H., Zu Siederdissen,C.H., Tafer,H., Flamm,C., Stadler,P.F., Hofacker,I.L. *et al.* (2011) ViennaRNA Package 2.0.. *Algorithms Mol. Biol.*, **6**, 26.
32. Mathews,D.H. and Turner,D.H. (2002) Experimentally derived nearest-neighbor parameters for the stability of RNA three-and four-way multibranch loops. *Biochemistry*, **41**, 869–880.
33. Lu,Z.J., Gloor,J.W. and Mathews,D.H. (2009) Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA*, **15**, 1805–1813.
34. Do,C.B., Woods,D.A. and Batzoglou,S. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**, e90–e98.
35. Aalberts,D.P. and Nandagopal,N. (2010) A two-length-scale polymer theory for RNA loop free energies and helix stacking. *RNA*, **16**, 1350–1355.
36. Bompfünewerer,A.F., Backofen,R., Bernhart,S.H., Hertel,J., Hofacker,I.L., Stadler,P.F. and Will,S. (2008) Variations on RNA folding and alignment: lessons from Benasque. *J. Math. Biol.*, **56**, 129–144.
37. Xu,Z., Almudevar,A. and Mathews,D.H. (2012) Statistical evaluation of improvement in RNA secondary structure prediction. *Nucleic Acids Res.*, **40**, e26.

38. Woodson,S.A. and Crothers,D.M. (1987) Proton nuclear magnetic resonance studies on bulge-containing DNA oligonucleotides from a mutational hot-spot sequence. *Biochemistry*, **26**, 904–912.
39. Gutell,R.R., Lee,J.C. and Cannone,J.J. (2002) The accuracy of ribosomal RNA comparative structure models. *Curr. Opin. Struct. Biol.*, **12**, 301–310.
40. Lyngsø,R.B. and Pedersen,C.N. (2000) RNA pseudoknot prediction in energy-based models. *J. Comp. Biol.*, **7**, 409–427.
41. Sankoff,D. (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, **45**, 810–825.
42. Lyngsø,R.B., Zuker,M. and Pedersen,C.N. (1999) Internal loops in RNA secondary structure prediction. In: *Proceedings of the Third Annual International Conference on Computational Molecular Biology*, ACM, pp. 260–267.
43. Jühling,F., Mörl,M., Hartmann,R.K., Sprinzl,M., Stadler,P.F. and Pütz,J. (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res.*, **37**(suppl. 1), D159–D162.
44. Rosenblad,M.A., Gorodkin,J., Knudsen,B., Zwieb,C. and Samuelsson,T. (2003) SRPDB: signal recognition particle database. *Nucleic Acids Res.*, **31**, 363–364.
45. Schultes,E.A. and Bartel,D.P. (2000) One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science*, **289**, 448–452.
46. Andronescu,M., Condon,A., Hoos,H.H., Mathews,D.H. and Murphy,K.P. (2007) Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics*, **23**, i19–i28.