Improved RNA secondary structure prediction by maximizing expected pair accuracy

ZHI JOHN LU,^{1,3} JASON W. GLOOR,^{1,3} and DAVID H. MATHEWS^{1,2,3}

¹Department of Biochemistry and Biophysics, University of Rochester Medical Center, Rochester, New York 14642, USA

²Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, New York 14642, USA

ABSTRACT

Free energy minimization has been the most popular method for RNA secondary structure prediction for decades. It is based on a set of empirical free energy change parameters derived from experiments using a nearest-neighbor model. In this study, a program, MaxExpect, that predicts RNA secondary structure by maximizing the expected base-pair accuracy, is reported. This approach was first pioneered in the program CONTRAfold, using pair probabilities predicted with a statistical learning method. Here, a partition function calculation that utilizes the free energy change nearest-neighbor parameters is used to predict base-pair probabilities as well as probabilities of nucleotides being single-stranded. MaxExpect predicts both the optimal structure (having highest expected pair accuracy) and suboptimal structures to serve as alternative hypotheses for the structure. Tested on a large database of different types of RNA, the maximum expected accuracy structures are, on average, of higher accuracy than minimum free energy structures. Accuracy is measured by sensitivity, the percentage of known base pairs correctly predicted, and positive predictive value (PPV), the percentage of predicted pairs that are in the known structure. By favoring double-strandedness or single-strandedness, a higher sensitivity or PPV of prediction can be favored, respectively. Using MaxExpect, the average PPV of optimal structure is improved from 66% to 68% at the same sensitivity level (73%) compared with free energy minimization.

Keywords: RNA secondary structure; free energy minimization; partition function; nearest-neighbor model

INTRODUCTION

Many classes of functional RNAs have been discovered, and the pace of discovery has accelerated over the last decade. These are termed noncoding RNAs (ncRNA) because they function directly and not via a protein product (Eddy 2002). The multiple functions of ncRNA have fundamentally changed the original formulation of the Central Dogma of Biology, where it was originally thought that proteins were the only final effectors of genetics. Both the discovery (Washietl et al. 2005; Ustilov et al. 2006; Torarinsson et al. 2008) and the analysis of function (Mathews and Turner 2006) of ncRNA depend on the determination of structure.

RNA secondary structure refers to the set of canonical base pairs (A-U, G-C, and G-U) in an RNA sequence. The

most popular methods for RNA secondary structure prediction from sequence are those methods that predict the most probable structure. One such approach is based on a biophysical model that predicts the preferred equilibrium structure, the structure with lowest Gibbs free energy change of folding. In free energy minimization, a nearest-neighbor model (Xia et al. 1998; Mathews et al. 1999, 2004; Lu et al. 2006) is used to predict the conformational stability of a given structure. The set of free energy and enthalpy change nearest-neighbor parameters was derived by linear regressions from a set of optical melting experiments on model systems (Mathews et al. 1999, 2004; Lu et al. 2006). An alternative to finding the most likely structure by free energy minimization is the approach in which the most probable structure is found using a stochastic context-free grammar (Dowell and Eddy 2004). The parameters underlying the stochastic context-free grammar are trained on the set of sequences with known structures. Recently, free energy minimization and knowledge-based structure prediction were combined (Andronescu et al. 2007). The product of this work is an alternative set of free energy change parameter values. These were derived from constraint satisfaction on the set of optical melting

³The authors have no conflict of interest.

Reprint requests to: David H. Mathews, Department of Biochemistry and Biophysics, University of Rochester Medical Center, 601 Elmwood Avenue, Box 712, Rochester, NY 14642, USA; e-mail: david_mathews@urmc.rochester.edu; fax: (585) 275-6007.

Article published online ahead of print. Article and publication date are at <http://www.rnajournal.org/cgi/doi/10.1261/rna.1643609>.

experiment data and a large database of RNA sequences with known secondary structures (Andronescu et al. 2007).

Secondary structure prediction can be benchmarked for accuracy using sensitivity and positive predictive value (PPV) for base-pair prediction. Sensitivity is the percentage of known pairs correctly predicted, and PPV is the percentage of predicted pairs that are in the known structure. These two statistics are calculated as:

$$\text{sensitivity} = \frac{\text{number of correctly predicted base pairs}}{\text{total number of known base pairs}},$$

$$\text{PPV} = \frac{\text{number of correctly predicted basepairs}}{\text{total number of predicted basepairs}}.$$

The sensitivity of free energy minimization has been benchmarked as high as 73% on a diverse database of RNA sequences with known structures of fewer than 700 nucleotides (Mathews et al. 2004; Mathews and Turner 2006), but the PPV of free energy minimization is only 66% (Mathews 2004). The lower PPV has two causes. First, there is a tendency to overpredict base pairs because the formation of pairs lowers the free energy change. Second, occasionally the database of known structures does not annotate all experimentally determined base pairs, and predicting a correct pair that is not annotated lowers PPV.

As an alternative method to finding the most probable structure, the structure with maximum expected accuracy can be predicted. Pseudo-knot-free structures are predicted by maximizing the sum of the base-paired and single-stranded nucleotide probabilities, called expected accuracy, where pairing probabilities can be weighted by a factor, γ (Do et al. 2006):

$$\text{Expected accuracy}(S) = \sum_{(i,j) \in \text{BP}} \gamma \cdot 2P_{\text{bp}}(i,j) + \sum_{k \in \text{SS}} P_{\text{ss}}(k), \quad (1)$$

where $P_{\text{bp}}(i, j)$ is the base-pair (bp) probability for a nucleotide at position i and a nucleotide at position j ; and $P_{\text{ss}}(k)$ is the single-stranded (SS) nucleotide probability for the nucleotide at position k . The sums are over all base pairs (BPs), the set of BPs, and all SS, the set of SS nucleotides in structure S . It makes intuitive sense that this is a measure of expected folding accuracy because it has previously been shown that base pairs with high base-pairing probabilities in the thermodynamic ensemble are more likely to be in the known structure (Mathews 2004). For example, in the lowest free energy structure, base pairs with pair probability >99% have a 91% PPV (Mathews 2004).

Do et al. (2006) first introduced the concept of predicting the maximum expected accuracy structure to RNA secondary structure prediction with their program

CONTRAFold. CONTRAFold uses probabilistic parameters learned from a set of RNA secondary structures to predict base-pair probabilities and then predicts structures using the maximum expected accuracy approach. Subsequently, maximum expected accuracy structure prediction was applied to predicting the secondary structure common to multiple, aligned RNA sequences (Kiryu et al. 2007). The underlying pair probabilities in that approach were calculated using a composite score of free energy change and co-variation (Hofacker et al. 2002).

This study explores the use of maximum expected accuracy in single sequence secondary structure prediction, where thermodynamics are utilized to predict the underlying base-pair probabilities. A partition function calculation (McCaskill 1990; Mathews 2004) predicts base-pair probabilities with the current free energy change parameters at 37°C (Mathews et al. 2004). These probabilities are then utilized by a dynamic programming algorithm, implemented in the program MaxExpect, to assemble the structure with maximum expected accuracy as in CONTRAFold. The method is tested on a diverse database of different types of RNA, with the weighting factor, γ , varied to show the trade-off between sensitivity and PPV. For γ equal to 1, the average positive predictive value (PPV) of MaxExpect is 68%, with a sensitivity the same as the free energy minimization method, 73%. Therefore, the predicted maximum expected accuracy structure is of higher average accuracy than the minimum free energy structure.

In addition to the optimal structure, other possible or competing structures, called suboptimal structures, can also be predicted by MaxExpect. Suboptimal structures with expected pair accuracy lower than the optimal structure are predicted using a heuristic method similar to that developed previously for free energy minimization (Steger et al. 1984; Zuker 1989). The predicted suboptimal structures serve as alternative hypotheses for the structure. To our knowledge, this is the first implementation of maximum expected accuracy structure prediction that implements suboptimal structure prediction.

RESULTS

Maximization of expected accuracy (implemented in the program MaxExpect) provides a best estimate for an RNA secondary structure by maximizing expected accuracy of base pairs and single-stranded nucleotides (see details in Materials and Methods). The base-pair probabilities and single-strand probabilities are calculated from a partition function calculation (McCaskill 1990; Mathews 2004), which can be constrained with experimental data (Mathews 2004), such as chemical modification (Ehresmann et al. 1987) or enzymatic cleavage (Knapp 1989). Here it is shown that the optimal structure predicted from expected accuracy maximization provides better average accuracy than the structure predicted from free energy minimization,

although both methods are based on the same nearest-neighbor parameters for folding free energy change (Mathews et al. 2004).

RNA secondary structure prediction accuracy using MaxExpect

MaxExpect was tested on a diverse database of RNA sequences with known secondary structures (see Materials and Methods). The optimal structure, the maximum expected accuracy structure, is predicted and compared with the known structure in the database, and the accuracy of prediction is reported as sensitivity and PPV. The sensitivity versus PPV trade-off parameter, γ from Equation 1, is varied from 10^{-5} to 10^6 to find the optimal preference of double strandedness versus single strandedness. The plot of sensitivity as a function of PPV is shown in Figure 1. In addition to optimal structure prediction, a maximum of 750 suboptimal structures were predicted. The performance of the best suboptimal structure, i.e., the structure having the highest sensitivity, is also shown in Figure 1. This structure can only be determined by knowing the correct structure, but it represents the best hypothesis generated by MaxExpect.

A simple method related to maximizing expected accuracy is to assemble structures composed of base pairs with base-pairing probability larger than a threshold. The predicted structure is a valid secondary structure (each nucleotide has only one pairing partner) if the threshold is

0.5 or higher. This has been previously shown to predict structures with higher PPV than free energy minimization (Mathews 2004). Furthermore, for a threshold of exactly 0.5, the approach is similar to the approach of predicting the ensemble centroid using statistical sampling by SFold (Ding and Lawrence 2003; Ding et al. 2005). In fact, in the limit of an infinite sample size, the two approaches are equivalent. As shown in Figure 1 for a variety of threshold levels, the threshold method can produce structures with similar high PPV to those generated with a large γ in maximization of expected accuracy. It cannot, however, predict structures with as high sensitivity because many true base pairs having base-pair probabilities of 0.5 and below are missed.

The prediction performance of free energy minimization (Mathews et al. 2004), CONTRAfold 1.10, and CONTRAfold 2.02 (Do et al. 2006) are compared with MaxExpect in Figure 1. As compared to free energy minimization, MaxExpect offers an improved PPV at roughly the same sensitivity of structure prediction when $\gamma = 1$. CONTRAfold 2.02 performs similarly to MaxExpect and is a significant improvement over CONTRAfold 1.10. The folding parameters used by CONTRAfold 1 were learned from 151 representative RNA structures (Do et al. 2006) chosen from the Rfam database (Griffiths-Jones et al. 2003). For CONTRAfold 2, parameters were learned from a much larger data set, the S-Processed data set, with 3439 structures (Andronescu et al. 2007).

Table 1 summarizes the accuracies of MaxExpect, free energy minimization, and CONTRAfold 2.02 for each type of RNA in the database. The MaxExpect accuracy is reported for $\gamma = 1$, which is a point at which PPV is improved over free energy minimization. For CONTRAfold, accuracy is reported for $\gamma = 6$, which is the default value of γ (Do et al. 2006).

There is overlap in the S-Processed data set used for training CONTRAfold 2.02 and the testing data set used for this work. To examine the effect of this overlap, a cross-validation approach was taken for additional training and testing of CONTRAfold. For this cross-validation, all instances in the S-Processed data set for a specific type of RNA were removed from the parameter training, and then those parameters were applied to predict the structures of that type. For example, all the RNase P structures were removed when training a set of parameters used to predict RNase P structures for testing performance on RNase P. Interestingly, CONTRAfold performance was marginally improved,

FIGURE 1. Performance of different prediction methods: optimal structure prediction using free energy minimization, optimal structure prediction using MaxExpect, suboptimal structure prediction using MaxExpect, CONTRAfold 1.10, CONTRAfold 2.02, and structure assembling by varying base-pair probability threshold. For MaxExpect and CONTRAfold, sensitivity and PPV are calculated as a function of γ , which was varied between 10^{-5} and 10^6 , with even spacing on a logarithmic scale. For MaxExpect, the point at which $\gamma = 1$ is indicated. For CONTRAfold, the point at which $\gamma = 6.3$ is indicated, which is the closest point to the default γ of 6.0. Using the threshold method, sensitivity and PPV are calculated as a function of threshold, where the threshold was varied from 0.50 to 0.665. For free energy minimization, there is a single point on the graph because there is no parameter available to favor either sensitivity or PPV. The sensitivity and PPV are the average values for different types of RNA, where small and large subunit rRNA sequences are divided into domains of fewer than 700 nt. The performance of suboptimal structure prediction is calculated from the best predicted structure, which is defined as the structure with the highest sensitivity.

TABLE 1. Sensitivity and PPV of prediction methods

Type of RNA	MaxExpect ^a		Free energy minimization ^a		CONTRAFold ^b	
	Sensitivity (%) ^c	PPV (%) ^c	Sensitivity (%)	PPV (%)	Sensitivity (%)	PPV (%)
SSU rRNA ^d	62.1 ± 23.1 (47.2 ± 14.1)	58.0 ± 25.0 (42.7 ± 14.7)	61.4 ± 23.7 (45.5 ± 15.2)	54.8 ± 25.3 (38.3 ± 14.9)	63.5 ± 25.9 (51.2 ± 16.7)	58.0 ± 26.3 (46.5 ± 16.6)
LSU rRNA ^d	74.6 ± 11.9 (56.6 ± 14.7)	68.4 ± 11.6 (51.6 ± 14.2)	72.4 ± 17.2 (55.1 ± 11.5)	65.0 ± 16.3 (47.1 ± 12.2)	79.5 ± 11.7 (63.3 ± 15.0)	72.0 ± 10.0 (56.4 ± 13.8)
5S rRNA	72.5 ± 26.4	65.3 ± 23.6	72.9 ± 26.6	63.9 ± 23.8	81.2 ± 22.7	69.2 ± 19.0
Group I intron	71.2 ± 13.9	68.0 ± 15.1	70.2 ± 13.6	63.4 ± 13.5	71.5 ± 13.5	69.9 ± 10.5
Group II intron	87.0 ± 5.0	84.9 ± 9.4	88.1 ± 2.2	82.7 ± 6.8	75.1 ± 16.7	65.1 ± 25.5
RNase P	63.3 ± 15.2	62.7 ± 15.3	64.4 ± 12.8	61.8 ± 12.0	61.6 ± 6.6	60.0 ± 8.2
SRP	65.9 ± 26.3	51.3 ± 22.1	68.9 ± 25.4	52.9 ± 22.2	68.1 ± 23.5	52.1 ± 20.0
tRNA	85.8 ± 17.9	84.9 ± 19.6	85.6 ± 19.6	83.6 ± 22.2	84.8 ± 19.2	81.5 ± 19.3
Average ^e	72.8 ± 9.5	67.9 ± 11.9	73.0 ± 9.4	66.0 ± 11.4	73.1 ± 8.4	65.9 ± 9.3

^aThe secondary structures of RNA sequences are predicted using the nearest-neighbor free energy parameters at 37°C; $\gamma=1$ for MaxExpect.

^bCONTRAFold is version 2.02 with $\gamma=6.0$, the default value.

^cThe sensitivity and PPV are both calculated using the scoring method that allows slipped base pairs as explained in Materials and Methods. The values higher than the ones predicted from the free energy minimization method are bold.

^dThe small and large subunits of rRNA sequences are divided into domains of fewer than 700 nt. The predictions of the whole length are listed in parentheses, but not included in the average.

^eThe total average sensitivity and PPV are the average of sensitivities of different types of RNA.

so that sensitivity is higher at a given PPV with the cross-validation testing strategy (Supplemental Fig. 1). For example, with the non-cross-validated CONTRAFold, at $\gamma = 6.3$, the average sensitivity is 73.3% with a PPV of 65.7%. Using the cross-validated approach, at a sensitivity of 73.7%, the PPV is 67.2% using $\gamma = 3.2$. Supplemental Table 1 shows a comparison of accuracies for all RNA types at $\gamma = 6$. This improvement may be because more iterations of training were allowed during the cross-validation strategy than for the determination of the standard parameter set. The results support the hypothesis that the parameter estimation by CONTRAFold is robust and that CONTRAFold is likely to do as well predicting the structure of a novel RNA type as it is in predicting structures for known RNA types.

Time benchmarks

CONTRAFold 2.02 runs faster than MaxExpect and free energy minimization (Table 2). The calculation for maximizing expected accuracy using thermodynamic parameters has two parts: calculating the partition function, which requires most of the computation time, and then MaxExpect. The computation time and memory cost are shown in Table 2. The computation time complexity is $O(N^3)$ and memory complexity is $O(N^2)$ for MaxExpect. A significant factor in the longer time required by free energy minimization, and the partition function used here is the explicit calculation of helical coaxial stacking, which was previously demonstrated to take about two-thirds of the total calculation time and is absent from the CONTRAFold calculation

(Mathews 2004). Additionally, MaxExpect calculates suboptimal structures, and the method that facilitates this requires twice the calculation time of finding the optimal structure alone (see Materials and Methods).

Differences between minimum free energy structures and maximum expected accuracy structures

The distributions of base-pair probabilities in optimal structures are shown in Table 3. As expected, the average base-pair probability in the optimal structure predicted from MaxExpect is, on average, 0.042 higher than from free energy minimization when the weighting factor, γ , for MaxExpect is 1. The portion of predicted base pairs with pairing probability larger than 0.50 is 7.8% higher using MaxExpect than using free energy minimization, although the difference is only 0.7% between the portion of base pairs with pairing probability >0.99 . This shows that most highly probable pairs are already incorporated into structures predicted by free energy minimization. As has been demonstrated in a previous study on minimum free energy structures (Mathews 2004), base pairs with higher pairing probability provide more confidence of prediction accuracy than base pairs with lower pairing probability. Therefore, one of the reasons MaxExpect outperforms free energy minimization is because maximum expected accuracy structures include similar number of base pairs with high pairing probability but fewer base pairs with low pairing probability than minimum free energy structure. This fact is illustrated in Figure 2, where an example of improved

TABLE 2. Calculation size and time for different RNA sequence lengths

Length (nt)	RNA type	MaxExpect ^a				Free energy minimization		CONTRAFold 2.02 ^b	
		Partition function		Maximizing expected accuracy		Time (h:min:sec)	Memory (MB)	Time (h:min:sec)	Memory (MB)
		Time (h:min:sec)	Memory (MB)	Time (h:min:sec)	Memory (MB)				
77	<i>E. coli</i> arginine tRNA	<0:00:01	—	<0:00:01	—	0:00:01	—	<0:00:01	—
268	<i>Bacillus subtilis</i> SRP	0:00:04	—	<0:00:01	—	0:00:04	—	0:00:01	—
433	<i>Tetrahymena thermophila</i> IVS LSU group I intron	0:00:19	34	0:00:02	—	0:00:15	9	0:00:3	—
631	<i>Saccharomyces cerevisiae</i> A5 group II intron	0:01:36	49	0:00:14	39	0:00:50	11	0:00:11	16
1542	<i>E. coli</i> small subunit rRNA	0:21:20	189	0:01:52	179	0:16:15	38	0:02:15	96
2904	<i>E. coli</i> large subunit rRNA	2:06:49	612	0:16:59	602	1:25:08	114	0:16:17	242

Calculation size and time using one core on a machine with two dual-core Opteron 270 processors and 2 GB of RAM, using the gcc (version 4.1.1) compiler and the Fedora 5 OS. CPU time was determined using the time command. Memory use is estimated by using top and is therefore not reported for calculations with short CPU time use.

^aThe calculation used by MaxExpect has two parts, calculating partition function and maximizing expected accuracy. Each component was run separately. MaxExpect was run with $\gamma = 1$.

^bCONTRAFold 2.02 was run with $\gamma = 6$.

structure prediction accuracy by maximizing expectation is illustrated for a 5S rRNA. The structure predicted by maximizing expected accuracy has a sensitivity of 91.4% and a PPV of 86.5%. Free energy minimization, however, has a sensitivity of 54.3% and a PPV of 55.9%.

DISCUSSION

Free energy minimization provides the single best guess of the secondary structure for an RNA sequence because it is the single most probable structure in an ensemble (Mathews et al. 2004). There are drawbacks to this analysis, however. It implicitly assumes a single conformation for the RNA at equilibrium even though multiple states of secondary structure can exist in the solution for one RNA sequence. This analysis also assumes that the underlying nearest-neighbor parameters have the required resolution to pick the correct conformation from other conformations.

The base-pair probabilities calculated from a partition function provide the information about base pairs shared by multiple secondary structures. The base pairs with higher base-pair probability are more likely to be correctly predicted than base pairs with lower base-pairing probability (Mathews 2004). Furthermore, base-pair probabilities are less prone to change due to errors in the thermodynamic parameters than is the identity of the lowest free energy structure (Layton and Bundschuh 2005). Base-pair probabilities can therefore overcome uncertainties arising from errors in the parameters. Maximization of expected accuracy (MaxExpect) utilizes this feature to predict RNA secondary structure by maximizing the weighted probability of base-pairing and single-strandedness. This is why MaxExpect can predict structures with the same sensitivity

as free energy minimization and also with higher positive predictive value (PPV). The free energy parameters in the nearest-neighbor model, which are derived from experiments, have errors because of the limited number of experiments that have been conducted compared to the possible sequence space and because there are some non-nearest-neighbor effects that are neglected.

The free energy parameters used in this study are derived at 37°C even though the tested RNA sequences are derived from organisms living at different temperatures. A method for predicting free energy change at temperatures other than 37°C has been developed in a previous study (Lu et al. 2006). RNA secondary structure prediction by maximizing expected accuracy should be improved by making predictions of base-pair probabilities at the organism's optimal growth temperature from which the sequence was derived.

Maximizing the expected accuracy of structure prediction improves the accuracy of structure prediction for all types of noncoding RNAs tested with the exception of signal recognition particle RNA (Table 2). Base-pair probability independence is assumed when predicting the maximum expected accuracy structure and may be a drawback because helix formation is cooperative. Signal recognition particle RNAs have a predominantly helical region connecting the Alu and S domains (composed of the helices labeled 5) (Zwieb et al. 2005). It is possible that this feature is not being correctly predicted because the cooperativity of helix formation is not included in maximization of expected accuracy other than during the calculation of individual base-pair probabilities in the partition function. A revised measure of expected accuracy that utilizes conditional stack probabilities (Bompmunewerer et al. 2008) could perhaps further improve the accuracy of structure prediction.

TABLE 3. The distribution of base pair probabilities (P_{bp}) in maximum expected accuracy structures and minimum free energy structures

Type of RNA	Average P_{bp}	Percent of base pairs in predicted optimal structure above pairing probability threshold				
		$p_{bp} > 0.99$	$p_{bp} > 0.95$	$p_{bp} > 0.9$	$p_{bp} > 0.7$	$p_{bp} > 0.5$
Maximum expected accuracy structure						
SSU rRNA ^a	0.814 ± 0.079 (0.770 ± 0.058)	25.2 ± 16.7 (16.2 ± 9.3)	41.0 ± 20.0 (30.3 ± 11.5)	49.9 ± 19.8 (40.5 ± 11.6)	71.3 ± 14.5 (64.0 ± 11.0)	89.7 ± 8.6 (85.0 ± 8.0)
LSU rRNA ^a	0.824 ± 0.065 (0.776 ± 0.048)	26.8 ± 13.6 (21.7 ± 6.9)	44.3 ± 16.6 (35.0 ± 8.1)	52.0 ± 15.6 (44.8 ± 7.5)	73.8 ± 12.1 (63.6 ± 11.1)	90.1 ± 7.6 (84.5 ± 6.8)
5S rRNA	0.846 ± 0.087	28.1 ± 17.6	44.1 ± 20.9	54.4 ± 22.8	77.8 ± 18.9	94.2 ± 9.1
Group I intron	0.840 ± 0.060	23.5 ± 13.4	43.0 ± 16.6	53.6 ± 16.7	77.7 ± 9.9	92.6 ± 6.4
Group II intron	0.863 ± 0.051	19.4 ± 7.0	48.9 ± 13.7	60.3 ± 16.9	82.3 ± 8.1	95.2 ± 4.5
RNase P	0.788 ± 0.061	22.1 ± 6.4	31.4 ± 8.6	42.1 ± 7.6	66.8 ± 17.1	88.4 ± 8.5
SRP	0.827 ± 0.090	26.6 ± 17.2	43.3 ± 20.8	51.3 ± 22.6	75.0 ± 17.6	91.7 ± 9.4
tRNA	0.870 ± 0.092	32.7 ± 21.5	52.4 ± 24.8	61.9 ± 25.2	81.2 ± 20.1	95.7 ± 9.2
Average	0.834 ± 0.027	25.6 ± 4.0	43.6 ± 6.1	53.2 ± 6.2	75.7 ± 5.1	92.2 ± 2.7
Minimum free energy structure						
SSU rRNA ^a	0.745 ± 0.137 (0.655 ± 0.123)	24.4 ± 16.6 (15.2 ± 9.0)	39.3 ± 20.1 (28.2 ± 10.9)	47.3 ± 20.3 (37.3 ± 11.4)	65.0 ± 18.2 (55.2 ± 13.7)	78.1 ± 17.2 (68.1 ± 16.1)
LSU rRNA ^a	0.770 ± 0.150 (0.647 ± 0.080)	26.3 ± 13.8 (20.4 ± 6.8)	43.2 ± 17.1 (32.5 ± 7.9)	50.6 ± 16.4 (39.8 ± 7.7)	69.6 ± 17.9 (53.8 ± 11.0)	80.8 ± 18.4 (64.5 ± 11.4)
5S rRNA	0.813 ± 0.132	27.4 ± 17.5	43.1 ± 20.9	53.2 ± 23.0	75.0 ± 20.9	87.4 ± 17.6
Group I intron	0.801 ± 0.077	22.7 ± 13.3	41.2 ± 16.6	51.4 ± 17.1	73.6 ± 12.1	84.5 ± 9.9
Group II intron	0.830 ± 0.064	18.7 ± 6.6	47.0 ± 12.7	58.0 ± 15.8	79.3 ± 8.0	89.9 ± 5.5
RNase P	0.756 ± 0.072	21.6 ± 7.1	30.7 ± 9.3	41.1 ± 8.7	65.2 ± 18.3	81.8 ± 10.2
SRP	0.770 ± 0.159	26.0 ± 17.3	42.0 ± 21.0	49.7 ± 23.0	70.1 ± 21.8	82.1 ± 19.2
tRNA	0.851 ± 0.117	32.3 ± 21.5	51.7 ± 25.0	61.1 ± 25.5	79.5 ± 21.2	90.8 ± 15.4
Average	0.792 ± 0.038	24.9 ± 4.1	42.3 ± 6.1	51.5 ± 6.2	72.2 ± 5.7	84.4 ± 4.6

The maximum expected accuracy structure is predicted using MaxExpect, where $\gamma = 1$. The minimum free energy structure is predicted using free energy minimization. The average is the mean for types of RNA. The boldface numbers are averages.

^aThe small and large subunits of rRNA sequences are divided into domains of fewer than 700 nt. The predictions of the whole length sequence are listed in parentheses, but not included in the average.

It is particularly interesting that CONTRAfold and MaxExpect predict secondary structures with similar overall accuracy. The nearest-neighbor models utilized by both approaches are similar, but the parameter values were determined with different methodologies. A recent study further optimized the parameters for a similar nearest-neighbor model by fitting both the set of optical melting data and the set of RNA sequences with known structure, demonstrating the potential for further accuracy improvement (Andronescu et al. 2007). Further breakthroughs in prediction accuracy will require innovation in the underlying stability models by discovery of features that affect stability that are currently unknown. These innovations will derive from new computational analyses (Parisien and Major 2008) and continued experimental determination of motif stabilities (Chen et al. 2005; Blose et al. 2007; Clanton-Arrowood et al. 2008; Davis and Znosko 2008).

MATERIALS AND METHODS

Nearest-neighbor parameters

The nearest-neighbor parameters for Watson–Crick helices are those determined by Xia et al. (1998); the parameters for GU pairs

are by Mathews et al. (1999); and those for loops are by Mathews et al. (2004), with one exception. The multibranch loop parameter for the bonus for every additional helix in the multibranch loop was previously modified from -0.6 to -0.9 (Mathews et al. 2004) to be within the range of experimental error (Mathews and Turner 2002) and provide the highest prediction sensitivity. For this study, the experimental value of a multibranch loop parameter (Diamond et al. 2001; Mathews and Turner 2002; Mathews et al. 2004) was used for both free energy minimization and the partition function calculations so that no optimization has been performed on the nearest-neighbor parameters.

Maximizing expected accuracy

The base-pair probability of an i - j pair, $P_{bp}(i, j)$, is predicted from a partition function calculation (McCaskill 1990; Mathews 2004) based on a biophysical model, the nearest-neighbor model (Xia et al. 1998; Mathews et al. 1999, 2004; Lu et al. 2006). The probability of nucleotide i being single-stranded is:

$$P_{ss}(i) = 1 - \sum_j P_{bp}(i, j).$$

After base-pair probabilities are calculated with the partition function calculation, MaxExpect utilizes a dynamic programming algorithm to find the structure with maximum expected accuracy

FIGURE 2. Example of structures predicted from MaxExpect and from free energy minimization. (A) The known structure for the *Methanococcus thermolithotrophicus* 5S rRNA in the database of 5S rRNA (Szymanski et al. 1998). (B) The predicted maximum expected accuracy structure. (C) The predicted minimum free energy structure. The probabilities of predicted pairs are color-annotated by base-pairing probability. The maximum expected accuracy structure (B) includes a similar number of base pairs with high pairing probability, i.e., $\geq 99\%$, as the minimum free energy structure (C), but fewer number of base pairs with low pairing probability, i.e., $\leq 50\%$. In each panel, base pairs that are in the known structure are drawn with a heavy line. The figures were made using XRNA (<http://rna.ucsc.edu/rnacenter/xrna/xrna.html>) and Adobe Illustrator (Adobe Systems Incorporated).

using Equation 1. This dynamic programming algorithm uses Nussinov-style recursions (Nussinov and Jacobson 1980), expanded to predict suboptimal solutions using the approach of Zuker (1989) and Steger et al. (1984). For a simpler set of Nussinov-style recursions that find only optimal solutions, the reader is referred to the prior work of Do et al. (2006) or Kiryu et al. (2007). Note also that there is a connection between the work of maximizing expected accuracy and recent work in finding pseudoknot free structures from pseudoknotted structures that retain the maximum number of pairs (Ponty 2006; Smit et al. 2008). For that work, the input to the same recursions used here is a probability of 1 for each pair that occurs in the pseudoknotted structure. The output is then the structure with the maximum number of non-pseudo-knotted pairs.

For an RNA sequence having N nucleotides, the following recursions compute the maximum expected accuracy with arrays $W(i, j)$ and $V(i, j)$:

$$W(i, j) = \text{Max} \begin{cases} P_{ss}(i), & \text{if } i = j, \\ P_{ss}(i) + W(i+1, j), \\ P_{ss}(j) + W(i, j-1), \\ V(i, j), \\ W(i, k) + W(k+1, j), \text{ for each } k \text{ when } i \leq k < j; \end{cases}$$

$$V(i, j) = \begin{cases} 0 & \text{if } j - i + 1 < \text{minimum hairpin length,} \\ \gamma \cdot 2P_{bp}(i, j) + W(i+1, j-1) & \text{if } i \text{ and } j \text{ can form a canonical pair,} \\ -\infty & \text{if } i \text{ and } j \text{ cannot form a canonical pair,} \end{cases}$$

where i is $\leq j$. $W(i, j)$ is the maximum expected accuracy for a sequence fragment from nucleotides i to j , inclusive. $V(i, j)$ is the maximum expected accuracy for a sequence fragment from nucleotides i to j , inclusive, given that i and j are base-paired. $W(i, k) + W(k+1, j)$ is computed for every k between i and j ($j > k \geq i$) to identify multibranch loops. The maximum expected accuracy for the whole sequence between nucleotide at position 1 and nucleotide at position N is equal to $W(1, N)$.

To facilitate suboptimal structure prediction, $V'(i, j)$ and $W'(i, j)$ are additionally computed. $W'(i, j)$ is the maximum expected accuracy for a sequence fragment including nucleotides 1 to i and j to N . $V'(i, j)$ is the maximum expected accuracy for a sequence fragment including nucleotides 1 to i and j to N , with i paired to j . The calculations of $V'(i, j)$ and $W'(i, j)$ are similar to $V(i, j)$ and $W(i, j)$.

Structures are determined using a trace-back procedure. The maximum expected accuracy possible for a structure, S , containing a pair between i and j is calculated according to:

$$\text{Maximum expected accuracy } S(i, j) = V(i, j) + V'(i, j) - P_{bp}(i, j).$$

Therefore, base pairs that are contained in structures with relatively high expected accuracy can be identified, and structures can be determined starting from a given pair of nucleotides that are in a structure with high expected accuracy. To ensure that predicted structures are sufficiently different from each other, a window parameter is used. When structures are generated, an array tracks the positions of base pairs that have been predicted. For an i - j pair, a square region is marked from nucleotides i - window to i + window and from j - window to j + window. When a new structure is determined, it is only provided to the

user if it contains at least window base pairs that have not been found in previously marked regions. In the case where a window is specified by the user as zero, at least one novel base pair must be found. In this study, at most, 750 suboptimal structures are generated for each RNA sequence with a window size of 0.

Database of RNA secondary structures

Secondary structures determined by comparative analysis were used as known structures. The database includes various types of RNA as used in previous benchmarks (Mathews et al. 1999, 2004; Mathews 2004), including small-subunit (Gutell 1994) and large-subunit rRNA (Gutell et al. 1993; Schnare et al. 1996), 5S rRNA (Szymanski et al. 1998), Signal Recognition Particle RNA (Larsen et al. 1998), Group I introns (Waring and Davies 1984; Damberger and Gutell 1994), Group II introns (Michel et al. 1989), RNase P RNA (Brown 1998), and tRNA (Sprinzl et al. 1998). The small-subunit rRNA (16S) sequences are divided into domains as defined by Jaeger et al. (1989). The large-subunit rRNA (23S) sequences are divided into domains of fewer than 700 nucleotides each (Mathews et al. 1999). Moreover, chemically modified nucleotides in tRNA that cannot be accommodated in a canonical helix are forced single-stranded for all structure prediction methods, as has been done previously (Mathews et al. 1999).

Method of scoring prediction accuracy

Because of the uncertainty of base-pair matches in comparative analysis, a base pair is considered to be correctly predicted even if it is slipped by one nucleotide on one strand. A base pair between i and j is considered to be correct if any of the following base pairs are predicted: i and j , i and $j - 1$, i and $j + 1$, $i - 1$ and j , and $i + 1$ and j . Base pairs between $i - 1$ and $j + 1$ or between $i + 1$ and $j - 1$ are not considered to be correct. This scoring method also reflects the possibility of dynamic behavior of base-pairing. Sensitivity and PPV values calculated with this slippage scheme are generally 2%–3% higher than when calculated from the exact base-pairing scheme (Lu et al. 2006); a table of accuracies without allowing slippage is provided as Supplemental Table 2. Furthermore, although CONTRAfold also predicts noncanonical base pairs, they are not counted for scoring the prediction accuracy. When reporting average sensitivity and PPV for a type of the RNA (Table 1), the mean is reported over the sensitivity and PPV of each sequence.

Availability

The source code of the Linux version of MaxExpect method is downloadable from the Mathews lab website (<http://rna.urmc.rochester.edu/>). The algorithm is also embedded in the RNAstructure package for Microsoft Windows, which is available for download at the Mathews lab website.

SUPPLEMENTAL MATERIAL

Supplemental material can be found at <http://www.rnajournal.org>.

ACKNOWLEDGMENTS

We thank Ivo L. Hofacker for helpful discussions on the prediction of maximum expected accuracy structures and two

anonymous reviewers for helpful comments. This study was supported by the National Institutes of Health through grant R01GM076485 to D.H.M.

NOTE ADDED IN PROOF

During the review of this manuscript, a study examining the use of thermodynamics in the prediction of maximum expected accuracy structures was published by Hamada et al. (2009).

Received March 10, 2009; accepted June 14, 2009.

REFERENCES

- Andronescu M, Condon A, Hoos HH, Mathews DH, Murphy KP. 2007. Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics* **23**: i19–i28.
- Blose JM, Manni ML, Klapek KA, Stranger-Jones Y, Zyra AC, Sim V, Griffith CA, Long JD, Serra MJ. 2007. Non-nearest-neighbor dependence of the stability for RNA bulge loops based on the complete set of group I single-nucleotide bulge loops. *Biochemistry* **46**: 15123–15135.
- Bompmfunerwer AF, Backofen R, Bernhart SH, Hertel J, Hofacker IL, Stadler PF, Will S. 2008. Variations on RNA folding and alignment: Lessons from Benasque. *J Math Biol* **56**: 129–144.
- Brown JW. 1998. The ribonuclease P database. *Nucleic Acids Res* **26**: 351–352.
- Chen G, Znosko BM, Kennedy SD, Krugh TR, Turner DH. 2005. Solution structure of an RNA internal loop with three consecutive sheared GA pairs. *Biochemistry* **44**: 2845–2856.
- Clanton-Arrowood K, McGurk J, Schroeder SJ. 2008. 3' terminal nucleotides determine thermodynamic stabilities of mismatches at the ends of RNA helices. *Biochemistry* **47**: 13418–13427.
- Damberger SH, Gutell RR. 1994. A comparative database of group I intron structures. *Nucleic Acids Res* **22**: 3508–3510.
- Davis AR, Znosko BM. 2008. Thermodynamic characterization of naturally occurring RNA single mismatches with G-U nearest neighbors. *Biochemistry* **47**: 10178–10187.
- Diamond JM, Turner DH, Mathews DH. 2001. Thermodynamics of three-way multibranch loops in RNA. *Biochemistry* **40**: 6971–6981.
- Ding Y, Lawrence CE. 2003. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res* **31**: 7280–7301.
- Ding Y, Chan CY, Lawrence CE. 2005. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA* **11**: 1157–1166.
- Do CB, Woods DA, Batzoglou S. 2006. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **22**: e90–e98.
- Dowell RD, Eddy SR. 2004. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics* **5**: 71. doi: 10.1186/1471-2105-5-71.
- Eddy SR. 2002. Computational genomics of noncoding RNA genes. *Cell* **109**: 137–140.
- Ehresmann C, Baudin F, Mougel M, Romby P, Ebel J, Ehresmann B. 1987. Probing the structure of RNAs in solution. *Nucleic Acids Res* **15**: 9109–9128.
- Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. 2003. Rfam: An RNA family database. *Nucleic Acids Res* **31**: 439–441.
- Gutell RR. 1994. Collection of small subunit (16S- and 16S-like-) ribosomal RNA structures. *Nucleic Acids Res* **22**: 3502–3507.
- Gutell RR, Gray MW, Schnare MN. 1993. A compilation of large subunit (23S- and 23S-like) ribosomal RNA structures. *Nucleic Acids Res* **21**: 3055–3074.
- Hamada M, Kiryu H, Sato K, Mituyama T, Asai K. 2009. Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics* **25**: 465–473.

- Hofacker IL, Fekete M, Stadler PF. 2002. Secondary structure prediction for aligned RNA sequences. *JMB* **319**: 1059–1066.
- Jaeger JA, Turner DH, Zuker M. 1989. Improved predictions of secondary structures for RNA. *Proc Natl Acad Sci* **86**: 7706–7710.
- Kiryu H, Kin T, Asai K. 2007. Robust prediction of consensus secondary structures using averaged base pairing probability matrices. *Bioinformatics* **23**: 434–441.
- Knapp G. 1989. Enzymatic approaches to probing RNA secondary and tertiary structure. *Methods Enzymol* **180**: 192–212.
- Larsen N, Samuelsson T, Zwieb C. 1998. The signal recognition particle database (SRPDB). *Nucleic Acids Res* **26**: 177–178.
- Layton DM, Bundschuh R. 2005. A statistical analysis of RNA folding algorithms through thermodynamic parameter perturbation. *Nucleic Acids Res* **33**: 519–524.
- Lu ZJ, Turner DH, Mathews DH. 2006. A set of nearest-neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. *Nucleic Acids Res* **34**: 4912–4924.
- Mathews DH. 2004. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA* **10**: 1178–1190.
- Mathews DH, Turner DH. 2002. Experimentally derived nearest-neighbor parameters for the stability of RNA three- and four-way multibranch loops. *Biochemistry* **41**: 869–880.
- Mathews DH, Turner DH. 2006. Prediction of RNA secondary structure by free energy minimization. *Curr Opin Struct Biol* **16**: 270–278.
- Mathews DH, Sabina J, Zuker M, Turner DH. 1999. Expanded sequence dependence of thermodynamic parameters provides improved prediction of RNA Secondary Structure. *J Mol Biol* **288**: 911–940.
- Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci* **101**: 7287–7292.
- McCaskill JS. 1990. The equilibrium partition function and base pair probabilities for RNA secondary structure. *Biopolymers* **29**: 1105–1119.
- Michel F, Umeson K, Ozeki H. 1989. Comparative and functional anatomy of group II catalytic introns—a review. *Gene* **82**: 5–30.
- Nussinov R, Jacobson AB. 1980. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc Natl Acad Sci* **77**: 6309–6313.
- Parisien M, Major F. 2008. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **452**: 51–55.
- Ponty Y. 2006. *Modélisation de séquences génomiques structurées, génération aléatoire et applications*. Université Paris-Sud, Paris.
- Schnare MN, Damberger SH, Gray MW, Gutell RR. 1996. Comprehensive comparison of structural characteristics in eukaryotic cytoplasmic large subunit (23S-like) ribosomal RNA. *J Mol Biol* **256**: 701–719.
- Smit S, Rother K, Heringa J, Knight R. 2008. From knotted to nested RNA structures: A variety of computational methods for pseudoknot removal. *RNA* **14**: 410–416.
- Sprinzel M, Horn C, Brown M, Ioudovitch A, Steinberg S. 1998. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res* **26**: 148–153.
- Steger G, Hofmann H, Fortsch J, Gross HJ, Randles JW, Sanger HL, Riesner D. 1984. Conformational transitions in viroids and virusoids: Comparison of results from energy minimization algorithm and from experimental data. *J Biomol Struct Dyn* **2**: 543–571.
- Szymanski M, Specht T, Barciszewska MZ, Barciszewski J, Erdmann VA. 1998. 5S rRNA data bank. *Nucleic Acids Res* **26**: 156–159.
- Torarinsson E, Yao Z, Wiklund ED, Bramsen JB, Hansen C, Kjems J, Tommerup N, Ruzzo WL, Gorodkin J. 2008. Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions. *Genome Res* **18**: 242–251.
- Uzilev AV, Keegan JM, Mathews DH. 2006. Detection of noncoding RNAs on the basis of predicted secondary structure formation free energy change. *BMC Bioinformatics* **7**: 173. doi: 10.1186/1471-2105-7-173.
- Waring RB, Davies RW. 1984. Assessment of a model for intron RNA secondary structure relevant to RNA self-splicing—a review. *Gene* **28**: 277–291.
- Washietl S, Hofacker IL, Stadler PF. 2005. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci* **102**: 2454–2459.
- Xia T, SantaLucia J Jr, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH. 1998. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick pairs. *Biochemistry* **37**: 14719–14735.
- Zuker M. 1989. On finding all suboptimal foldings of an RNA molecule. *Science* **244**: 48–52.
- Zwieb C, van Nues RW, Rosenblad MA, Brown JD, Samuelsson T. 2005. A nomenclature for all signal recognition particle RNAs. *RNA* **11**: 7–13.

RNA

A PUBLICATION OF THE RNA SOCIETY

Improved RNA secondary structure prediction by maximizing expected pair accuracy

Zhi John Lu, Jason W. Gloor and David H. Mathews

RNA 2009 15: 1805-1813 originally published online August 24, 2009

Access the most recent version at doi:[10.1261/rna.1643609](https://doi.org/10.1261/rna.1643609)

Supplemental Material

<http://rnajournal.cshlp.org/content/suppl/2009/08/11/rna.1643609.DC1>

References

This article cites 47 articles, 10 of which can be accessed free at:
<http://rnajournal.cshlp.org/content/15/10/1805.full.html#ref-list-1>

License

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

To subscribe to *RNA* go to:

<http://rnajournal.cshlp.org/subscriptions>