Bioinformatics, 32(7), 2016, 1033–1039
doi: 10.1093/bioinformatics/btv682
Advance Access Publication Date: 20 November 2015
Original Paper

Structural bioinformatics

AccessFold: predicting RNA–RNA interactions with consideration for competing self-structure

Laura DiChiacchio[1], Michael F. Sloma[1] and David H. Mathews[1,2,*]

[1]Department of Biochemistry and Biophysics and Center for RNA Biology and [2]Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY 14642, USA

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on September 2, 2014; revised on October 13, 2015; accepted on November 15, 2015

Abstract

Motivation: There are numerous examples of RNA–RNA complexes, including microRNA–mRNA and small RNA–mRNA duplexes for regulation of translation, guide RNA interactions with target RNA for post-transcriptional modification and small nuclear RNA duplexes for splicing. Predicting the base pairs formed between two interacting sequences remains difficult, at least in part because of the competition between unimolecular and bimolecular structure.

Results: Two algorithms were developed for improved prediction of bimolecular RNA structure that consider the competition between self-structure and bimolecular structure. These algorithms utilize two novel approaches to evaluate accessibility: free energy density minimization and pseudo-energy minimization. Free energy density minimization minimizes the folding free energy change per nucleotide involved in an intermolecular secondary structure. Pseudo-energy minimization (called AccessFold) minimizes the sum of free energy change and a pseudo-free energy penalty for bimolecular pairing of nucleotides that are unlikely to be accessible for bimolecular structure. The pseudo-free energy, derived from unimolecular pairing probabilities, is applied per nucleotide in bimolecular pairs, and this approach is able to predict binding sites that are split by unimolecular structures. A benchmark set of 17 bimolecular RNA structures was assembled to assess structure prediction. Pseudo-energy minimization provides a statistically significant improvement in sensitivity over the method that was found in a benchmark to be the most accurate previously available method, with an improvement from 36.8% to 57.8% in mean sensitivity for base pair prediction.

Availability and implementation: Pseudo-energy minimization is available for download as AccessFold, under an open-source license and as part of the RNAstructure package, at: http://rna.urmc.rochester.edu/RNAstructure.html.

Contact: david_mathews@urmc.rochester.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

RNA has been increasingly recognized as an active player in cellular biology. RNA molecules catalyze many reactions, are involved in transcript splicing and editing and serve as signals for protein localization (Doudna and Cech, 2002; Gesteland *et al.*, 2005; Walter and Blobel, 1982). The function of an RNA molecule is intimately related to its structure, and therefore the prediction of RNA structure is an important tool for understanding mechanism. There are a number of computational methods for predicting RNA secondary structure for a single sequence (Nussinov and Jacobson, 1980; Seetin and Mathews, 2012).

Functional RNA–RNA complexes are abundant in nature and are also used in multiple techniques for characterizing gene function *in vitro* and *in vivo*. MicroRNAs (miRNAs), small interfering RNAs (siRNAs) and small RNAs (sRNAs) are short RNA sequences that bind mRNA transcripts to regulate gene expression; guide RNAs (gRNAs) bind mRNA targets to direct post-transcriptional modifications and many spliceosomal RNAs form functional duplexes for splicing of introns from mRNA (Birnstiel, 1988; Houseley and Tollervey, 2009; Wu and Belasco, 2008).

Predicting RNA binding sites, especially for longer RNAs such as mRNA transcripts, is an important step in understanding the function of existing and newly discovered non-coding RNAs (Alkan *et al.*, 2006a; Pervouchine, 2004). It is also a tool for designing siRNAs used to study gene function or as therapeutics (Heale *et al.*, 2005; Lu and Mathews, 2007; Tafer *et al.*, 2008).

Computational prediction of RNA–RNA binding sites is a difficult problem because intermolecular structure competes with the formation of self-structure in a concentration-dependent manner. Algorithms for prediction of intermolecular structure generally build upon single-stranded folding algorithms and come in two main types: free energy minimization and partition function calculations. These utilize dynamic programming algorithms to compute the thermodynamics of folding using parameters fit to experiments, most popularly the Turner nearest neighbor parameters (Mathews *et al.*, 2004, 1999a; Xia *et al.*, 1998). Free energy minimization schemes compute the structure with the lowest folding free energy and are guaranteed to output the optimal structure given the input parameters and energy model. Partition function calculations consider the entire ensemble of possible structures and are used to determine probabilities for formation of each structure. In this way, a partition function calculation gives the probability that each possible base pair occurs given its prevalence within the ensemble (Mathews, 2004; McCaskill, 1990). These algorithms are successful at single-stranded structure prediction but perform with lower accuracy on bimolecular structures.

To get around these challenges, there has been a focus on the development of algorithms that predict only specific types of RNA–RNA interactions. An example of a well-characterized biological phenomenon for which there are multiple algorithms that perform with relative success is the hybridization of miRNAs and their mRNA targets. These are endogenous, 19–25 nt strands of RNA that base pair with a target mRNA, most often at its 3′-UTR and most often to repress translation via the recruitment of RNA-induced silencing complex, RISC (Chaudhuri and Chatterjee, 2007). Computational methods for the prediction of miRNA–mRNA targets rely on the well-characterized binding structures that have been previously determined: the miRNA must be complementary to the 3′-UTR of its target mRNA; must have a strong ‘seed’ binding site at the 5′-end of the interaction but not the 3′-end; must form a thermodynamically stable duplex; the sequence of the target mRNA must be conserved in the genomes of related species and cannot be impeded by strong secondary structure in the target mRNA (Chaudhuri and Chatterjee, 2007; Watanabe *et al.*, 2007). It is clear from this example that these algorithms are designed to test miRNA-target structure specifically. The principles of targeting the 3′-UTR and having a more thermodynamically stable 5′-end of the duplex do not apply to general prediction of bimolecular structure and would lead to false positives and negatives when applied to other types of binding problems.

It is necessary to develop a generalized algorithm for the reliable prediction of intermolecular binding sites of unknown structure class. This article introduces three algorithms for predicting RNA–RNA interactions. Each of these methods predicts only bimolecular helices, i.e. base pairs composed of one nucleotide from each of the two input sequences, where the helices are interrupted by unpaired nucleotides in internal or bulge loops. The first algorithm, DensityMin, computes the minimum free energy density structure or the structure with the lowest ratio of folding free energy change to number of nucleotides in the bimolecular structure. It was hypothesized that minimizing density would lead to shorter, more stable helices between strands that would be able to outcompete the potential intramolecular pairing partners. The second algorithm, AccessFold, uses a heuristic to utilize information from unimolecular partition function calculations to consider accessibility of nucleotides for pairing. In this approach, an energy, which is the sum of folding free energy change and a pseudo-energy that accounts for accessibility, is minimized. The pseudo-energy scales with the probability that a single nucleotide is involved in an intramolecular base pair as determined by the single-stranded partition function calculation. The pseudo-energy term serves to account for the propensity to form self-structure, excluding inaccessible nucleotides from base pairing in the bimolecular structure. This pseudo-energy penalty model for accounting for accessibility was also incorporated into the free energy density minimization algorithm described above to test any synergistic improvement. This algorithm is referred to as E-DensityMin, i.e. Extended-DensityMin.

A set of 17 systems with known bimolecular structures was assembled for benchmarking, and the performance of pseudo-energy minimization was compared to multiple variations of free energy minimization, partition function calculation and combination algorithms designed to predict RNA–RNA interactions: BiFold, DuplexFold, Bipartition and MaxExpect from RNAstructure (Lu *et al.*, 2009; Mathews, 2004; Mathews *et al.*, 1999b); RNAplex and RNAup from the Vienna package (Mückstein *et al.*, 2006; Tafer and Hofacker, 2008) and BINDIGO from the Aalberts lab (Hodas and Aalberts, 2004). BiFold, DuplexFold and RNAplex are variations of free energy minimization algorithms. BiFold predicts the minimum free energy structure including self-structure and treats unimolecular and bimolecular pairs equally (Mathews *et al.*, 1999b). DuplexFold, RNAplex and Bindigo do not allow self-pairing to occur. RNAplex and Bindigo are faster because they use modified functions for computing loop energies (Hodas and Aalberts, 2004; Tafer and Hofacker, 2008). Bipartition and MaxExpect are full bimolecular partition function calculations that do not consider intramolecular base pairs (Lu *et al.*, 2009; Mathews, 2004). Bipartition assembles structures by selecting only those base pairs that meet a minimum pairing probability threshold as determined by the user. MaxExpect computes the identical partition function, but the predicted structure is the structure with maximum expected accuracy (Lu *et al.*, 2009). RNAup uses a partition function calculation to determine the accessibility of an interval of nucleotides in each strand and then computes the minimum free energy structure involving those intervals that are accessible (Mückstein *et al.*, 2006) and is used as the current gold standard for using accessibility information. In all cases except BiFold, these are programs that predict the structures of interacting RNA sequences but only predict the bimolecular pairs, i.e. those pairs that involve one nucleotide from each strand. For BiFold, both unimolecular and bimolecular pairs can be predicted, but the class of structures is restricted to those that are not equivalent to pseudoknots, e.g. kissing loop interactions are not considered (Mathews *et al.*, 1999b).

2 Methods

2.1 Overview

Given two sequences of nucleotides, A, C, G and U, the overall goal is to predict the collection of canonical base pairs AU, GC, GU that occur between them. This is achieved using a nearest neighbor model in which the folding free energy change parameters were fit to experimentally determined stabilities for stacking base pairs, internal loops, bulge loops and multibranch loops. These algorithms use the 2004 Turner nearest neighbor parameters (Mathews *et al.*, 2004). The only exception is that the per helix term in multibranch loops is set at -0.6 kcal/mol, a value more consistent with experiments than the value found by optimizing structure prediction accuracy (Lu *et al.*, 2009; Mathews and Turner, 2002).

2.2 Free energy density minimization

In the free energy density minimization algorithm, DensityMin, the optimal structure is computed by finding the intermolecular structure with the lowest folding free energy change per nucleotide in the structure. This is implemented using a dynamic programming algorithm, and the recursions are provided in the Supplementary Materials. The number of nucleotides in the structure is often larger than twice the sum of base pairs, as it includes internal loops and bulges that intervene intermolecular pairs. This adds a third dimension to the usual dynamic programming algorithm storage tables, in which the length of the interaction must be tracked in addition to the folding energies for each subsequence. This method was first described for single-stranded folding (Alkan *et al.*, 2006b), and it scales better for bimolecular pair prediction (in the absence of unimolecular structure as implemented here) because there are no structure branches to consider. The scaling is $O(NMSL)$ where N is the length of sequence 1, M is the length of sequence 2, S is the maximum loop size and L is the maximum number of nucleotides allowed in bimolecular structure. DensityMin was hypothesized to be more amenable to the RNA–RNA interaction problem by modeling the competition between self-structure and bimolecular structure formation.

There are several user-defined parameters utilized in DensityMin (and E-DensityMin). The minimum and maximum number of nucleotides allowed in the bimolecular structure are user-specified. This length includes internal loops, bulges and dangling ends in both strands. The performance reported was chosen as the best among multiple combinations of minimum and maximum lengths: 2 or 50 nucleotides as minimum length with a maximum length of 100 or a minimum length of 100 with maximum length of 200. An additional user-defined parameter specifies the maximum number of unpaired nucleotides in an internal loop or bulge loop. This is set to the standard default of 30 in each case. Results for DensityMin are reported here using minimum and maximum lengths of interaction of 50 nts and 200 nts, respectively, with a maximum number of unpaired nucleotides of 30.

2.3 Pseudo-energy minimization

In pseudo-free energy minimization, an accessibility term for each sequence is incorporated into free energy minimization. The partition function calculation for each sequence is first computed independently. The probability that each nucleotide i is unpaired, P_i, is then computed by subtracting the sum of probability that i is base paired with any possible pairing partner. This probability is then converted into an energetic penalty using:

$$\Psi_i = -\gamma RT \ln P_i$$

The pseudo-energy penalty is Ψ_i for nucleotide i, where γ is a scaling factor, R is the Boltzmann constant and T is the absolute temperature. In predicting the bimolecular base pairs, Ψ_i is added to the folding free energy for each nucleotide, i, involved in a base pair. This accounts for accessibility but treats each nucleotide as though its ensemble pairing probability were independent of other nucleotides. Although this is a simplification, it has been used successfully in methods that predict structures conserved in two or more sequences (Harmanci *et al.*, 2008, 2011; Hofacker *et al.*, 2004; Will *et al.*, 2007) and in single-sequence structure prediction (Lu *et al.*, 2009; Tabei *et al.*, 2008). The scaling factor, γ, was optimized using the dataset of experimentally determined control structures. The total pseudo-energy, denoted by $\Psi\Delta G^\circ$, is the sum of the folding free energy change calculated using nearest neighbor parameters (Mathews *et al.*, 2004; Xia *et al.*, 1998) and Ψ_i for each nucleotide involved in a bimolecular base pair.

$\Psi\Delta G^\circ$ is minimized using a standard free energy minimization scheme with a dynamic programming algorithm where only bimolecular pairs involving a nucleotide from each of the two sequences are included. Loops are limited in size, a default of 30 unpaired nucleotides, to reduce the scaling of the algorithm to $O(NMS)$, where N is the length of sequence 1, M is the length of sequence 2 and S is the maximum number of unpaired nucleotides in loops. Note that this scaling is for AccessFold only, and this requires a prior set of partition function calculations, which scale $O(N^3+M^3)$. The recursions are provided in the Supplementary Materials.

2.4 E-DensityMin

Performance for E-DensityMin is reported using a minimum and maximum length of interaction of 100 nts and 200 nts, respectively, and a maximum number of unpaired nucleotides in an internal or bulge loop of 30.

2.5 Structure prediction benchmarks

The algorithms were scored using a dataset of 17 known structures assembled for this work, including 4 two-piece tmRNAs (Gaudin *et al.*, 2002; Keiler *et al.*, 2000; Sharkady and Williams, 2004), four archaeal split pre-tRNAs (Chan *et al.*, 2011), three human miRNA–mRNA duplexes (Papadopoulos *et al.*, 2009), four bacterial sRNA–mRNA duplexes (Cao *et al.*, 2010), one human small nuclear RNA duplex (Birnstiel, 1988) and one gRNA–mRNA duplex from *Trypanosoma bruceii* (Yu and Koslowsky, 2006). The inclusion criteria for known structures was stringent, where only structures determined by chemical mapping, comparative sequence analysis, mutational analyses with compensating mutations or a combination of these techniques were included. The split tRNAs and tmRNAs were determined by a combination of comparative sequence analysis and chemical mapping. The gRNA–mRNA duplex was determined by *in vitro* mutational analyses. The human snRNA duplex was determined by comparative sequence analysis. The miRNA- and sRNA-target duplexes came from mirTarbase and sRNATarbase, respectively, with the filter 'Luciferase assay – mutant MRE – mutant miRNA' for miRNAs and either 'Footprint Analysis', 'Point Mutation in sRNA' or 'Point Mutation in mRNA' for sRNAs with further manual filtering for duplexes demonstrating recovery of activity with the mutated sRNA/mRNA by a compensating change in its pairing partner. The set of sequences and structures is provided as Supplementary Information.

Sensitivity and positive predictive value (PPV) were used to measure success. Sensitivity is the fraction of known pairs correctly predicted. PPV is the fraction of predicted pairs in the known structure. A base pair is considered correct if it slipped by a single nucleotide on one strand (Mathews *et al.*, 1999a). For example, if positions i and j are paired in the known structure, a predicted pair between i and j, $i-1$ and j or i and $j+1$ are considered correctly predicted. A pair between $i-1$ and $j+1$ would not be counted as a true positive pair. False-negative pairs are those pairs that are present in the known structure but not predicted, and false-positive pairs are those that are predicted but are not in the known structure.

Finally, to determine statistical significance in differences between algorithms, a two-tailed, paired *t*-test was used to compare performance pairwise by algorithm using both sensitivity and PPV (Xu *et al.*, 2011). α was set to 0.05. Pairing was done on RNA families, e.g. pairing mean performance of tmRNAs.

2.6 Optimizing γ

Because of the small number of test sequences, the delete-1 jackknife method was used to optimize γ for E-DensityMin and AccessFold independently. For each of the jackknife calculations, a set of calculations were run with γ from 0 to 1, stepping in increments of 0.1. Sensitivity was used as a measure for optimizing accuracy, but it was found that optimizing by PPV produces identical results. For AccessFold, $\gamma = 0.4$ produces optimal result for every test case. E-DensityMin performed with an optimal sensitivity with $\gamma = 0.8$ and performed with near-optimal PPV at this setting as well. Results are reported for each program with these optimal values for γ, respectively.

2.7 Benchmarks of other software

The Vienna Package version 2.1.9 was used to run RNAup and RNAplex. RNAup was run with consideration for structure in both strands rather than the default of target-strand only accessibility using the flag –b. RNAup scores are reported for default parameters for interaction length, where the maximal length of unstructured nucleotides is 4 and the maximal length of interaction is 25. Additional calculations allowing for larger unstructured regions up to 30 nts as well as longer regions of interaction up to 200 nts did not improve average performance over the defaults. RNAplex was run using default parameters of maximal length for unpaired loops (40 nts) and energy penalty for extending duplexes (0 kcal/mol). Performance improved with the application of a 0.3 kcal/mol/nt penalty for extending duplexes using the flag –c 30 (Supplementary Materials). The 2012 version of BINDIGO was run using default parameters. BiFold, Maximum Expected Accuracy and Bipartition were run from RNAstructure version 5.8 (Reuter and Mathews, 2010). BiFold, MaxExpect and Bipartition were run with default parameters, and the optimal Bipartition structure reported here is assembled using a threshold of 50% probability of base pairing.

2.8 Benchmarks of time

Time benchmarks were run on a single core of a compute node with two 12-core Intel Xeon E5-2695 processors and 60 GB of RAM. The ‘time’ command was used to time the program and the real time is reported.

3 Results

3.1 Accuracy

Performance of each algorithm was assessed using a set of 17 known bimolecular structures including miRNA–mRNA duplexes, sRNA–mRNA duplexes, one gRNA–mRNA duplex, one snRNA duplex and split tmRNA and tRNA structures. Tables 1 and 2 summarize the results by RNA family in average sensitivity and PPV, respectively. AccessFold has the second-highest average sensitivity, after Bifold and the second-highest average PPV, after RNAup. Tables 1 and 2 list the performance of RNAplex using its default parameters. Other tests were performed including accessibility in the RNAplex calculation. The overall performance was similar, and those benchmark results are provided in the Supplementary Materials.

Statistical analysis using a paired *t*-test shows that the increase in sensitivity by AccessFold over RNAplex and RNAup is statistically significant (Supplementary Materials). Bifold also performed well, providing a statistically significant increase in sensitivity over every other program tested except for AccessFold. The improved PPV of AccessFold over RNAplex is also statistically significant. RNAup had the highest mean PPV of the programs tested, and the difference in PPV between RNAup and every other program except for AccessFold and densitymin is statistically significant. There is

Table 1. Performance by sensitivity

RNA Family	BINDIGO (%)	RNAplex (%)	DuplexFold (%)	Bifold (%)	DensityMin (%)	Bipartition (%)	MaxExpect (%)	RNAup (%)	E-Density (%)	AccessFold (%)
miRNA	0	0	0	17.6	0	0	0	0	0	0
gRNA	52.0	0	52.0	100.0	48.0	44.0	48.0	48.0	0	52.0
snRNA	0	0	0	66.7	66.7	0	0	47.6	47.6	66.7
tmRNA	7.4	0	0	28.2	0	11.6	18.3	20.2	13.7	47.7
tRNA	86.7	63.5	86.7	86.7	61.1	86.7	86.7	36.0	56.7	82.9
sRNA	0	0	0	91.0	23.1	0	0	69.2	25.0	97.5
Mean[a]	24.4	10.6	23.1	65.1	33.1	23.7	25.5	36.8	19.7	57.8

[a] Mean is the arithmetic average of scores over families of RNA.

Table 2. Performance by PPV

RNA Family	BINDIGO (%)	RNAplex (%)	DuplexFold (%)	Bifold (%)	DensityMin (%)	Bipartition (%)	MaxExpect (%)	RNAup (%)	E-Density (%)	AccessFold (%)
miRNA	0	0	0	0.2	0	0	0	0	0	0
gRNA	32.5	0	30.2	51.0	66.7	29.7	30.8	100	0	39.4
snRNA	0	0	0	19.4	73.7	0	0	100	55.6	60.9
tmRNA	7.1	0	0.0	9.1	0	7.5	9.6	75.0	14.3	39.6
tRNA	62.1	62.1	59.9	62.8	81.3	62.1	62.1	68.5	66.8	72.0
sRNA	0	0	0	6.0	14.3	0	0	75.0	0	79.5
Mean[a]	17.0	10.4	15.0	24.7	39.3	16.5	17.1	69.7	22.8	48.6

[a] Mean is the arithmetic average of scores over families of RNA.

insufficient power in a benchmark of this size to conclude that the difference in PPV between AccessFold and RNAup is significant.

These results demonstrate that pseudo-energy minimization provides a statistically significant increase in sensitivity when compared to the best known algorithm for including self-structure, RNAup. The average sensitivity for RNAup on this dataset was 36.8% and improved to 57.8% using pseudo-energy minimization in AccessFold. The average PPV for RNAup was 69.7% and for AccessFold was 48.6%, but this difference was not statistically significant. Minimizing free energy density (DensityMin) provides an increase in PPV over standard free energy minimization (BiFold) and bimolecular partition function calculations (Bipartition) but lowers the sensitivity. Incorporating the pseudo-energy penalty into free energy density minimization, using E-DensityMin, decreased the performance in both sensitivity and PPV. The free energy density approach does not perform as well as AccessFold in either sensitivity or PPV.

3.2 Time

Computational time was benchmarked for each new algorithm and RNAup. Table 3 lists the times for representative calculations.

With shorter sequence lengths, the computation time is comparable; however, with sequence lengths greater than 500 nucleotides, AccessFold is slower than RNAup. This is due to the computation of a full partition function for each sequence with AccessFold as compared to partition function calculations on shorter, overlapping intervals in RNAup.

3.3 Example

Figure 1 shows a sample structure prediction, the betaproteobacterial tmRNA. This structure illustrates a generalization of the structures that can be predicted by AccessFold as compared to RNAup, which requires a continuously accessible interval for binding. In this example, an intervening hairpin stem-loop divides the accessible region. AccessFold does not limit the prediction to a single highly accessible interval but is still able to perform comparatively well on cases that truly do represent a single contiguous binding site. Rather than rigidly setting a binding site and only considering base pairs in this location, AccessFold implicitly considers the propensity for the hairpin stem to form and so does not consider it a likely site for intermolecular base pairs to form. The presence of this highly penalized intervening sequence, however, does not preclude the identification of accessible pairing sites adjacent to the hairpin. These intervening hairpins are found in the split tmRNA and split tRNA sequences in the benchmark dataset.

Table 3. Computation time

Sequence + length	RNAup	AccessFold	E-DensityMin	Bipartition
tRNA: 48 + 65 nts	0.1 s	0.1 s	1.7 s	0.1 s
gRNA: 98 + 59 nts	0.3 s	0.2 s	4.0 s	0.4 s
snRNA: 130 + 125 nts	0.4 s	0.4 s	16.8 s	1.2 s
tmRNA: 207 + 78 nts	0.8 s	0.9 s	19.2 s	1.6 s
sRNA: 87 + 414 nts	1 m 15.7 s	5.3 s	45.7 s	7.5 s
miRNA: 5397 + 22 nts	14 m 12 s	257 m 10 s	545 m 28 s	172 m 32 s

AccessFold is able to consider the generalization of intervening structures between binding sites. The accessible sites could, e.g. be bulge or internal loops interrupted by intramolecular helices or could be hairpin loops as in kissing hairpin loop complexes. One limitation, however, is that the accessible regions must be closer than the maximum internal or bulge loop size (default of 30 nucleotides). For example, the human small nuclear RNA duplex (Birnstiel, 1988) has a split binding site, but the two sites are separated by over thirty nucleotides, so only one of the two binding regions can be predicted by AccessFold.

4 Discussion

A major barrier to the development of general algorithms addressing the RNA–RNA interaction prediction problem is a limited number of experimentally determined bimolecular structures. In many experimental applications, it is only important to know that the interfering RNA binds the target, as determined by expression profiles. In developing a general algorithm for prediction of bimolecular structure; however, it is important to use only those structures whose exact base pairs are determined as a test set to measure success. In this work, a general algorithm for bimolecular structure prediction was developed and benchmarked using a small set of known structures. Accuracy could be improved when a larger set of known structures is available for testing. Another issue is that the small set of types of RNA–RNA interactions provides limited power to show that differences in program performances are statistically significant. A larger set of known structures would additionally improve this.

Although algorithms addressing specific biological problems have been successful, particularly in predicting miRNA binding sites, it is necessary to develop a general algorithm for newly discovered ncRNAs. An algorithm for predicting RNA–RNA interactions that

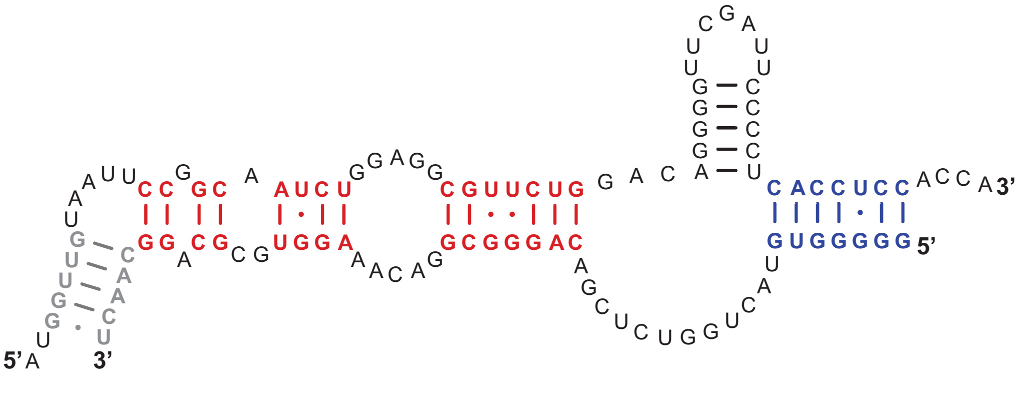

Fig. 1. tmRNA structure. The accepted structure in the region of interaction of the split tmRNA from *D. aromatica.* The base pairs colored in red were correctly predicted by AccessFold, while the base pairs colored in blue were correctly predicted for both AccessFold and RNAup. The base pairs in gray were not correctly identified by either algorithm. Note that the hairpin stem loop in the top strand is not able to be predicted by either program because it involves pairs from one strand only. This hairpin stem loop interrupts the accessible region, resulting in two regions in the top strand that are accessible to binding

uses information from bimolecular free energy minimization as well as unimolecular partition function calculations was developed here and called AccessFold. In this approach, a pseudo-energy term is incorporated into the free energy minimization algorithm. This term scales with the probability that a single nucleotide is involved in an intramolecular base pair as determined by the single-stranded partition function calculation. The pseudo-energy term serves to penalize the propensity to form self-structure, excluding inaccessible nucleotides from base pairing in the bimolecular structure. A set of 17 known bimolecular structures was compiled for testing, and the performance of pseudo-energy minimization was compared to multiple variations of free energy minimization, partition function calculation and combination algorithms designed to predict RNA–RNA interactions. Pseudo-energy minimization provides a statistically significant increase in sensitivity when compared to the prior best available algorithm for considering self-structure, RNAup. This result is interesting because RNAup rigorously calculates target structure accessibility, whereas AccessFold uses a heuristic that assumes the probability of forming a pair is independent of the probability the neighboring nucleotide is unpaired. Figure 1 shows that this heuristic can be advantageous in situations where a binding site is discontinuous.

One aspect of bimolecular structure prediction that is inadequately addressed is the concentration dependence of interaction. The equilibrium for formation of bimolecular pairs is concentration dependent, but the equilibrium for unimolecular pairs is independent of concentration. It is possible that the scaling factor, γ, is contributing to the improved performance by AccessFold by independently weighting the unimolecular and bimolecular components of structure. Other programs treat intra- and inter-molecular pairs equally, which does not model interactions appropriately. On the other hand, the scaling factor in AccessFold is also not rigorously accounting for concentration because a biophysical model would penalize the interaction once per bimolecular complex, not per base pair, and also because the concentrations of strands vary for RNA types. One possible avenue for further improving the accuracy of bimolecular structure prediction methods would be to explicitly account for effect of concentration on the interaction between strands. The scaling factor, γ, also accounts for the fact that the pseudo-free energy is treating the unpairing probabilities of each nucleotide as independent. This means that the pseudo-free energy would tend to over-penalize the opening of regions for binding.

The poorer performance of DensityMin and E-DensityMin, programs that minimize the density of folding free energy change, as compared to AccessFold indicates that predicting compact bimolecular structures is not sufficient to model the competition provided by unimolecular structure. The results are provided here as an alternative method for comparison.

Interestingly, for miRNA–mRNA interactions, all of the free energy minimization approaches failed to identify the experimentally verified target locations, with the exception of one interaction identified by Bifold (Table 1). This suggests that when there is specific knowledge about interactions, such as the requirement for complementarity in the seed region that is often used in miRNA target prediction, tailored applications should to be used. These results also support the prevailing consensus that miRNA target selection is kinetically controlled.

Funding

This work was supported by National Institutes of Health (NIH) grant R01 GM076485 to D.H.M.; NIH grant T32 GM068411 to L.D. and NIH grant T32 GM07356 to L.D.

Conflict of Interest: none declared.

References

Alkan,C. *et al.* (2006a) RNA-RNA interaction prediction and antisense RNA target search. *J. Comput. Biol.*, **2**, 267–282.

Alkan,C. *et al.* (2006b) RNA secondary structure prediction via energy density minimization. *RECOMB* 2006, LNBI 130–142.

Birnstiel,M.E. (1988) *Structure and Function of Major and Minor Small Nuclear Ribonucleoprotein Particles*. Springer-Verlag, Berlin, Germany.

Cao,Y. *et al.* (2010) sRNATarBase: a comprehensive database of bacterial sRNA targets verified by experiments. *RNA*, **16**, 2051–2057.

Chan,P.P. *et al.* (2011) Discovery of permuted and recently split transfer RNAs in Archaea. *Genome Biol.*, **12**, R38.

Chaudhuri,K. and Chatterjee,R. (2007) MicroRNA detection and target prediction: integration of computational and experimental approaches. *DNA Cell Biol.*, **5**, 321–337.

Doudna,J.A. and Cech,T.R. (2002) The chemical repertoire of natural ribozymes. *Nature*, **418**, 222–228.

Gaudin,C. *et al.* (2002) Two-piece tmRNA in cyanobacteria and it's structural analysis. *Nucleic Acids Res.*, **30**, 2018–2024.

Gesteland,R.F. *et al.* (2005) *The RNA World*, 3rd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Harmanci,A.O. *et al.* (2008) PARTS: probabilistic alignment for RNA joint secondary structure prediction. *Nucleic Acids Res.*, **36**, 2406–2417.

Harmanci,A.O. *et al.* (2011) TurboFold: iterative probabilistic estimation of secondary structures for multiple RNA sequences. *BMC Bioinformatics*, **12**, 108.

Heale,B.S. *et al.* (2005) siRNA target site secondary structure predictions using local stable substructures. *Nucleic Acids Res.*, **33**, e30.

Hodas,N.O. and Aalberts,D.P. (2004) Efficient computation of optimal oligo-RNA binding. *Nucleic Acids Res.*, **32**, 6636–6642.

Hofacker,I.L. *et al.* (2004) Alignment of RNA base pairing probability matrices. *Bioinformatics*, **20**, 2222–2227.

Houseley,J. and Tollervey,D. (2009) The many pathways of RNA degradation. *Cell*, **136**, 763–776.

Keiler,K.C. *et al.* (2000) tmRNAs that encode proteolysis-inducing tags are found in all known bacterial genomes: a two-piece tmRNA functions in Caulobacter. *Proc. Natl Acad. Sci. USA*, **97**, 7778–7783.

Lu,Z.J. and Mathews,D.H. (2007) Efficient siRNA selection using hybridization thermodynamics. *Nucleic Acids Res.*, **36**, 640–647.

Lu,Z.J. *et al.* (2009) Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA*, **15**, 1805–1813.

Mathews,D.H. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, **10**, 1178–1190.

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.

Mathews,D.H. *et al.* (1999a) Predicting oligonucleotide affinity to nucleic acid targets. *RNA*, **5**, 1458–1469.

Mathews,D.H. *et al.* (1999b) Expanded sequence dependence of thermodynamic parameters provides improved prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.

Mathews,D.H. *et al.* (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl Acad. Sci. USA*, **101**, 7287–7292.

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

Mückstein,U. *et al.* (2006) Thermodynamics of RNA-RNA binding. *Bioinformatics*, **22**, 1177–1178.

Nussinov,R. and Jacobson,A.B. (1980) Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl Acad. Sci. USA*, **77**, 6309–6313.

Papadopoulos,G.L. *et al.* (2009) The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res.*, **37**, D155–D158.

Pervouchine,D.D. (2004) IRIS: intermolecular RNA interaction search. *Genome Inform.*, **2**, 92–101.

Reuter,J.S. and Mathews,D.H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, **11**, 129.

Seetin,M.G. and Mathews,D.H. (2012) RNA structure prediction: an overview of methods. *Methods Mol. Biol.*, **905**, 99–122.

Sharkady,S.M. and Williams,K.P. (2004) A third lineage with two-piece tmRNA. *Nucleic Acids Res.*, **32**, 4531–4538.

Tabei,Y. *et al.* (2008) A fast structural multiple alignment method for long RNA sequences. *BMC Bioinformatics*, **9**, 33.

Tafer,H. and Hofacker,I.L. (2008) RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics*, **24**, 2657–63.

Tafer,H. *et al.* (2008) The impact of target site accessibility on the design of effective siRNAs. *Nat. Biotechnol.*, **26**, 578–583.

Walter,P. and Blobel,G. (1982) Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. *Nature*, **299**, 691–698.

Watanabe,Y. *et al.* (2007) Computational methods for microRNA target prediction. *Methods Enzymol.*, **427**, 65–86.

Will,S. *et al.* (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, **3**, e65.

Wu,L. and Belasco,J.G. (2008) Let me count the ways: mechanisms of gene regulation by miRNAs and siRNAs. *Mol. Cell*, **29**, 1–7.

Xia,T. *et al.* (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick pairs. *Biochemistry*, **37**, 14719–14735.

Xu,Z. *et al.* (2011) Statistical evaluation of improvement in RNA secondary structure prediction. *Nucleic Acids Res.*, **40**, e26.

Yu,L.E. and Koslowsky,D.J. (2006) Interactions of mRNAs and gRNAs involved in trypanosome mitochondrial RNA editing: structure probing of a gRNA bound to its cognate mRNA. *RNA*, **12**, 1050–1060.