Published online 3 February 2012 *Nucleic Acids Research, 2012, Vol. 40, No. 10* ***4681–4690***
doi:10.1093/nar/gks052

Predicting kissing interactions in microRNA–target complex and assessment of microRNA activity

Song Cao[1,2] and Shi-Jie Chen[1,2]

[1]Department of Physics and [2]Department of Biochemistry, University of Missouri, Columbia, MO 65211, USA

ABSTRACT

MicroRNAs (miRNAs) are a class of short RNA molecules that play an important role in post-transcriptional gene regulation. Computational prediction of the miRNA target sites in mRNA is crucial for understanding the mechanism of miRNA-mRNA interactions. We here develop a new computational model that allows us to treat a variety of miRNA-mRNA kissing interactions, which have been ignored in the currently existing miRNA target prediction algorithms. By including all the different inter- and intra-molecular base pairs, this new model can predict both the structural accessibility of the target sites and the binding affinity (free energy). Applications of the model to a test set of 105 miRNA-gene systems show a notably improved success rate of 83/105. We found that although the binding affinity alone predicts the miRNA repression efficiency with a high success rate of 73/105, the structure in the seed region can significantly influence the miRNA activity. The method also allows us to efficiently search for the potent miRNA from a pool of miRNA candidates for any given gene target. Furthermore, extension of the method may enable predictions of the three-dimensional (3D) structures of miRNA/mRNA complexes.

INTRODUCTION

microRNAs (miRNAs) are short single-stranded non-coding RNAs (~22 nt). In eukaryotic cells, miRNAs bind to the 3′-untranslated region (UTR) of the target messenger RNA transcripts (mRNAs) (1–3) and cause silencing of a specific sequence and result in translational repression. miRNAs play crucial roles in gene expression, development and human diseases such as cancer. Since the discovery of the first miRNA (*lin-4*) in *Caenorhabditis elegans* (4), to date, over 16 000 miRNAs (including over 1400 in humans) have been identified (http://www.mirbase.org/) (5,6). A large number of these miRNAs have been found to be crucial for the normal cell development. Down-regulated expression of miRNAs have been related to several diseases such as heart hypertrophy and cancer in human. Recent evidences indicate that mir-34a (7) and mir-26a (8) can suppress tumor growth. Such miRNAs could lead to promising anti-cancer drug in the future. To understand how miRNAs function, we need the structural information about the target sites and miRNA/mRNA complexes. Given the fact that few 3D structures have been determined in experiments (9,10), computational predictions of the target sites and the miRNA/mRNA structures become highly needed (11,12).

Many current computational predictions for miRNA targets are based on either sequence-match/RNA secondary structures, sequence/site conservation or a combination of the structural and sequence features (13–27). For example, one of the first miRNA target predicting programs, TargetScan (17), requires orthologous 3′-UTR sequence and target site conservation in multiple organisms as well as sequence complementarity at the 'seed' region of the UTR (17). The algorithm is mainly based on the sequence-match method and does not explicitly account for the conformational distribution for the miRNA and the target mRNA. Other algorithms are based on the energetics for miRNA–target binding. For example, RNAhybrid (16) ranks the target sites according to the binding affinities. However, RNAhybrid does not treat complex structural motifs such as kissing complexes and does not account for the accessibility of the target, which has been suggested to be potentially important for miRNA–target interaction.

In order to form a stable miRNA/mRNA complex (Figure 1), the intramolecular base pairs inside miRNA and around the target sites are completely unzipped. Disruption of these intramolecular base pairs allows for the formation of new intermolecular base pairs (usually ~20 bps). Thus, both the site accessibility and the binding affinity between miRNA and target sites are important. Computational predictions based on models such as STarMir, PITA and mirWIP (13,14,26), which can account for the site accessibility and the binding affinity, have suggested that including the site accessibility led to improvement in the prediction of the target (13,14,26,28,29).

*To whom correspondence should be addressed. Tel: +1 573 882 6626; Fax: +1 573 882 4195; Email: chenshi@missouri.edu

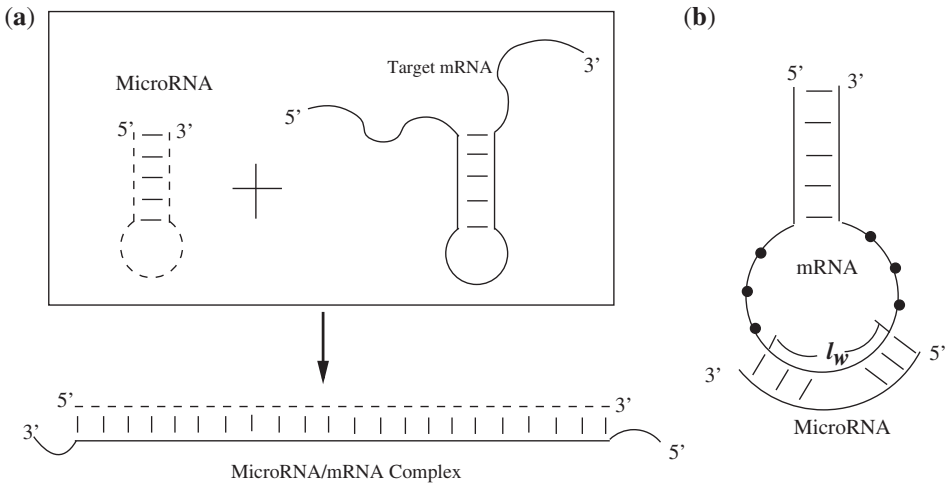

Figure 1. (**a**) The binding process between a microRNA and a target mRNA. The binding often involves the disruption of the intramolecular base pairs inside microRNA and mRNA. (**b**) A kissing interaction between miRNA and mRNA, in which miRNA binds to the hairpin or internal loop of the structural mRNA.

Despite the recent advances in the predictions of miRNA target sites, several crucial problems remain. One of the problems is the accurate calculation of the binding affinity for miRNA–target interaction (see Figure 1b). The current algorithms do not treat binding-induced redistribution of the conformational ensemble of the miRNA–target system (13,14,26). An important issue here is how to evaluate the entropy and free energy changes of the system upon binding. Previous studies on the kissing complexes and other RNA folding systems such as pseudoknots suggested that a reliable estimation for the entropy is indispensable for folding predictions (30–36). In addition, few miRNA/mRNA complex structures have been experimentally determined (9), which highlights the necessity of developing a computational model that can predict miRNA/target structure and folding stability. In the present study, we develop such a model based on the statistical mechanical analysis of the system. In the model, we consider explicitly the entropy change associated with the formation of the miRNA/mRNA complex. This model distinguishes from the other existing algorithms through the physics-based direct computation of the entropy and the binding free energy, especially for the different kissing complexes between miRNA and mRNA. Statistical mechanical approach requires the enumeration of all the possible structures. For a miRNA–mRNA complex which can be a large system, exhaustive enumeration for the complete conformational ensemble (the original statistical mechanics method) is not viable due to the required exceedingly long computational time. We develop a probabilistic domain-based method to dissect the full structure of the miRNA–mRNA complex into the miRNA–mRNA binding domain and the 5′ and 3′ unbound domains; see 'Materials and Methods' for details. Comparisons between the structure and free energy predictions from the domain-based method and from the original statistical mechanical method (based on the exact conformational enumeration for the full miRNA–mRNA system) show that the domain-based method is quite accurate. Furthermore, based on a recently developed 3D RNA structure prediction model (37), the current model enables predictions for the 3D structures for miRNA–mRNA complexes, which would provide the highly needed structural details and mechanistic insights into miRNA–mRNA interactions.

MATERIALS AND METHODS

Site accessibility

In the miRNA–mRNA binding process, the intramolecular base pairs between miRNA and mRNA often requires the disruption of the intramolecular base pairs in miRNA and mRNA around the target sites (Figure 1). Such site accessibility combined with the binding free energy together affect the miRNA–target binding and miRNA activity. As follows, we describe a statistical mechanical model that explicitly accounts for those effects.

We previously developed a virtual bond-based RNA folding model (called the 'Vfold' model) (38). The model provides an effective method for direct and complete conformational sampling. Extensive tests with the experimental data suggest that the model may be quite reliable (38). The Vfold model can treat both intramolecular and intermolecular base pairing and predict the free energy change (ΔG_{bind}) upon binding (38):

$$\Delta G_{\text{bind}} = \Delta G_{\text{miRNA/mRNA}} - (\Delta G_{\text{miRNA}} + \Delta G_{\text{mRNA}}) + \Delta G_{\text{init}} = -k_B T \ln\left(\frac{Q_{\text{miRNA/mRNA}}}{Q_{\text{miRNA}} \cdot Q_{\text{mRNA}}}\right) + \Delta G_{\text{init}} \quad (1)$$

where $Q_{\text{miRNA/mRNA}}$, Q_{miRNA} and Q_{mRNA} are the partition functions for the microRNA–RNA complex and the single-stranded (free) microRNA and mRNA, respectively. $k_B = 0.002$ kcal/mol/K is the Boltzmann constant and T is the temperature. ΔG_{init} is the free energy change associated with the nucleation of the two single

strands (miRNA and mRNA, respectively). For the two strands at equal concentration, we can calculate ΔG_{init} from the following formula: $\Delta G_{\text{init}} = -k_B T \ln(C_T/2)$, where C_T is the concentration of microRNA or mRNA.

In the calculations for the partition functions, we sum over all the possible structures (base pairing patterns) for the miRNA–mRNA complex, the free miRNA and the mRNA. Therefore, the algorithm accounts for the binding-induced changes in the conformational distribution. Moreover, in each miRNA–mRNA complex structure, inter-molecular base pairs compete with intra-molecular base pairs because a nucleotide is allowed to participate only one base pair in a structure. Therefore, the theory can effectively account for the accessibility of the target site.

ΔG_{bind} is an important criteria to determine microRNA–target binding. We use the experimental data for small interfering RNA (siRNA)–target binding and activity such as cleavage efficiency to test the ΔG_{bind}-activity correlation. A siRNA is a close analogy of miRNA though they may regulate gene expression through different mechanisms. A siRNA interferes with the expression of a specific gene through base-pairing with and cleaving the specific target in mRNA. As shown in Supplementary Figure S1a, the predicted ΔG_{bind} (from Equation 1) indeed shows an excellent correlation with the cleavage efficiency (39). From Supplementary Figure S1a, we can extract an analytical relationship between the cleavage efficiency η_{cleavage} and ΔG_{bind}:

$$\eta_{\text{cleavage}} = e^{0.0189(\Delta G_{\text{bind}}/k_B T + 6.48)} - 1.$$

In the calculation, the ion concentration is assumed to be 1M Na^+, the strand concentration for siRNA and the target RNA is equal to 1 nM and the temperature is 42°C (39). We do not consider the kinetic effect because the system has reached the thermal equilibrium in the experiment (39).

In addition to the sequences in Supplementary Figure S1a, we also find a good correlation between the Luciferase expression and ΔG_{bind} for other sequences. For example, for HIV(40), we find the Luciferase expression is inversely correlated to the ΔG_{bind} (see Supplementary Figure S1b), which is consistent with the above correlation between the cleavage efficiency and the ΔG_{bind}. A large ΔG_{bind} indicates a high binding affinity between siRNA and HIV targets and a lower Luciferase expression. Supplementary Figure S1b also yields an analytical expression between the Luciferase expression (η_{luci}) and the free energy change ΔG_{bind}:

$$\eta_{\text{luci}} = 2 - e^{0.0164(\Delta G_{\text{bind}}/k_B T + 19.42)}.$$

All the sequences in Supplementary Figure S1a and b have the same target sites, which can form the complementary base pairs with siRNAs. Different target structures result in very different cleavage efficiency and Luciferase expression. The two tested examples show the importance of considering the site accessibility in predicting siRNA–target binding and cleavage efficiency. The conclusion is consistent with the recent computational studies on siRNA and miRNA (13,14,41).

A new computational model for predicting the target sites

In the previous study (38), we developed a computational model for predicting the free energy landscape and folding thermodynamics of RNA–RNA complex up to hundreds of nucleotides. However, the length of the 3′-UTR mRNA sequence for a specific gene can reach thousands of nucleotides. Thus, direct application of the previous folding model to miRNA and mRNA interaction is not feasible. Here, we develop a new computational model that allows us to treat long RNA sequences.

In the Vfold model, the inter-molecular base pairs are inferred from the base pairing probability p_{ij} between the nucleotide i in miRNA and the nucleotide j in mRNA. In the statistical mechanical framework, p_{ij} is computed from the partition function:

$$p_{ij} = \frac{\alpha\, Q_{\text{miRNA/mRNA}}(i,j)}{Q_{\text{tot}}}; \qquad Q_{\text{tot}} = (\alpha\, Q_{\text{miRNA/mRNA}} + Q_{\text{miRNA}} \cdot Q_{\text{mRNA}}) \tag{2}$$

where $\alpha = e^{-\Delta G_{\text{init}}/k_B T}$ and $Q_{\text{miRNA/mRNA}}(i,j)$ is the conditional partition function of all the conformations that contain base pair (i, j). $Q_{\text{miRNA/mRNA}}(i, j)$ can be calculated from the method described in Ref. (38). Q_{tot} is the total partition function for the system that consists of the free miRNA, the free mRNA and the miRNA–mRNA complex. In the above equation, α represents the initiation penalty for miRNA–mRNA association. Thus, the computational time for calculating all the possible base pairing probabilities scales with the sequence lengths as $l_{\text{miRNA}} \cdot l_{\text{mRNA}} \cdot t_{\text{unit}}$, where l_{miRNA} and l_{mRNA} are the lengths of miRNA and mRNA, respectively, and t_{unit} is the computational time for calculating a partition function (such as Q_{tot} or $Q_{\text{miRNA/mRNA}}(i, j)$ for a given (i, j)).

In the new computational model, for structures without kissing interaction, we dissect the mRNA sequence into three domains, namely, (1, $i_w - 1$), (i_w, $i_w + l_w - 1$) and ($i_w + l_w$, l_m) (see Supplementary Figure S2a). (i_w, $i_w + l_w - 1$) is the domain for miRNA–mRNA binding. l_w is the width of the binding window. i_w is the starting point of the binding site and l_m is the length of the mRNA. For this type of structures, there is no interaction between the domains outside the binding site region, thus the probability for miRNA binding to the binding domain (i_w, $i_w + l_w - 1$) of the mRNA is determined by the following equation:

$$P_b(i_w, l_w) = \frac{\alpha\, Q_{\text{mRNA}}^{(1,i_w-1)} Q_{\text{miRNA/mRNA}}^{(i_w,i_w+l_w-1)} Q_{\text{mRNA}}^{(i_w+l_w,l_m)}}{Q_{tot}}.$$

Here $Q_{\text{mRNA}}^{(1,i_w-1)}$ and $Q_{\text{mRNA}}^{(i_w+l_w,l_m)}$ are the partition functions for the mRNA from nucleotides 1 to $i_w - 1$ and from nucleotides $i_w + l_w$ to l_m, respectively, and $Q_{\text{miRNA/mRNA}}^{(i_w,i_w+l_w-1)}$ is the partition function for the miRNA–mRNA complex formed from nucleotides i_w to $i_w + l_w - 1$ (in the mRNA).

For structures with kissing interactions outside the miRNA–mRNA binding region (see the color-shaded region in Supplementary Figure S2b), we divide the mRNA sequence into four parts: the colored region with inter-domain interactions and the other three domains $(x+1, i_w-1)$, (i_w, i_w+l_w-1) and $(i_w+l_w, y-1)$. The partition function $Q_{\mathrm{miRNA/mRNA}}$ for the miRNA-mRNA complex can be calculated as the following:

$$Q_{\mathrm{miRNA/mRNA}} = \alpha\, Q_{\mathrm{mRNA}}^{(x+1,i_w-1)} Q_{\mathrm{miRNA/mRNA}}^{(i_w,i_w+l_w-1)} Q_{\mathrm{mRNA}}^{(i_w+l_w,y-1)} Q_I(x,y)\, e^{\Delta S_2(l_w,l_{eff})/k_B}$$

where $Q_{\mathrm{mRNA}}^{(x+1,i_w-1)}$ and $Q_{\mathrm{mRNA}}^{(i_w+l_w,y-1)}$ are the partition functions for the mRNA from nucleotides $x+1$ to i_w-1 and from nucleotides i_w+l_w to $y-1$, respectively. In the calculation of $Q_{\mathrm{mRNA}}^{(x+1,i_w-1)}$ and $Q_{\mathrm{mRNA}}^{(i_w+l_w,y-1)}$, we allow the formation of all the possible stem-loop structures (not shown in the figure) in domains $(x+1, i_w-1)$ and $(i_w+l_w, y-1)$. $\Delta S_2(l_w, l_{eff})$ is the loop entropy change upon the formation of the kissing interaction (base pairing). $\Delta S_2(l_w, l_{eff})$ is dependent on the length of the binding site (l_w) and the effective loop length (l_{eff}). To calculate l_{eff}, we replace the stem closed by the base pair (x, y) with 1 nt. l_{eff} is equal to the number of unpaired nucleotide from $x+1$ to i_w-1 and from i_w+l_w to $y-1$ plus 1. In practice, ΔS_2 can be pre-calculated and tabulated so that the entropy parameters can be directly read out from the table [such as Table 1 and the supplementary material in Cao and Chen (42)].

$Q_I(x, y)$ is the partition function for the kissing region (color-shaded in the figure), i.e. the complex formed by strands $s1$ and $s2$ (Supplementary Figure S2c). Here $s1$ and $s2$ are the chain segments (y, l_m) and $(1, x)$, respectively, and $Q_I(x, y)$ is calculated from the method in Cao and Chen (38).

For a fixed window width l_w, we vary i_w from 1 to l_m-l_w+1 and for each l_w, we calculate the binding probability $P_b(i_w, l_w)$. We set l_w to vary from 7 to 30 nt. Here $l_w = 7$ corresponds to the minimal requirement to form a viable miRNA/mRNA complex (15,17) and $l_w = 30$ is a reasonable maximum length for the region of the known target site (17).

The purpose of dividing the whole mRNA sequence into domains is to parse the conformational enumeration in the partition function calculation into the shorter chain segments whose conformational enumerations are computationally less intensive. The algorithm causes the total conformational count to be an additive (instead of multiplicative) combination of the conformational count for the chain segments. Thus, the algorithm significantly improves the computational efficiency. Specifically, the computational time t_{tot} for predicting the different $P_b(i_w, l_w)$'s is on the same order of magnitude as t_{unit} and this new algorithm can reduce the computational time by a factor of $l_{\mathrm{miRNA}} \cdot l_{\mathrm{mRNA}}$ compared to the previous method (38). Supplementary Figure S3 shows the computational time for the current new model (rectangle) and the original statistical mechanical method (circles). The results show that the new method is much faster than the original statistical mechanical method. The new method can treat long sequence around 1400 nt in a few days on an Intel(R) Xeon(R) CPU 5150 @ 2.66G Hz on Dell EM64T cluster system.

Inclusion of the entropy parameter

The Vfold model provides an effective computational tool to enumerate the conformations from which we can evaluate the conformational entropy and the partition function. The partition function gives the free energy of the system. In particular, the model can give the conformational entropy and an estimation for the free energy for the different kissing complexes between miRNA and mRNA (Figure 1b). In addition, the Vfold can also predict the partition function and free energies for the free mRNA and miRNA (43). For example, before miRNA–mRNA binding, the hairpin loop entropy (ΔS_1) can be obtained from the computational model (43) and the empirical thermodynamic parameter (53). After binding, the entropy of the constrained hairpin loop (ΔS_2) is dependent on the length of the binding region (l_w) and the number of unpaired nucleotides in the constrained hairpin loop (see filled circles in Figure 1b and Supplementary Figure S2b). The benchmark test in Supplementary Figure S3 shows that inclusion of an accurate entropy parameter for the kissing interaction does not significantly slow down the computational speed.

RESULTS

Computational prediction of target sites

siRNA/HIV complex. Westerhout and Berkhout (40) perform a systematic study on how the target structure affects siRNA function. It was found that siRNA can completely disrupt the target structure and tightly bind to the target sites. We here use one of the HIV mutants, T4, to show the structural change in the binding process. Experimental studies indicate that the siRNA is a potent repressor for the gene expression of T4. The sequence lengths of siRNA and T4 are 19 and 47 nt, respectively. The short lengths of the sequences allow us to exhaustively enumerate all the possible conformations for the miRNA–mRNA complex and use the original statistical mechanical method to predict the structure of the single-stranded T4 sequence and the siRNA–T4 complex. Figure 2a and b show that the stem of T4 is completely disrupted upon siRNA binding at the target site. Meanwhile, we find that the nucleotides in the 3′ tail can refold into a new hairpin-like structure.

We have also applied the domain-based method to this system. Comparisons with the original statistical mechanical method show that the two methods give consistent structure and binding affinity for the complex. The result supports the validity of the domain-based method. Figure 2c shows the binding probability $P_b(i_w, l_w)$ as predicted from the domain-based method for miRNA binding to an l_w-nt stretch in mRNA starting from nucleotide i_w. $P_b(i_w, l_w)$ is sharply peaked at ($i_w = 1$, $l_w = 19$). The result agrees with the predicted secondary structure (Figure 2b) predicted from the original statistical

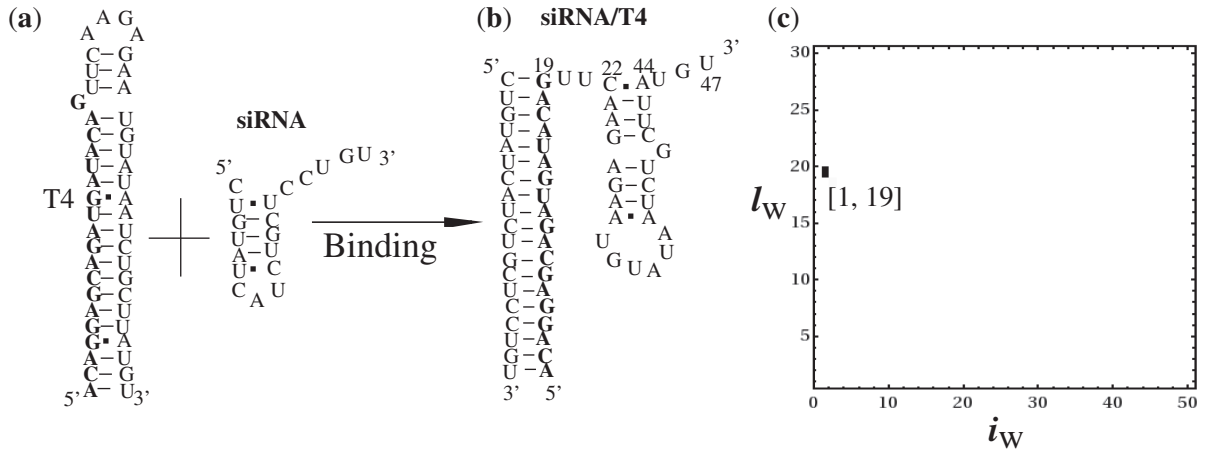

Figure 2. The conformational change caused by the siRNA binding to a HIV-1 mutant (T4). siRNA can induce the complete unzipping of T4 and T4 refolds into a new structure.

mechanical method. From the test case, we find that the domain-based method can indeed correctly predict the target site.

Drosophila melanogaster: to further validate the new computational model, we predict the binding sites for several experimentally confirmed systems in *D. melanogaster*. Figure 3 shows the predicted binding sites for mir-4/bagpipe, mir-2/grim, mir-7/hairy and mir-2/rpr. We draw the density plots for the binding probability function $P_b(i_w, l_w)$. The darkest dot in the figure indicates the most probable binding site. From Figure 3, which shows the predicted binding sites for mir-4/bagpipe, mir-2/grim, mir-7/hairy and mir-2/rpr are [93, 109], [59, 82], [441, 465] and [181, 202], we find that the predicted sites are consistent with the experimental results (14,44). In addition, Supplementary Figure S4 (upper panel) shows the predicted binding sites of three other experimentally studied systems. The binding regions are [34, 58], [363, 382] and [230, 245] for mir-2b/sickle, mir-9a/sens and mir-278/expanded, respectively. The predicted sites again agree with the suggested target sites from the experimental data (18,45,46).

Homo sapiens: we further tested the new computational model using the experimental data for miRNA binding to *H. sapiens*. Supplementary Figure S4 (lower) shows the predicted binding sites for three systems in *H. sapiens*. It has been found in the experiment that mir-29b can regulate the gene expression of Tcl1, which is related to the prognosis and progression of chronic lymphocytic leukemia (47). Our theory predicts that mir-29b tightly binds to Tcl1 with a binding affinity of $e^{-\Delta G_{\text{bind}}/k_B T} \sim 2115.8$. In the calculation, we use 4.1 kcal/mol for the initiation free energy ΔG_{init} (38,48) for the association of miRNA and mRNA (Equation 1). In addition, the predicted binding site is in agreement with the experiment (47). Moreover, application of the theory to other systems, such as the mir-196a/hoxb8 and mir-126/vcam-1 complexes, also shows good agreement with the previously reported results from sequence alignment among different species (49) for mir-196a/hoxb8 and the experimental data for the mir-126/vcam-1 complex (50).

Prediction of the functional miRNAs that tightly bind to the target

The above studies aim to predict the targets for a given miRNA. An equally important problem is to predict the miRNAs for a given target. The ability to identify the miRNA from a pool of miRNAs for any given gene target is highly needed for efficient therapeutic design through the strategy of miRNA-regulated gene expression. Figure 4 shows the predicted binding affinity between gene rpr and the available 163 miRNAs from 'http://www.microrna.org/microrna/home.do'. rpr is a central regulator of apoptosis in *D. Melanogaster*. The computational screening based on the binding affinity ranks mir-2a as the top candidate. The calculated binding affinity for mir-2a is 1.2×10^9. The high affinity is consistent with the experimental findings that mir-2a can efficiently repress the gene expression of rpr (14). This example on rpr indicates that the functional miRNAs tightly bind to the target sites and the computational approach can indeed identify the functional miRNAs from the predicted binding affinities.

Assessment of miRNA activity

The activity of a miRNA is determined not only by the binding affinity (13,14), but also by the structure of the target sites (44). The miRNA function is also influenced by other factors as shown by several experimentally deduced rules. For example, the complementarity between nucleotides 2 and 8 of miRNAs (the 'seed' region) and the target counterpart is also critical for target recognition for a functional miRNA (44). Previous studies showed that the combination of binding affinity and seed-pairing rule can lead to improved predictions for miRNA activities (14). However, lacking a physical model for the entropies and the binding free energies for the miRNA–mRNA system, especially for the key intermolecular interactions such as the kissing complexes, would adversely impact the reliability of the computational predictions (13,14,16,28). Here, as shown below, a more rigorous physical modeling for the the inter- and intra-molecular interactions (such as kissing complexes) and the conformational redistributions

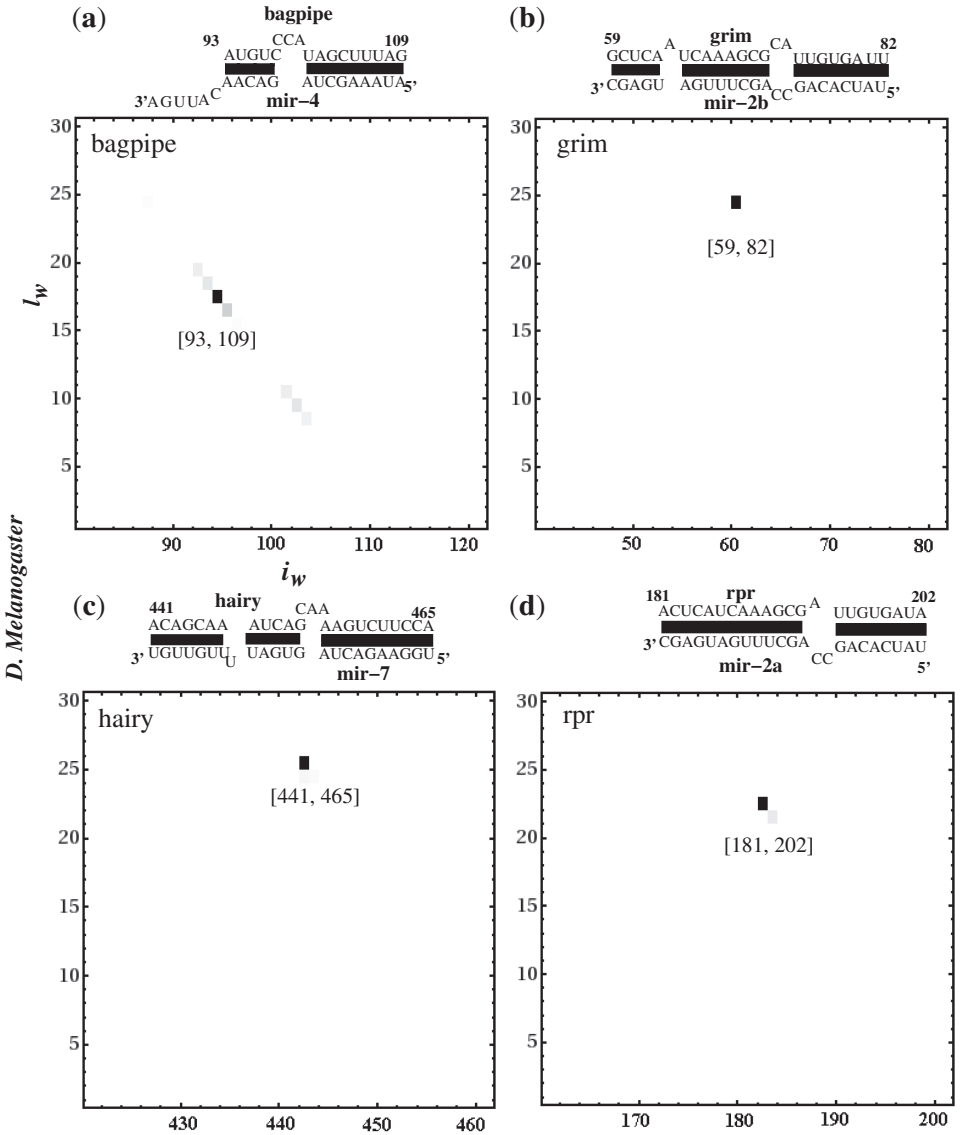

Figure 3. The predicted target sites for (**a**) mir-4/bagpipe, (**b**) mir-2/grim, (**c**) mir-7/hairy and (**d**) mir-2/rpr in *D. melanogaster*. The *x*-axis represents the position of the first binding nucleotide for each gene. The *y*-axis represents the window width of the binding domain. The predicted target sites are in agreement with the experiments (14,44). For example, mir-2 binds to the region (i_s = 59, i_w = 24) in (b) and the predicted target site is [59, 82], which is in a good agreement with the experiment (44).

upon miRNA–mRNA binding can indeed lead to improved predictions for miRNA activities.

To test our method, we predict the miRNA activities for the 105 test cases in Ref. (14). The selected 105 test cases satisfy two criteria: (i) the length of gene is shorter than 1400 nt, and (ii) the sequence of the gene can be found in the database (http://flybase.org/). In the calculation, we allow two types of seed sites, namely, the canonical seed site and non-canonical seed site. For the canonical seed sites, we allow only WC or GU base pairs in the seed site. For the non-canonical seed sites, we allow mismatches or single-nucleotide bulges in the seed site.

Supplementary Table S1 shows the predicted results for the test cases. With a 7.48/28.67 cutoff for the binding affinities for the canonical/non-canonical sites, we can correctly predict the activities for 73 out of 105 miRNA–gene target pairs. The 105−73 = 32 failed cases are mostly due to false positive predictions. However, the binding affinity does not provide information about the structures of the miRNA–mRNA complexes, especially in the seed region. We further applied our method to predict the structures for the (totally 97) complexes that have non-zero binding affinities; see Supplementary Figure S5. A close examination of the structures in Supplementary Figure S5 indicates three types of 'non-classical seed sites': (i) a bulge loop longer than 1 (e.g. rtGEF and htt genes), (ii) a single mismatch or unpaired nucleotide in positions 2, 3 and 4 (e.g. CG18662, CG4484 and sd genes), and (iii) a binding site that is too close to the coding gene ($\leq$8 nt) (e.g. yellow-c and boss genes). According to the rule of the miRNA–mRNA sequence complementarity in the seed region, we treat the above non-classical seed sites as

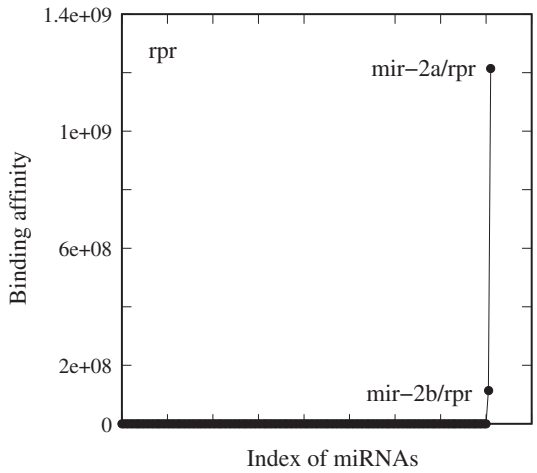

Figure 4. The predicted binding affinity between rpr and 163 miRNAs in *D. melanogaster* (http://www.mirbase.org/). In the calculation, we use 4.1 kcal/mol (48) value for the initiation (nucleation) energy for the association of the the miRNA and the mRNA. The experimentally validated functional miRNA (mir-2a) is ranked top based on our calculated binding affinity (14).

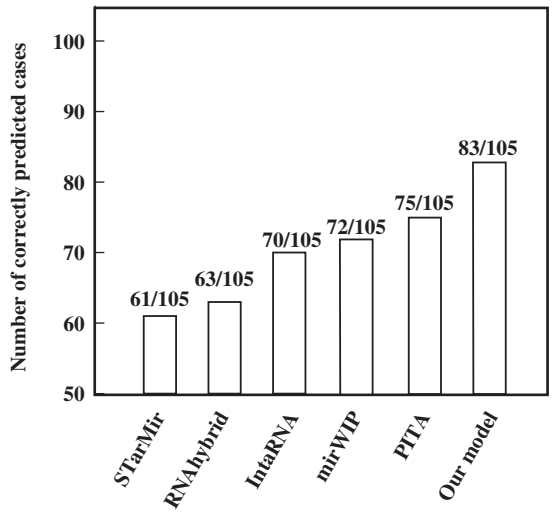

Figure 5. A comparison of the success rate between our model and other models: STarMir (13), RNAhybrid (16), IntaRNA (52), mirWIP (26) and PITA (14).

non-functional. The consideration of such structural requirement leads to further improved results. Supplementary Table S1 lists the predicted activity solely based on the predicted structures around the target sites. For a miRNA being functional, we require the miRNA–mRNA complex to pass both the binding affinity (the 7.48/28.67 cutoff) and the structure criteria (see the the above three rules for the non-functional seed sites). Comparisons with the experimental results give a success rate of 83 out of 105 cases for our model. This suggests an improved accuracy of the model as compared to other existing models (see Figure 5). We attribute the improved success rate to the accurate free energy model for the kissing interactions between the miRNA and the target as well as the more detailed structural studies for the target site. For the 105 test cases, we found that the predicted binding sites for 17 cases involve kissing interactions (Figure 1b): mir-279/SP555, mir-310/imd, mir-124/Gli, mir-287/DIP1, mir-7/hairy, mir-2b/skl, mir-2a/rpr, mir-7/Brd, mir-7/Tom, mir-14/wg, mir-278/Lar, mir-278/CG18815, mir-2b/CG1969, mir-2b/CG4269, mir-8/disp, mir-2a/scyl and mir-9a/brat. As an example, in Supplementary Figure S6, we show the predicted structure for the mir-7/Brd complex. We find that mir-7 forms kissing interactions with Brd through binding to a long internal loop (see the nucleotides marked by green color in the figure).

Prediction of 3D structure for microRNA/mRNA complex

The 3D structure for the microRNA/mRNA complex and for the free miRNA and mRNA are highly needed for understanding the binding energetics. Moreover, the 3D structures provide the direct information about the formation of the miRNA–mRNA complex and the interactions between the complex and other surrounding cofactors (51). We recently developed a free energy-based method to predict the 3D structure from the RNA sequence (37). For any given sequence, we first predict the 2D structures (base pairs) from the free energy model. For each predicted 2D structure, we construct a 3D scaffold by using the fragment templates selected from the PDB database. In the final step, using the 3D scaffold as the initial structure, we run the all-atom energy minimization and predict the all-atom 3D structure. The use of the physical model for the free energies, especially for structures with cross-linked loops, and the use of a novel method for template selection from the PDB database lead to an improved accuracy in the structure prediction (37).

To show the applicability of the 3D structure prediction method to miRNA–target systems, we predict the 3D structure of let-7/lin-41 complex. we chose this structure because it is one of the few available 3D structures for miRNA–mRNA complex that have been experimentally determined (9). Figure 6a shows the predicted 2D structure for let-7/lin-41 complex, which agrees with the experiment exactly. In the experiment (9), Cevec and Plavec *et al.* designed a hairpin structure (Figure 6b) to mimick the structure of let-7/lin-41 complex. The two structures contain the same internal loop (UUA-AU). Figure 6c shows the comparison between the measured NMR structure and the predicted structure. The overall rmsd is 1.9 Å, which shows a good agreement with the experimental result.

DISCUSSION

We developed and applied a new method to identify the gene–target site and the miRNA activity. Furthermore, to improve the computational efficiency, we developed a domain-based reduction method for the miRNA–target structure prediction. Compared to our previously developed domain-based model (42), the current model has two advantages. First, the current model can account for (long-range) inter-domain interactions (base pairing)

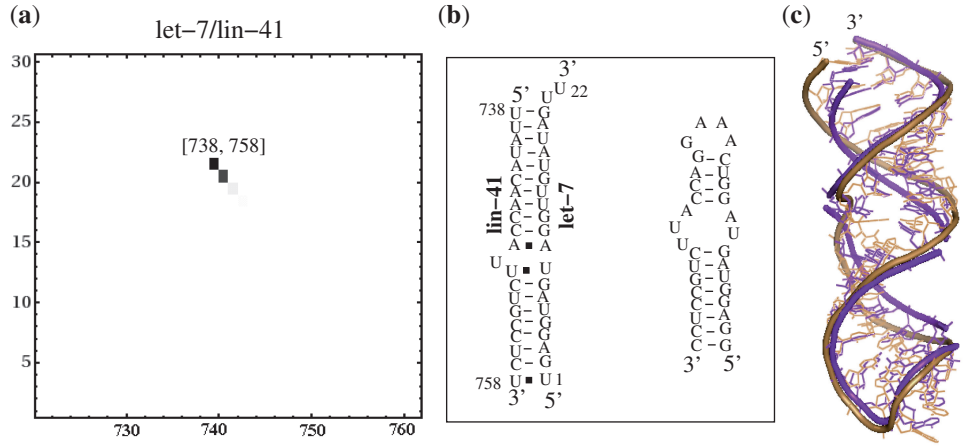

Figure 6. (**a**) The predicted target sites for let-7/lin-41 by the Vfold model. (**b**) The structures of the let-7/lin-41 complex and the hairpin structure used to mimic the complex structure (9). (**c**) The predicted 3D structure (purpleblue) and the experimental NMR structure (sand) for the hairpin structure in (b). The pdb id is 2jxv. The RMSD between the predicted structure and the experimental structure is 1.9 Å.

outside the target sites (see Supplementary Figure S2b). Second, the previous domain-based model is for monomeric RNAs while the current model can treat RNA–RNA complexes such as miRNA–mRNA complexes.

Extensive tests of the theory showed improved success rate as compared with other target-finding algorithms. For example, for 105 test cases in *Drosophila*, the model can correctly predict 83 cases, which shows improved success rate than other existing models (13,14,16,52). The better performance stems from two main improvements in the model. First, our method accounts for the different types of kissing contacts between miRNA and the target sites. The entropies and free energies for the interactions are evaluated with the explicit consideration of the excluded volume between different structural elements. Second, the model is based on the complete ensemble of all the possible inter- and intra-molecular base pairs, thus, the model effectively accounts for the target site accessibility and the conformational re-distribution of mRNA upon miRNA binding.

Our analysis shows that miRNA activity is largely (with a rate of 73/105) determined by the miRNA–mRNA binding affinity. However, the fact that the affinity alone can lead to many false positives indicates the insufficiency of using the binding affinity alone as the only indicator of miRNA activity. Consideration of the structure in the seed region of the miRNA–mRNA complex leads to much improved predictions with success rate increased from 73/105 to 83/105. The result suggests that both the binding energetics (binding affinity) and the structure in the seed region are important factors responsible for the miRNA activity.

Moreover, our algorithm also provides a reliable method for selecting the functional miRNA for a given gene. For instance, we find that the experimentally validated mir-2a, which is predicted to have the highest binding affinity to rpr gene, is correctly identified by our method. Furthermore, based on a recently developed 3D structure prediction model (37), we can predict the 3D structure for the different miRNA–mRNA complex.

The model, however, has several limitations. First, the model cannot consider the effect of the cofactors such as the surrounding proteins. Second, the current new model is based on the assumption that miRNA and mRNA interact mainly in the target site region and the length of the target site is <30 nt. The validity of such an approximation should be further examined for large structures involving distant contacts, especially with the presence of cofactors. Third, the current model can only treat genes with lengths <1400 nt. For a longer gene sequences, we need to develop a computationally more efficient algorithm. Finally, a web-based software will be needed and will be set up in the near future for predicting miRNA target sites and activity.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1 and Supplementary figures 1–6.

ACKNOWLEDGMENTS

Most of computations involved in this research were performed on the HPC resources at the University of Missouri Bioinformatics Consortium (UMBC).

FUNDING

National Institutes of Health (GM-063732); NSF grants (MCB-0920067 and MCB-0920411). Funding for open access charge: National Institutes of Health, National Science Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Neilson,J.R. and Sharp,P.A. (2008) Small RNA regulators of gene expression. *Cell*, **134**, 899–902.
2. Bartel,D.P. (2009) MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, **136**, 215–233.

3. Zhao,Y., Zhao,Y., He,S., Liu,C., Ru,S., Zhao,H., Yang,Z., Yang,P., Yuan,X., Sun,S. *et al.* (2008) MicroRNA regulation of messenger-like noncoding RNAs: a network of mutual microRNA control. *Trends Gene.*, **24**, 323–327.
4. Lee,R.C., Feinbaum,R.L. and Ambros,V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.
5. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
6. Ruby,J.G., Stark,A., Johnston,W.K., Kellis,M., Bartel,D.P. and Lai,E.C. (2007) Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Res.*, **17**, 1850–1864.
7. Fontana,L., Sorrentino,A., Condorelli,G. and Peschle,C. (2008) Role of microRNAs in haemopoiesis, heart hypertrophy and cancer. *Biochem. Soc. Trans.*, **36**, 1206–1210.
8. Kota,J., Chivukula,R.R., O'Donnell,K.A., Wentzel,E.A., Montgomery,C.L., Hwang,H.-W., Chang,T.-C., Vivekanandan,P., Torbenson,M., Clark,K.R. *et al.* (2009) Therapeutic microRNA delivery suppresses tumorigenesis in a murine liver cancer model. *Cell*, **137**, 1005–1017.
9. Cevec,M., Thibaudeau,C. and Plavec,J. (2008) Solution structure of a let-7 miRNA: Lin-41 mRNA complex from C. elegans. *Nucleic Acids Res.*, **36**, 2330–2337.
10. Sashital,D.G. and Doudna,J.A. (2010) Structural insights into RNA interference. *Curr. Opin. Struct. Biol.*, **20**, 90–97.
11. Sethupathy,P., Megraw,M. and Hatzigeorgiou,A.G. (2006) A guide through present computational approaches for the identification of mammalian microRNA targets. *Nat. Methods*, **3**, 881–886.
12. Ritchie,W., Flamant,S. and Rasko,J.E.J. (2009) Predicting microRNA targets and functions: traps for the unwary. *Nat. Methods*, **6**, 397–398.
13. Long,D., Lee,R., Williams,P., Chan,C.Y., Ambros,V. and Ding,Y. (2007) Potent effect of target structure on microRNA function. *Nat. Struct. Mol. Biol.*, **14**, 287–294.
14. Kertesz,M., Iovino,N., Unnerstall,U., Gaul,U. and Segal,E. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.*, **39**, 1278–1284.
15. Rajewsky,N. (2006) MicroRNA target predictions in animals. *Nat. Genet.*, **39**, S8–S13.
16. Rehmsmeier,M., Steffen,P., Hchsmann,M. and Giegerich,R. (2004) Fast and effective prediction of microRNA/target duplexes. *RNA*, **10**, 1507–1517.
17. Lewis,B.P., Shih,I.-H., Jones-Rhoades,M.W., Bartel,D.P. and Burge,C.B. (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787–798.
18. Stark,A., Brennecke,J., Russell,R.B. and Cohen,S.M. (2003) Identification of Drosophila microRNA targets. *PLoS. Biol.*, **1**, 397–409.
19. Enright,A.J., John,B., Gaul,U., Tuschl,T., Sander,C. and Marks,D.S. (2003) MicroRNA targets in Drosophila. *Genome Biol.*, **5**, R1.
20. John,B., Enright,A.J., Aravin,A., Tuschl,T., Sander,C. and Marks,D.S. (2004) Human microRNA targets. *PLoS Biol.*, **2**, 1862–1879.
21. Kiriakidou,M., Nelson,P.T., Kouranov,A., Fitziev,P., Bouyioukos,C., Mourelatos,Z. and Hatzigeorgiou,A. (2004) A combined computational-experimental approach predicts human miRNA targets. *Genes Dev.*, **18**, 1165–1178.
22. Krek,A., Grün,D., Poy,M.N., Wolf,R., Rosenberg,L., Epstein,E.J., MacMenamin,P., Da Piedade,I., Gunsalus,K.C., Stoffel,M. *et al.* (2005) Combinatorial microRNA target predictions. *Nat. Genet.*, **37**, 495–500.
23. Grun,D., Wang,Y.-L., Langenberger,D., Gunsalus,K.C. and Rajewsky,N. (2005) MicroRNA target predictions across seven Drosophila species and comparison to mammalian targets. *PLoS Comput. Biol.*, **1**, 51–66.
24. Saetrom,O., Snove,O. Jr and Saetrom,P. (2005) Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms. *RNA*, **11**, 995–1003.
25. Miranda,K.C., Huynh,T., Tay,Y., Ang,Y.-S., Tam,W.-L., Thomson,A.M., Lim,B. and Rigoutsos,I. (2006) A pattern-based method for the identification of microRNA binding sites and their corresponding heteroduplexes. *Cell*, **126**, 1203–1217.
26. Hammell,M., Long,D., Zhang,L., Lee,A., Carmack,C.S., Han,M., Ding,Y. and Ambros,V. (2008) mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein-enriched transcripts. *Nat. Methods*, **5**, 813–819.
27. Marin,R. and Vanicek,J. (2011) Efficient use of accessibility in microRNA target prediction. *Nucleic Acids. Res.*, **39**, 19–29.
28. Hofacker,I.L. (2007) How microRNAs choose their targets. *Nat. Genet.*, **39**, 1191–1192.
29. Obernosterer,G., Tafer,H. and Martinez,J. (2008) Target site effects in the RNA interference and microRNA pathways. *Biochem Soc. Trans.*, **36**, 1216–1219.
30. Cao,S. and Chen,S.-J. (2006) Predicting RNA pseudoknot folding thermodynamics. *Nucleic Acids Res.*, **34**, 2634–2652.
31. Cao,S. and Chen,S.-J. (2009) Predicting structures and stabilities for H-type pseudoknots with interhelix loops. *RNA*, **15**, 696–706.
32. Andronescu,M.S., Pop,C. and Condon,A. (2010) Improved free energy parameters for RNA pseudoknotted secondary structure prediction. *RNA*, **16**, 26–42.
33. Andronescu,M., Condon,A., Hoos,H.H., Mathews,D.H. and Murphy,K.P. (2010) Computational approaches for RNA energy parameter estimation. *RNA*, **16**, 2304–2318.
34. Sperschneider,J., Datta,A. and Wise,M.J. (2011) Heuristic RNA pseudoknot prediction including intramolecular kissing hairpins. *RNA*, **17**, 27–38.
35. Abraham,M., Dror,O., Nussinov,R. and Wolfson,H.J. (2008) Analysis and classification of RNA tertiary structures. *RNA*, **14**, 2274–2289.
36. Seetin,M.G. and Mathews,D.H. (2011) Automated RNA tertiary structure prediction from secondary structure and low-resolution restraints. *J. Comput. Chem.*, **32**, 2232–2244.
37. Cao,S. and Chen,S.-J. (2011) Physics-based de novo prediction of RNA 3D structures. *J. Phys. Chem. B*, **115**, 4216–4226.
38. Cao,S. and Chen,S.-J. (2006) Free energy landscapes of RNA/RNA complexes: with applications to snRNA complexes in spliceosomes. *J. Mol. Biol.*, **357**, 292–312.
39. Ameres,S.L., Martinez,J. and Schroeder,R. (2007) Molecular basis for target RNA recognition and cleavage by human RISC. *Cell*, **130**, 101–112.
40. Westerhout,E.M. and Berkhout,B. (2007) A systematic analysis of the effect of target RNA structure on RNA interference. *Nucleic Acids Res.*, **35**, 4322–4330.
41. Tafer,H., Ameres,S.L., Obernosterer,G., Gebeshuber,C.A., Schroeder,R., Martinez,J. and Hofacker,I.L. (2008) The impact of target site accessibility on the design of effective siRNAs. *Nat. Biotechnol.*, **26**, 578–583.
42. Cao,S. and Chen,S.-J. (2012) A domain-based model for predicting large and complex pseudoknotted structures. *RNA Biol.*, **9**, 201–212.
43. Cao,S. and Chen,S.-J. (2005) Predicting RNA folding thermodynamics with a reduced chain representation model. *RNA*, **11**, 1884–1897.
44. Brennecke,J., Stark,A., Russell,R.B. and Cohen,S.M. (2005) Principles of microRNA-target recognition. *PLoS Biol.*, **3**, 0404–0418.
45. Li,Y., Wang,F., Lee,J.-A. and Gao,F.-B. (2006) MicroRNA-9a ensures the precise specification of sensory organ precursors in Drosophila. *Genes Dev.*, **20**, 2793–2805.
46. Teleman,A.A., Maitra,S. and Cohen,S.M. (2006) Drosophila lacking microRNA miR-278 are defective in energy homeostasis. *Genes Dev.*, **20**, 417–422.
47. Pekarsky,Y., Santanam,U., Cimmino,A., Palamarchuk,A., Efanov,A., Maximov,V., Volinia,S., Alder,H., Liu,C.-G., Rassenti,L. *et al.* (2006) Tcl1 expression in chronic lymphocytic leukemia is regulated by miR-29 and miR-181. *Cancer Res.*, **66**, 11590–11593.
48. 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.

49. Yekta,S., Shih,I.-H. and Bartel,D.P. (2004) microRNA-directed cleavage of HOXB8 mRNA. *Science*, **304**, 594–596.
50. Harris,T.A., Yamakuchi,M., Ferlito,M., Mendell,J.T. and Lowenstein,C.J. (2008) MicroRNA-126 regulates endothelial expression of vascular cell adhesion molecule 1. *Proc. Natl Acad. Sci. USA*, **105**, 1516–1521.
51. Wang,H.-W., Noland,C., Siridechadilok,B., Taylor,D.W., Ma,E., Felderer,K., Doudna,J.A. and Nogales,E. (2009) Structural insights into RNA processing by the human RISC-loading complex. *Nat. Struct. Mol. Biol.*, **16**, 1148–1153.
52. Busch,A., Richter,A.S. and Backofen,R. (2008) IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, **24**, 2849–2856.
53. Serra,M.J. and Turner,D.H. (1995) Predicting thermodynamic properties of RNA. *Methods Enzymol.*, **259**, 242–261.