Methods 85 (2015) 90–99

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Quantifying the strength of miRNA–target interactions

Jeremie Breda, Andrzej J. Rzepiela, Rafal Gumienny, Erik van Nimwegen, Mihaela Zavolan *

Biozentrum, University of Basel and Swiss Institute of Bioinformatics, Klingelbergstrasse 50–70, 4056 Basel, Switzerland

ARTICLE INFO

Article history:
Received 6 February 2015
Received in revised form 9 April 2015
Accepted 10 April 2015
Available online 16 April 2015

Keywords:
miRNA
MIRZA
CLIP
CLASH
Non-canonical miRNA binding
miRNA target prediction

ABSTRACT

We quantify the strength of miRNA–target interactions with MIRZA, a recently introduced biophysical model. We show that computationally predicted energies of interaction correlate strongly with the energies of interaction estimated from biochemical measurements of Michaelis–Menten constants. We further show that the accuracy of the MIRZA model can be improved taking into account recently emerged experimental data types. In particular, we use chimeric miRNA–mRNA sequences to infer a MIRZA–CHIMERA model and we provide a framework for inferring a similar model from measurements of rate constants of miRNA–mRNA interaction in the context of Argonaute proteins. Finally, based on a simple model of miRNA-based regulation, we discuss the importance of interaction energy and its variability between targets for the modulation of miRNA target expression *in vivo*.

1. Introduction

MicroRNAs (miRNAs) have emerged as important regulators of gene expression across a wide range of species. They are endogenously encoded small RNAs that are incorporated in ribonucleoprotein complexes also containing an Argonaute (Ago) protein, which they guide to other RNA targets to modulate their expression [1]. Although comparative genomic analyses indicate that a miRNA has on average hundreds of targets [2], how these predicted targets respond to changes in miRNA concentration is not entirely clear. The best-documented outcome of miRNA–target interaction is target destabilization [3], which is typically modest, but can give rise to interesting behaviors of miRNA-containing regulatory networks. These include the 'threshold–linear' response of miRNA targets to their transcriptional induction [4,5] and the ultrasensitivity of target expression to the miRNA concentration [6]. The steady-state level of a given mRNA reflects the balance between transcription and decay. If the mRNA decay rate were constant, not modulated by miRNAs, the mRNA level would be expected to increase linearly with the transcription rate. However, if transcriptional induction occurs in the presence of a cognate miRNA, the target is expected to respond in a 'threshold–linear' manner: when the transcription rate is low, the few mRNA molecules that are produced are bound by the cognate miRNA and degraded. Once the transcription rate is sufficiently high for the mRNAs to saturate the miRNA–Ago complexes, the mRNAs escape the miRNA-induced repression and accumulate at a rate proportional to their transcription rate. The location of the threshold depends on the abundance of miRNA–Ago complexes, while the steepness of the transition between the two regimes depends additionally on the affinity of miRNA–target interaction.

We can illustrate these concepts with a simple model that focuses on the interaction of a single miRNA target with the miRNA and on the effect of this interaction on the rate of target decay, ignoring the possible effect of miRNAs on translation, the possible competition between targets for miRNAs and vice versa, other secondary effects such as feedbacks on target transcription rates, etc. Although these aspects most likely are relevant in *in vivo* situations, they go beyond the scope of our present study. Let us consider a miRNA target that is transcribed at rate α [mol s^{-1}] and decays with rate δ [s^{-1}]. The free miRNA target F [mol] associates at rate β [mol^{-1} s^{-1}] with miRNA–Ago complexes whose total concentration in a cell we assume to be constant, Σ [mol]. This leads to the formation of ternary target–miRNA–Ago complexes whose concentration we denote by A [mol], which can either dissociate into their components with rate ϱ [s^{-1}], or fall apart due to the degradation of the miRNA target, which occurs at rate $d\delta$ [s^{-1}]. The dynamics of these molecular species can then be described by the following equations:

$$\frac{dF}{dt} = \alpha - \delta F - \beta(\Sigma - A)F + \rho A \quad (1)$$

$$\frac{dA}{dt} = \beta(\Sigma - A)F - \rho A - d\delta A \quad (2)$$

* Corresponding author.
E-mail address: mihaela.zavolan@unibas.ch (M. Zavolan).

http://dx.doi.org/10.1016/j.ymeth.2015.04.012
1046-2023/

Solving this system of differential equations we obtain the dependency between the concentration of the free (and total) target and its transcription rate, which has the threshold–linear form. Fig. 1 shows how the concentration of the free mRNA target responds to changes in target transcription rate, assuming values for the parameters $\delta = 0.1$ 1/h and $d = 1.55$, which we have recently estimated [7]. To illustrate the expected behavior of high and low affinity targets we use two distinct values of the rate of ternary complex formation β, namely 0.24 and 2.4 cell/molecule/hour, and two distinct values of the rate of ternary complex dissociation ρ, namely 2.16 and 21.6 1/h. To further explore the behavior of targets of low, intermediate and high abundance miRNAs, we consider three total concentrations Σ of miRNA–Ago complexes, namely 10, 100 and 1000 molecules/cell. Our model thus assumes that the total concentration of miRNA–Ago complexes (free or bound to targets) is constant and does not respond to changes in miRNA target concentration. Although it remains unclear whether this assumption holds *in vivo*, data showing that the targets of endogenous miRNAs are up-regulated in response to transfection of exogenous siRNAs [8] suggest that at least the number of Argonaute molecules in a cell does not scale with the number of small RNAs that are present in cells. It can be observed that the transcription rate at which the target escapes miRNA regulation and accumulates rapidly depends on the total concentration of miRNA–Ago complexes, and that the transition is sharper for targets that have a higher rate of association with miRNA–Ago complexes. These behaviors have been observed in experiments with reporter constructs [5,9].

So far we discussed the expected behavior of an individual miRNA target. However, because a miRNA probably has hundreds of targets, one of the strongly debated questions in the field is whether changes in expression of one of these targets affects the expression of the others by modulating their interaction with the common targeting miRNA. Computational studies have shown that the targets of a miRNA are expected to respond in an asymmetrical manner, changes in expression of high-affinity targets affecting the binding of the lower affinity targets but not the other way around [10,11]. Whether these behaviors indeed occur *in vivo* is largely unknown. Rather, it has become clear that progress in understanding the impact of miRNAs on gene expression requires accurate measurements of miRNA abundance in single cells, estimates of the number of binding sites that a miRNA typically accesses within a cell, and estimates of the affinity of interaction between a miRNA and its multiple targets.

The abundance of individual miRNAs in mammalian cells varies over orders of magnitude (see for e.g. [12]). MiR-122, a highly-expressed, hepatocyte-specific miRNA can reach 66,000 copies per cell in mouse liver cells and 135,000 in primary human hepatocytes [13]. The more typical range for well-expressed miRNAs is 1000–10,000 molecules per cell [12], which can probably be accommodated by the population of Ago proteins, whose abundance per cell has been estimated to be ~140,000–170,000 molecules (in a mouse epidermis and a human melanoma cell) [14].

The number of target sites that a miRNA can access within an individual cell remains hotly debated [9]. Recently developed methods have enabled quantification of mRNA species within single cells, although the mRNA capture rate appears to be low, around 10% [15]. A cursory analysis of the published mouse embryonic stem cell (ESC) single cell data shows that among the mRNAs that were captured, miRNA targets occur in a handful of copies such that the top 100 predicted targets of individual miRNAs yield a few hundred captured target molecules per cell (Fig. 2). The targets of the mouse ESC-specific miR-294 are less abundant, ~1 captured mRNA per cell, compared to targets of the ubiquitously expressed miR-16 and of some miRNAs that are expressed in differentiated tissues (e.g. the general differentiation marker let-7, the neuron-specific miR-124, the muscle-specific miR-1 and the epithelia-specific miR-200a), which were captured in 2–5 copies, on average. Assuming a capture rate of 10%, a mouse ESC thus expresses on average 10–50 molecules per miRNA target. The argument can be made that our estimation ignores the fact that ESCs already contain miRNAs which have reduced the levels of their targets and that we have thus underestimated the number of miRNA targets. Indeed, to improve these estimates we would need to quantify mRNA abundance in ESCs devoid of miRNAs (Drosha/Dicer knock-out ESCs). However, many studies in which miRNAs have been transfected in cells in which they were not previously expressed found only modest changes (less than 2-fold) in target levels and thereby decay rates (see for e.g. [7]). If a miRNA does target over a hundred distinct mRNA species, binding to perhaps multiple sites within a mRNA, the number of putative binding sites of a miRNA in a single cell can reach 10^3–10^4. Precise estimates of the number of binding sites and the ratio of binding sites to miRNA–Ago molecules are essential for understanding the behavior of the targets *in vivo*, in individual cells.

2. Inferring the strength of miRNA–target interactions from experimentally-determined target sites; theory

An important breakthrough in the experimental identification of miRNA targets came with the development of methods based on the crosslinking and immunoprecipitation of Argonaute proteins (Ago-CLIP) [17,18], which enabled the capture of *in vivo* miRNA targets in high-throughput. The basic principle is to crosslink proteins to RNAs *in vivo* with ultraviolet light,

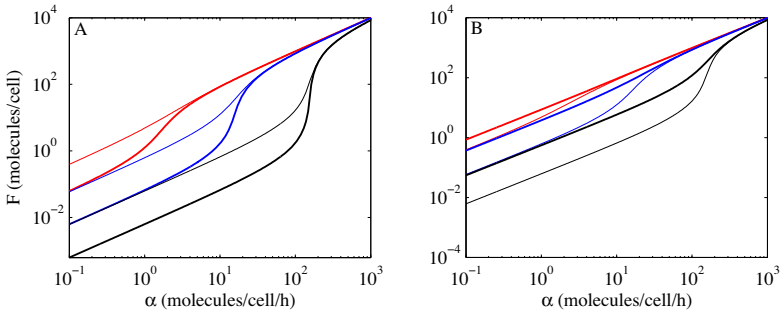

Fig. 1. Accumulation of miRNA targets as a result of increasing transcription, in the presence of miRNAs, based on the steady state solution of Eqs. (1) and (2). The three colors correspond to three total concentrations of miRNA–Ago complexes of 10 (red), 100 (blue) and 1000 (black) molecules/cell. (A) Thin lines correspond to low rates of target–miRNA–Ago association $\beta = 0.24$ cell/molecules/hour, and thick lines to 10-fold higher association rates, $\beta = 2.4$ cell/molecules/hour, with $\rho = 2.16$ 1/hour. (B) Thin lines correspond to low rates of target–miRNA–Ago dissociation of $\rho = 2.16$ 1/h, and thick lines to 10-fold higher dissociation rates, 21.6 1/h, with $\beta = 0.24$ cell/molecules/hour.

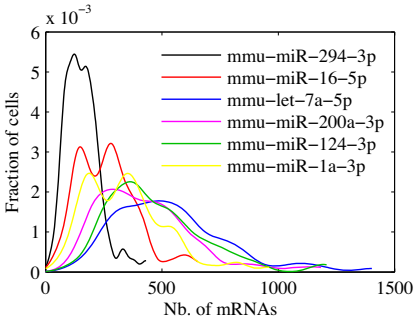

Fig. 2. Distribution of the number of targets of individual miRNAs that were captured from individual ESCs [15]. For each miRNA, the number of molecules of top 100 targets that were predicted with the seed-MIRZA-G-C miRNA target prediction program [16] were counted. The actual number of molecules was probably 10-fold higher (assuming that the capture rate of mRNA molecules in mRNA-seq is ~10%).

immunoprecipitate the protein of interest and associated RNAs with a specific antibody, and prepare the protein-bound RNA fragments for deep sequencing. The resulting reads can be used not only to identify the mRNAs that were bound by miRNA-guided Argonaute proteins, but also to learn more about how miRNAs interact with their targets. For example, to describe this interaction, in previous work we introduced a model (MIRZA) that includes besides parameters for A–U, G–C, and G–U base pairs, for symmetrical and asymmetrical loops, a set of parameters corresponding to miRNA position-dependent contributions to the interaction energy [19]. The latter could result from the interaction taking place within the context of the Argonaute protein (Fig. 3). Parameter values were inferred within a probabilistic framework, by maximizing the likelihood of the CLIP data. They confirmed the known importance of the miRNA 5′ end (also known as 'seed' [2]) in the interaction with the target. However, application of the model to the CLIP sites suggested that many are bound in a 'non-canonical' manner (i.e. without perfect complementarity to the miRNA seed) and that the proportion of non-canonical sites that were captured for a given miRNA with CLIP increased with the abundance of the miRNA [19]. Because MIRZA provides a quantitative measure of the strength of interaction of miRNAs with target sites, it can be used not only for genome-wide prediction of binding sites but also to study miRNA-dependent regulation in deeper quantitative detail. In a parallel development, a next step in the experimental identification of miRNA target sites has been taken with the simultaneous capture of interacting miRNAs and target sites as chimeric sequence reads [20,21]. Initial analysis of these data suggested that miRNAs may differ in their mode of interaction with the targets.

Thus, important open questions for the quantitative modeling of miRNA–target interactions are: what approach yields the most predictive model; what structure does this model have; are miRNA-specific models necessary to explain the experimental data? In the following we describe the miRNA–target interaction models that we inferred with the MIRZA approach from various types of high-throughput data, and we evaluate their ability to identify functional miRNA targets, that are destabilized upon transfection of the cognate miRNA.

2.1. Input data: Argonaute-bound RNA fragments. Output: general model of miRNA–target interaction MIRZA–CLIP

A target site m of a miRNA μ can be in one of two states, namely bound or unbound to the miRNA. Denoting the energies of the bound and unbound states by E_B and $E_{\underline{B}}$, the probability to find the site in bound state will be given by $P_B = \frac{e^{E_B}}{e^{E_B}+e^{E_{\underline{B}}}}$. The 'bound' state consists in fact of all ways in which the miRNA is hybridized with the target in the context of the Ago protein. Denoting by $E(m, \mu, \sigma)$ the energy of the state in which site m is bound to miRNA μ in configuration σ, e^{E_B} is proportional to $\Sigma_\sigma e^{E(m,\mu,\sigma)}$. Similar to the standard model of RNA–RNA interaction [23], $E(m, \mu, \sigma)$ can be written in terms of a small number of parameters such as the energy of A–U, G–C and G–U base pairs, the energy for opening a loop in the miRNA–target hybrid, energies for extending a loop by a nucleotide in the miRNA, or in the mRNA, or by two unpaired nucleotides in the miRNA and target. In addition, specific to the

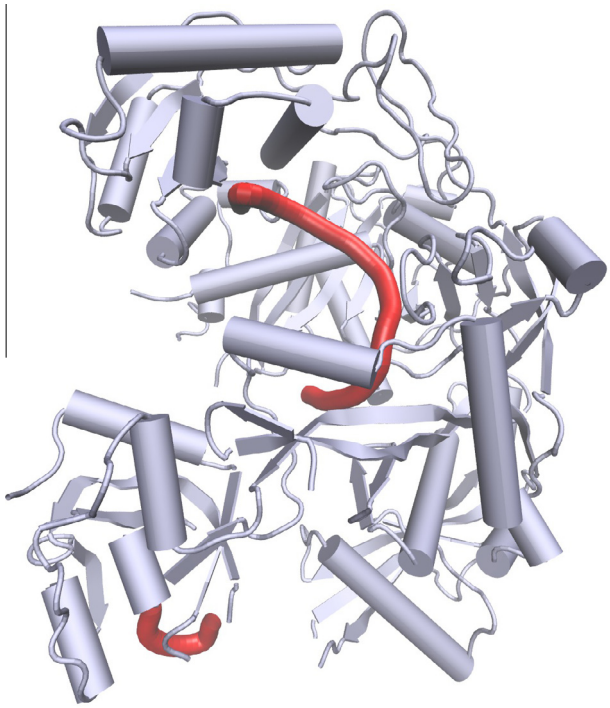

Fig. 3. Crystal structure of the human AGO-2 protein (silver) in complex with miR-20a (red) [22]. The 'seed' nucleotides are visible in the structure because the conformational entropy of the miRNA 5′ end in the binding pocket of AGO-2 is limited. The residues 11–16 of the miRNA are not resolved due to their conformational freedom. The terminal 3′ end nucleotides, that contribute to the anchoring of the miRNA within AGO-2, are again visible.

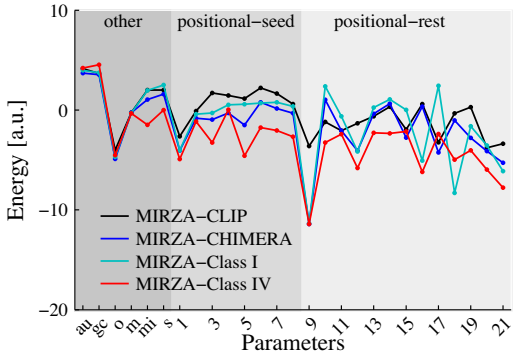

Fig. 4. The 27 parameters of various MIRZA model variants. From left to right, base-pair parameters (A–U, G–C, G–U = 0), loop parameters (o: opening a loop, m: looped out mRNA nucleotide, mi: looped out miRNA nucleotide, s: symmetrical loop) and the 21 positional parameters are shown. The parameters of the MIRZA–CLIP model are shown in black, those of the MIRZA–CHIMERA model in blue, those of the MIRZA-Class I model in cyan and those of the MIRZA-Class IV model in red.

MIRZA model of miRNA–target interaction [19] is a set of miRNA-position-specific energies (Fig. 4). The logarithm of the 'quality score' of a site for a miRNA that MIRZA computes can be viewed as the energy of interaction between the miRNA and the target. An efficient dynamic programming algorithm for computing target quality scores has been proposed [19]. This enables one to infer the parameters of the MIRZA model by maximizing the likelihood of the Ago-CLIP data. Here we have repeated the analysis of the ~3000 Ago2-CLIP sites that were reproducibly isolated in multiple CLIP experiments [19,24] to derived the baseline MIRZA–CLIP model shown in Fig. 4.

2.2. Input data: chimeric miRNA–mRNA sequence reads. Output: general model of miRNA–target interaction MIRZA–CHIMERA

As mentioned in the Introduction, Helwak et al. [20] designed the Crosslinking and Sequencing of Hybrids approach (CLASH), in which the interacting RNAs are ligated prior to sequencing, thereby enabling the simultaneous capture of interacting miRNAs and target sites. These appear as "chimeric reads" each composed partly of a miRNA and partly of the miRNA target. Grosswendt et al. [21] subsequently reported that a substantial number of ligated miRNA–target site chimeras can be found even in samples prepared with a standard CLIP protocol. In contrast to Ago-CLIP, in these data sets there is no uncertainty about the miRNA that guided the interaction with each target site captured in the chimeras. Thus, in maximizing the likelihood of the data to infer a MIRZA-type model, one only needs to sum over all the ways in which the miRNA and target site in each chimera hybridizes with each other (and not over the miRNAs that could have interacted with the target site, as in the case of Ago-CLIP sites). We used the miRNA-target site pairs that were inferred by Grosswendt et al. from various PAR-CLIP and HITS-CLIP experiments (Table 1 and Supplementary Table 3 in [21]) to construct a general model that could explain all these interactions. We called this model MIRZA–CHIMERA. Compared to the MIRZA–CLIP model that we inferred from Ago-CLIP data, MIRZA–CHIMERA seems to put less emphasis on the miRNA seed (Fig. 4). The functional relevance of these differences will be discussed in the following sections.

2.3. Input data: chimera of a specific miRNA with target sites. Output: miRNA-specific model of interaction with the target

The CLASH study reported that some miRNAs, such as miR-92a and miR-181b, interact with their targets predominantly through their 3′ rather than the 5′ end, yielding 'Class IV' chimeras [20]. Other miRNAs such as those of the let-7 family were captured rather in 'Class I' chimeras, in which the miRNA presumably interacted through the 'seed'. These observations suggest that the accuracy of miRNA target prediction could be improved through the use of miRNA-specific models of interaction. We decided to test this hypothesis here. However, because the available data sets [20,21] contain a limited number of distinct target sites ligated to individual miRNAs, we inferred 'Class'-specific rather than miRNA-specific models. Concretely, from the data of Grosswendt et al. [21] we selected a total 2589 chimeras of 24 miRNAs (those that yielded predominantly Class I chimeras in the data of Helwak et al. [20]) to train the "MIRZA-Class I" model and 949 chimeras of 8 miRNAs (those that yielded predominantly Class IV chimeras) to train the "MIRZA-Class IV" model. The corresponding miRNAs are listed in Table 1. The parameters of these models, shown in Fig. 4, indicate a positive contribution of the seed positional parameters in the MIRZA-Class I model, but not in the MIRZA-Class IV model. However, Fig. 4 also shows a trend of positional parameters to progressively decrease from the seed to the 3′ end in the MIRZA-Class IV model, but not in the MIRZA-Class I model. We test the functional relevance of these differences in a subsequent section.

Table 1
Chimeras of the indicated miRNAs, obtained from the data set of Grosswendt et al. [21] were used to infer MIRZA-Class I and MIRZA-Class IV models.

MIRZA-Class I	let-7a-5p, let-7e-5p, let-7f-5p, miR-10a-5p, miR-10b-5p, miR-125a-5p, miR-125b-5p, miR-1260b, miR-1301-3p, miR-130b-3p, miR-15b-5p, miR-17-5p, miR-183-5p, miR-185-5p, miR-23a-3p, miR-27b-3p, miR-31-5p, miR-324-3p, miR-339-5p, miR-34a-5p, miR-423-5p, miR-455-3p, miR-484, miR-744-5p
MIRZA-Class IV	miR-181b-5p, miR-221-3p, miR-30c-5p, miR-30d-5p, miR-320a, miR-361-5p, miR-92a-3p, miR-92b-3p

It has been recently observed that the miRNAs that were reported to form Class IV hybrids have G/C-rich 3′ ends [25]. We reproduced these observations here (Fig. 5). Furthermore, we found that the proportion of Class I hybrids that were captured for a miRNA decreases with the G/C content of the miRNA 3′ end, while the proportion of Class IV hybrids shows the opposite trend (Fig. 5). A possible explanation behind the different propensities of different miRNAs to yield Class I or Class IV chimeras is that the G/C-content of the miRNA 3′ end stabilizes the interaction with the target site, facilitates ligation and leads to an over-representation of this type of sites among the chimeric sequences. This possibility would need to be investigated in more detail before miRNA-specific modes of interaction are inferred from chimera data.

3. Results

3.1. Evaluating the models on biochemical data

The 'quality score' assigned to a site by the MIRZA model takes into account all possible configurations in which the miRNA can hybridize to the target site within the ternary miRNA–target site–Ago complex, and provides an estimate of the binding energy between the miRNA and the target site. Thus, if the model is accurate, it should be able to predict the free energy of interaction determined with biochemical approaches. The dissociation constant K_D, which is the ratio of the rates of dissociation (k_{off}) and association (k_{on}) of molecules in a complex, $K_D = \frac{k_{off}}{k_{on}}$, should be related to the Gibbs free energy of interaction through the relationship $\Delta G = -k_B T \log\left(\frac{1}{K_D}\right)$, where k_B is the Boltzmann constant and T is the temperature. Although only few measurements of miRNA–target dissociation constants are available, particularly for mammalian systems, Wee et al. [26] measured a related constant, namely the Michaelis–Menten constant. This is defined as $K_M = \frac{k_{cat}+k_{off}}{k_{on}}$, thus including besides the dissociation and association rates the rate with which the miRNA catalyzes the target cleavage. Wee et al. measured for K_M's for perfectly complementary sequences (*PM*) and for sequences that have mismatches at different positions along the miRNA (*MM*) in the context of Argonaute 1 protein of Drosophila melanogaster [26] and then correlated $\log \frac{K_M^{PM}}{K_M^{MM}}$ with the difference in the free energy of interaction of the perfectly matched and mismatched hybrids given by the RNAStructure software [27]. Computing this correlation separately for duplexes in which mismatches were located at the 5′ and 3′ ends of the miRNA, respectively, Wee et al. concluded that the standard base pairing rules apply to miRNA–Ago2–target complexes [26]. We thus sought to use the measurements of Wee et al. [26] to further validate the MIRZA models that we inferred from CLIP data sets.

First, we compared the energy differences inferred from measurements of K_M's with those predicted with the current version

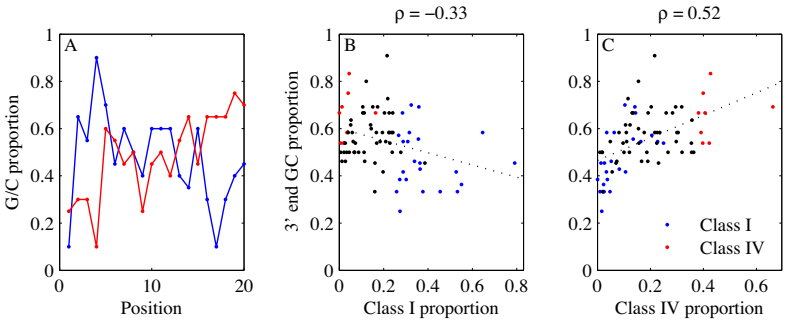

Fig. 5. Relationship between the nucleotide composition of the miRNA and the type of hybrids in which the miRNA was captured. The miRNAs used to infer the MIRZA-Class I model are shown in blue, the miRNAs used to infer the MIRZA-Class IV model are shown in red and other miRNAs are shown in black. Data for analysis taken from Helwak et al. [20]. (A) Proportion of G/C nucleotides at different positions along miRNAs that yield predominantly Class I and IV hybrids/chimeric reads in the data set of Helwak et al. [20]. (B) Correlation between the proportion of G/C nucleotides at the 3′ end of a miRNA and the proportion of captured Class I chimeras. (C) Correlation between the proportion of G/C nucleotides at the 3′ end of a miRNA and the proportion of detected Class IV chimeras.

(5.7) of the RNAStructure software and with those predicted with MIRZA-type models. As described by Wee et al. [26], we found relatively good correlations between RNAStructure-based predictions and experimental measurements, if we consider separately hybrids with mismatches in the miRNA seed region (Spearman correlation coefficient $\rho = 0.81$, p-value = 0.015) and in the miRNA 3′ end (Spearman correlation coefficient $\rho = 0.4$, p-value = 0.20). However, considering all the hybrids together, the correlation is rather poor (Spearman correlation coefficient $\rho = 0.20$), presumably because the nearest neighbor model implemented in RNAStructure does not appropriately describe interactions that take place within RNA–protein complexes, where different nucleotides in the RNA can have disproportionate contributions to the energy of interaction.

In contrast, evaluating *all* of the hybrids within the MIRZA–CLIP model yields predictions that are strongly correlated with the experimental results (Spearman correlation coefficient $\rho = 0.85$, p-value = 3.6e−09, 95% confidence interval = [0.71, 0.93]). Interestingly, the MIRZA–CHIMERA model gives a slightly higher correlation with the experimental data (Spearman correlation coefficient $\rho = 0.87$, p-value = 3e−09, 95% confidence interval = [0.73, 0.94]), although the difference is not significant. Thus, these two models, that were inferred from different types of sequenced miRNA target sites, predict remarkably well the energies of interaction between miRNAs and target sites that are inferred from biochemical measurements (Fig. 6).

3.2. Genome-wide prediction of miRNA targets

One of the main applications of these models is in the genome-wide prediction of miRNA binding sites. However, the predicted energy of interaction between a miRNA and a target site is only one of the factors that contributes to a functional interaction. Other features of the target site have also been shown to be important [28]. Thus, in recent work we sought to build on MIRZA and develop a model that is suitable for accurate prediction of miRNA binding sites genome-wide. The resulting MIRZA-G model combines the MIRZA target quality score with the accessibility of the target site, the G/U content of the region in which the site is embedded, the relative location of the site in the transcript and, optionally, with the degree of evolutionary conservation of the putative target site (Fig. 7). MIRZA-G is trained by fitting a generalized linear model with a logit function to discriminate between miRNA-complementary sites located in mRNAs that do and mRNAs that do not respond to the transfection of the cognate miRNAs [16]. Furthermore, because high-throughput studies evaluate the effects of miRNAs at the level of transcripts and genes rather than individual sites, MIRZA-G computes transcript/gene scores, summing up the probabilities that individual target sites have a functional impact. Using different MIRZA variants to compute target quality scores for the MIRZA-G model we can test the ability of these variants to predict which transcripts are most affected by the transfection of individual miRNAs. Thus, we employed the MIRZA–CLIP/CHIMERA/Class I/Class IV models individually within the MIRZA-G framework to predict and rank targets of individual miRNAs. Because different MIRZA variants yield different distributions of target quality scores and in the genome-wide prediction of target sites we only consider putative sites with a minimal ‘target quality’ score, we have used different thresholds for different models. The weight of different features of target sites within the MIRZA-G model were kept unchanged.

To determine a target quality score threshold for different MIRZA variants we noted that ‘canonical’ interactions that involve perfect pairing of the miRNA seed have the highest scores with all MIRZA variants. Thus, we employed the procedure that we used before for MIRZA–CLIP [16]. That is, with each MIRZA variant, we assigned to each of the 2998 CLIPed sites from Khorshid et al. [19] the most likely guiding miRNA. This was the miRNA with the highest target quality score for the site given under the considered MIRZA model. We then predicted the structure of the most likely hybrid between the target site and the guiding miRNA, and divided the sites into canonical – those with perfect base-pairing over nucleotides 2–8 of the miRNA or perfect pairing over nucleotides 2–7 followed by an adenine (opposite position 1 in the miRNA) – and non-canonical – all other sites. Based on the cumulative distribution of target quality scores for canonical and non-canonical sites, we set a threshold that allowed us to capture the majority of canonical sites without including too many non-canonical sites, that may be artifactually captured. For MIRZA–CLIP a threshold of 50 captures 91% of canonical sites and 18% non-canonical sites, for MIRZA–CHIMERA a threshold of 20 captures 97% canonical and 20% of non-canonical sites, for MIRZA-Class I a threshold of 30 leads to the capture of 94% of the canonical and 18% of non-canonical sites, while for MIRZA-Class IV a threshold of 20 captures 94% of canonical target sites and 20% of the non-canonical target sites.

3.3. Wide range of MIRZA quality scores across the targets of a given miRNA

Although we do not focus on this aspect here, it has been proposed that differences in affinity between targets may underlie

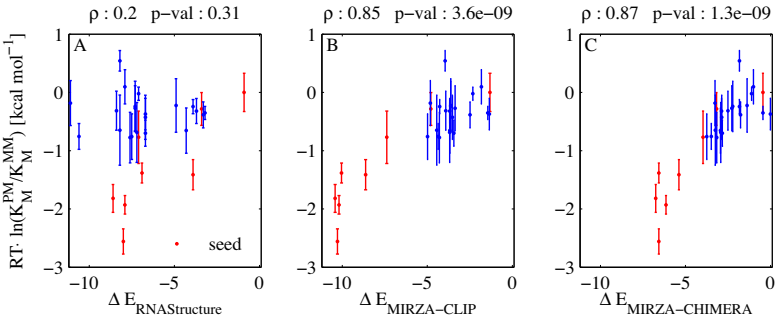

Fig. 6. Ratio of binding free energies of mismatched and perfectly matched hybrids. The Spearman correlation was computed between the values estimated based on biochemical measurements (energy of interaction $\sim \ln(1/K_M)$) and values predicted with three distinct models: RNAStructure 5.7 (left), MIRZA–CLIP (center) and MIRZA–CHIMERA (right). Data points in red correspond to hybrids with mismatches in the miRNA seed region, those in blue to hybrids with mismatches in the 3′ region.

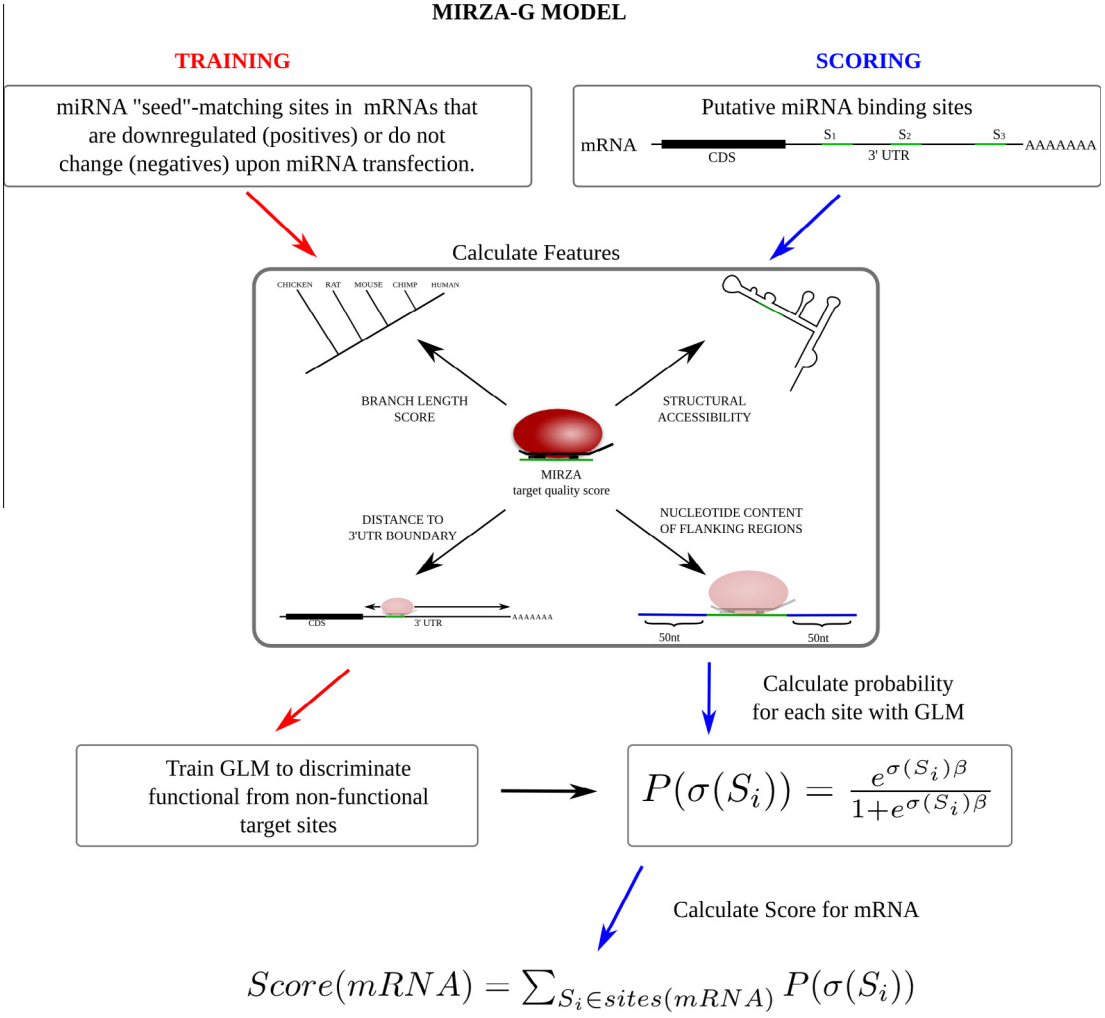

Fig. 7. Diagram of the approach for predicting miRNA targets with MIRZA-G.

asymmetries in the crosstalk of mRNAs that bind the same miRNAs [10]. Thus, having shown that the target quality scores computed with MIRZA models correlate very well with the affinities of miRNA–target interactions measured with biochemical methods, we wondered how much variation there is in the affinity of different target sites for a miRNA. Therefore, we determined the MIRZA target quality score for all the sites of all miRNAs that were considered in the genome-wide predictions with MIRZA-G. These had a probability of being functional of at least 0.12 (see [16] for details). For each miRNA we have divided the 0–10 range of MIRZA target quality scores into bins of 0.2 and have shown the distribution of the target sites of each miRNA as a heat-map, which each line corresponding to a miRNA and the intensity of the color indicating the density of target sites within a bin (Fig. 8). It can be seen that the target sites of an individual miRNA span a range of ∼4 log units or they can differ by ∼50-fold in the predicted affinity.

3.4. Evaluation of the MIRZA models on miRNA transfection data

miRNAs have been reported to destabilize their mRNA targets, inhibit their translation [1], and even to increase transcript stability under specific circumstances [29]. Of these, perhaps the least controversial is mRNA destabilization, which has been argued to be the dominant mechanism behind the repressive effect of miRNA, with translational repression playing a small, perhaps more transient role [3]. The importance of this mechanism is further underscored by observations that miRNA-complementary sites that are conserved in evolution and sites that induce strongest downregulation of their host transcripts upon miRNA transfection have similar properties [28]. Furthermore, acting through the miRNA pathway, small interfering RNAs (siRNA) also destabilize many transcripts (the so-called "off-target" mRNAs) [30]. Thus, it is reasonable to expect that the extent of mRNA destabilization upon miRNA transfection is a robust measure of the strength of interaction between a miRNA and the mRNA. Consequently, the ranking assigned by a computational miRNA target prediction method to mRNAs should correlate well with their change in expression upon miRNA transfection. This is the assumption that we make in discussing the relative performance of various models for miRNA target prediction.

First, we tested whether the models can predict the mRNA expression changes that were induced by individual transfections of miRNAs. To this end, we used data corresponding to 26 miRNA transfections into human cells and one transfection into mouse cells (Table 2). The processing of the transfection data was described extensively in [16]. For each type of MIRZA model of miRNA–target interaction we used two variants of the genome-wide MIRZA-G prediction model to predict sites. One of these considered the evolutionary conservation of the sites and the other did not [16] (see Fig. 7). We sorted targets predicted by each of these models in the order of their prediction score. We then computed the median log2 fold-change of the top N predicted transcripts as a function of the number N of top targets considered. The average profiles, computed over the 26 data sets, are shown in Fig. 9A–B. We found that all four models perform as expected in predicting miRNA targets genome-wide. Consistent with its slightly better performance in predicting the *in vitro*-measured free energy of interaction between miRNAs and target sites, the targets predicted by the MIRZA–CHIMERA model are somewhat more downregulated compared to the targets predicted with MIRZA–CLIP, particularly when the evolutionary conservation of the sites is not taken into account.

Next we asked whether Class I and Class IV-specific models are more accurate in predicting targets of miRNAs that have been found to yield predominantly Class I and Class IV chimeras, respectively. As representatives of the first we chose the let-7 family of miRNAs and as a representative of the latter the miR-92a. Because we did not find transfection data for Class IV-chimera forming human miRNAs, we used a data set obtained from mouse cells transfected with the mouse miR-92a. The results, shown in Fig. 9, panels C–D for let-7 and E–F for miR-92a, clearly indicate that the general MIRZA–CLIP and MIRZA–CHIMERA models are more accurate in predicting transcript downregulation upon miRNA transfection than Class I/IV-specific models. Together with the fact that the sites that are predicted with these models tend to be canonical sites, these results indicate that the origin and relevance of Class IV hybrids needs to be further investigated. As mentioned above, a possibility that needs to be ruled out is that the experimental procedure for isolating miRNA–target hybrids via chimeric sequences enriches for non-canonical hybrids that have increased stability prior to ligation.

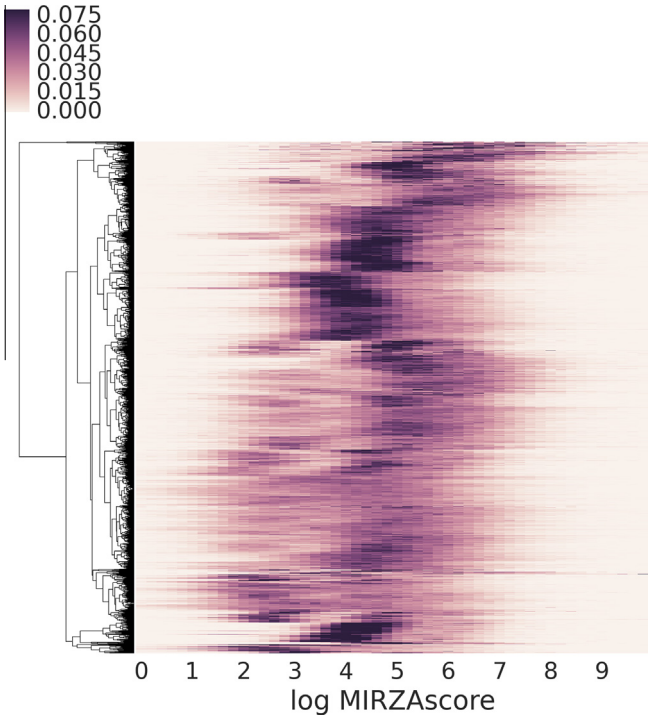

Fig. 8. Distribution of the MIRZA quality scores of target sites of individual miRNAs. Each line corresponds to one miRNA and the intensity of the color indicates the density of target sites within a particular range of target quality scores, computed with MIRZA–CLIP.

Table 2
Data sets of mRNA expression changes following miRNA transfection that were used to test the MIRZA models.

References	Data source (Gene Expression Omnibus (GEO) accession/URL)	miRNAs in the data set
Dahiya et al. [31]	GSE10150	miR-200c, miR-98
Frankel et al. [32]	GSE31397	miR-101
Gennarino et al. [33]	GSE12100	miR-26b, miR-98
Hudson et al. [34]	GSE34893	miR-106b
Leivonen et al. [35]	GSE14847	miR-206, miR-18a, mir-193b, miR-302c
Linsley et al. [36]	GSE683	miR-103, miR-215, miR-17, miR-192, let-7c, miR-106b, miR-16, miR-20, miR-15a, miR-141, miR-200a
Selbach et al. [37]	http://psilac.mdc-berlin.de/download/	miR-155, let-7b, miR-30a, miR-1, miR-16
Olive et al. [38]	GSE53225	miR-92a

3.5. Inferring a MIRZA model from biochemical data

The results presented above indicate that the MIRZA–CLIP/CHIMERA models explain well both the biochemical data as well as the response of mRNAs to miRNA transfection. However,

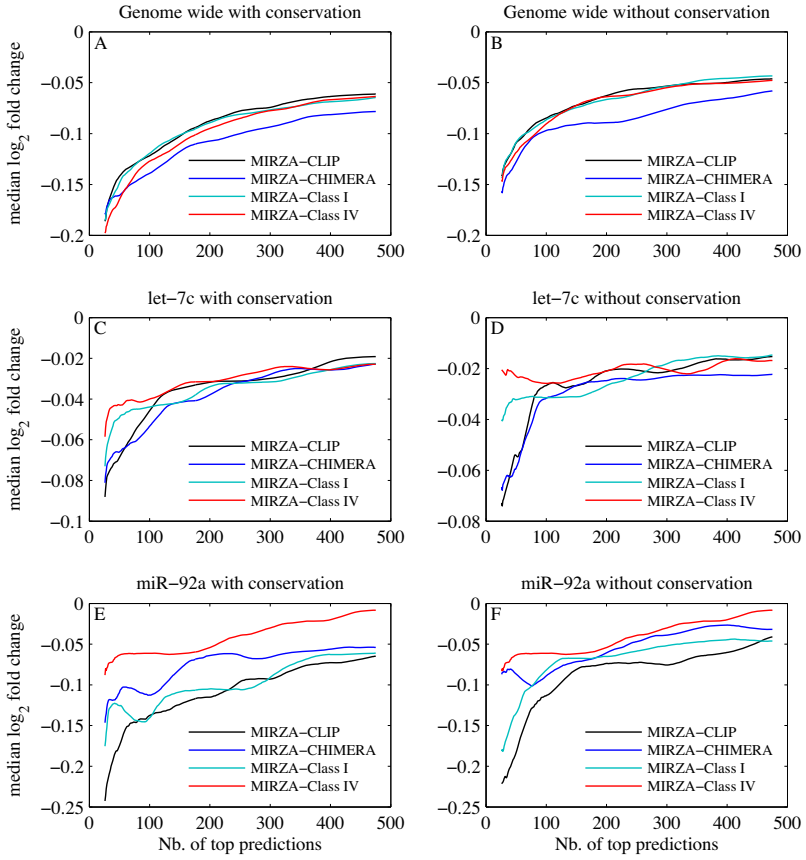

Fig. 9. Relationship between prediction score and the extent of mRNA downregulation. Genome-wide target predictions were carried out with the MIRZA-G generalized linear model [16], within which the target quality scores were calculated with different MIRZA variants: MIRZA–CLIP, MIRZA–CHIMERA, MIRZA-Class I and MIRZA-Class IV. Measurements of mRNA expression in control and miRNA-transfected cells were used to determine the log2 fold-changes of predicted miRNA targets. (A) Median log2 fold-change of the top *N* targets of the transfected miRNA, in function of *N*, were averaged over a data set of 26 miRNA transfection experiments. (C) Same procedure, but showing the median log2 fold-change of predicted let-7 targets upon let-7 transfection (Table 2, data from [36] and [37]). (E) Same procedure, but showing the median log2 fold-change of predicted targets of the mouse miR-92a upon miR-92a transfection in mouse cells (Table 2, data from [38]). For (A), (C) and (E), genome-wide predictions were carried out including evolutionary conservation whereas for (B), (D) and (F), without [16].

given the complexity of CLIP experiments and the indirect nature of the resulting data, one wonders whether an even more accurate model of miRNA–target interaction could not be derived from *in vitro* measurements of interaction affinity as obtained in the study of Wee et al. [26]. To gain further insight into the design of an efficient experiment, we generated synthetic data sets of hybrids, computed their pseudo-energies of interaction with MIRZA–CLIP, and then asked how our ability to recover the model parameters from the synthetic data sets depends on the number and type of hybrids and the accuracy of the provided pseudo-energies.

First, we simulated the experimental design of Wee et al. [26], in which energies of interaction between close variants of a single miRNA (let-7) and their perfectly complementary sequences were measured. There are 1890 possible two point-mutants of let-7, from which we sampled datasets of different sizes. An alternative design is to measure the energies of interaction between 'random' small RNAs and their partially complementary sequences. In this approach the small RNA is an entirely 'random' sequence whereas the interacting site is a sequence whose complementarity to the small RNA varies. To construct it, we first chose the average number of complementary nucleotides. With probabilities of complementarity chosen uniformly between 0.25 and 1, we can simulate from interactions of random RNA fragments to interactions of perfectly complementary sequences. This second approach is meant to provide datasets containing more information in terms of pairs of interacting nucleotides than the first approach. For both methods, while constructing subsets of various sizes, we aimed to cover uniformly the space of interaction energies and of nucleotide positions involved in the binding. Finally, we considered the possibilities that the measurements are not entirely accurate. To simulate this, we added gaussian noise to the computed interaction energy for each hybrid with a standard deviation of 0 (no noise), 1%, 5% and 10% of the predicted energy of interaction. For each data set size and each noise level we generated 100 synthetic data sets. To each synthetic data set we applied the simulated annealing procedure that was described in Khorshid et al. [19] to recover the parameters of the MIRZA model used to generate the pseudo-energies. The results, averaged over the 100 replicates of each setting, are shown in Fig. 10. They indicate that if the measurement noise is less than 10%, ~250 hybrids, chosen from across the entire range of expected affinities would be sufficient to recover the model parameters with reasonable accuracy (root mean square difference, RMSD, between recovered and input parameters < 1). If the measurements were very precise (relative error of a few percent), the number of hybrids necessary to recover a model with RMSD < 1 is considerable smaller, ~100, which is within reach with the technology available today. The experimental design of measuring closely related variants of a single miRNA does not yield equally accurate parameter values from a comparable number of hybrids, presumably due to the limited sampling of nucleotide/position combinations.

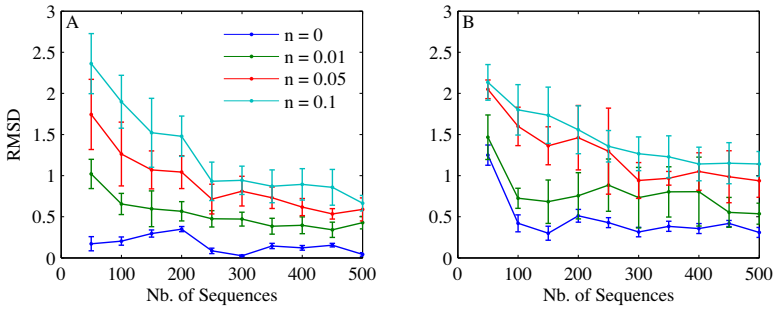

Fig. 10. Root mean square difference (RMSD) between the MIRZA parameters used to generate the training set and the MIRZA parameters inferred from the training data, as a function of the size of the training set. The colors correspond to the noise added to the training set data (0%, 1%, 5% and 10% of the predicted energy value). For (A), the data sets were generated with the 'randomized' procedure, whereas for (B), the data sets were generated through mutations of the let-7 miRNA.

4. Discussion and perspective

That miRNAs are important for the proper development and function in a large number of species is undisputed. Similar to transcription regulation by transcription factors, miRNA-dependent regulation is 'combinatorial'. That is, a regulator typically has many targets and a target is affected by many regulators. In contrast to transcription factors, miRNAs induce milder changes in target expression, which makes it more difficult to distinguish *bona fide* regulatory effects from biological or experimental variability. Consequently, a number of distinct directions are pursued in the field. Many groups have started to explore functional consequences of miRNA–target interaction that go beyond the repression of a single miRNA target into dynamical aspects of the response of a larger network, containing multiple miRNAs and multiple targets [9–11,39,40]. Such a network is quite complex and can exhibit very rich behaviors. For example, a recent study emphasized that even an increased expression of some miRNA targets can be expected in response to the increased expression of a miRNA. This could happen if miRNAs with different efficiencies in target down-regulation compete for the same sites on the target, because over-expression of the miRNA that is less effective in repressing the target could lead to the displacement of the miRNA that is more effective and thus to a net increase in target expression [41]. Additional experiments are necessary to determine whether this behavior occurs *in vivo*.

More generally, given the wide range of behaviors that computational models can predict, it is important to sufficiently constrain them with accurate parameters. Indeed, as described in previous sections, recent studies have started to provide measurements of the concentrations and the rate of interactions between the relevant molecular players. Our work shares this aim. Up to this point we used high-throughput data sets of miRNA binding sites that were derived with various approaches to parameterize a model of miRNA–target interaction. This model allows us to compute the energy of interaction between miRNAs and arbitrary target sites and to carry out genome-wide predictions of miRNA targets. We have shown that the model inferred from sequenced Argonaute/miRNA binding sites predicts quite accurately hybrid energies that are measured with biochemical methods *in vitro*. Furthermore, we have proposed a strategy for deriving a MIRZA-like model from biochemical measurements that can be obtained with the technology available today.

Although on its own, the energy of miRNA–target interaction is not sufficiently predictive of functional interactions, it is one of several informative features that together enable fairly accurate transcriptome-wide predictions. These additional features reflect the secondary structure of the target mRNA, its interactions with RNA-binding proteins, as well as other factors that are yet not understood but can be captured in the degree of evolutionary conservation of the putative miRNA binding site. Dynamical changes in the miRNA targetome between cell types or cell states will remain difficult to model computationally, but they may be important for the interpretability of experimental data. For example, it has been shown that taking into account tissue/condition-specific isoform expression can improve the prediction of miRNA targets [42], because alternative polyadenylation can change the susceptibility of transcripts to miRNA regulation. Conversely, miRNA stability is also subject to regulation, e.g. by addition of nucleotides (especially of uridine and adenine) at the 3′ end [43]. Argonaute protein modifications, mainly phosphorylation, provide another layer of regulation, relieving target repression or changing the subcellular localization [44]. Nevertheless, the approach that we presented here provides the basis on which more complex, context-specific and even dynamical models describing the impact of miRNA regulation on cellular function can be developed.

Acknowledgments

Jeremie Breda is a Werner-Siemens fellow at Biozentrum and Rafal Gumienny is supported by the Marie Curie Initial Training Network RNPnet project (#289007) from the European Commission. This work was also supported by SystemsX.ch, the systems biology initiative in Switzerland through the RTD project StoNets.

References

[1] E. Huntzinger, E. Izaurralde, Nat. Rev. Genet. 12 (2011) 99–110.
[2] B.P. Lewis, C.B. Burge, D.P. Bartel, Cell 120 (2005) 15–20.
[3] S.W. Eichhorn, H. Guo, S.E. McGeary, R.A. Rodriguez-Mias, C. Shin, D. Baek, et al., Mol. Cell 56 (2014) 104–115.
[4] E. Levine, Z. Zhang, T. Kuhlman, T. Hwa, PLoS Biol. 5 (2007) e229.
[5] S. Mukherji, M.S. Ebert, G.X.Y. Zheng, J.S. Tsang, P.A. Sharp, A. van Oudenaarden, Nat. Genet. 43 (2011) 854–859.
[6] N.E. Buchler, M. Louis, J. Mol. Biol. 384 (2008) 1106–1119.
[7] J. Hausser, M. Zavolan, Nat. Rev. Genet. 15 (2014) 599–612.
[8] A.A. Khan, D. Betel, M.L. Miller, C. Sander, C.S. Leslie, D.S. Marks, Nat. Biotechnol. 27 (2009) 549–555.
[9] A.D. Bosson, J.R. Zamudio, P.A. Sharp, Mol. Cell 56 (2014) 347–359.
[10] M. Figliuzzi, E. Marinari, A. De Martino, Biophys. J. 104 (2013) 1203–1213.
[11] C. Bosia, A. Pagnani, R. Zecchina, PLoS One 8 (2013) e66609.
[12] U. Bissels, S. Wild, S. Tomiuk, A. Holste, M. Hafner, T. Tuschl, et al., RNA 15 (2009) 2375–2384.
[13] J. Chang, E. Nicolas, D. Marks, C. Sander, A. Lerro, M.A. Buendia, et al., RNA Biol. 1 (2004) 106–113.
[14] D. Wang, Z. Zhang, E. O'Loughlin, T. Lee, S. Houel, D. O'Carroll, et al., Genes Dev. 26 (2012) 693–704.
[15] D. Grün, L. Kester, A. van Oudenaarden, Nat. Methods 11 (2014) 637–640.
[16] R. Gumienny, M. Zavolan, Nucleic Acids Res. (2015).
[17] S.W. Chi, J.B. Zang, A. Mele, R.B. Darnell, Nature 460 (2009) 479–486.
[18] M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, et al., Cell 141 (2010) 129–141.

[19] M. Khorshid, J. Hausser, M. Zavolan, E. van Nimwegen, Nat. Methods 10 (2013) 253–255.
[20] A. Helwak, G. Kudla, T. Dudnakova, D. Tollervey, Cell 153 (2013) 654–665.
[21] S. Grosswendt, A. Filipchyk, M. Manzano, F. Klironomos, M. Schilling, M. Herzog, et al., Mol. Cell 54 (2014) 1042–1054.
[22] E. Elkayam, C.-D. Kuhn, A. Tocilj, A.D. Haase, E.M. Greene, G.J. Hannon, et al., Cell 150 (2012) 100–110.
[23] T. Xia, J. SantaLucia Jr., M.E. Burkard, R. Kierzek, S.J. Schroeder, X. Jiao, et al., Biochemistry 37 (1998) 14719–14735.
[24] S. Kishore, L. Jaskiewicz, L. Burger, J. Hausser, M. Khorshid, M. Zavolan, Nat. Methods 8 (2011) 559–564.
[25] X. Wang, Bioinformatics 30 (2014) 1377–1383.
[26] L.M. Wee, C.F. Flores-Jasso, W.E. Salomon, P.D. Zamore, Cell 151 (2012) 1055–1067.
[27] J.S. Reuter, D.H. Mathews, BMC Bioinformatics 11 (2010) 129.
[28] J. Hausser, M. Landthaler, L. Jaskiewicz, D. Gaidatzis, M. Zavolan, Genome Res. 19 (2009) 2009–2020.
[29] S. Vasudevan, Y. Tong, J.A. Steitz, Science 318 (2007) 1931–1934.
[30] A.L. Jackson, J. Burchard, J. Schelter, B.N. Chau, M. Cleary, L. Lim, et al., RNA 12 (2006) 1179–1187.
[31] N. Dahiya, C.A. Sherman-Baust, T.-L. Wang, B. Davidson, I.-M. Shih, Y. Zhang, et al., PLoS One 3 (2008) e2436.
[32] L.B. Frankel, J. Wen, M. Lees, M. Høyer-Hansen, T. Farkas, A. Krogh, et al., EMBO J. 30 (2011) 4628–4641.
[33] V.A. Gennarino, M. Sardiello, R. Avellino, N. Meola, V. Maselli, S. Anand, et al., Genome Res. 19 (2008) 481–490.
[34] R.S. Hudson, M. Yi, D. Esposito, S.A. Glynn, A.M. Starks, Y. Yang, et al., Oncogene 32 (2012) 4139–4147.
[35] S.-K. Leivonen, R. Mäkelä, P. Ostling, P. Kohonen, S. Haapa-Paananen, K. Kleivi, et al., Oncogene 28 (2009) 3926–3936.
[36] P.S. Linsley, J. Schelter, J. Burchard, M. Kibukawa, M.M. Martin, S.R. Bartz, et al., Mol. Cell. Biol. 27 (2007) 2240–2252.
[37] M. Selbach, B. Schwanhäusser, N. Thierfelder, Z. Fang, R. Khanin, N. Rajewsky, Nature 455 (2008) 58–63.
[38] V. Olive, E. Sabio, M.J. Bennett, C.S. De Jong, A. Biton, J.C. McGann, et al., Elife 2 (2013) e00822.
[39] L. Poliseno, L. Salmena, J. Zhang, B. Carver, W.J. Haveman, P.P. Pandolfi, Nature 465 (2010) 1033–1038.
[40] R. Denzler, V. Agarwal, J. Stefano, D.P. Bartel, M. Stoffel, Mol. Cell 54 (2014) 766–776.
[41] D. Nyayanit, C.J. Gadgil, RNA (2015).
[42] J.-W. Nam, O.S. Rissland, D. Koppstein, C. Abreu-Goodger, C.H. Jan, V. Agarwal, et al., Mol. Cell 53 (2014) 1031–1043.
[43] Y.-K. Kim, I. Heo, V.N. Kim, Cell 143 (2010) 703–709.
[44] M. Ha, V.N. Kim, Nat. Rev. Mol. Cell Biol. 15 (2014) 509–524.