Incorporating structure to predict microRNA targets

Harlan Robins*†, Ying Li‡, and Richard W. Padgett‡

*Institute for Advanced Study, Olden Lane, Princeton, NJ 08540; and ‡Department of Molecular Biology and Biochemistry, Waksman Institute, Cancer Institute of New Jersey, Rutgers, The State University of New Jersey, Piscataway, NJ 08854-1020

Communicated by Arnold J. Levine, Institute for Advanced Study, Princeton, NJ, January 31 2005 (received for review December 13, 2004)

MicroRNAs (miRNAs) are a recently discovered set of regulatory genes that constitute up to an estimated 1% of the total number of genes in animal genomes, including *Caenorhabditis elegans*, *Drosophila*, mouse, and humans [Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. (2001) *Science* 294, 853–858; Lai, E. C., Tomancak, P., Williams, R. W. & Rubin, G.M. (2003) *Genome Biol.* 4, R42; Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. (2001) *Science* 294, 858–862; Lee, R. C. & Ambros, V. (2001) *Science* 294, 862-8644; and Lee, R. C., Feinbaum, R. L. & Ambros, V. (1993) *Cell* 115, 787–798]. In animals, miRNAs regulate genes by attenuating protein translation through imperfect base pair binding to 3′ UTR sequences of target genes. A major challenge in understanding the regulatory role of miRNAs is to accurately predict regulated targets. We have developed an algorithm for predicting targets that does not rely on evolutionary conservation. As one of the features of this algorithm, we incorporate the folded structure of mRNA. By using *Drosophila* miRNAs as a test case, we have validated our predictions in 10 of 15 genes tested. One of these validated genes is *mad* as a target for *bantam*. Furthermore, our computational and experimental data suggest that miRNAs have fewer targets than previously reported.

mRNA structure | target prediction

MicroRNAs are a class of small, ≈22-nt RNAs that share properties with silencing RNAs (1). In plants, most microRNA (miRNA) genes bind sequences perfectly and lead to mRNA degradation (2, 3). However, in animals, with a notable exception (4), they function by preventing translation without mRNA degradation (5, 6). The mechanism by which the bound miRNA down-regulates translation of its target mRNA remains unknown. Currently, only a handful of miRNAs have experimentally determined function *in vivo*. These miRNAs include *lin-4* and *let-7* in *Caenorhabditis elegans*, *bantam*, and *mir-14* in *Drosophila*, and *mir-23* in humans, playing vital roles in development and apoptosis (7–14). Even this modest set of data has some discernable common features that partially determine a set of rules governing the binding of miRNAs to their targets. It has been observed that toward the 5′ end of the miRNA there is a perfect Watson–Crick base pair matching of at least seven consecutive nucleotides (15). Recent experimental evidence has added more insights into the 3′ UTR-binding rules (1, 16), but a complete understanding of miRNA–target interactions is not known. Because miRNA genes control many cellular processes, it is important to identify their targets with high accuracy.

Incorporating the experimentally determined features and deduced rules, we developed an algorithm for predicting miRNA targets in animals that significantly reduces dependence on evolutionary homology without sacrificing accuracy. The algorithm consists of four parts; (*i*) the 5′ seven nucleotides, (*ii*) scoring the match of the entire miRNA, (*iii*) incorporating 3′ UTR structure of the target, and (*iv*) combining scores for multiple sites in the targets. Applying the algorithm to *Drosophila melanogaster*, we analyzed 73 miRNAs from the MiRNA registry (which can be accessed at www.sanger.ac.uk/Software/Rfam/mirna/index.shtml) and the 3′ UTRs of 9,230 transcripts from Ensembl's Ensmart (which can be accessed at www.ensembl.org). A list of miRNAs and their predicted targets in the order ranked by our algorithm is in Table 2, which is published as supporting information on the PNAS web site.

Materials and Methods

To calculate the *P* value giving the probability that the correlation between free bases and real binding is random, we first folded the 3′ UTR from *C. elegans lin-28*, *lin-41*, *lin-14*, *daf-12*, and *Drosophila Hid*. Then, we counted the possible binding positions in all of these genes that would give an overlap of three or more bases between the seven seed nucleotides and a region of free bases (in a loop or bubble). Dividing the total number of positions by the total number of nucleotides in the 3′ UTRs gives a probability that one random seed would overlap a freebase region. This probability is 0.228. Because 12 of the 19 binding sites we are considering have seeds that overlap free bases, we compute the probability of getting 12 or more of 19, given a probability of 0.228 for each event. The result is the $P = 0.0002$.

To validate the predicted targets of *Drosophila* miRNAs, reporter assay *Drosophila* S2 cells are used to monitor changes in gene expression. First, we constructed a sensor for each target gene by replacing the 3′ UTR of firefly (*Photinus pyralis*) *luciferase* (*Pp-luc*) with the 3′ UTR of the target gene under the control of *Drosophila actin* promoter. *Pp-luc* alone in the same expression vector was used as negative control. To generate the miRNA expression constructs, miRNA genes and 100–200 bp of flanking DNA were amplified from *Drosophila* genomic DNA by PCR and cloned into vectors. Expression of the miRNA genes was induced by the *Drosophila actin* promoter. All of the miRNA gene constructs were confirmed by sequencing.

Transient transfections into S2 were used to determine the effect of the miRNA gene on the expression levels of the firefly *luciferase*. The ratio between the firefly and the *renilla reniformis luciferase* (*Rr-luc*) was used as an internal control for transfection efficiency. Three days after transfection, the activities of *Pp-luc* and *Rr-luc* were determined by the Dual-Glo luciferase assay (Promega). Each experiment was repeated three times, and the averages were used in comparisons.

Results and Discussion

Observing the experimentally determined miRNA target sites in *lin-14*, *daf-12*, and *lin-41* in *C. elegans* and *hid* in *Drosophila*, it was noticed that at the 5′ end of the miRNA there is a perfect match of at least seven consecutive nucleotides (dubbed the seed). The necessity of this match in the functionality of a target site has been confirmed through direct experiment (16). For each miRNA, we use the reverse complement of the sets of seven nucleotides in a row that end within the last three bases of the miRNA. This seed is used to establish a first cut of possible targets by searching the set of 3′ UTRs from *Drosophila* for matches to these seeds.

Drawing again on both the observation of known target sites and recent direct experimental tests, we wrote a recursive program to score the entire binding site. The nonseed part of the miRNAs bind imperfectly to their targets but contribute to the

Abbreviation: miRNA, microRNA.

†To whom correspondence should be addressed. E-mail: hrobins@ias.edu.

overall stability. Given the small binding window of the miRNA, the known target sites form many more Watson–Crick base pairs than randomly expected. However, we cannot simply rank binding sites according to lowest binding energy for a couple reasons. First, the paper of Doench and Sharp (16) provides evidence that G-U pairs do not contribute to the effectiveness of a binding site outside the seed and significantly reduce the binding if a G-U pair is found within the seed. Second, the known binding sites have not evolved to minimize binding energy. Therefore, we set up a scoring algorithm that weights A-U and G-C pairs positively, treats G-U pairs as neutral, and penalizes mismatches and gaps. Applying this scoring algorithm to the known target sites, we then choose a cutoff such that all these sites score robustly above the cutoff. We define robust to mean that a single change in the binding site, outside of the seed, would not be able to move any of the known sites below the cutoff.

The above two criteria reduce the list of targets to a few hundred. Additional reductions in the target list can be made by examining the structure of the target 3′ UTR. Folded mRNA consists of nucleotides that are base-paired and those that are free. We hypothesize that single-stranded miRNAs can only search stretches of free mRNA for potential target sites. According to Boltzmann's rules, the binding probability is proportional to the exponential of the difference in binding energies of the two states. If a stretch of RNA is unbound in one state and bound in the other, the probability of binding is relatively high. On the other hand, if the mRNA is folded so that the site of interest is based paired with another part of the mRNA, then the energy difference between the two states is smaller, and the binding probability is smaller. Of course, there are proteins wrapping the miRNA that could potentially play a role in recognition. However, there is no evidence that the relevant proteins recognize either sequence or structure of the mRNA targets.

To test this hypothesis, we folded the 3′ UTRs of the known targets in *C. elegans* and *Drosophila* and calculated the probability that the known target sites were correlated with the free nucleotides from the folded target. For the known targets of *C. elegans*, we used those listed in Banerjee and Slack (6) that meet the above two criteria, and, for *Drosophila*, we used the five bantam targets in *hid* listed in Brennecke *et al.* (14). Specifically, we required that of the seven seed nucleotides in the miRNA, at least three consecutive bases paired with free bases from the 3′ UTR. We chose three bases for the following two reasons. First, the minimal length of an RNA hairpin loop is three nucleotides, which is a physical constraint from the limited flexibility of RNA. Second, recognition for base pairing of free strands require three consecutive complementary bases, and a string of two matches, then a mismatch, will not form a double strand. Because the folding algorithm is prone to error on a global scale for long sequences, we focus on the local stems produced from the folding. We restricted our set of free bases to those found in the loop at the end of a stem or the bubble located at the base of the stem. Given this restriction, we calculate the P value of 0.0002 as the probability that the correlation between known seeds and free bases is random (see *Materials and Methods*). This structural requirement for our target sites removes 80% of the false sites, whereas we lose only one-third of our real binding sites. These statistics are determined from the experimentally verified targets mentioned above. This is a substantial gain in accuracy because most real target mRNAs have multiple sites and we improve by a factor of three for each site; so, our algorithm folds the 3′ UTRs of all of the *Drosophila* genes by using the VIENNA FOLDING package and then throws out all potential targets that do not have an overlap with free bases as described above (17). We need only fold the 3′ UTR because we are looking at local structure (not global), and the performance of the folding algorithm decreases dramatically as the sequence length increases. There are many other ways in which we can take structure into account such as considering alternative foldings. However, we would require more known targets to get solid statistics. We hope to improve the use of structure as more targets are discovered.

Table 1. Tested miRNA targets

Target gene	miRNA	Rank	Repressed
MAD	Bantam	4	Yes
Hid	Bantam	1	Yes
CRMP	Mir-287	2	Yes
HLHm5	Mir-7	1	Yes
SP555	Mir-279	1	Yes
Imd	Mir-310	3	Yes
Tutl	Mir-1	1	Yes
Su(z) 12	Mir-34	3	Yes
Rt	Mir-12	1	Yes
Gli	Mir-124	1	Yes
Fng	Mir-7	3	Yes
DIP1	Mir-287	1	No
CG14991	Mir-303	1	No
tup	Mir-278	2	No
Yellow-c	Mir-317	3	No
CG13380	Mir-318	2	No
Boss	Mir-286	5	No
CG32057	Mir-288	8	No
Ke1	Mir-276b	6	No
la2	Mir-316	30	No

The final part of the algorithm ranks the remaining targets by computing a combined score for multiple sites within one 3′ UTR. The known targets have multiple binding sites in their 3′ UTR, and experimental evidence supports cooperative effects with multiple sites in each 3′ UTR (1). Fitting to the experimentally generated curves from Doench *et al.* (1), we sum the scores and then take the result to the power of 1.2.

Having partially based our algorithm on observations of known targets, it is required that these targets score highly when our algorithm is applied. We met this consistency check successfully. In *C. elegans*, *lin-14* and *daf-12* were two of the top three ranking targets of miRNA *let-7*, whereas *lin-14* was also the top ranking target of miRNA *lin-4*. In *Drosophila*, *hid* ranked first as a target for the miRNA *bantam*.

We tested 19 potential targets predicted by our algorithm by use of a reporter gene in *Drosophila* S2 cells. The 3′ UTR of the firefly *luciferase* gene was replaced with the 3′ UTRs of the *Drosophila* targets and transfected into *Drosophila* cells (see *Materials and Methods*). Each experiment was repeated three times. Table 1 contains our algorithm's predictions regarding these 19 targets. Fifteen of the 19 targets were high-scoring targets that were chosen to represent the group of targets that scored in the top four for some miRNA. These 15 targets tested the validity of the algorithm. Ten of the 15 targets showed significant repression when the corresponding miRNA was expressed (Fig. 1). For the five targets that failed, we tested the miRNA constructs to confirm that they were functioning. We used the *bantam*/*hid* pair as a control because this result has been verified *in vivo*. Our result is that the algorithm predicts the top four targets for each miRNA with ≈67% accuracy. Because the validation is in cell lines, the positive results provide evidence that the regulation has a functional role in live animals.

Three of the remaining four tested targets were chosen randomly from the group that ranked between 5 and 10 for their respective miRNAs and the final target ranked 30th for its miRNA. Experiments in cell culture showed no effect of the miRNA on the presumed targets. By using Fisher's exact test, we

CELL BIOLOGY

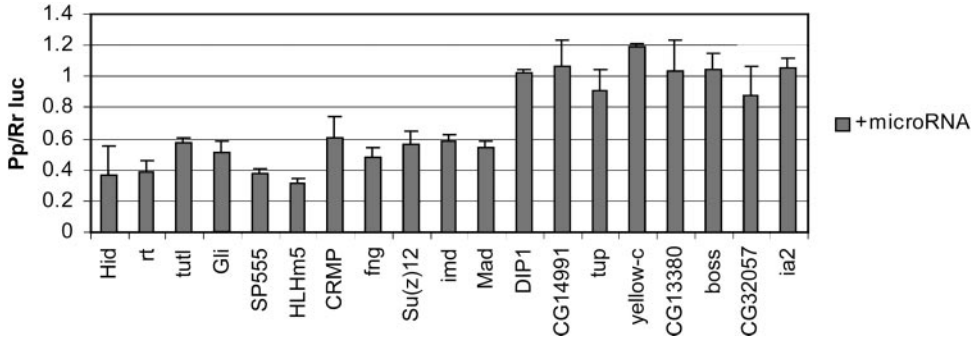

Fig. 1. A graph of *luciferase* reporter intensity from miRNA target genes. The 3′ UTR targets are from the genes listed on the *x* axis. The particular miRNA that pairs with each gene is found in Table 1. As a control, we use *hid* as the target of *bantam*. The luciferase activity before expressing the miRNAs were normalized to 1 for all cells, so the values in the bar graph are the fraction of luciferase intensity with the miRNA expressed. Each experiment was repeated three times, given the error bars. The 11 targets on the left are regulated by a miRNA, and the 10 on the right are not.

can say with 93% certainty that the median number of targets for each miRNA is 10 or fewer. Additionally, we can say with 97% confidence that the median number is <30. These experiments address the question of how many targets a given miRNA is likely to have. Our results suggest that the number is smaller than previously thought.

Because our accuracy is sufficiently high, we were able to avoid requiring a cut on homology. Although homology can help improve accuracy, it comes at the expense of losing real targets. Because the *pseudoobscura* genome has not been completely annotated, we run into two major problems when trying to apply homology. The first problem is that the length of the 3′ UTRs are not known, so we can only approximate the length. This problems is difficult because *Drosophila* 3′ UTRs vary widely. Second, approximately one-fifth of the time the ESTs cut in the middle of a 3′ UTR, prohibiting us from checking homology. Given these two limitations, we lose a substantial percentage of real targets.

As a representative example, one of the targets we validated is *mad* regulated by *bantam*, which we would not have found had we required a homology cut. The miRNA *bantam* has been shown function in two processes (14). It prevents apoptosis by down-regulating the apoptotic gene *hid*. Also, mutants in *bantam* increase cell proliferation, but the target gene that interfaces with cell-cycle control is unknown. In our studies, we found that *mad* is a target of *bantam*. Although bantam represses the *mad* reporter to the same extent that it represses *hid* in our control, we wanted to confirm that the cause of repression was because of the *bantam*-binding sites in the *mad* 3′ UTR. We made point mutations in the fourth and fifth positions (as read from 3′ to 5′) of the two *bantam*-binding sites in the 3′ UTR of *mad*. Doench *et al.* (1) showed that mutating the fourth and fifth positions in a target site was sufficient to eliminate binding. Transfecting the points mutants into *Drosophila* S2 cells as described above, we find that mutating one site partially restores the level of luciferase in the presence of *bantam*, whereas mutating both binding sites completely restores the level (Fig. 2).

Because *mad* is involved in propagating *decapentaplegic* signals, which promote proliferation in the fly, it is unlikely that the *bantam*/*mad* interactions are involved in the cell-cycle regulation observed for *bantam*. Possibly more than one of the seven TGFβ-like ligands signal through *mad*, raising the possibility that the *bantam*/*mad* interaction affects a different TGFβ-like pathway. Alternatively, the *bantam*/*mad* interaction may function through *decapentaplegic*, but in a different developmental process. Further *in vivo* experiments are warranted to examine this interaction.

To date, three algorithms are published for finding miRNA targets from a whole genome: two in *Drosophila* and one in vertebrates (15, 18, 19). Two other algorithms have been applied to specific genes or miRNAs (20, 21). Only Lewis *et al.* (15) estimated and tested a false-positive rate for targets. Lewis *et al.* (15) tested their algorithm in humans and established a success rate of approximately two-thirds, aiming for an accurate, as opposed to comprehensive, list of targets. Their success hinged strongly on homology, limiting their targets to those where homologous genes in both mouse and rat were also predicted as strong targets of homologous miRNAs. Using either mouse or rat, but not both, dramatically drops the success rate of their algorithm.

For the two *Drosophila* algorithms, we are able to directly compare results. To do the comparison, we focused on the experimentally validated genes and their corresponding miRNAs. The Enright *et al.* (18) algorithm has almost no overlapping results with our predictions. In particular, their algorithm scored only one of the 10 targets we validated experimentally in their list of the top 10 for their partner miRNAs. Because they did not experimentally validate any of their results, we are unable to run the comparison in the other direction.

The Stark *et al.* (19) algorithm provides a large list of targets of for each miRNA without determining accuracy. They chose six targets to validate partially based on their algorithm and partially based on biological intuition. Of their six validated targets, our algorithm ranks three of them in the top 10 for their partner miRNAs and two others in the top 20. One of their targets allows us to demonstrate the gain that we achieve from our structure cut. Stark *et al.* (19) validated *reaper* as a target for

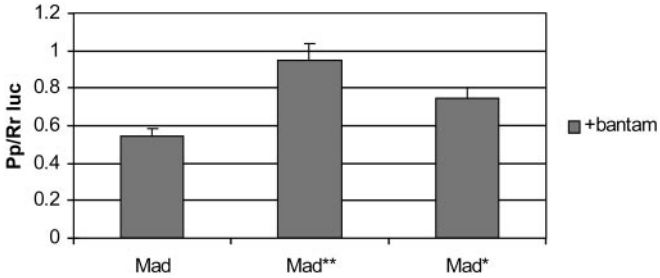

Fig. 2. Both of the two *bantam*-binding sites on the 3′ UTR of *mad* are shown to contribute to repression. Positions 4 and 5 counting from the 5′end of the binding site are mutated in one (*) and two (**) bantam-binding sites, knocking out the sites. Knocking out one binding site partially restores the activity of mad (*), whereas knocking out two binding sites completely restores activity (**).

mir-2a. Our algorithm ranks *reaper* as the number one target for *mir-2a*. If we run our algorithm without including the structure cut, *reaper* drops to 25th. Next, we compared their predictions with our validated results. From their list of the top 100 targets for each miRNA, 2 of our 10 targets scored well (in the top three), 2 scored in the 20s, and the 5 others did not make their top 60 (the final target was *mad*, and they did not publish their *bantam* targets).

We have presented an algorithm that provides a substantial increase in accuracy for predicting miRNA targets. As more experimental data becomes available to elucidate the binding rules of miRNAs to their targets, we expect to improve our algorithm.

H.R. thanks Hagar Barak and Arnold Levine for many useful discussions and helpful comments regarding the project. R.W.P. thanks Jizong Gao and Huang Wang for excellent technical assistance. We also thank Andrei Ruckenstein for useful discussions in the initial phases of this project. This work was supported by a grant from the National Institutes of Health and an Academic Excellence Grant from Rutgers University (to R.W.P.), a Busch Predoctoral Fellowship (to Y.L.), and in part by the Shelby White and Leon Levy Initiatives Fund and the David and Lucile Packard Foundation (to H.R).

1. Doench, J. G., Petersen, C. P. & Sharp, P. A. (2003) *Genes Dev.* **17,** 438–442.
2. Bartel, D. P. (2004) *Cell* **116,** 281–297.
3. Chen, X. (2004) *Science* **303,** 2022–2025.
4. Yekta, S., Shih, I. H. & Bartel, D. P. (2004) *Science* **304,** 594–596.
5. Ambros, V. (2001) *Cell* **107,** 823–826.
6. Banerjee, D. & Slack, F. (2002) *BioEssays* **24,** 119–129.
7. Lee, R. C. & Ambros, V. (2001) *Science* **294,** 862–864.
8. Wightman, B., Ha, I. & Ruvkun, G. (1993) *Cell* **75,** 855–862.
9. Moss, E. G., Lee, R. C. & Ambros, V. (1997) *Cell* **88,** 637–646.
10. Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., Horvitz, H. R. & Ruvkun, G. (2000) *Nature* **403,** 901–906.
11. Abrahante, J. E., Daul, A. L., Li, M., Volk, M. L., Tennessen, J. M., Miller, E. A. & Rougvie, A. E. (2003) *Dev. Cell* **4,** 625–637.
12. Lin, S.Y., Johnson, S.M., Abraham, M., Vella, M.C., Pasquinelli, A., Gamberi, C., Gottlieb, E. & Slack, F.J. (2003) *Dev. Cell* **4,** 639–650.
13. Xu, P., Vernooy, S. Y., Guo, M. & Hay, B. A. (2003) *Curr. Biol.* **13,** 790–795.
14. Brennecke, J., Hipfner, D. R., Stark, A., Russell, R. B. & Cohen, S. M. (2003) *Cell* **113,** 25–36.
15. Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. (2003) *Cell* **115,** 787–798.
16. Doench, J. G. & Sharp, P. A. (2004) *Genes Dev.* **18,** 504–511.
17. Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M. & Schuster, P. (1994) *Monatsh. Chem.* **125,** 167–188.
18. Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C. & Marks, D. S. (2003) *Genome Biol.* **5,** R1.
19. Stark, A., Brennecke, J., Russell, R. B. & Cohen, S. M. (2003) *PloS Biol.* **1,** 1–3.
20. Rajewsky, N. & Socci, D. (2004) *Dev. Biol.* **267,** 529–535.
21. Kiriakidou, M., Nelson, P. T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z. & Hatzigeorgiou, A. (2004) *Genes Dev.* **18,** 1165–1178.