Deposited research article

Computational identification of microRNA targets

Nikolaus Rajewsky* and Nicholas D Socci†‡

Addresses: * Department of Biology, New York University 1009 Main Building, 100 Washington Square East, New York, NY 10003-6688, USA. †Department of Pathology, and Seaver Foundation for Bioinformatics, Albert Einstein College of Medicine, 1300 Morris Park Ave, Bronx, NY 10461, USA. ‡Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, NY 10021, USA.

Correspondence: Nikolaus Rajewsky. E-mail: nikolaus.rajewsky@nyu.edu

Posted: 14 January 2004

Genome ***Biology*** 2004, **5:**P5

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/2/P5

Received: 12 January 2004

This is the first version of this article to be made available publicly, and no other version is available at present.

comment
reviews
reports
deposited research
refereed research
interactions
information

Computational identification of microRNA targets

Nikolaus Rajewsky[1,*] and Nicholas D. Socci[2,3]

[1] Department of Biology, New York University

1009 Main Building, 100 Washington Square East, New York, NY 10003-6688, USA

[2]Department of Pathology, and Seaver Foundation for Bioinformatics

Albert Einstein College of Medicine, 1300 Morris Park Ave, Bronx, NY 10461, USA

[3]Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, NY 10021, USA

[*] for correspondence: email nikolaus.rajewsky@nyu.edu, phone (212) 998-8271, fax (212) 995-4015

Submitted to Developmental Biology 12 Oct 2003, accepted for publication 1 Dec 2003 (*in press*).

ABSTRACT

Recent experiments have shown that the genomes of organisms such as worm, fly, human and mouse encode hundreds of microRNA genes. Many of these microRNAs are thought to regulate the translational expression of other genes by binding to partially complementary sites in messenger RNAs. Phenotypic and expression analysis suggest an important role of microRNAs during development. Therefore, it is of fundamental importance to identify microRNA targets. However, no experimental or computational high-throughput method for target site identification in animals has been published yet. Our main result is a new computational method which is designed to identify microRNA target sites. This method recovers with high specificity known microRNA target sites which previously have been defined experimentally. Based on these results, we present a simple model for the mechanism of microRNA target site recognition. Our model incorporates both kinetic and thermodynamic components of target recognition. When we applied our method to a set of 74 *Drosophila melanogaster* microRNAs, searching 3' UTR sequences of a predefined set of fly mRNAs for target sites which were evolutionary conserved between *Drosophila melanogaster* and *Drosophila pseudoobscura*, we found that a number of key developmental body patterning genes such as *hairy* and *fushi-tarazu* are likely to be translationally regulated by microRNAs.

Submitted to Developmental Biology 12 October 2003, accepted for publication 1 December 2003 (*in press*).

Keywords

microRNA, miRNA, computational, translational regulation, *cis*-regulatory sites, target sites, development, body patterning.

Introduction

MicroRNA (miRNA) genes are a new and large class of genes which do not encode proteins. They produce roughly 22 nucleotides long transcripts that in many cases are thought to function as antisense regulators of other mRNAs (Ambros, 2001). By now, hundreds of miRNAs in human, mouse, worm, and fly have been identified using molecular and computational approaches (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001; Ambros et al., 2003; Lim et al., 2003b,a; Lai et al., 2003; Aravin et al., 2003). So far, however, the biological function of miRNAs has been elucidated in only a few examples (lin-4 and let-7 in C. elegans, bantam and mir-14 in fly, and microRNA-23 in human). lin-4 and let-7 are involved in the timing of developmental processes (heterochronic genes), bantam has been shown to affect cell proliferation and death (Brennecke et al., 2003), mir-14 regulates the expression of the cell death pathway and fat metabolism (Xu et al., 2003). These results suggest a broad range of possible functions for miRNAs. The overall importance of miRNAs for development has been further established by the notion that many miRNAs appear to have temporal or tissue-specific patterns of gene expression (Houbaviy et al., 2003; Lau et al., 2001; Lagos-Quintana et al., 2002; Lim et al., 2003b; Ambros et al., 2003; Aravin et al., 2003).

In most known cases in animals, miRNAs regulate the *translational expression* of genes. miRNAs are thought to bind partially complementary sites in $3'$ untranslated regions of mature target mRNA in the cytoplasm and, by unknown mechanisms, to modulate (repress) translation of the target mRNA (for reviews, see Ambros, 2001; Moss and Poethig, 2002; Moss, 2002; Ambros, 2003; Carrington and Ambros, 2003 and Banerjee and Slack, 2002). To understand the biological function of miRNAs, it is necessary to identify their targets. No high-throughput experimental techniques for target site identification have been reported yet. Computational approaches have been successful in plants, where known target sites tend to be almost perfectly complementary to miRNAs (Rhoades et al., 2002) and where miRNAs are thought to promote *degradation* of the target mRNA (for a review, see Carrington and Ambros, 2003). In animals, however, the miRNA:mRNA base pairing appears to be less than perfect, which has greatly hindered computational approaches for target site identification.

Here, we report on a first computational method for miRNA target site identification in animals. We set out by compiling a set of experimentally reasonably well established target sites from the literature. Our dataset comprised 25 target sites (training set) for lin-4 and let-7 in *C. elegans*. We then found an algorithm which recovers most of these sites with high specificity when compared to random sites. A simple and intuitive model for kinetic and thermodynamic aspects of target site recognition is consistent with this method.

We applied our algorithm to a previously cloned and sequenced set of miRNAs in fly (Aravin et al., 2003; Lai et al., 2003). We find highly scoring, conserved putative target sites in several key developmental body patterning transcription factors such as *fushi-tarazu* and *hairy.* Further computational analysis with existing tools (Rajewsky et al., 2002) suggests that some of the miRNA genes such as mir-263b may have enhancers with binding sites for subsets of the body-patterning transcription factors.

Results

A new algorithm for the computational identification of microRNA target sites recovers known target sites with high specificity

A careful examination of our set of known miRNA binding sites revealed in most cases the presence of a GC rich string of *consecutive* base pairings with the miRNA. Based on this, we designed a simple scoring scheme that detects this "binding nucleus". The score for the nucleus is the weighted sum of consecutive basepairs (GC, AU, and GU). We fit these three parameters by maximizing the difference of the mean scores between the training set and random background sequences ("signal", see Materials & methods) divided by the standard deviation of the background scores ("noise"). We refer to these values as the Z-scores (see figure 1). The best fit was obtained with the weights $w_{GC} = 5$, $w_{AU} = 2$, $w_{GU} = 0$. These values can be scaled by an arbitrary factor. When varying w_{GU} only slightly, the Z-score decreases dramatically, while varying the ratio of w_{GC}/w_{AU} between 2–3 yields comparable Z-scores.

These fitted weights give us the best discrimination between the training set and the background. However, a cutoff value of the model score needs to be determined to set the threshold level for target site detection. Naturally, there is a tradeoff between sensitivity and specificity when setting this threshold. For example, for the training set a threshold value of 25 recovers 84% (21 of 25) of the training data, while detecting one false positive per 4000 bases of scanned target sequence. At a higher threshold level of 27 we recover half of the training set, and obtain only one false positive per 11000 bases of target sequence. However, the threshold level chosen for the miRNAs in the training set may not be optimal for other miRNAs. The optimal level may depend on the GC content of the microRNA as well as on other features of sequence composition. Therefore, the threshold score level is set for each miRNA independently by running it over a random sequence and recording the distribution of the scores. From this distribution one can compute p-values for scores. The threshold is then set by cutting off at a desired p-value. Figure 2 shows the score histogram for one of the training miRNAs (*lin-4*) and demonstrates the specificity of the nucleus score for recovering known target sites.

The size of the nucleus is typically 6-8 bases long and therefore represents less than half the total length of the miRNA. One might expect that one could improve the discrimination between target sites and random sites by incorporating sequence homology between a target site and the miRNA beyond the nucleus. Indeed, *in silico* hybridization of our training target sites to their miRNAs via MFOLD (see Materials & methods) suggests that for most of the RNA:RNA duplexes, a larger fraction of the miRNA is involved in base pairing. Thus, after the binding nucleus was located, a window of 40 bases was extracted from the target sequence and hybridized to the miRNA. The computed binding free energy value is then used to further filter for potential target sites. For example, the combination of nucleus score and free energy at a free energy cutoff of -17.4 kcal/mol further reduced the number of false positives by roughly 10% when detecting half of the known target sites. A few RNA:RNA duplexes in the training set appear to have exceptionally low free energies (lower than roughly -27 kcal/mol). Thus, the free energy can also be used to flag outstanding candidates. However, we note that the main contribution to the specificity of our algorithm stems from the nucleus score. We also note that we tried a great variety of different alignment procedures and screened the parameter space for each one to increase the specificity of the algorithm. None of these approaches outperformed the nucleus score or could efficiently replace MFOLD as a postprocessing step after the nucleus scoring.

The nucleus model recovers known fly *bantam* targets, and a functionally relevant base in *lin-4*

As a test of our nucleus score, we searched the 4017 nucleotides long *D. melanogaster hid* transcript (*head involution defective*, CG5123-RA, http://rail.bio.indiana.edu:7084/) for target sites for the microRNA *bantam* (Brennecke et al., 2003). In (Brennecke et al., 2003) five target sites had been verified experimentally. Our nucleus score detected four of them while predicting no other sites. As an additional test, we mutated *in silico* the cytosine residue of *lin-4* which has been shown to be essential for post-transcriptional regulation of *lin-14* (Ha et al., 1996). All *lin-14* target sites scored very poorly for this mutation of *lin-4*.

A simple model for the mechanism of miRNA target recognition

The ratios of our optimal weights for the nucleus score (which is just the sum over these weights) turn out to correspond well to the experimentally known RNA:RNA basepairing energy ratios. Thus, our nucleus can be interpreted as a *kinetic* component of target site recognition: the miRNA needs to be presented with sequence that allows for sufficiently many energetically favorable and consecutive basepairings in order to rapidly zip up and thereby overcome thermal diffusion.

The second phase of our algorithm models the *thermodynamic annealing* of the entire miRNA to to the target. According to our results, the free energy of the target site:miRNA duplex needs to be lower than roughly half of the free energy of a target site with perfect complementarity to the miRNA. We had found that most of the ability of our algorithm to discriminate target sites from random sites comes from the nucleus. Thus, according to our model, most of the target recognition seems to take place during the kinetic phase of miRNA to target binding. The model can also explain the observation (Lai, 2002) that some fly miRNA sequences have *substrings* which are complementary to known 3' UTR sequence motifs that mediate translational repression. However, we remark that the nucleus may not necessarily appear as consecutive base pairs in the predicted secondary structure of the mRNA/miRNA duplex.

Searching for new miRNA targets in fly

We applied our algorithm to a set of 74 *D. melanogaster* miRNA genes which have been recently identified (see Materials & methods). The efficiency of the algorithm would allow a genome wide search for targets. However, searching genome wide for targets of a large set of often differentially expressed miRNAs is likely to produce results which are difficult to interpret. Therefore, we decided to focus on a set of 31 well characterized developmental genes, the body patterning genes (Materials & methods), which are central to a large regulatory network during development. We reasoned that these genes (many of which are key regulatory genes thoughout development) are likely to be targets of certain miRNAs. To further reduce the number of false positives, we limited our search to the 3' UTRs of the genes in our dataset since all known miRNA target sites reside in 3' UTRs. Finally, we filtered all predictions for sites which are, for each *D. melanogaster* 3'UTR, also at least present once in the orthologous *D. pseudoobscura* 3' UTR, reasoning that these sites are more likely to be functional. In our cross-species analysis we do not make the assumption that target sites reside in a conserved chunk of RNA or that the order of multiple sites in a 3' UTR is the same in both species since we know very little about the evolutionary mechanims behind 3' UTR sequence evolution. Setting the nucleus score p-value and the RNA:RNA duplex minimal free energy such that we recover 84% of the known targets in our dataset (see above), we found 39 high scoring *melanogaster* putative target sites (see table 1) which were also present in each case in the orthologous UTR in *pseudoobscura*. Figures 3-5 present the predicted secondary structures for some of these hits and the position of each nucleus.

Detailed analysis of the significance and validation of these data should be accompanied by experiments and is not in the scope of this paper. We will discuss one of the most interesting cases.

The miRNA genes *mir-309, mir-318, mir-263b*, and *mir-3* all hit the pair-rule genes *fushi-tarazu* and *odd-skipped* and nothing else. Conversely, *fushi-tarazu* appears not to be targeted by any other miRNAs from our dataset, only *odd-skipped* has additional target sites for *mir-5* and *mir-8* which in turn do not hit any other gene. The position of the nucleus relative to the miRNA is almost perfectly constant for each miRNA across all its hits (for example, all mir-309 nuclei are at the 5' end of the miRNA at position 2, all mir-3 nuclei are at the 5' end of the miRNA at position 1-4), indicating that the same *cis*-regulatory motif may be used to coordinate the action of a miRNA across different genes. Again, this observation is consistent with (Lai, 2002) where it was shown that certain miRNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation.

Discussion

We found that key developmental genes in fly such as the pair-rule genes *fushi-tarazu, odd skipped* and *hairy* are possibly targeted by miRNAs (*mir-309, mir-318, mir-263b, mir-3*) and *mir-7*. However, ultimately the function of miRNAs has to be elucidated in the context of their own expression. Some of the above miRNA genes (*mir-3, mir-309, mir-7*) are indeed known to be expressed during fly development (Aravin et al., 2003). Furthermore, since it seems possible that miRNA genes are transcriptionally regulated similarly to most protein coding genes (Johnson et al., 2003), we have computationally searched the fly genomes for enhancers in the vicinity of miRNA gene loci. More precisely, we searched the neighborhood of miRNA gene loci for clusters of binding sites for body-patterning transcription factors using an existing algorithm (Rajewsky et al., 2002; Mallela et al., 2003). We found high scoring clusters very close to some miRNA genes (for example *mir-263b*, figure 6) and will test some of these predictions in the future. Hopefully, we will thus understand more about how transcriptional and translational gene regulation are intertwined.

Our algorithm for miRNA target site detection can almost certainly be improved. Since the algorithm is based on experimental knowledge for only two miRNA genes, it seems clear that as new targets for other miRNA genes will be discovered, our understanding and modeling of the target recognition process will improve and reduce the number of false positives. Futhermore, based on the observation that most genes which are targeted by miRNAs appear to have multiple, co-clustered binding sites for multiple miRNAs in their mRNA and that miRNA genes are thus likely to act combinatorially on target genes (see also Doench et al., 2003), the most substantial improvement could perhaps be made by incorporating searches for clusters of binding sites into the algorithm. Information about expression profiles of miRNA genes will then help to filter for meaningful combinations of miRNA binding sites. Key to further improvements may be to understand more about the evolution of miRNA binding sites and thus to improve cross-species analysis and our understanding how gene regulatory networks evolve.

We have demonstrated that our algorithm can detect miRNA target sites with high specificity, that it leads to a simple model for the mechanisms behind miRNA target site recognition, and that it can be applied to existing data to make testable predictions about miRNA function. Thus, we believe that it will help to shed more light on the unfolding, exciting universe of miRNA genes.

Materials and methods

A set of 25 experimentally defined miRNA binding sites (Training set)

The training set consisted of 25 experimentally defined target sites (see Banerjee and Slack, 2002; Lin et al., 2003 and references therein) of the *C. elegans* miRNAs *lin-4* and *let-7*. We padded these sites with the corresponding genomic sequences such that each site had a length of 30 nucleotides. Our 25 pairs are (3 pairs lin-14:let-7, 7 pairs lin-14:lin-4, lin-28:let-7, lin-28:lin-4, 2 pairs of lin-41:let-7, lin-41:lin-4, 2 pairs hbl:lin-4, 8 pairs of hbl:let-7).

Random sequences

Random sequences were produced by site independent sampling of specified `ACGU` background frequencies. We used background frequencies of ($p_A = 0.34$, $p_C = 0.19$, $p_G = 0.18$, $p_U = 0.29$). These frequencies are consistent with the sequence composition of the *C. elegans* 3'UTRs of the target genes in our training set, and they match the base frequencies of 3'UTRs from the set of all known full length *D. melanogaster* cDNAs. We checked that we obtained very similar base frequencies when mapping the 3'UTRs to *D. pseudoobscura* (see below).

Using MFOLD to predict the secondary structure and free energy of RNA:RNA duplexes

The second step of our algorithm involves *in silico* hybridization of the mature miRNA to mRNA using the MFOLD RNA folding program (Zuker, 2003). Note that just joining the two RNA sequences with some linker residues and then running MFOLD would produce unreliable results since the linker residues would be treated like an interior loop and would introduce incorrect contributions to the free energy. However, the new MFOLD software allowed us to overcome this problem (Zuker, 2003). First the two RNA sequences were joined together with an artificial linking segment of non-nucleotide elements (represented by the symbol `L`). For example if one sequence was `5'-ACGTACGT-3'` and the other was `5'-GCATGCAT-3'` the resulting artificial single sequence is `5'-ACGTACGTLLLGCATGCAT-3'`. The remaining step is to prevent any base pairing within the original two sequences. Following (Zuker, 2003) this was accomplished by passing a special configuration file to the MFOLD program which prohibits pairing within a given range of bases. The MFOLD program was run with a temperature setting of 20^{o}C and default parameters otherwise.

A set of 74 *Drosophila* miRNA genes

The miRNAs used in this study came from two sources. A set of 62 miRNAs were found experimentally via the cloning of small RNAs in *Drosophila melanogaster* (Aravin et al., 2003) that were kindly provided to us by the Tom Tuschl prior to publication. A second set came from a computational study of Lai et al (Lai et al., 2003). This set was identified by searching both *melanogaster* and *pseudoobscura* genomic sequence for short conserved sequences with an extended stem-loop structure and a given pattern of divergence between the two species. We BLASTed the second set against the first and found 12 non-redundant miRNA's (*miR-274, miR-219, miR-276a, miR-33, miR-280, miR-281a, miR-282, miR-284, miR-263a, miR-289, miR-287, miR-288*). We added these 12 to the first set and obtained our final dataset. We did not exclude from our dataset the roughly 25 % of miRNA genes which were detected in adult animals or testes only, reasoning that the expression assays used are certainly not perfectly sensitive, and also reasoning that we can always backtrack our results.

Set of genes important for fly body patterning

Our set comprises all well known key early genes (*nanos, oskar, vasa, tudor, Pumilio, Staufen, Fat facets*), gap (*hunchback, bicoid, tailless, caudal, Kruppel, giant, knirps, sloppy paired 1, sloppy paired 2, buttonhead, collier, crocodile, empty spiracles, huckebein, orthodenticle, cap'n'collar*) , and pair-rule genes (*eve, hairy, ftz, runt, odd-paired, paired, Tenascin major, odd-skipped*).

Extraction of 3'UTR sequences in *D. melanogaster and D. pseudoobscura*

The $3'$ UTRs for *Drosophila melanogaster* were extracted from the BDGP genome annotation release 3.1 at `www.fruitfly.org/sequence/download.html`. The *crocodile* gene from our dataset was the only gene which did not have a $3'$ UTR of at least 50 basepairs length. In almost all cases where a gene had multiple transcripts the corresponding $3'$ UTR sequences were still the same. However in one case (for the gene *collier*) this was not true and we chose the transcript CG10197-RB which mapped to the experimentally known cDNA BcDNA:RE03728 of this gene. We note that for roughly 50 % of the genes in our dataset the $3'$ UTR annotations from release 3.1 are directly supported by the drosophila cDNA library. To define the homologous *Drosophila pseudoobscura* $3'$ UTR sequences we used the Berkeley genome pipeline (Couronne et al., 2003; Bray et al., 2003) website (`pipeline.lbl.gov/pseudo/`) which presents a genomic alignment of the two species. We looked up the corresponding alignments for each *melanogaster* $3'$ UTR region and extracted the *pseudoobscura* sequence from the alignment with the highest percentage identity. We checked that in each case the coding region of the gene was in the same alignment. The only genes for which we found not enough homology to unambiguously define the *pseudoobscura* $3'$ UTR were *giant* and *Tenascin major*. These genes were excluded from our dataset. The $3'$ UTR pairs in our dataset have an average percentage identity of 0.52.

Acknowledgments

We are indebted to Tom Tuschl (Rockefeller University) who shared his unpublished miRNA data with us prior to publication. We thank Steve Small, Frank Slack, Ralf Bundschuh and Mihaela Zavolan for stimulating discussions. We also thank Michael Zuker for help with MFOLD. The source code for our algorithm will be available upon request on the authors.

Note added in proof: While writing this manuscript, we were informed about a similar computational approach by Stephen Cohen's group (EMBL, Heidelberg) and another paper on a related topic by AJ Enright, B John, U Gual, T Tuschl, C Sander and D Marks (Genome Biology 2003 5:R1).

References

Ambros, V., 2001. microRNAs: tiny regulators with great potential. Cell 107 (7), 823–826.

Ambros, V., 2003. MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing. Cell 113 (6), 673–676.

Ambros, V., Lee, R. C., Lavanway, A., Williams, P. T., Jewell, D., 2003. MicroRNAs and other tiny endogenous RNAs in C. elegans. Curr. Biol. 13 (10), 807–818.

Aravin, A. A., Lagos-Quintana, M., Yalcin, A., Zavolan, M., Marks, D., Snyder, B., Gaasterland, T., Meyer, J., Tuschl, T., 2003. The small RNA profile during Drosophila melanogaster development. Dev. Cell 5 (2), 337–350.

Banerjee, D., Slack, F., 2002. Control of developmental timing by small temporal RNAs: a paradigm for RNA-mediated regulation of gene expression. Bioessays 24 (2), 119–129.

Bray, N., Dubchak, I., Pachter, L., 2003. Avid: A global alignment program. Genome Res. 13, 97.

Brennecke, J., Hipfner, D. R., Stark, A., Russell, R. B., Cohen, S. M., 2003. bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene *hid* in Drosophila. Cell 113 (1), 25–36.

Carrington, J. C., Ambros, V., 2003. Role of microRNAs in plant and animal development. Science 301 (5631), 336–338.

Couronne, O., Poliakov, A., Bray, N., Ishkhanov, T., Ryaboy, D., Rubin, E., Pachter, L., Dubchak, I., 2003. Strategies and tools for whole-genome alignments. Genome Res. 13 (1), 73–80.

Doench, J. G., Petersen, C. P., Sharp, P. A., 2003. siRNAs can function as miRNAs. Genes Dev. 17 (4), 438–442.

Ha, I., Wightman, B., Ruvkun, G., 1996. A bulged lin-4/lin-14 rna duplex is sufficient for Caenorhabditis elegans lin-14 temporal gradient formation. Genes Dev 10 (23), 3041–3050.

Houbaviy, H. B., Murray, M. F., Sharp, P. A., 2003. Embryonic stem cell-specific MicroRNAs. Dev. Cell 5 (2), 351–358.

Johnson, S. M., Lin, S. Y., Slack, F. J., 2003. The time of appearance of the C. elegans let-7 microRNA is transcriptionally controlled utilizing a temporal regulatory element in its promoter. Dev. Biol 259 (2), 364–379.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., Tuschl, T., 2001. Identification of novel genes coding for small expressed RNAs. Science 294 (5543), 853–858.

Lagos-Quintana, M., Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W., Tuschl, T., 2002. Identification of tissue-specific microRNAs from mouse. Curr. Biol. 12 (9), 735–739.

Lai, E. C., 2002. Micro rnas are complementary to 3' utr sequence motifs that mediate negative post-transcriptional regulation. Nat. Genet. 30 (4), 363–364.

Lai, E. C., Tomancak, P., Williams, R. W., Rubin, G. M., 2003. Computational identification of drosophila microRNA genes. Genome Biol. 4 (7), R42.

Lau, N. C., Lim, L. P., Weinstein, E. G., Bartel, D. P., 2001. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science 294 (5543), 858–862.

Lee, R. C., Ambros, V., 2001. An extensive class of small RNAs in Caenorhabditis elegans. Science 294 (5543), 862–864.

Lim, L. P., Glasner, M. E., Yekta, S., Burge, C. B., Bartel, D. P., 2003a. Vertebrate microRNA genes. Science 299 (5612), 1540.

Lim, L. P., Lau, N. C., Weinstein, E. G., Abdelhakim, A., Yekta, S., Rhoades, M. W., Burge, C. B., Bartel, D. P., 2003b. The microRNAs of caenorhabditis elegans. Genes Dev. 17 (8), 991–1008.

Lin, S.-Y., Johnson, S. M., Abraham, M., Vella, M. C., Pasquinelli, A., Gamberi, C., Gottlieb, E., Slack, F. J., 2003. The *C elegans* hunchback homolog, hbl-1, controls temporal patterning and is a probable microRNA target. Dev. Cell 4 (5), 639–650.

Mallela, J., Kacmarczyk, T., Bonavia, A., Rajewsky, N., 2003. The ahab webserver. http://gaspard.bio.nyu.edu/Ahab.html, unpublished.

Moss, E. G., 2002. MicroRNAs: hidden in the genome. Curr Biol 12 (4), R138–R140.

Moss, E. G., Poethig, R. S., 2002. MicroRNAs: something new under the sun. Curr Biol 12 (20), R688–R690.

Rajewsky, N., Vergassola, M., Gaul, U., Siggia, E. D., 2002. Computational detection of genomic cis-regulatory modules applied to body patterning in the early drosophila embryo. BMC Bioinformatics 3 (1), 30.

Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B., Bartel, D. P., 2002. Prediction of plant microRNA targets. Cell 110 (4), 513–520.

Xu, P., Vernooy, S. Y., Guo, M., Hay, B. A., 2003. The drosophila microRNA mir-14 suppresses cell death and is required for normal fat metabolism. Curr Biol 13 (9), 790–795.

Zuker, M., 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31 (13), 3406–3415.

Figure 1: Pictorial representation of the Z-score. The Z-score is a comparison of the typical value in the target (training) set with the distribution of scores from the random background. The background distribution is on the left and is parametrized by its mean (μ) and its width (variance σ). The mean (x) of the training set is on the right. The Z-score which represents an approximate signal to noise ratio measurement is given by $Z = (x - \mu)/\sigma$. It indicates how far above the background the target signal is.

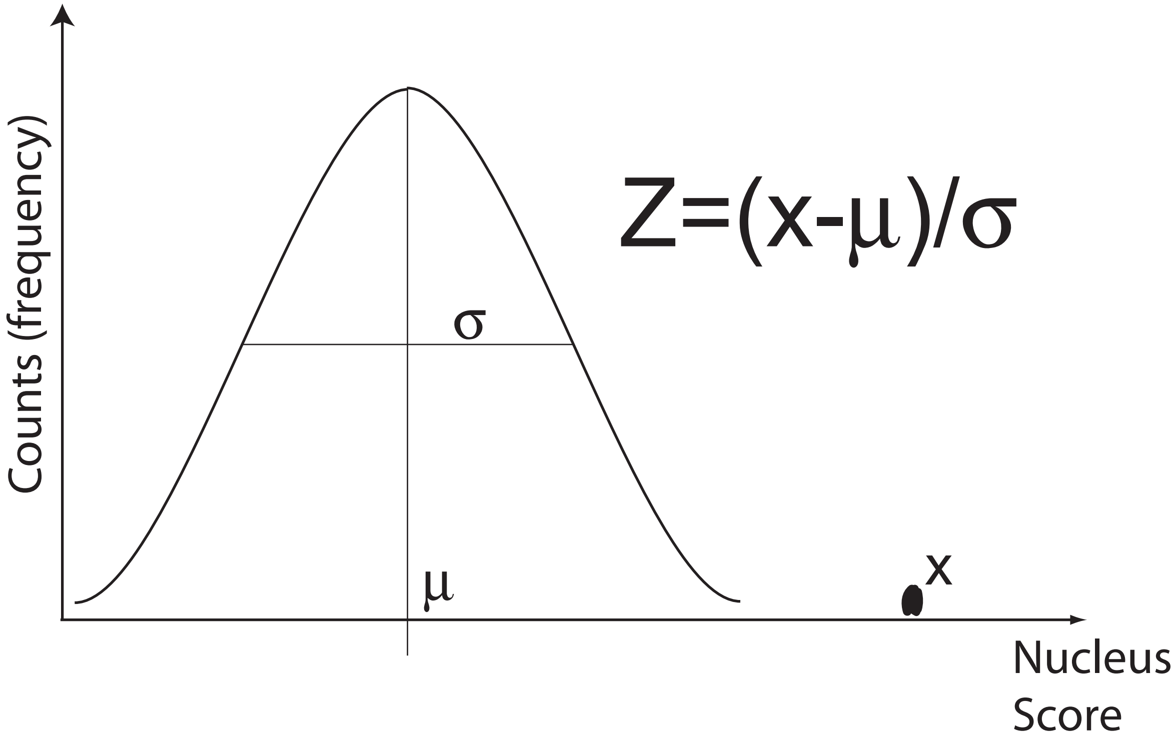
Counts (frequency)
Z=(x-μ)/σ
σ
μ
x
Nucleus
Score

Figure 2: Nucleus score histogram for the *C. elegans* miRNA *lin-4*. A search window of 30 bases was shifted in steps of 10 over random sequence of length 1,000,000 bases. At each position, the nucleus score was recorded. A score threshold of 25 (27) will recover 84% (50%) of the known *lin-4* target sites in our training set, but is only rarely exceeded by scores obtained from the random sequence. Thus, the nucleus score recovers with high specificity the known targets.

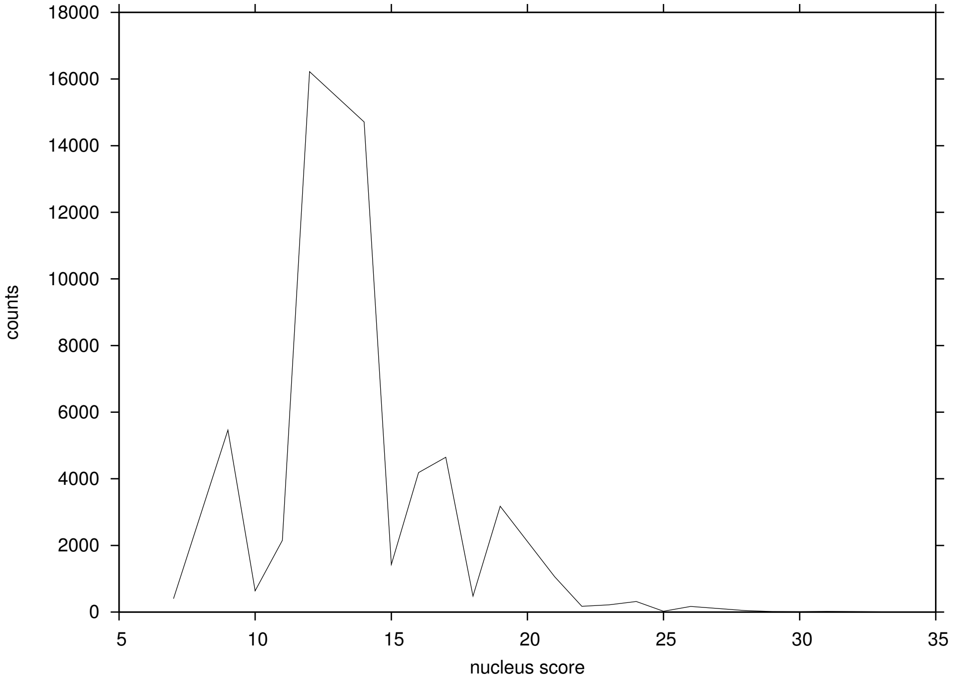
counts
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
5
10
15
20
25
30
35
nucleus score

Figure 3: Predicted targeting of the *Drosophila melanogaster* gene *fushi-tarazu* by *mir-309*. Shown is the mRNA:miRNA duplex as predicted by MFOLD. The free energy is -23.4 kcal/mol, the nucleus has a score of 31 (p value 0.0001) and is located at the 5' end of the miRNA and 271 bases downstream of the stop codon. The *fushi-tarazu* ortholog in *pseudoobscura* also has a predicted target site for *mir-309*. All nuclei are found close to the 5' end of the miRNA. GC basepairings are marked in red, AU and GU in blue.

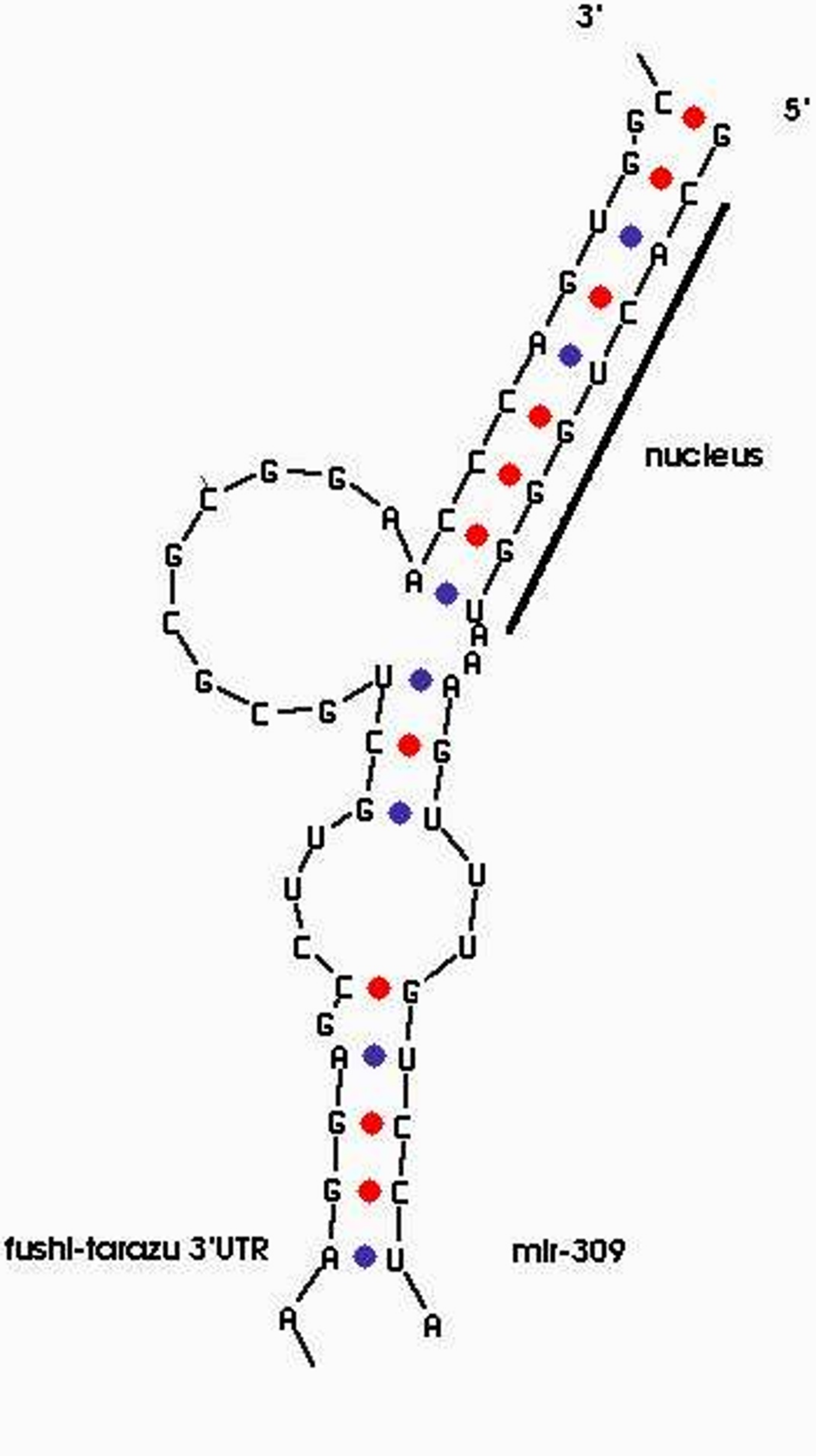
3'
5'
nucleus
fushi-tarazu 3'UTR
mir-309

Figure 4: Predicted targeting of the *Drosophila melanogaster* gene *fushi-tarazu* by *mir-3*. Shown is the mRNA:miRNA duplex as predicted by MFOLD. The free energy is -30.8 kcal/mol, the nucleus has a score of 28 (p value 0.0001) and is located at the 5' end of the miRNA and 173 bases downstream of the stop codon. Another putative target site for *mir-3* is located 99 bases further downstream. The *fushi-tarazu* ortholog in *pseudoobscura* also has a predicted target site for *mir-3*. All nuclei are found close to the 5' end of the miRNA. GC basepairings are marked in red, AU and GU in blue.

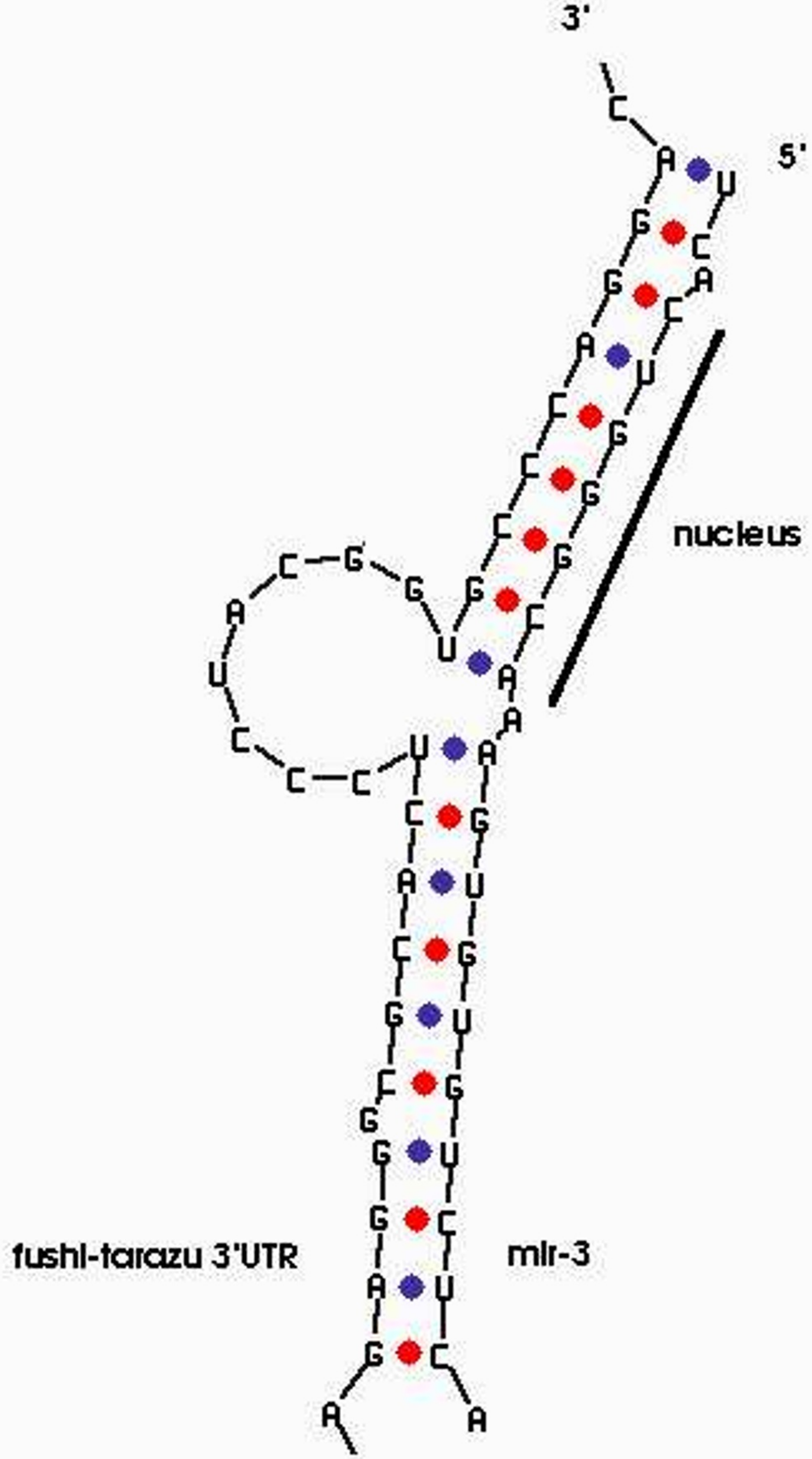
3'
5'
nucleus
fushi-tarazu 3'UTR
mir-3
C
A
G
G
A
C
C
C
G
U
U
C
A
C
G
C
G
G
G
A
G
A
U
C
A
C
U
G
G
G
C
A
A
A
G
U
G
U
G
U
C
U
C
A
C
G
A
U
C
C
G

Figure 5: Predicted targeting of the *Drosophila melanogaster* gene *hairy* by *mir-7.* Shown is the mRNA:miRNA duplex as predicted by MFOLD. The free energy is -30.6 kcal/mol, the nucleus has a score of 30 (p value 0.0001) and is located at the 5' end of the miRNA. The target site is 438 bases downstream of the stop codon. The *hairy* ortholog in *pseudoobscura* also has a predicted target site for *mir-7.* All nuclei are found close to the 5' end of the miRNA. GC basepairings are marked in red, AU and GU in blue.

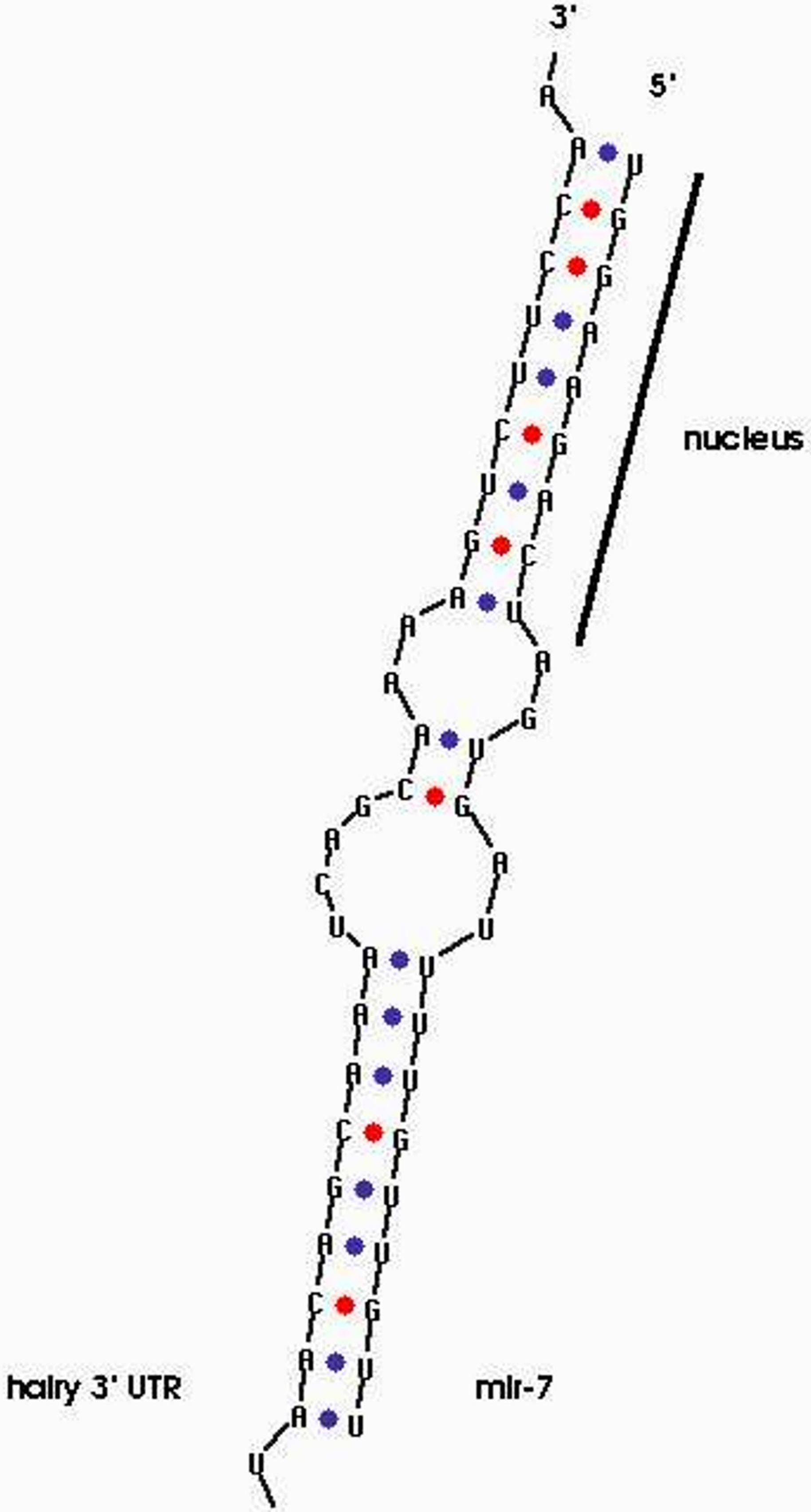
3'
5'
nucleus
hairy 3' UTR
mir-7

Figure 6: The *D. melanogaster mir-263b* locus was searched for clusters of binding sites for the body patterning transcription factors *kruppel, caudal, bicoid, hunchback, tailless, torRE, bicoid* following (Rajewsky et al., 2002). We used the Ahab webserver (Mallela et al., 2003). The score for finding a cluster is shown as a function of position in the locus (in nucleotides). Two peaks in the score bracket the position of the miRNA gene.

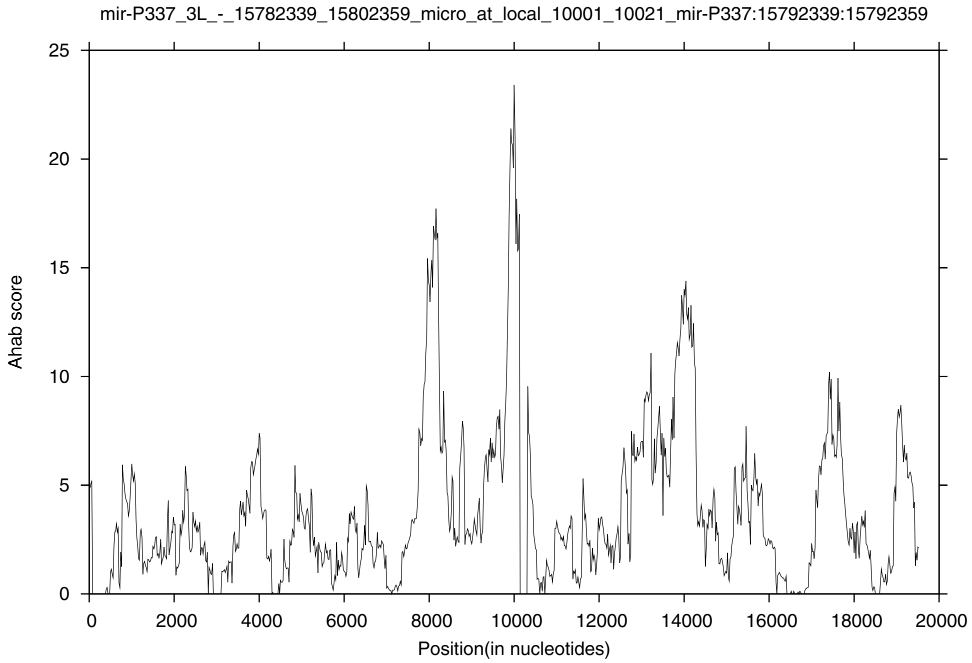
mir-P337_3L_-_15782339_15802359_micro_at_local_10001_10021_mir-P337:15792339:15792359
Ahab score
0
5
10
15
20
25
Position(in nucleotides)
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000

Table 1: Putative targets of miRNAs within the set of patterning genes (see Material and methods). The free energy of the predicted RNA:RNA duplexes is given (in kcal/mol) as well as the position of the nucleus in the mRNA (in nucleotides downstream of the stop codon).

Gene	miRNA	Free Eng,Position
kn	mir-312	-21.7,107
kn	mir-313	-22.3,106
kn	mir-92b	-17,28
kn	mir-2a-2:mir-2a-1	-18.8,1078
opa	mir-8	-23.7,139
oc	mir-317	-23.5,171 -17.8,199 -17.8,205
oc	mir-133	-18.6,172
btd	mir-7	-18.3,455
tll	mir-6-1:mir-6-2:mir-6-3	-21.3,61
tll	miR-219	-27.4,86
slp1	mir-79	-16.1,36
slp1	mir-8	-19,144
cnc	mir-315	-18.5,983
cnc	mir-279	-18.1,989 -18.1,991
run	miR-287	-20,132 -19.5,334 -18.4,438
ftz	mir-318	-21.5,272
ftz	mir-309	-23.4,271
ftz	mir-263b	-19,272
ftz	mir275	-23.2,263
ftz	mir-3	-30.8,173 -25.8,272

Gene	miRNA	Free Eng,Position
ems	mir-312	-20,565
ems	mir-133	-18,546
ems	miR-263a	-20.8,422
vas	mir-P323-1:mir-P323-2	-18.3,668
vas	mir275	-23.8,637
odd	mir-318	-20.7,182
odd	mir-263b	-18.7,182
odd	mir-309	-18.3,181
odd	mir-5	-22,673
odd	mir-8	-19.7,22
odd	mir-3	-20.7,182
kni	miR-284	-27.8,144
nos	mir-124	-26.5,174
stau	miR-280	-16.4,373
stau	mir305	-29.4,589
h	miR-289	-22.9,102
h	let-7	-23.2,381
h	mir-7	-30.6,460 -30.6,461
hkb	let-7	-22.4,157