Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry Letters

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

Predicting potential miRNA target sites within gene promoters

Scott T. Younger[a], Alexander Pertsemlidis[b,*], David R. Corey[a,*]

[a] Departments of Pharmacology and Biochemistry, UT Southwestern Medical Center at Dallas, Dallas, 6001 Forest Park Road, TX 75390, USA
[b] Eugene McDermott Center for Human Growth and Development, 5323 Harry Hines Boulevard, Dallas TX 75390-8591, USA

ARTICLE INFO

Article history:
Received 18 February 2009
Revised 9 April 2009
Accepted 10 April 2009
Available online 18 April 2009

Keywords:
RNAi
miRNA
MicroRNA
Gene promoter
Antigene
Noncoding RNA
Duplex RNA
Hybridization

ABSTRACT

Synthetic small duplex RNAs that are complementary to gene promoters can activate or inhibit target gene expression. The potency and robustness of gene modulation by these RNAs suggests that natural mechanisms may exist to facilitate recognition of sequences within gene promoters by endogenous small RNAs. Here, we describe computational methods for identifying potential miRNA target sites within gene promoters. These methods will facilitate investigations of whether miRNAs interact with sequences outside of 3′-untranslated regions and suggest new targets for the design of synthetic modulators of gene expression.

Synthetic small duplex RNAs complementary to gene promoters within chromosomal DNA have been reported to be potent inhibitors or activators of target gene expression in mammalian cells.[1–5] We refer to these synthetic RNAs as antigene RNAs (agRNAs) to distinguish them from small duplex RNAs that target mRNA. agRNAs recruit members of the argonaute (AGO) protein family to RNA transcripts that originate from the target gene promoter in either the sense or antisense direction.[6–9] Data suggest that recognition of the target RNA occurs in close proximity to the chromosome, resulting in transcriptional modulation of the target gene.

One remarkable feature of the synthetic agRNAs that we have examined is the potency and robustness of their activity when they are introduced into cells. This potency, coupled with the presence of protein machinery that facilitates their function, suggests that endogenous small RNAs may possess the ability to recognize gene promoters. If RNA could direct proteins to specific gene promoters, such RNA-mediated modulation of transcription might have evolutionary advantages relative to the development of gene-specific protein transcription factors.

Synthetic duplex RNAs that are complementary to mRNA (small interfering RNAs or siRNAs) are also potent and robust agents for modulating gene expression.[10] siRNAs are known to have endogenous analogs that regulate gene expression called microRNAs (miRNAs).[11] miRNAs are processed inside the cell from RNA precursors that contain stem-loop structures. These stem-loop structures are processed by the double-stranded nucleases Drosha and Dicer to produce mature miRNAs.

As of the current release of the miRNA repository (miRBase v12.0), 866 human miRNAs have been annotated, but this number continues to increase. Several miRNAs that recognize sequences within the 3′-untranslated regions (3′UTR) of mRNA transcripts have been characterized. Many miRNAs, however, have no known targets[12,13] while some can recognize multiple mRNAs[13], suggesting that the determinants of miRNA interactions are complex and poorly understood.

Two reports based on computational analyses have suggested that miRNAs can modulate gene expression through promoter recognition. Dahiya and co-workers used publicaly available software (RegRNA) to search for potential miRNA target sites within the promoter of the E-cadherin gene.[14] They identified one potential binding site for miR-373 within the E-cadherin promoter and reported that introduction of a synthetic miR-373 mimic increased expression of the gene by sixfold at the level of the mRNA. Rossi and co-workers searched for perfect complementarity between miRNAs and gene promoters.[15] Their analysis suggested that miR-320 targets the genomic location from which it is transcribed and showed that expression of miR-320 and the adjacent gene, POLR3D, are anti-correlated.

The above-mentioned studies either analyzed a single gene promoter or used highly stringent sequence comparison criteria. These

* Corresponding authors.
E-mail addresses: alexander.pertsemlidis@utsouthwestern.edu (A. Pertsemlidis), david.corey@utsouthwestern.edu (D.R. Corey).

0960-894X/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmcl.2009.04.032

approaches were not intended to assess the broader potential for miRNAs to recognize gene promoters, warranting a more thorough evaluation of the relationship between miRNAs and promoter sequences.

A practical justification for more comprehensive studies is that validating natural gene targets of miRNAs is a complex and difficult process. The development of systematic and efficient methods for identifying promoter sequences that may be miRNA targets is essential for prioritizing predictions and efficiently allocating experimental resources towards validating the most promising targets. Here we examine computational methods for predicting potential miRNA targets within gene promoters and demonstrate that promoters are strong candidates for miRNA regulation.

Sequence acquisition. To identify putative promoter-targeting miRNAs we constructed a database comprised of miRNA and gene promoter sequences from public sequence repositories. Promoter sequences were acquired from the UCSC genome browser (hg 18) and consisted of the 200 nucleotides immediately 5′ to the annotated transcription start site for each gene[16,17]. We chose 200 base sequences (−200 to −1) for initial evaluations but larger promoter regions can also be examined. Mature miRNA sequences were obtained from miRBase (Build 12.0), which contains sequences of experimentally determined precursor and mature miRNAs.[18–20]

Analysis of seed sequence matches. Synthetic promoter-targeting RNAs recognize non-coding (ncRNA) transcripts that overlap gene promoters. We used promoter DNA sequences to construct datasets representing potential ncRNA transcripts in both the sense and antisense direction for each gene promoter as we hypothesize that endogenous small RNAs would also recognize these ncRNA transcripts. For comparison we also obtained the sequences of the 5′UTR, coding sequences (CDS), and 3′UTR for each gene (Fig. 1A).

A basic requirement for target recognition by miRNAs is perfect complementarity between the target sequences and bases 2–8 of the mature miRNA sequence, called the seed sequence. We determined the number of seed matches within potential sense and antisense transcripts that overlap gene promoters and compared them to seed matches within the 3′UTR region of mRNAs (Fig. 1B). We found that seed matches within promoter-overlapping transcripts occur 80% as frequently as seed matches within 3′UTRs, indicating that gene promoter sequences have the potential to be miRNA targets (Fig. 2A). Our analysis detected the previously reported complementarity between miR-320 and the POLR3D promoter.[15]

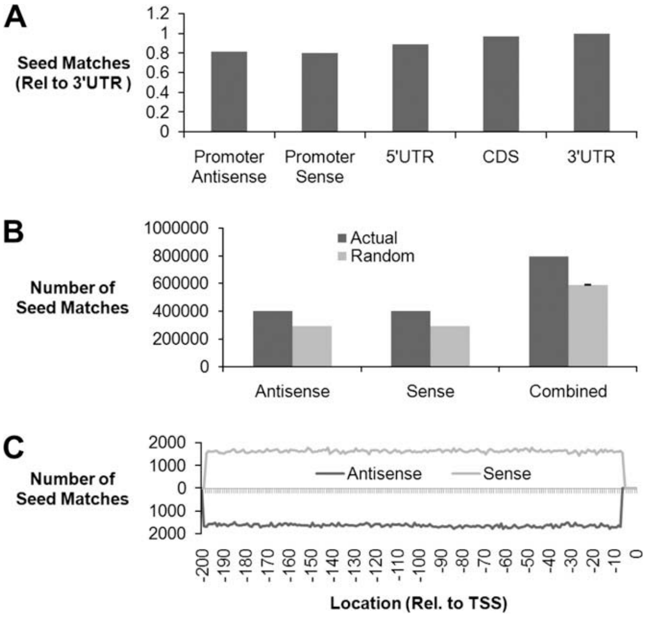


Figure 2. Gene promoters contain predicted miRNA target sites. (A) The frequency of seed sequence matches in promoter regions, 5′UTRs, coding regions, and 3′UTRs. (B) Comparison of seed matches within promoter-overlapping transcripts versus randomized promoter sequences. ($p < .01$) (C) Distribution of seed match locations within sense and antisense transcripts that overlap gene promoters from −1 to −200 relative to the +1 transcription start site.

To evaluate the statistical significance of seed matches within gene promoter sequences we tabulated the frequency of occurrences of seed matches in 100 randomizations of each promoter sequence. We found that seed matches occur 75% as frequently within randomized as opposed to actual promoter sequences (Fig. 2B). The excess of observed to expected seed sequence matches within promoter sequences was similar for both putative sense and antisense transcripts. This result implies that promoter sequences are enriched for potential targets for recognition by miRNAs. Matches are equally distributed throughout the 200 base gene promoter segments surveyed, suggesting that no particular region of a gene promoter is more likely than another to contain a predicted miRNA target site (Fig. 2C).

Ranking matches. Our analysis identified nearly 800,000 miRNA seed matches within 27,345 gene promoter sequences (Fig. 2B).

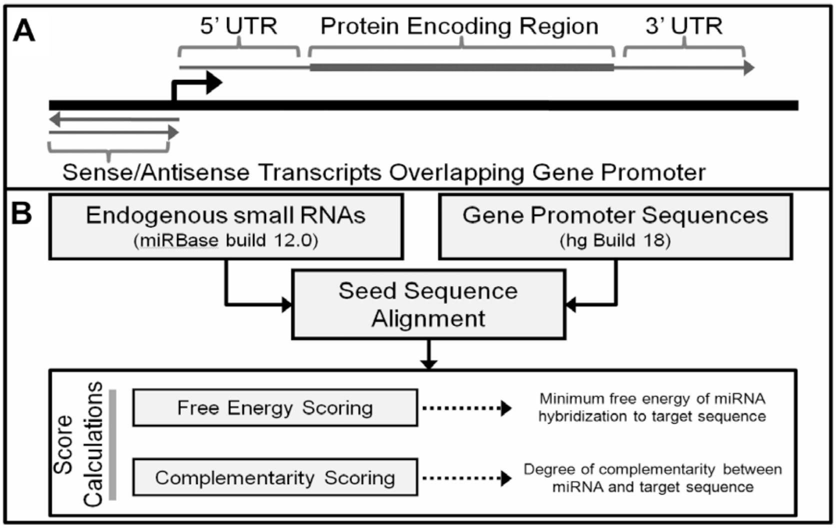


Figure 1. (A) Diagram of sequences that are potential miRNA targets. (B) Schematic of algorithm used to predict miRNA targets within gene promoters.

This large number required investigation of additional factors to prioritize target predictions. Although not necessarily a prerequisite for miRNA function, the minimum free energy (MFE) of hybridization between miRNAs and their predicted target sites have been successfully used to predict miRNA target sites within 3′UTRs[21]. We reasoned that MFE values may also be useful for prioritizing miRNA target predictions within gene promoters.

The MFE values were calculated for miRNA hybridization to predicted target sites (based on seed sequence matches, hereafter simply referred to as predictions) within putative promoter-overlapping transcripts and within 10 randomizations of promoter sequences. We found that predictions with lower MFE values occurred more frequently in actual promoter sequences than in randomized sequences (Fig. 3A). The difference between the distributions of MFE values demonstrates that predictions with low MFE values occur more often than would be expected at random, implying that these predictions are more likely to be biologically significant and that MFE values will be useful criteria for prioritizing target predictions.

During the course of the MFE analysis we identified several miRNA target predictions within gene promoters that had notably low MFE values. These observations prompted us to compare the MFE values for target predictions within gene promoters to target predictions within 3′UTRs (Fig. 3B). We calculated the mean MFE value for all predictions within gene promoters to be −24.27 kcal/mol and −24.32 kcal/mol for putative sense and antisense promoter-overlapping transcripts, respectively. The mean MFE for all predictions within 3′UTRs was −20.57 kcal/mol, more than 3.5 kcal/mol higher than predictions within promoters. The difference in mean MFE values suggests that, on average, miRNA recognition of sequences at gene promoters would be more energetically favorable than recognition of 3′UTR sequences.

To further evaluate the differences between target predictions within gene promoters and 3′UTRs, we examined the distribution of MFE values for all predictions within the different sequence datasets. As previously indicated by the mean MFE values, roughly 50% of target predictions within gene promoters had MFE values below −24.3 kcal/mol. Interestingly, only 22% of predictions within 3′UTRs had MFE values below −24.3 kcal/mol (Fig. 3B). The difference in MFE value distributions demonstrates that gene promoters are enriched relative to 3′UTRs for predicted target sites with low free energies of hybridization and may actually represent more favorable miRNA targets than 3′UTRs.

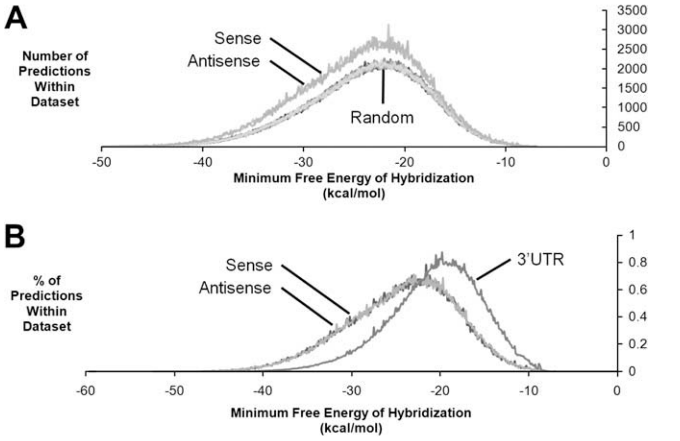
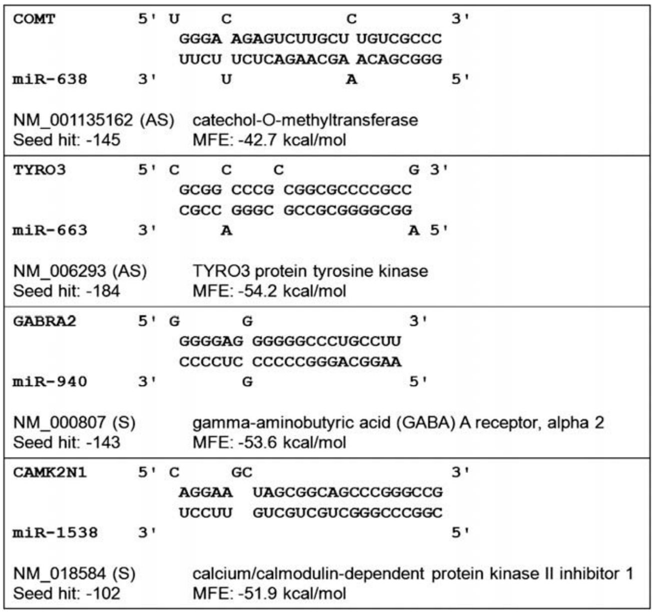


Figure 3. MFE properties of miRNA target predictions. (A) Distribution of MFE values for predictions within sense and antisense overlapping transcripts as compared to randomized promoter sequences. (B) Distribution of MFE values for predictions within sense and antisense overlapping transcripts as compared to 3′UTRs.



```
COMT        5' U    C           C        3'
                GGGA AGAGUCUUGCU UGUCGCCC
                UUCU UCUCAGAACGA ACAGCGGG
miR-638     3'      U           A        5'

NM_001135162 (AS)   catechol-O-methyltransferase
Seed hit: -145      MFE: -42.7 kcal/mol

TYRO3       5' C    C    C            G 3'
                GCGG CCCG CGGCGCCCCGCC
                CGCC GGGC GCCGCGGGGCGG
miR-663     3'      A                 A 5'

NM_006293 (AS)      TYRO3 protein tyrosine kinase
Seed hit: -184      MFE: -54.2 kcal/mol

GABRA2      5' G      G              3'
                GGGGAG GGGGGCCCUGCCUU
                CCCCUC CCCCCGGGACGGAA
miR-940     3'        G              5'

NM_000807 (S)       gamma-aminobutyric acid (GABA) A receptor, alpha 2
Seed hit: -143      MFE: -53.6 kcal/mol

CAMK2N1     5' C     GC                  3'
                AGGAA  UAGCGGCAGCCCGGGCCG
                UCCUU  GUCGUCGUCGGGCCCGGC
miR-1538    3'                           5'

NM_018584 (S)       calcium/calmodulin-dependent protein kinase II inhibitor 1
Seed hit: -102      MFE: -51.9 kcal/mol
```

Figure 4. Examples of predicted miRNA targets within sequences of promoter-overlapping transcripts.

Another criterion used in miRNA target prediction is sequence complementarity. Sequence complementarity alone has been used successfully to predict miRNA target sites within 3′UTRs.[22] We used the Needleman–Wunsch algorithm[23] to evaluate the degree of sequence complementarity between miRNAs and predicted target sites within gene promoters (Fig. 1B). We identified over 200 individual miRNAs with near perfect complementarity to their predicted target sites within gene promoters. A selected subset of these predictions is listed in Figure 4. The high degree of complementarity between miRNAs and gene promoters further demonstrates that gene promoters are promising candidates for miRNA targets.

Strong evidence that gene expression can be modulated using synthetic duplex RNAs that are complementary to gene promoters suggests that natural gene regulation may include recognition of gene promoters by miRNAs. Such recognition would have evolutionary advantages, given the large difference between protein transcription factors and miRNAs in their efficiency of generating new selectivity for gene promoters through mutation.

Our computational algorithm can be used to identify promising miRNA target sites within gene promoters. We identify many seed sequence matches within promoters and demonstrate that they are almost as common as those within 3′UTRs. We also identify many miRNA/promoter pairs that have unusually strong complementarity. These results can be used to prioritize miRNA/promoter pairs for the demanding studies necessary to validate whether these interactions are biologically significant.

Acknowledgments

This work was supported by grants from the National Institutes of Health (GM 77253 to D.R.C. and CA 129632 to A.P.), The Robert A. Welch Foundation (I-1244), and an NIH Pharmacological Sciences Training Grant (GM07062 to STY). We thank A. Guillory for technical assistance.

References and notes

1. Morris, K. V.; Chan, S. W.; Jacobsen, S. E.; Looney, D. J. *Science* **2004**, *305*, 1289.
2. Ting, A. H.; Schuebel, K. E.; Herman, J. G.; Baylin, S. B. *Nat. Genet.* **2005**, *37*, 906.

3. Janowski, B. A.; Huffman, K. E.; Schwartz, J. C.; Ram, R.; Hardy, D. B.; Shames, D. S.; Minna, J. D.; Corey, D. R. *Nat. Chem. Biol.* **2005**, *1*, 216.
4. Li, L. C.; Okino, S. T.; Zhao, H.; Pookot, D.; Urakami, S.; Enokida, H.; Dahiya, R. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 17337.
5. Janowski, B. A.; Younger, S. T.; Hardy, D. B.; Ram, R.; Huffman, K. E.; Corey, D. R. *Nat. Chem. Biol.* **2007**, *3*, 166.
6. Janowski, B. A.; Huffman, K. E.; Schwartz, J. C.; Ram, R.; Nordsell, R.; Shames, D. S.; Minna, J. D.; Corey, D. R. *Nat. Struct. Mol. Biol.* **2006**, *13*, 787.
7. Kim, D. H.; Villeneuve, L. M.; Morris, K. V.; Rossi, J. J. *Nat. Struct. Mol. Biol.* **2006**, *13*, 793.
8. Han, J.; Kim, D.; Morris, K. V. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 12422.
9. Schwartz, J. C.; Younger, S. T.; Nguyen, N_B.; Hardy, D. B.; Corey, D. R. *Nat. Struct. Mol. Biol.* **2008**, *15*, 842.
10. Fire, A.; Xu, S.; Montgomery, M. K.; Kostas, S. A.; Driver, S. E.; Mello, C. C. *Nature* **1998**, *391*, 806.
11. Lagos-Quintana, M.; Rauhut, R.; Lendeckel, W.; Tuschl, T. *Science* **2001**, *294*, 85.
12. Lee, R. C.; Feinbaum, R. L.; Ambros, V. *Cell* **1993**, *75*, 843.
13. John, B.; Enright, A. J.; Aravin, A.; Tuschl, T.; Sander, C.; Marks, D. S. *PLoS Biol.* **2004**, *2*, e363.
14. Place, R. F.; Li, L. C.; Pookot, D.; Noonan, E. J.; Dahiya, R. *Proc. Natl. Acad. Sci U.S.A.* **2008**, *105*, 1608.
15. Kim, D. H.; Saetrom, P.; Snøve, O. Jr.; Rossi, J. J. *Proc. Natl. Aad. Sci U.S.A.* **2008**, *105*, 16230.
16. International Human Genome Sequencing Consortium. *Nature* **2001**, *409*, 860.
17. Kent, W. J.; Sugnet, C. W.; Furey, T. S.; Roskin, K. M.; Pringle, T. H.; Zahler, A. M.; Haussler, D. *Genome Res.* **2002**, *12*, 996.
18. Griffiths-Jones, S.; Saini, H. K.; van Dongen, S.; Enright, A. J. *NAR* **2008**, *36*, D154.
19. Griffiths-Jones, S.; Grocock, R. J.; van Dongen, S.; Bateman, A.; Enright, A. J. *NAR* **2006**, *34*, D140.
20. Griffiths-Jones, S. *NAR* **2004**, *32*, D109.
21. Stark, A.; Brennecke, J.; Russell, R. B.; Cohen, S. M. *PLoS Biol.* **2003**, *1*, 1.
22. Lai, E. C. *Nat. Genet.* **2002**, *30*, 363.
23. Needleman, S. B.; Wunsch, C. D. *J. Mol. Biol.* **1970**, *48*, 443.