Antagomirbase: A putative antagomir database

Sayak Ganguli*, Sanga Mitra, Abhijit Datta

DBT-Centre for Bioinformatics, Presidency University, Kolkata – 700073; Sayak Ganguli – Email: sayakbif@yahoo.com; *Corresponding author

Abstract:
The accurate prediction of a comprehensive set of messenger putative antagomirs against microRNAs (miRNAs) remains an open problem. In particular, a set of putative antagomirs against human miRNA is predicted in this current version of database. We have developed Antagomir database, based on putative antagomirs-miRNA heterodimers. In this work, the human miRNA dataset was used as template to design putative antagomirs, using GC content and secondary structures as parameters. The algorithm used predicted the free energy of unbound antagomirs. Although in its infancy the development of antagomirs, that can target cell specific genes or families of genes, may pave the way forward for the generation of a new class of therapeutics, to treat complex inflammatory diseases. Future versions need to incorporate further sequences from other mammalian homologues for designing of antagomirs for aid in research.

Availability: http://bioinfopresidencycollegekolkata.edu.in/antagomirs.html

Background:
MicroRNAs (miRNAs) are a class of endogenous, small regulatory RNA averaging 22 nucleotides in length that mediate the post-transcriptional regulation of messenger RNAs. They bind to target messages in a sequence-specific manner, and induce translational repression or endonucleolytic cleavage. Recently, a novel class of chemically engineered oligonucleotides, termed antagomirs, is efficient and specific silencers of miRNAs. Moreover, the recent employment of synthetic analogues of these small RNA molecules termed 'antagomirs' has shown that microRNAs of interest can be specifically targeted **[1, 2]**. Functions have only been experimentally assigned to a small fraction of the experimentally designed antagomirs. These putative antagomirs are 20 nucleotides in length and complementary to the mature target human-miRNA. They specifically silenced miRNA expression (miR-122) up-regulating expression of hundreds of genes predicted to be repressed by miR **[3]**. Paradoxically, antagomir treatment also revealed a significant number of down-regulated genes that may be activated as opposed to repressed by miR **[4, 5, 6]**. Computational approaches are thus likely to remain an important means of studying antagomir targets for the foreseeable future, not least as a means of directing wet-lab experiments. Several algorithms have been used to predict antagomir targets in animal species;

these are listed in **Table 1 (see Supplementary material)**. We built an experimental antagomir in a way that it could adhere to several humiR, respectively. Interactions are viewed in a global context by predicting folds for the entire miRNA, rather than just its 3'- UTR or seed match. The stem-loop sequences of humiR have been used in our experiment.

Figure 1: Screenshot of the database main page

The probability profile displays predicted accessible sites on the target RNA. Because an accessible site can be targeted by a number of antisense oligos, selection of the "optimal" one can be based on binding energy, together with other empirical rules such as GC content, avoidance of GGGG (or more stringent GGG) motifs, etc. Stronger binding is indicated by smaller binding energy (stacking energies are *negatively valued*). For example, an antisense oligo with a binding energy of -10 kcal/mol is more effective than an oligo with a binding energy of -5 kcal/mol. The antisense oligo binding energy is a weighted sum of the antagomir/miRNA stacking energies for

the hybrid formed by the antagomir and the targeted sequence. For a base-pair stack, the weight for the sum is calculated by the probability of the unpaired dinucleotide in the target sequence that is involved in the stack. This weighting scheme accounts for the structural variation at the target site among the structures in the sample (**Figure 1**).

Implementation:
Database:
The antagomir database is based on the following assumptions.
(1) Based on the humiR sequence the following details about the putative antagomirs are obtained with the help of Sfold.
Column 1: target position (starting - ending)
Column 2: target sequence (5'→3')
Column 3: antisense oligo (5'→3')
Column 4: GC content
Column 5: oligo binding energy (kcal/mol)
(2) The database also includes the secondary structure of antagomirs along with the other details mentioned above. The secondary structures have been designed with the help of mfold.

Tool:
A tool has been integrated in this database. The tool accepts a sequence of 20-25 nuleotides and if it finds a match with the existing putative antagomirs in the database then it returns the respective target i.e. humiR ID along with the secondary structure of antagomirs.

Statistical analysis of predicted targets:
Negative normalized free energy:
The occurrence of favourable hybridizations of short antagomirs with long miRNAs can frequently be attributed to chance: the longer the miRNA, the more likely the incidence. In order to eliminate the effect of sequence length on our measure of free energy **[7, 8]**, we define the negative normalized free energy where m is the length of the target sequence searched, and n is the length of the antagomir as shown in equation 1:
$g_n = -(g/\log(mn))$.

Extreme value statistics:
Extreme value distributions (EVDs) are limiting distributions that describe the minimum or maximum of independent random variables **[9]**. If we consider the antagomir-miRNA duplex energy estimation to be essentially an optimization procedure that produces a minimum, the negative normalized free energy described above is a corresponding maximum, and can be described by an EVD having a distribution function of the form; $P[G \leq t] = D(t) = \exp(-\exp(a-t/b))$ (equation 2). A transformation then converts this distribution function into a straight line: $\log(-\log(D)) = (a-t)/b = (-1/b)t + a/b$ (equation 3). By scanning for targets of random antagomir sequences in the miRNA sequences in the dataset, we obtain a set of negative normalized free energies, which we expect will follow an EVD. We then transform the distribution function of the empirical EVD into a straight line, as in Equation 3, and estimate the parameters of the EVD by a linear least squares fit to the line $y =$

$mx + c$, obtaining b=-1/m (equation 4) and a=cb (equation 5). We can now compute, for each predicted antagomir-miRNA duplex, a p-value, the probability that the same or a more favorable free energy is observed due to chance: $P[Z \geq g_n] = 1 - \exp(-\exp((a- g_n)/b))$ (equation 6), where a and b are estimated EVD parameters, and g_n is the negative normalized free energy from equation 1 **[7]**.

Results:
Antagomir database is the central online repository for antagomirs sequence data, annotation and target prediction. The current release (ver.1) contains 22 humiRNA from *Homo sapiens*, expressing 53 distinct putative antagomir sequences **(Table 2, see Supplementary material)**. The humiRID and the secondary structure of antagomir are included in the database. The best four matches of the secondary structure with respect to free energy are given in **Figure 2**.
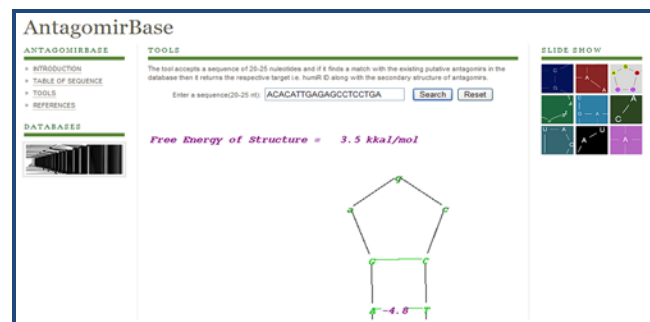

Figure 2: Screenshot of the database showing the UID (User Interface Design)

Conclusion:
The prediction of antagomir targets is an open and difficult problem in spite of a few years of the existing research. Antagomir Database designed based on human miR targets, is shown to provide predictions characterized by favorable sensitivity, which comes at a price of an increased number of predictions. Our future work will concentrate in the inclusion of more miRNAs from different species as targets of antagomirs.

References:
[1] Thadani R & Tammi MT. *BMC Bioinformatics*. 2006 **7**(Suppl 5): S20 [PMID: 17254305]
[2] Krützfeldt J *et al. Nature* 2005 **438**: 685 [PMID: 16258535]
[3] Cheng AM *et al. Nucleic Acids Res*. 2005 **33**: 1290 [PMID: 15741182]
[4] Weiler J *et al. Gene Ther*. 2006 **13**: 496 [PMID: 16195701]
[5] Meister G *et al. RNA*. 2004 **10**: 544 [PMID: 14970398]
[6] Mattes J *et al. Am J Respir Cell Mol Biol*. 2007 **36**: 8 [PMID: 16917074]
[7] Rehmsmeier M *et al. RNA*. 2004 **10**: 1507 [PMID: 15383676]
[8] Karlin S & Altschul SF. *Proc Natl Acad Sci U S A*. 1990 **87**: 2264 [PMID: 2315319]
[9] Gumbel EJ: *Statistics of Extremes New York: Columbia University Press*; 1958

BIOINFORMATION

Supplementary material:

Table 1: Algorithms used to predict antagomir targets

Software used	Purpose
Sfold	To determine the antagomir sequence
mfold	Designing the secondary structure of antagomirs

Table 2: List of putative antagomirs and their target sequence

Seq No.	Target Position	Target Sequence	Putative Antagomirs	Binding energy	GC Content
1	27- 46	UCAGGAGGCUCUCAAUGUGU	ACACATTGAGAGCCTCCTGA	-8.7	50.0%
2	1-20	AGGAAGCUUCUGGAGAUCCU	AGGATCTCCAGAAGCTTCCT	-11.3	50.0%
3	35- 54	GAUGUUUACCGUUUAAAUCC	GGATTTAAACGGTAAACATC	-4.6	35.0%
4	35- 54	CCCUUAGAUGCUAAGAUGGG	CCCATCTTAGCATCTAAGGG	-8.7	50.0%
5 a	34- 53	UGUGAGGUAUUUUAGUAACA	TGTTACTAAAATACCTCACA	-2.6	30.0%
5 b	35- 54	GUGAGGUAUUUUAGUAACAU	ATGTTACTAAAATACCTCAC	-2.6	30.0%
5 c	36- 55	UGAGGUAUUUUAGUAACAUC	GATGTTACTAAAATACCTCA	-2.6	30.0%
5 d	37- 56	GAGGUAUUUUAGUAACAUCA	TGATGTTACTAAAATACCTC	-2.6	30.0%
6 a	25- 44	GUCUAGUAUGUCUCAUCCCC	GGGGATGAGACATACTAGAC	-8.0	50.0%
6 b	33- 52	UGUCUCAUCCCCUACUAGAC	GTCTAGTAGGGGATGAGACA	-8.0	50.0%
7	79- 98	UGGGUUUGAAUAUCGACAGC	GCTGTCGATATTCAAACCCA	-12.2	45.0%
8 a	23- 42	UGGGAUUUGCUUCCUGUCAC	GTGACAGGAAGCAAATCCCA	-7.0	50.0%
8 b	24- 43	GGGAUUUGCUUCCUGUCACA	TGTGACAGGAAGCAAATCCC	-7.0	50.0%
9	31- 50	GGUAGCUUUUUGAGAUCUAC	GTAGATCTCAAAAAGCTACC	-0.8	40.0%
10 a	28- 47	UUGCCUUGUGGCAUGGAGUU	AACTCCATGCCACAAGGCAA	-4.4	50.0%
10 b	31- 50	CCUUGUGGCAUGGAGUUCAA	TTGAACTCCATGCCACAAGG	-4.4	50.0%
10 c	32- 51	CUUGUGGCAUGGAGUUCAAG	CTTGAACTCCATGCCACAAG	-4.4	50.0%
10 d	33- 52	UUGUGGCAUGGAGUUCAAGC	GCTTGAACTCCATGCCACAA	-4.4	50.0%
10 e	33- 52	UGUGGCAUGGAGUUCAAGCA	TGCTTGAACTCCATGCCACA	-4.4	50.0%
11 a	25- 44	UGGGCCUGGGUUUCUGAACA	TGTTCAGAAACCCAGGCCCA	-8.7	50.0%
11 b	27- 46	GGCCUGGGUUUCUGAACACA	TGTGTTCAGAAACCCAGGCC	-8.7	50.0%
12	29- 48	UCCUCUCAACAGUAGUCAGG	CCTGACTACTGTTGAGAGGA	-6.2	50.0%
13 a	50- 69	UAGAUAACUAAAUCGACAAC	GTTGTCGATTTAGTTATCTA	-1.4	30.0%
13 b	51- 70	AGAUAACUAAAUCGACAACA	TGTTGTCGATTTAGTTATCT	-1.4	30.0%
13 c	52- 71	GAUAACUAAAUCGACAACAA	TTGTTGTCGATTTAGTTATC	-1.4	30.0%
14 a	50- 69	GUUGUUGUCUUACUGCGCUC	GAGCGCAGTAAGACAACAAC	-9.9	50.0%
14 b	53- 72	GUUGUCUUACUGCGCUCAAC	GTTGAGCGCAGTAAGACAAC	-9.9	50.0%
14 c	56- 75	GUCUUACUGCGCUCAACAAC	GTTGTTGAGCGCAGTAAGAC	-9.9	50.0%
15 a	48- 67	UGUUGUUCUGAUGUACUACG	CGTAGTACATCAGAACAACA	-5.4	40.0%
15 b	49- 68	GUUGUUCUGAUGUACUACGA	TCGTAGTACATCAGAACAAC	-5.4	40.0%
15 c	50- 69	UUGUUCUGAUGUACUACGAC	GTCGTAGTACATCAGAACAA	-5.4	40.0%
15 d	51- 70	UGUUCUGAUGUACUACGACA	TGTCGTAGTACATCAGAACA	-5.4	40.0%
15 e	52- 71	GUUCUGAUGUACUACGACAA	TTGTCGTAGTACATCAGAAC	-5.4	40.0%
15 f	53- 72	UUCUGAUGUACUACGACAAC	GTTGTCGTAGTACATCAGAA	-5.4	40.0%
15 g	54- 73	UCUGAUGUACUACGACAACA	TGTTGTCGTAGTACATCAGA	-5.4	40.0%
16	91- 110	UGACAUAAACACUCCGCUCU	AGAGCGGAGTGTTTATGTCA	-11.5	45.0%
17 a	46- 65	UGUGUGGUAUCCGUAUAGUC	GACTATACGGATACCACACA	-6.8	45.0%
17 b	47- 66	GUGUGGUAUCCGUAUAGUCA	TGACTATACGGATACCACAC	-6.8	45.0%
17 c	48- 67	UGUGGUAUCCGUAUAGUCAC	GTGACTATACGGATACCACA	-6.8	45.0%
17 d	49- 68	GUGGUAUCCGUAUAGUCACA	TGTGACTATACGGATACCAC	-6.8	45.0%
17 e	50- 69	UGGUAUCCGUAUAGUCACAG	CTGTGACTATACGGATACCA	-6.8	45.0%
18	44- 63	UGUGAGCAAUAGUAAGGAAG	CTTCCTTACTATTGCTCACA	-10.3	40.0%
19	55- 74	GGUGAGUUUGGGAUUUGAAA	TTTCAAATCCCAAACTCACC	-10.1	40.0%
20 a	54- 73	UGGGUUGAACUGUGUGGACA	TGTCCACACAGTTCAACCCA	-6.8	50.0%
20 b	55- 74	GGGUUGAACUGUGUGGACAA	TTGTCCACACAGTTCAACCC	-6.8	50.0%
20 c	56- 75	GGUUGAACUGUGUGGACAAG	CTTGTCCACACAGTTCAACC	-6.8	50.0%
20 d	57- 76	GUUGAACUGUGUGGACAAGC	GCTTGTCCACACAGTTCAAC	-6.8	50.0%
20 e	59- 78	UGAACUGUGUGGACAAGCUC	GAGCTTGTCCACACAGTTCA	-6.8	50.0%
21	48- 67	UUUGGGCAGCUCAGGCAAAC	GTTTGCCTGAGCTGCCCAAA	-8.4	55.0%
22 a	43- 62	CACUGGUGAGGUAACAGGAU	ATCCTGTTACCTCACCAGTG	-12.9	50.0%
22 b	44- 63	ACUGGUGAGGUAACAGGAUC	GATCCTGTTACCTCACCAGT	-12.9	50.0%