Supplementary Materials

RIsearch: Fast RNA-RNA Interaction Search using a Simplified Nearest-Neighbor Energy Model

Anne Wenzel, Erdinç Akbaşlı and Jan Gorodkin

August 10, 2012

1 The RIsearch scoring scheme

![Figure 1: The scoring matrix. Free energy contributions in kcal/mol are derived from Turner’s nearest neighbor parameters \[1\], and multiplied by -100 to build the score. This has the advantage that we get integers and can maximize over all possible local ‘alignments’. Letters in the first column denote consecutive nucleotides in the query sequence (5’ to 3’). Letters in the first row refer to the target sequence (3’ to 5’). Reading example: The energy contribution of stacking a CG pair onto an AU pair (-2.2 kcal/mol) can be found in the row ‘AC’ and column ‘UC’, resulting in 220. ∞ prevents alignments ending with a mismatch, it is set to -2000 in the implementation. * denotes cases that the algorithm itself forbids, such as having a gap in one sequence followed by a gap in the other (insertion after deletion). It is set to an arbitrary value, as it is never read.](image-url)
Table 1: Possibilities and limitations of the scoring scheme. The table shows the free energies in kcal/mol as computed by different methods for the specified duplexes. Example duplexes with corresponding energies had to be extended with generic WC-helices to be predicted by the different tools. Upper part of table: Tools based on Turner 1999 parameters, lower part: tools based on 2004 parameter set. The top rows ('TurnerNNDBxx') should be considered as reference, methods deviate from this because of rounding and because some of the more complex rules are not implemented in the algorithms. (*) The same predictions as given by RNAduplex, were also reported by RNAcofold, BINDIGO (web server only), and PairFold (web server). Apparently, RNAhybrid does not apply the intermolecular initiation energy that is typically given with +4.1 kcal/mol, thus yielding much lower energies.

The interaction in the first column is a self-complementary duplex, but none of the methods seems to implement the suggested symmetry correction. The third duplex shows the stack of GU followed by UG in two different contexts, stabilizing and destabilizing. As we only look at one previous position we cannot incorporate these scores. Similarly, look-up tables for the $2 \times 2$ internal loop in duplex number six are not used by RIsearch.

| Duplex  | TurnerNNDB99 | RNAduplex* | RNAplex | RNAhybrid [2] | RIsearch t99 | TurnerNNDB04 | DuplexFold [3] | RIsearch t04 |
|---------|--------------|------------|---------|---------------|--------------|--------------|---------------|--------------|
| 5’AGCGCU 3’UCGCGA | -7.94 | -8.30 | -8.30 | -13.40 | -8.41 | -7.94 | -8.30 | -8.41 |
| 5’GCACG 3’CGUGC | -7.64 | -8.00 | -8.00 | -10.10 | -6.01 | -7.64 | -8.00 | -6.01 |
| 5’GGUCGUGU 3’CUGGUGCG | -7.52 | -5.80 | -5.80 | -6.30 | -1.16 | -7.52 | -5.80 | -1.16 |
| 5’GGCCGUC 3’CCG-CAG | -7.52 | -10.50 | -10.50 | -9.90 | -4.81 | -7.52 | -10.50 | -4.81 |
| 5’GCGAACUGCACG 3’CGCU---CGUGC | -7.52 | -4.20 | -4.20 | -15.10 | -10.56 | -7.52 | -4.20 | -10.56 |
| 5’CAGACG 3’GUAGGC | -7.52 | -15.10 | -15.10 | -11.00 | -0.20 | -7.52 | -15.10 | -0.20 |
| 5’GCGAACUGCACG 3’CGCU---CGUGC | -7.52 | -11.00 | -11.00 | -10.60 | -6.84 | -7.52 | -11.00 | -6.84 |

Figure 2: Accuracy on simulated data (next page). Variations of Figure 3 from the main paper for sequences of different length and GC-content. The Pearson product-moment correlation coefficient $r$ and Spearman’s rank correlation coefficient $\rho$ are given in panel (b).
Figure 2: Caption on previous page.
2 Speed and memory benchmarks

2.1 RIssearch vs. RNAplex simple form

All measurements in this section were done with the GNU time command. Run times are given in seconds, the total amount of CPU time spent in user and kernel mode (%U + %S). For memory consumption we report the maximum resident set size (RSS) in Kilobytes (%M). There is a known bug in time, so this value is four times too large. However, the relative memory reduction is not affected. Values given in the following tables are what is reported by time without correction.

For RNAplex we specified parameter \(-f 2\), so it uses the simple energy model in the backtracking as well, instead of the full energy model. For both tools, RIssearch and RNAplex, we set the per-nucleotide penalty to 0.3 kcal/mol. All other parameters were left at their default values.

| machine | RNAplex time | RNAplex mem | RIssearch time | RIssearch mem | comparison time | comparison mem |
|---------|--------------|-------------|---------------|--------------|----------------|----------------|
| laptop  | 24.55        | 5600        | 8.81          | 2288         | 2.79           | 2.45           |
| server  | 12.64        | 5744        | 5.28          | 2448         | 2.39           | 2.35           |

Table 2: Comparison of runtime and memory requirements. Here we just repeated the ‘Single Sequence Runtime benchmark’ as given in the Supplementary Material to [7]. Query a set of 19 bacterial sRNAs (average length 131nt) in 100 target sequences of length 1200nt. time is runtime in seconds, mem is max RSS [kB], both as reported by time (U+S / M). For ‘comparison’ it is the speedup, respectively the memory reduction achieved by RIssearch in comparison to RNAplex. RIssearch was compiled with the default \(-O3\); RNAplex was compiled with default parameters. Additionally, RNAplex was compiled specifying \(-O3\), but performance did not improve. The laptop used has an Intel® Core™2 Duo CPU P8700 with 2.53 GHz, the server a Intel® Xeon® CPU X5570 with 2.93 GHz.

| #query | RNAplex    | RIssearch  | speedup |
|--------|------------|------------|---------|
| 1      | 17201.00   | 94.91      | 181.23  |
| 5      | 18547.07   | 472.18     | 39.28   |
| 100    | 39988.77   | 9093.04    | 4.40    |
| 500    | 130301.59  | 45598.25   | 2.86    |
| 1223   | 288182.65  | 112141.20  | 2.57    |

Table 3: Runtime on large genomic sequence. The complete human chromosome 1 without any filtering (249,250,621 nt) was used as target. The first column shows the number of miRNA sequences in the query set, up to 1223 which comprises all human mature miRNAs in miRBase 16. The second and third column give the runtime in seconds, the relative speedup of RIssearch is given in the last column. Memory consumption was reduced by a factor of 1.44 in all cases. From the data it seems that RNAplex uses much more time for the initialization. If cleaned for that, simply by subtracting the runtimes for a single microRNA, the speedup ranges between 2.4 and 3.6. Both tools compiled with \(-O3\) optimization.
| seq lengths[nt] | RNAplex | RIsearch | comparison |
|----------------|---------|----------|------------|
|                | query   | target   |            | time      | mem      | time      | mem      |
| 10             | 1E+3    | 0.00     | 0.00       | 2304      | NaN      | 2.40      |          |
| 10             | 1E+4    | 0.00     | 0.00       | 2624      | NaN      | 2.23      |          |
| 10             | 1E+5    | 0.04     | 0.01       | 6176      | 4.00     | 1.71      |          |
| 10             | 1E+6    | 0.68     | 0.18       | 37952     | 3.78     | 1.49      |          |
| 10             | 1E+7    | 29.77    | 1.89       | 354336    | 15.75    | 1.45      |          |
| 10             | 1E+8    | 2870.63  | 19.16      | 3518416   | 149.82   | 1.44      |          |
| 25             | 1E+3    | 0.00     | 0.00       | 2304      | NaN      | 2.22      |          |
| 25             | 1E+4    | 0.01     | 0.00       | 2640      | NaN      | 2.22      |          |
| 25             | 1E+5    | 0.11     | 0.04       | 6176      | 2.75     | 1.70      |          |
| 25             | 1E+6    | 1.36     | 0.44       | 37936     | 3.09     | 1.49      |          |
| 25             | 1E+7    | 36.70    | 4.50       | 354336    | 8.16     | 1.45      |          |
| 25             | 1E+8    | 2843.67  | 45.35      | 3518432   | 62.70    | 1.44      |          |
| 25             | 1E+9    | 274312.68| 437.99     | 35159040  | 626.30   | 1.44      |          |
| 100            | 1E+3    | 0.00     | 0.00       | 2384      | NaN      | 2.36      |          |
| 100            | 1E+4    | 0.04     | 0.01       | 2672      | 4.00     | 2.21      |          |
| 100            | 1E+5    | 0.41     | 0.17       | 6160      | 2.41     | 1.71      |          |
| 100            | 1E+6    | 4.41     | 1.73       | 37952     | 2.55     | 1.49      |          |
| 100            | 1E+7    | 67.41    | 17.47      | 354352    | 3.86     | 1.45      |          |
| 100            | 1E+8    | 3158.54  | 176.02     | 3518416   | 17.94    | 1.44      |          |
| 100            | 1E+9    | 274929.19| 1693.63    | 35159040  | 162.33   | 1.44      |          |
| 1000           | 1E+3    | 0.04     | 0.01       | 2480      | 4.00     | 2.35      |          |
| 1000           | 1E+4    | 0.41     | 0.17       | 2800      | 2.41     | 2.19      |          |
| 1000           | 1E+5    | 4.07     | 1.70       | 6176      | 2.39     | 1.71      |          |
| 1000           | 1E+6    | 40.93    | 17.06      | 37952     | 2.40     | 1.49      |          |
| 1000           | 1E+7    | 444.54   | 170.71     | 354336    | 2.60     | 1.45      |          |
| 1000           | 1E+8    | 7330.30  | 1671.38    | 3518416   | 4.39     | 1.44      |          |
| 1000           | 1E+9    | 320057.08| 16612.16   | 35159040  | 19.27    | 1.44      |          |

Table 4: Speed and memory benchmark on randomly generated sequences of different lengths. Columns 1 and 2 give the length (in nucleotides) of the query and target sequences respectively. Columns 3 and 4 list time \((user + system)\) in seconds and memory requirements (maximum RSS [kB]), both as reported by the \texttt{time} command for RNAplex. Columns 5 and 6 show the same for RIsearch. Columns 7 and 8 show the improvement of RIsearch over RNAplex. RIsearch is at least 2.39 times as fast in all measurable cases.

Here, RNAplex performed better with the default compiler flags again, i.e., not specifying \texttt{-O 3}.

### 2.2 Using accessibility information

In order to run RNAplex -a, accessibility profiles need to be computed with RNAplfold. As memory requirements for the larger chromosomes exceeded our resources, we used human chromosome 21 here (the smallest one). With the recommended settings, RNAplfold runs more than 26 hours and uses more than 24 GiB to compute the accessibility profiles. The subsequent run of RNAplex in its current implementation also is more resource-demanding when making use of this information in comparison to the simple version. Screening this target with 5 miRNAs as query (as in Supplementary Table 3) takes:

|                  | time (minutes) | mem (GiB) |
|------------------|----------------|-----------|
| RNAplex -a (with accessibility) | 27.1          | 8.1       |
| RNAplex -c 30 (w/o accessibility) | 13.6          | 0.6       |
| RIsearch -d 30   | 1.7            | 0.4       |
3 Performance on known bacterial sRNA interactions
| sRNA-mRNA     | Sensitivity | PPV       | F-measure | MCC       |
|--------------|-------------|-----------|-----------|-----------|
|               | plex-a | plex-c | RIs99 | RIs04 | plex-a | plex-c | RIs99 | RIs04 | plex-a | plex-c | RIs99 | RIs04 | plex-a | plex-c | RIs99 | RIs04 |
| GcvB-gltI*   | 0.923 | 0.846 | 1.000 | 1.000 | 1.000 | 0.393 | 0.448 | 0.448 | 0.961 | 0.537 | 0.619 | 0.619 | 0.961 | 0.537 | 0.619 | 0.619 |
| GcvB-argT*   | 0.875 | 1.000 | 1.000 | 1.000 | 0.824 | 0.800 | 1.000 | 1.000 | 0.848 | 0.889 | 1.000 | 1.000 | 0.848 | 0.889 | 1.000 | 1.000 |
| GcvB-dppA    | 1.000 | 0.941 | 0.941 | 0.941 | 0.515 | 0.485 | 0.500 | 0.485 | 0.680 | 0.640 | 0.653 | 0.640 | 0.955 | 1.000 | 1.000 | 1.000 |
| GcvB-livJ    | 0.955 | 1.000 | 1.000 | 1.000 | 0.955 | 1.000 | 1.000 | 1.000 | 0.955 | 1.000 | 1.000 | 1.000 | 0.955 | 1.000 | 1.000 | 1.000 |
| GcvB-livK*   | 1.000 | 1.000 | 1.000 | 1.000 | 0.565 | 0.565 | 0.481 | 0.929 | 0.722 | 0.722 | 0.650 | 0.963 | 0.752 | 0.752 | 0.694 | 0.964 |
| GcvB-oppA    | 1.000 | 1.000 | 1.000 | 1.000 | 0.500 | 0.581 | 1.000 | 1.000 | 0.556 | 0.640 | 1.000 | 1.000 | 0.556 | 0.640 | 1.000 | 1.000 |
| GcvB-STM4351 | 0.889 | 1.000 | 1.000 | 1.000 | 0.821 | 0.902 | 0.902 | 0.902 | 0.300 | 0.761 | 0.761 | 0.761 | 0.300 | 0.761 | 0.761 | 0.761 |
| MicA-lamB    | 1.000 | 1.000 | 1.000 | 1.000 | 0.821 | 0.902 | 0.902 | 0.902 | 0.300 | 0.761 | 0.761 | 0.761 | 0.300 | 0.761 | 0.761 | 0.761 |
| MicA-ompA    | 1.000 | 0.938 | 0.938 | 0.938 | 0.565 | 0.565 | 0.481 | 0.929 | 0.722 | 0.722 | 0.650 | 0.963 | 0.752 | 0.752 | 0.694 | 0.964 |
| DsrA-rpoS*   | 0.571 | 0.778 | 0.778 | 0.778 | 0.615 | 0.875 | 0.875 | 0.875 | 0.617 | 0.882 | 0.882 | 0.882 | 0.617 | 0.882 | 0.882 | 0.882 |
| RprA-rpoS    | 0.316 | 0.733 | 0.733 | 0.733 | 0.300 | 0.733 | 0.733 | 0.733 | 0.300 | 0.761 | 0.761 | 0.761 | 0.300 | 0.761 | 0.761 | 0.761 |
| IstR-tisA*   | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.902 | 0.902 | 0.902 | 1.000 | 0.906 | 0.906 | 0.906 | 1.000 | 0.906 | 0.906 | 0.906 |
| MicC-ompC    | 0.727 | 0.727 | 0.727 | 0.727 | 0.842 | 0.842 | 0.842 | 0.842 | 0.853 | 0.853 | 0.853 | 0.853 | 0.853 | 0.853 | 0.853 | 0.853 |
| MicF-ompF    | 0.800 | 0.800 | 0.800 | 0.800 | 0.952 | 0.960 | 0.960 | 0.960 | 0.873 | 0.960 | 0.960 | 0.960 | 0.873 | 0.960 | 0.960 | 0.960 |
| RyhB-sdhD    | 0.588 | 0.754 | 0.754 | 0.754 | 0.741 | 0.871 | 0.754 | 0.754 | 0.767 | 0.875 | 0.759 | 0.759 | 0.767 | 0.875 | 0.759 | 0.759 |
| RyhB-sodB    | 1.000 | 1.000 | 1.000 | 1.000 | 0.621 | 0.900 | 1.000 | 1.000 | 0.671 | 0.905 | 1.000 | 1.000 | 0.671 | 0.905 | 1.000 | 1.000 |
| SgrS-ptsG    | 0.739 | 0.739 | 0.739 | 0.739 | 0.850 | 0.850 | 0.850 | 0.850 | 0.860 | 0.860 | 0.860 | 0.860 | 0.860 | 0.860 | 0.860 | 0.860 |

| sRNA-mRNA     | SF, −subopt | SF, +subopt | LF, −subopt | LF, +subopt |
|--------------|-------------|-------------|-------------|-------------|
| GcvB-argT*   | 0.787 | 0.639 | 0.655 | 0.656 | 0.736 | 0.599 | 0.581 | 0.641 | 0.734 | 0.603 | 0.597 | 0.639 | 0.748 | 0.611 | 0.607 | 0.644 |
| GcvB-dppA    | 0.846 | 0.917 | 0.906 | 0.919 | 0.785 | 0.821 | 0.873 | 0.898 | 0.787 | 0.840 | 0.866 | 0.891 | 0.801 | 0.854 | 0.877 | 0.899 |

Table 5: Prediction accuracy. Sensitivity (also recall or true positive rate (TPR)), positive predictive value (PPV) or precision, F-measure, and MCC (harmonic and geometric mean of the first two) were calculated for the set of 17 experimentally verified sRNA-mRNA interactions. Averages are given in the last four lines. We tested RNAplex using precomputed accessibility profiles (plex-a), as well as the basic version (plex-c) with per-nucleotide penalty \(-c 30\). RIs99 and RIs04 stand for RIsearch using 1999 and 2004 parameter set, respectively, also with per-nucleotide penalty of 30 (\(= 0.3\) kcal/mol). Numbers shown in gray italics refer to interactions that have not been found as the single best-scoring, but only when taking into account suboptimal solutions. For interactions marked with an asterisk (*), we have extracted a longer and a shorter version from the literature. For example, [8] identified residues that were protected in \textit{in vitro} footprinting experiments and extended the target sites by biocomputational predictions. For this table we used the shorter forms (SF, with boundaries as given in Main Table 1). When instead using the longer forms (LF, maximum numbers of pairs shown in the original papers), we get the average measures as reported in the last two lines. When excluding suboptimal solutions (−subopt), RNAplex with accessibility misses only one interaction, while the other methods miss five each. When the top prediction does not share a base pair with the experimentally verified location, they contribute with 0 to the average. In all these cases it is enough to look at the three best suboptimal solutions in order to find one that overlaps the verified location. When allowing these suboptimal solutions (+subopt), the values as printed in gray italics contribute to the average.
4 Identifying human miRNA targets on chromosome-scale

4.1 Ranking known targets
| interaction pair | chromosome | mRNA name | miRNA name | GUUGle  | GUUGle* | TargetScanS | miRanda | GUUGle* ∩ | Rank product | Relative hit score |
|------------------|------------|-----------|------------|---------|---------|-------------|---------|------------|--------------|------------------|
| AGTR1            | miR-155    | 3+        | 97,261,371 | 13.25   | 113,671 | 14.37       | 180,386 | 8          | 570,135      | 4.59  5.66 |
| BCL2             | miR-16     | 18−       | 39,669,911 | 18.48   | 718     | 18.90       | 3,843   | 10         | 15,346       | 6.17  2.15 |
| SLC7A1           | miR-122    | 13−       | 49,987,579 | 22.88   | 125     | 23.80       | 290     | 9          | 233,723      | 22.48 20.62 |
| TPP3             | miR-16     | 16−       | 39,076,387 | 19.41   | 584     | 20.80       | 1,505   | 7          | 295,580      | 3.81  3.81 |
| CLOCK            | miR-141    | 4−        | 92,263,291 | 14.93   | 211     | 19.20       | 33,745  | 9          | 189,863      | 16.40 14.60 |
| CXCL12           | miR-23a    | 10−       | 67,660,041 | 18.01   | 12,188  | 19.80       | 12,591  | 8          | 345,446      | 16.40 14.60 |
| CYP1B1           | miR-127b   | 2−        | 123,169,357| 26.94   | 27      | 28.20       | 32      | 12         | 8,253         | 16.40 14.60 |
| E2F3             | miR-34a    | 6+        | 85,195,913 | 18.53   | 45,062  | 17.20       | 226,049 | 9          | 219,194      | 16.40 14.60 |
| EZH2             | miR-101    | 7−        | 77,667,101 | 15.45   | 3,659   | 16.90       | 7,342   | 9          | 69,204       | 16.40 14.60 |
| PARP8            | miR-145    | 5−        | 88,834,294 | 20.01   | 28,064  | 21.80       | 32,349  | 10         | 52,652       | 16.40 14.60 |
| FSTL1            | miR-206    | 3−        | 97,261,371 | 15.08   | 67,128  | 18.40       | 31,741  | 8          | 345,446      | 16.40 14.60 |
| GJA1             | miR-1      | 6+        | 85,195,913 | 12.49   | 237,213 | 15.03       | 180,291 | 8          | 1,087,518    | 16.40 14.60 |
| HAND2            | miR-1      | 4−        | 85,263,291 | 12.14   | 29,151  | 12.20       | 106,885 | 8          | 868,628      | 16.40 14.60 |
| HOXA1            | miR-10a    | 7−        | 77,667,101 | 12.29   | 162,987 | 15.93       | 75,673  | 8          | 285,837      | 16.40 14.60 |
| KIT              | miR-221    | 4−        | 92,263,291 | 14.95   | 3,659   | 16.90       | 7,342   | 9          | 667,674      | 16.40 14.60 |
| KIT              | miR-222    | 4−        | 92,263,291 | 15.20   | 134,277 | 15.25       | 301,413 | 7          | 725,891      | 16.40 14.60 |
| KRAS             | let-7a     | 12−       | 63,199,786 | 14.71   | 46,793  | 16.30       | 59,593  | 7          | 1,815,721    | 16.40 14.60 |
| LIN28A           | let-7b     | 1+        | 111,179,527| 25.61   | 444     | 27.00       | 670     | 14         | 8,845        | 16.40 14.60 |
| MAPK14           | miR-24     | 6+        | 85,195,913 | 27.07   | 154     | 27.10       | 653     | 10         | 290,052      | 16.40 14.60 |
| MYCN             | miR-101    | 2+        | 123,169,357| 13.25   | 58,119  | 13.85       | 90,266  | 9          | 111,230      | 16.40 14.60 |
| NRAS             | let-7a     | 1−        | 111,179,527| 13.20   | 216,020 | 17.70       | 50,049  | 9          | 1,812,612    | 16.40 14.60 |
| PTEN             | miR-19a    | 10+       | 67,660,041 | 16.38   | 1,769   | 17.70       | 3,779   | 10         | 11,427       | 16.40 14.60 |
| ARHGAP32         | miR-132    | 11−       | 64,323,812 | 18.11   | 11,583  | 18.80       | 38,484  | 10         | 52,347       | 16.40 14.60 |
| SMC1A            | let-7e     | X−        | 53,731,681 | 21.33   | 1,629   | 22.20       | 3,416   | 12         | 33,335       | 16.40 14.60 |
| TMSB4X           | miR-1      | X+        | 53,731,681 | 16.00   | 925     | 16.90       | 2,168   | 10         | 171,143      | 16.40 14.60 |
| TPM1             | miR-21     | 15+       | 41,621,622 | 13.28   | 30,154  | 15.60       | 29,478  | 9          | 50,347       | 16.40 14.60 |

Table 6: The table shows the interacting gene and miRNA, with chromosome information for target site location (name, strand, and the number of bases that have been masked). For each tool, we report a threshold (thr) found by looking for the highest scoring hit that overlaps a verified interaction site of this mRNA-miRNA pair. For RISearch and RNAplex this threshold is the ΔG, for GUUGle the match length, for TargetScanS the context+ score, for miRanda again the energy. As count we report the number of hits (of the given miRNA within this chromosome, direction) that fulfill this threshold, i.e., all predictions that score at least as good as the best verified interaction for that pair. GUUGle* uses only the seed of the miRNA as query (nt 1–8) instead of the whole mature miRNA (in GUUGle). For RISearch and RNAplex we additionally intersected the hits with those from GUUGle* and present the counts in the last two columns (number of predictions that overlap complete GUUGle seed matches and fulfill the energy threshold as applied for RISearch and RNAplex respectively). NF stands for ‘not found’. We use two different methods to evaluate performance and give the results in the last two rows where the best result is highlighted in bold.
### 4.2 Efficacy of RIsearch as filter

| interaction pair | mRNA   | TargetScanS | miRanda | 
|------------------|--------|-------------|---------| 
|                 |        | unfiltered  | G*      | RIs     | G*RIs     | unfiltered  | G*      | RIs     | G*RIs     |
|                  |        | G*∩RIs     |         |         |           |           |         |         |         | 
| AGTR1            | miR-155 | 13,575      | 0.00%  | 57.51%  | 57.67%    | 16,753    | 18.24%  | 49.41%  | 58.82%    |
| BCL2             | miR-16  | 6,212       | 0.00%  | 16.85%  | 17.06%    | 7,147     | 9.49%   | 20.99%  | 27.12%    |
| SLC7A1           | miR-122 | 4,752       | 0.00%  | 5.35%   | 5.89%     | 6,139     | 9.09%   | 7.31%   | 17.66%    |
| TPPP3            | miR-16  | 7,384       | 0.00%  | 15.28%  | 15.59%    | 7,599     | 8.19%   | 17.70%  | 23.41%    |
| CLOCK            | miR-141 | 15,922      | 0.00%  | 47.46%  | 47.61%    | 14,890    | 18.66%  | 38.72%  | 50.64%    |
| CXCL12           | miR-23a | 14,275      | 28.13% | 56.39%  | 64.57%    | 11,865    | 8.42%   | 41.93%  | 47.91%    |
| CYP1B1           | miR-27b | 23,204      | 0.00%  | 23.59%  | 23.83%    |           |         |         |           |
| E2F3             | miR-34a | 10,783      | 0.00%  | 1.28%   | 2.14%     | 12,938    | 9.90%   | 4.04%   | 15.82%    |
| EZH2             | miR-101 | 8,447       | 0.00%  | 68.06%  | 68.08%    | 6,139     | 9.09%   | 7.31%   | 17.66%    |
| PARP8            | miR-145 | 15,476      | 11.45% | 25.85%  | 35.67%    | 15,476    | 11.45%  | 25.85%  | 35.67%    |
| FSTL1            | miR-206 | 15,069      | 0.00%  | 25.78%  | 25.99%    | 19,702    | 9.00%   | 30.56%  | 38.80%    |
| GJA1             | miR-1   | 12,865      | 0.00%  | 67.40%  | 67.44%    | 18,334    | 9.64%   | 60.36%  | 65.04%    |
| GJA1             | miR-206 | 12,865      | 0.00%  | 67.40%  | 67.44%    | 18,334    | 9.64%   | 60.36%  | 65.04%    |
| HAND2            | miR-1   | 13,614      | 0.00%  | 65.60%  | 65.65%    | 11,865    | 8.42%   | 41.93%  | 47.91%    |
| HOXA1            | miR-10a | 7,548       | 0.00%  | 10.57%  | 10.73%    | 20,788    | 10.03%  | 9.04%   | 18.30%    |
| KIT              | miR-221 | 9,406       | 35.73% | 22.02%  | 44.78%    | 11,952    | 21.81%  | 16.60%  | 31.41%    |
| KIT              | miR-222 | 9,406       | 35.73% | 29.13%  | 51.57%    | 7,777     | 10.17%  | 28.79%  | 35.64%    |
| KRAS             | let-7a  | NF          | NF      | NF      | NF        | 20,788    | 10.03%  | 9.04%   | 18.30%    |
| LIN28A           | let-7b  | 11,539      | 0.00%  | 7.04%   | 8.28%     | 11,080    | 8.03%   | 7.58%   | 16.25%    |
| MAPK14           | miR-24  | 10,229      | 0.00%  | 2.38%   | 2.39%     | 11,080    | 8.03%   | 7.58%   | 16.25%    |
| MYCN             | miR-101 | 13,408      | 0.00%  | 67.62%  | 67.70%    | 10,340    | 14.72%  | 34.61%  | 45.34%    |
| NRSAS            | let-7a  | 11,446      | 0.00%  | 13.68%  | 13.91%    | 20,290    | 9.23%   | 13.03%  | 21.45%    |
| PTEN             | miR-19a | 8,299       | 0.00%  | 54.40%  | 54.58%    | 10,680    | 19.58%  | 56.77%  | 65.38%    |
| ARHGAP32         | miR-132 | 7,378       | 0.00%  | 50.76%  | 50.95%    | 6,139     | 9.09%   | 7.31%   | 17.66%    |
| SMG1A            | let-7e  | 6,288       | 0.00%  | 9.27%   | 9.83%     | 11,564    | 10.78%  | 12.92%  | 23.74%    |
| TMSB4X           | miR-1   | 9,227       | 0.00%  | 67.42%  | 67.49%    | 12,742    | 10.38%  | 60.57%  | 65.24%    |
| TPM1             | miR-21  | 3,968       | 0.00%  | 70.14%  | 70.21%    | 12,692    | 12.38%  | 27.40%  | 36.60%    |

| G full | Average | 10,684 | 3.98% | 35.25% | 37.62% | 12,692 | 12.38% | 27.40% | 36.60% |
|--------|---------|--------|------|--------|--------|--------|--------|--------|--------|
|        |         | 3.98%  | 36.93% | 2.56%  | 29.83% |

**Table 7: RIsearch as pre-filter.** For each of the interactions, we report the unfiltered number of predictions made by TargetScanS and miRanda, together with the relative reduction achieved by different (combination of) tools. G*: GUUGle* as described in previous table; RI: RIsearch with a threshold of -11 kcal/mol; G∩RI: combination of both. Interactions that are denoted as not found (NF), have been found by the filter, but not by the respective method. Results are summarized as their averages below. The last row shows the according averages, when using GUUGle instead of GUUGle*.

These results show that GUUGle can hardly reduce TargetScanS candidates. This is because TargetScanS uses an even stricter seed requirement, all candidates identified by TargetScanS are also found by GUUGle. The only exception is the “7mer-la” criterion, an exact match to positions 2–7 of the mature miRNA (the seed) followed by an ‘A’. In these cases, a perfect complementary stretch of six nucleotides is sufficient to be considered as candidate for TargetScanS. There are only three interactions where GUUGle in fact reduces the number of candidates.

For miRanda the trend is not as strong, but also here we see that RIsearch alone achieves a bigger candidate reduction than GUUGle alone. As both tools filter out different candidates, their combined effect is strongest.
The reductions that can be achieved by RIssearch as a pre-filter differ widely (see Suppl. Table e.g. for TargetScanS relative reduction varies between 1% and 70%). Part of the explanation is the difference in GC-content of the mature miRNA sequences. The higher the GC-content, the more likely are low binding energies. With the conservative threshold of -11 kcal/mol, the list of candidates can not be reduced substantially in those cases. One could address this, by choosing a stricter cut-off for miRNAs with a potentially stronger interaction.

This relation can be seen in the figure to the right. The Pearson correlation coefficient $r$ between the GC-content and the reduction in miRanda hits is -0.6595 (p-value: 0.00046) and for TargetScanS -0.6996 (p-value: 9.953e-5).

Figure 3: Effect of GC-content of the miRNA sequence on relative reduction achieved.

References

[1] D. H. Mathews, M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner, “Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure,” Proc Natl Acad Sci U S A, vol. 101, pp. 7287–7292, May 2004.

[2] M. Rehmsmeier, P. Steffen, M. Höchsmann, and R. Giegerich, “Fast and effective prediction of microRNA/target duplexes,” RNA, vol. 10, no. 10, pp. 1507–17, 2004.

[3] J. S. Reuter and D. H. Mathews, “RNAstructure: software for RNA secondary structure prediction and analysis,” BMC Bioinformatics, vol. 11, p. 129, 2010.

[4] S. H. Bernhart, H. Tafer, U. Mückstein, C. Flamm, P. F. Stadler, and I. L. Hofacker, “Partition function and base pairing probabilities of RNA heterodimers,” Algorithms Mol Biol, vol. 1, no. 1, p. 3, 2006.

[5] N. O. Hodas and D. P. Aalberts, “Efficient computation of optimal oligo-RNA binding,” Nucleic Acids Res, vol. 32, no. 22, pp. 6636–6642, 2004.

[6] M. Andronescu, R. Aguirre-Hernández, A. Condon, and H. H. Hoos, “RNAsoft: A suite of RNA secondary structure prediction and design software tools,” Nucleic Acids Res, vol. 31, pp. 3416–3422, Jul 2003.

[7] H. Tafer, F. Amman, F. Egggenhofer, P. F. Stadler, and I. L. Hofacker, “Fast accessibility-based prediction of RNA–RNA interactions,” Bioinformatics, vol. 27, no. 14, pp. 1934–1940, 2011.

[8] C. M. Sharma, F. Darfeuille, T. H. Plantinga, and J. Vogel, “A small RNA regulates multiple ABC transporter miRNAs by targeting C/A-rich elements inside and upstream of ribosome-binding sites,” Genes Dev, vol. 21, pp. 2804–2817, Nov 2007.