BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 22 2008, pages 2657–2663
doi:10.1093/bioinformatics/btn193

Sequence analysis

RNAplex: a fast tool for RNA–RNA interaction search

Hakim Tafer and Ivo L. Hofacker*

Institute for Theoretical Chemistry, University of Vienna, Währingerstrasse 17, A-1090 Vienna, Austria

Received on November 5, 2007; revised on April 11, 2008; accepted on April 15, 2008

Advance Access publication April 23, 2008

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: Regulatory RNAs often unfold their action via RNA-RNA interaction. Transcriptional gene silencing by means of siRNAs and miRNA as well as snoRNA directed RNA editing rely on this mechanism. Additionally ncRNA regulation in bacteria is mainly based upon RNA duplex formation. Finding putative target sites for newly discovered ncRNAs is a lengthy task as tools for cofolding RNA molecules like `RNAcofold` and `RNAup` are too slow for genome-wide search. Tools like `RNAhybrid` that neglects intramolecular interactions have runtimes proportional to $\mathcal{O}(m \cdot n)$, albeit with a large prefactor. Still in many cases the need for even faster methods exists.

Results: We present a new program, `RNAplex`, especially designed to quickly find possible hybridization sites for a query RNA in large RNA databases. `RNAplex` uses a slightly different energy model which reduces the computational time by a factor 10–27 compared to `RNAhybrid`. In addition a length penalty allows to focus the target search on short highly stable interactions.

Availability: RNAplex can be downloaded at http://www.tbi.univie.ac.at/~htafer/

Contact: ivo@tbi.univie.ac.at

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

For decades, RNA molecules were dismissed as simple cell servants quietly transmitting genetic information from DNA and converting it into proteins. However, the discovery that double-stranded non-coding RNAs (dsRNAs) can efficiently inhibit gene expression by hybridizing to a target mRNA aroused strong interest in the scientific community. Recent studies have shown that many RNA–RNA interactions play a crucial role in different cellular processes. RNA–RNA interactions mediate pseudouridylation and methylation of rRNA (Bachellerie *et al.*, 2002), splicing of pre-mRNA (Zorio *et al.*, 1997), nucleotide insertion into mRNAs (Benne, 1992), transcription and translation control (siRNA, miRNA, stRNA) (Banerjee and Slack, 2002; Fire *et al.*, 1998; Kugel and Goodrich, 2007) or plasmid replication control (Eguchi, 1990). While siRNAs are often fully complementary to their targets, most of the ncRNAs interact in a more intricate manner which does not involve perfect hybridization. For example in *Escherichia Coli*, *OxyS*, which is involved in oxidative stress response, interacts with its target mRNA, *fhlA*, through a two sites kissing complex formation (Argaman and Altuvia, 2000).

Systematic target prediction for the plethora of genomic information brought by ncRNA detection programs and high-throughput sequencing is a challenging problem (Washietl *et al.*, 2007) and different kinds of tools are available to solve it. Purely sequence-based methods like BLAST (Altschul *et al.*, 1990) or FASTA (Pearson and Lipman, 1988) search for long stretches of perfect complementarity between a query and a target sequence. GUUGle (Gerlach and Giegerich, 2006) can efficiently locate potential complementary regions and, in contrast to BLAST, also allows to consider G–U pairs. A typical application for these programs is, for example, siRNA target search. Their main drawback is that they do not exploit information about the thermodynamics of the interaction between the query and the target RNA. Moreover, their lack of sensitivity is a real issue when looking for more complex interactions found, for example, between miRNA and their targets.

RNA-folding algorithms based on free energy minimization are at present among the most accurate and most generally applicable approaches for RNA folding (Turner and Sugimoto, 1988; Zuker, 2000; Zuker and Stiegler, 1981). They are based upon a large number of measurements performed on small RNAs and the assumption that stacking base pairs and loop entropies contribute, additively to the free energy of RNA secondary structures (Mathews, 2004; Mathews *et al.*, 1999). A straightforward approach to folding two RNA molecules is to concatenate the two sequences and apply a slightly modified RNA-folding algorithm. This approach is used, for example, by the `RNAcofold` (Hofacker *et al.*, 1994; Bernhart *et al.*, 2006), `pairfold` (Andronescu *et al.*, 2005) and `NUPACK` (Ren *et al.*, 2005) programs. However, the restriction to pseudo-knot free structures in standard folding algorithms is a more serious issue when dealing with RNA duplexes, as many known RNA–RNA interactions are mediated by, e.g., 'kissing hairpins' or other structure motifs that appear as pseudo-knots when the sequences are artificially concatenated.

As in the case of single sequences (Akutsu, 2000) inclusion of pseudo-knots makes the problem Non-deterministic Polynomial time (NP)-complete (Alkan *et al.*, 2006) in the unrestricted case. Polynomial time complexity can be achieved like in Alkan *et al.* (2006) and Pervouchine (2004), where intramolecular structures of each molecule are pseudo-knot free and intermolecular binding pairs are not allowed to cross. While these algorithms

*To whom correspondence should be addressed.

can predict complicated interaction motifs, such as the bacterial OxyS–fhlA system, they run in $\mathcal{O}(n^3 \cdot m^3)$ making them prohibitively expensive for most applications. Moreover, these algorithms suffer from a lack of experimentally measured energy parameters: little is known about the energetics of more complicated loop types, so that predicted optimal structures will often not correspond to reality.

Pseudo-knot free hybrid structures as in the case of RNAcofold can be computed in $\mathcal{O}((n+m)^3)$ time. However, the exclusion of pseudo-knots essentially means that interactions can happen only in the exterior loop of the concatenated sequences. Mückstein *et al.* (2006) recently considered an asymmetric model in which the base pairing is unrestricted in a large target RNA, while the interaction partner is restricted to intermolecular base pairs. RNAup works by modeling the total binding energy as a sum of two contributions, the energy needed to make the target site accessible (by breaking intramolecular pairs) and the energy gained through the RNA–RNA interaction. In contrast to RNAcofold, RNAup allows binding to an unpaired region in any kind of loop, the main limitation is that the interaction is confined to a single binding site.

A further reduction in computational complexity is achieved by omitting the computation of secondary structures within the monomers. This idea was first introduced by RNAhybrid (Rehmsmeier *et al.*, 2004) and is also implemented by RNAduplex from the *Vienna RNA* package. It is the simplest and fastest approach with a theoretical time complexity scaling as $\mathcal{O}(m^2 \cdot n^2)$ which can be reduced to $\mathcal{O}(m \cdot n \cdot L^2)$ by restricting the maximum loop length to L.

These programs are fast enough, e.g. to search for possible targets of a miRNA. However, for applications where target predictions have to be performed for a large number of small RNAs or when all pairwise comparisons between many RNAs need to be computed, the need for even faster methods still exists. Here we present a new version of RNAduplex, RNAplex, which is based on a slight simplification of the energy model. In this model, the loop energy is an affine function of the loop size instead of a logarithmic function. This approach reduces the time complexity to $\mathcal{O}(m \cdot n)$ resulting in a speedup factor of 10–27 when compared to RNAhybrid. In particular, the relative energy difference between both energy models remains <7% for all known miRNA/mRNA interactions. It is worth noting that for special problems, such as miRNA target prediction, further optimizations are possible, e.g. by exploiting the near-perfect complementarity of miRNA seed region and target, a feature used RNAhybrid. Since our aim is to provide a general tool for any kind of target search, we have currently not implemented such features.

As an example application we use RNAplex to predict targets of bacterial small RNAs, and show how to combine RNAplex with a more precise but more CPU intensive method, here RNAup, for fast and accurate target search.

2 METHODS

2.1 Energy model

RNAduplex/RNAhybrid are essentially equivalent to the classic RNA-folding algorithm of Zuker and Stiegler (1981) when only interior loops are allowed. As such they have a time complexity of $\mathcal{O}((n \cdot m)^2)$ in the naive implementation, where n and m represent the length of the interacting nucleotide sequences. It is a common practice to speed up these algorithms by restricting the loop size to L leading to $\mathcal{O}(n \cdot m \cdot L^2)$, where $L = 30$ in the case of RNAduplex. Here we use a simplified energy model that allows us to get rid of the constant but fairly large prefactor L^2.

Since we are neglecting intramolecular structure here, the only loop types that can appear are stacked pairs, bulge loops and interior loops. The Turner energy parameters provide look-up tables for the free energies of stacked pairs as well as for small interior loops (1x1, 2x1 and 2x2 loops). These look-up tables are used in RNAplex without change. Likewise, bulge loops of length 1 are treated exactly as in the full-energy model, namely by adding the stacking energy of the two pairs closing the loop plus a sequence independent penalty. Larger bulge loops are normally assigned a length-dependent penalty that grows logarithmically for large loops. In RNAplex this bulge energy is approximated by an affine function. Similarly, large interior loops are normally modeled by a size dependent term, an asymmetry penalty, and sequence dependent 'terminal mismatches'. Here again, we replace the size dependent loop energy by an affine function. Finally the asymmetry term is approximated by penalizing asymmetrical extension of interior loops [See Equation (1)]. The resulting energy model is exact for small loops and slightly overestimates the loop energies of large interior, bulge loops as well as strongly asymmetric loops (See Fig. 1).

2.2 Recursion

The structure of RNA duplexes predicted by our model can be decomposed into stacking pairs, interior loops and bulges. Our dynamic programming algorithm therefore employs four tables representing substructures that end in a base pair C, interior loop I and bulge on the first or second sequence, B^x,B^y, respectively. The central quantity $C_{i,j}$ stores the best energy of interaction between subsequence $x_1 \ldots x_i$ and $y_j \ldots y_m$. Similarly $B^{x,y}_{i,j}$ store the best energy of interaction given that residue y_j, respectively x_i, is aligned to a bulge. Finally, $I_{i,j}$ stores the best energy of interaction given that x_i and y_j are in an interior loop. The asymmetry penalty is modeled by allowing symmetrical extension of the interior loops as well as asymmetrical, penalized, interior loop extension [See Equation (1)]. Based on these matrices the recursion relation can be written as:

$$C_{i,j} = \min \begin{cases} C_{i-1,j+1} + S(i,j;i-1,j+1) \\ C_{i-1,j+2} + S(i,j;i-1,j+2) + P_{\text{bulge}} \\ C_{i-2,j+1} + S(i,j;i-2,j+1) + P_{\text{bulge}} \\ C_{i-2,j+2} + I(i,j;i-2,j+2) \\ C_{i-3,j+2} + I(i,j;i-3,j+2) \\ C_{i-2,j+3} + I(i,j;i-2,j+3) \\ C_{i-3,j+3} + I(i,j;i-3,j+3) \\ I_{i-1,j+1} + M(i,j;i-1,j+1) \\ B^x_{i-1,j+1} \\ B^y_{i-1,j+1} \end{cases} \quad (1)$$

$$I_{i,j} = \min \begin{cases} C_{i-1,j+1} + M(i-1,j+1;i,j) + g^I_{\text{open}} + 2g^I_{\text{ext}} \\ I_{i-1,j} + g^I_{\text{ext}} + A \\ I_{i-1,j+1} + 2g^I_{\text{ext}} \\ I_{i,j+1} + g^I_{\text{ext}} + A \end{cases} \quad (2)$$

$$B^x_{i,j} = \min \begin{cases} C_{i-1,j} + g^B_{\text{open}} + g^B_{\text{ext}} \\ B^x_{i-1,j} + g^B_{\text{ext}} \end{cases} \quad (3)$$

$$B^y_{i,j} = \min \{ C_{i,j+1} + g^B_{\text{open}} + g^B_{\text{ext}} \quad (4)$$

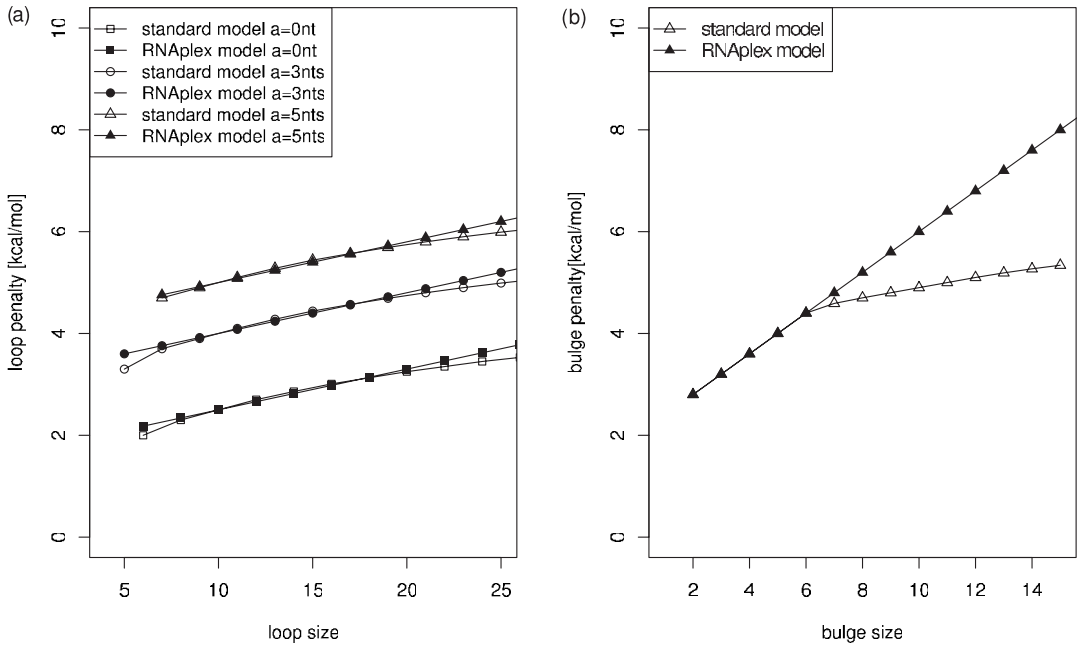

Fig. 1. Comparison of the RNAplex energy model against the Turner energy model for bulges and interior loops. (**a**) Plot of the interior loop penalty against the total loop size for three different values of asymmetry. The model used in RNAplex slightly overestimates the loop energies. (**b**) Plot representing the bulge loop penalty against the bulge size. Our model agree exactly with the Thurner model for bulge size up to 6 nt.

where $\mathcal{S}(i, j, k, l)$ represents the energy gained by stacking the $x_i \cdot y_j$ base pair onto the $x_k \cdot y_l$ base pair. As usual, bulges of length 1 are modeled as the sum of a bulge penalty P_{bulge} plus the stacking energy of the adjacent base pairs. $\mathcal{M}(i, j; i-1, j+1)$ represents the 'mismatch' energy of the unpaired nucleotides $(i-1,\ j+1)$ adjacent to the pair $(i,\ j)$. $\mathcal{I}$ represents the energy contribution of the small interior loops. Furthermore, $g^{B,I}_{\text{open}}$ and $g^{B,I}_{\text{ext}}$ represent the parameters of the affine loop energy function that approximates the conventional Turner loop energies. These parameters were gained by linearly fitting the loop energy model. Finally, A represents the asymmetry penalty that approximates the extra destabilizing energy of asymmetrical loops. The above recursion is graphically represented in Figure 2.

In our model a duplex starts with two stacked pairs $(i,\ j) \cdot (i-1,\ j+1)$. The initialization of the recursion matrices should ensure that all structural element has to start and end inside the recursion matrices. This means that no interior loops and no bulges on the target sequence may be closed before $i=3$. Moreover, no bulge and no interior loop on the query sequence may be closed before $j=m-2$. Finally, $C_{1,0}$ is set to 0. As a consequence the matrices are initialized in the following way

$$I_{1,j} = I_{2,j} = \infty \forall j$$
$$B^x_{1,j} = B^x_{2,j} = \infty \forall j$$
$$I_{i,m} = I_{i,m-1} = \infty \forall i$$
$$B^y_{i,m} = B^y_{i,m-1} = \infty \forall i$$

When comparing an RNA of length m against a large database of length $n \gg m$ the optimal interaction typically spans the full length of the shorter RNA m. However, long interactions, extending over many helical turns, are sterically hindered, and moreover have to compete with the tendency to form intramolecular structure. Therefore, hits consisting of a short but stable duplex should be preferable over interactions that attain a good score only by adding many weak interactions over a long region. To counter this effect, RNAplex contains an option that introduces a per nucleotide penalty to the interaction energy. Especially for longer queries, this results in shorter and statistically more significant interactions.

2.3 Memory usage

In order to reduce the memory consumption of RNAplex we do not store the full matrices from the recursion from Section 2.2. It is easy to see that each position (i, j) in the matrices $C_{i,j}$, $I_{i,j}$, $B^x_{i,j}$ and $B^y_{i,j}$ can be computed from the previous three columns. We therefore store only the most recent four columns of each matrix. In addition, the maximum interaction energy of each column as well as its location on both RNA strands are stored in linear tables of length n, which is sufficient to locate all positions in the query and the target sequences where an interaction with score higher than a given threshold T ends. This reduces memory usage to $\cdot\ \mathcal{O}(16 \cdot m + 3 \cdot n)$ with m the length of the query sequence and n the length of the target sequence. To obtain the actual structure we then recompute the local alignment of the substring of the query sequence to the substring of the target sequence. In this step we use the full-energy model rather than the simplified model used in Equation (1–4). Here the memory consumption is $\mathcal{O}(l^2)$ where l represents the maximum hybridization length. Surprisingly, the improved memory usage led to a reduction of computation time by a factor two, presumably because of better cache utilization.

2.4 Sensitivity assessement

To test whether the simplified energy model affects the sensitivity of RNAplex, we assessed how well RNAplex, RNAhybrid and RNAduplex recovered experimentally confirmed miRNA–mRNA interactions. We used 27 interactions taken from TarBase (Sethupathy *et al.*, 2006) involving 25 mRNAs and 22 miRNAs. For each of the reported interactions, the hybridization energy of the reported target site with its cognate miRNA was computed with RNAplex, RNAduplex and RNAhybrid. Moreover for each miRNA–mRNA pair, the 10 best binding sites were identified using

RNAplex, RNAduplex and RNAhybrid. For RNAhybrid we constrainted the hybridization to target sites which were fully complementary to the miRNA seed region, since this gave the highest sensitivity in the test. The experimentally confirmed binding site was then reported as recovered if it overlapped with any of the 10 best hits (Table 1). All three programs performed similarly well with RNAduplex retrieving 22 out of 27 interactions, while RNAplex and RNAhybrid each recovered 20 interactions.

2.5 RNAplex speedup

When comparing search speed to RNAhybrid, we found that the speedup varied with sequence length and program options, but was at least 10 in all cases. RNAhybrid performed best for miRNA target search when limiting the search to targets with a perfect seed matches, i.e. Watson–Crick pairs only at miRNA positions 2 to 7. Without this constraint RNAplex speedup increased from 10- to 20-fold. Furthermore the speedup increased slightly for longer query sequences, reaching 27 for query sequences of length 320. In the tests above, we searched only for the single most stable interaction site. While RNAplex can return suboptimal interaction sites without a speed penalty, RNAhybrid needs to repeat the whole dynamic programming procedure for each desired suboptimal, making it accordingly expensive.

We also tested whether the computation time of RNAplex was reduced further by identifying stretches of complementarity before attempting the more time consuming dynamic programming procedure. We used GUUGle, which locates potential helical regions under RNA base pairing rules with the help of suffix arrays to find these highly complementary regions. The trade-off between speed and sensitivity is controlled by the *ktup* parameter, which specifies the size of complementarity to search for (word size). We compared the CPU time and sensitivity of RNAplex and GUUGle+RNAplex when searching for experimentally verified miRNA targets. Up to a word size of seven RNAplex is faster than GUUGle+RNAplex, while the sensitivities of both programs are the same. For larger word size, GUUGle+RNAplex performs better than RNAplex however at the cost of a reduced sensitivity. RNAplex+GUUGle may prove to be useful for searching of gapped interactions with complementary regions longer than 7 nt.

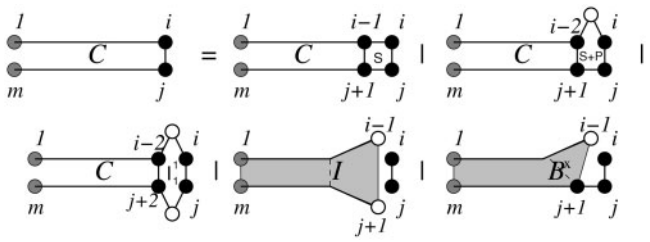

Fig. 2. Simplified representation of the structure decomposition used in RNAplex. For clarity only the decomposition of the closed structure terms [see Equation (1)] is shown. Black dots represent paired bases. White dots denote unpaired bases. Given that x_i and y_j are paired, C stores the best energy of interaction between $x_1 \ldots x_i$ and $y_j \ldots y_m$. $\mathcal{S}$ is the stacking energy of two pairs of nucleotides. P is the bulge penalty to add to 1×0 bulges. I is the matrix holding the best energy of interaction given that x_i and y_j are in an interior loop. I_1^1 is the destabilizing energy of a 1×1 interior loop (1×2, 2×1 and 2×2 cases not shown) and B^x represents the matrix storing the best energy of interaction given that residue y_j is aligned to a bulge. The cases where x_i and y_j do not pair (interior loop and bulge extension and/or creation) are not shown.

3 RESULTS

In order to test the usability of RNAplex, we consider the problem of predicting mRNA targets of bacterial small RNAs. We used a dataset consisting of eight small downregulating RNAs from *E.coli*, for which mRNA targets as well as the known position of interaction on the targets were mostly taken from Urban *et al.* (2007). In all cases the binding sites were located in the vicinity of the start codon of the respective mRNAs. We first looked at how well RNAplex and RNAhybrid retrieve the known binding sites. For each target gene, a subsequence of 401 nt centered around the start codon was retrieved. For each pair of interactions, both RNAplex and RNAhybrid computed the positions of the most stable interactions. Bacterial small RNAs are typically over 80 nt in length, but usually interact only within a region of 10–40 nt, much smaller than the length of the RNA. We therefore made use of the ability in RNAplex to favor short stable interactions by setting a per nucleotide penalty of 0.3 kcal/mol. The value 0.3 kcal/mol was chosen, because it corresponds to the average per nucleotide duplex energy between two unrelated RNA sequences.

Table 1. Binding site summary for 27 functional miRNA–mRNA interactions in Human, taken from TarBase Sethupathy *et al.*, (2006)

mRNA	miRNA	$\Delta G_{\text{RNAduplex}}$	$\Delta G_{\text{RNAplex}}$	$\Delta G_{\text{RNAhybrid}}$
AGTR1	miR-155	−11.50(NF)	−11.50(NF)	−17.2(NF)
BCL2	miR-16	−18.90(2)	−18.90(1)	−24.1(1)
CAT-1	miR-122	−23.80(1)	−23.80(1)	−29.0(1)
CGI-38	miR-16	−20.80(2)	−20.80(2)	−26.0(NF)
Clock	miR-141	−16.40(1)	−16.40(1)	−22.1(1)
CXCL12	miR-23a	−8.90 (NF)	−8.90 (NF)	−14.0(5)
CYP1B1	miR-27b	−28.20(1)	−28.20(1)	−33.6(1)
E2F3	miR-34a	−19.10(2)	−19.10(NF)	−25.1(1)
Enx-1	miR-101	−16.90(1)	−16.90(1)	−22.4(NF)
FLJ2130	miR-145	−21.80(1)	−21.80(1)	−27.4(NF)
Fstl1	miR-206	−18.40(6)	−18.40(NF)	−23.2(2)
GJA1	miR-1	−14.30(3)	−14.30(4)	−20.6(2)
GJA1	miR-206	−14.53(10)	−14.53(10)	−20.5(4)
Hand2	miR-1	−12.20(1)	−12.20(1)	−18.1(2)
HOXA1	miR-10a	−15.93(3)	−15.93(5)	−22.7(1)
KIT	miR-221	−17.70(3)	−17.70(NF)	−23.4(2)
KIT	miR-222	−14.70(NF)	−14.70(NF)	−19.8(3)
KRAS	let-7a	−14.10(NF)	−14.10(7)	−18.9(6)
Lin28	let-7b	−27.40(1)	−27.40(1)	−33.5(1)
MAPK14	miR-24	−27.10(1)	−27.10(1)	−32.2(1)
MYCN	miR-101	−13.80(NF)	−13.80(NF)	−20.7(1)
NRAS	let-7a	−16.10(2)	−16.10(3)	−21.1(NF)
PTEN	miR-19a	−17.70(1)	−17.70(1)	−23.2(1)
RICS	miR-132	−18.80(1)	−18.80(1)	−25.1(1)
SMC1L1	let-7e	−22.20(1)	−22.20(1)	−27.5(1)
TMSB4X	miR-1	−16.90(1)	−16.90(1)	−21.9(NF)
TPM1	miR-21	−15.60(1)	−15.60(1)	−19.6(NF)

Columns 1 and 2 contain the name of the mRNA and miRNA, respectively. The Column 3–5 contain the interaction energy for the reported miRNA–mRNA interactions as computed by RNAduplex, RNAplex and RNAhybrid, respectively. The number in parenthesis represent the rank of the experimental target site where 1 stands for the most stable interaction and 10 for the 10th best interaction. NF means that the reported target site was not found among the 10 best interaction sites and are shown in red.

With these settings RNAplex was able to precisely locate seven out of seven interactions, with a maximal difference of 30 nt (Table 2). In two cases RNAplex failed to predict the correct target site. In contrast, RNAhybrid maximized the length of hybridization, leading to substantially longer target sites. In six out of nine cases, experimental and predicted target sites overlapped. However, the size of the predicted interaction did not allow clear localization of the proper target boundary.

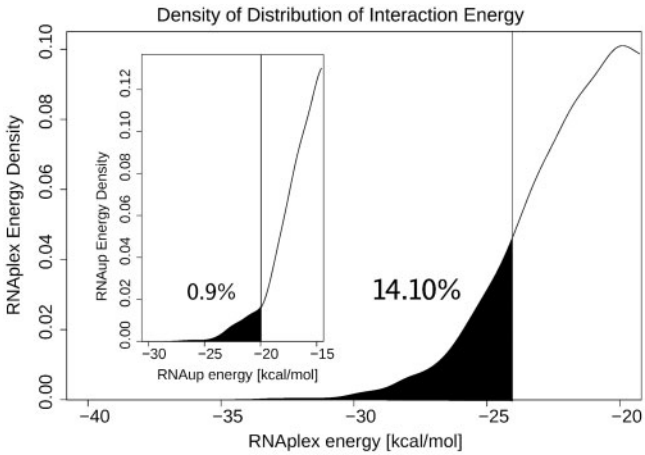

Fig. 3. Density of distribution of interaction energy as computed by RNAplex for miR-134 against all mouse 3′UTRs. The vertical line represents the energy of the experimentally confirmed miR-134/Limk1 interaction as computed by RNAplex. The black area represents the proportion of 3′UTRs having a higher energy of interaction than the experimentally predicted one. The number in parenthesis represents the percentage of the entire distribution falling below the threshold defined binding energy, as computed by RNAplex, of the experimentally verified interaction. The inset shows the density of distribution of interaction energy as computed by RNAup for miR-134 against the mouse 3′UTRs. The vertical line represents the energy of binding as computed by RNAup for the experimentally confirmed miR-134/Limk1 interaction. The number in parenthesis represents the percentage of the entire distribution falling below the threshold defined by the binding energy, as computed by RNAup, of the experimentally verified interaction.

In a further step we assessed the specificity of both programs for finding putative sRNA targets. As before, we extracted sequences of 401 nt from all mRNAs in the *E.coli* genome (4463 genes). For each mRNA–sRNA pair, the binding location and the best interaction energy was kept. Then for each sRNA we recorded the number of mRNAs that had a better interaction energy than the known target and whose interaction site overlapped a 40 nt region centered around the start codon (the most frequently targeted mRNA region). The average rank of the known target was 107 for RNAplex and 996 for RNAhybrid. Furthermore, RNAplex finished the computation in 103 s on an 2.4 GHz intel E6600-based machine, 26 times faster than RNAhybrid (used with default parameters).

The high number of false positives of both programs is not unexpected, since the competition between intra- and intermolecular base pairing is completely ignored. The significant role of target site accessibility has also been stressed in recent studies on RNA interference (Ameres *et al*., 2007; Kertesz *et al*., 2007; Long *et al*., 2007). Programs that fully include the effects of target site accessibility, such as RNAup (Mückstein *et al*., 2006) from the *Vienna RNA* package, have much better specificity at the expense of much higher computational cost. In fact, for the examples shown, RNAup ranked the correct sRNA–RNA combination higher and was able to correctly predict all target site positions. However, it needs about 52 CPU days to compute all sRNA–mRNA pairs on an 2.4 GHz Intel Core Duo (Mückstein *et al*., 2008).

We therefore recommend to use RNAplex in a target detection pipeline, where each candidate binding site reported by RNAplex are inspected by a more CPU intensive method. The results in Table 2 show that RNAplex can be used as a filter to greatly reduce the number of wrong interaction candidates, with only a slight loss in sensitivity. Applying RNAup on all reported target sites would take ~7 h, 200 times less than using RNAup alone.

As a further example we searched with a similar method for target sites of mouse miR-134, an miRNA involved in regulating dendritic development and the differentiation of

Table 2. Binding site summary for the 10 functional interactions from Urban *et al.* (2007)

mRNA	sRNA	$\Delta G_{\text{RNAplex}}$	Position RNAplex		$\Delta G_{\text{RNAhybrid}}$	Position RNAhybrid		Pos.lit.		cite
RyhB	sodB	−25.20(87)	+183	+162	−58.4(1247)	−239	−100	−4	+5	Geissmann and Touati (2004)
DsrA	hns	−21.90(128)	+1	+20	−49.0(1296)	−170	−61	+7	+19	Lease *et al.* (1998)
MicA	ompA	−23.90(67)	−22	−5	−54.2(58)	−87	+30	−21	−6	Rasmussen *et al.* (2005)
MicC	ompC	−22.00(97)	−31	−14	−71.1(120)	−86	+36	−30	−15	Chen *et al.* (2004)
MicF	ompF	−26.80(34)	−27	+10	−47.5(1010)	−150	−61	−16	+10	Chen *et al.* (2004)
Spot42	galK	−29.30(38)	+4	+37	−79.4(28)	−112	+40	−19	+21	Mller *et al.* (2002)
SgrS	ptsG	−23.30(170)	+150	+171	−139.0(1938)	−68	+200	−28	+4	Kawamoto *et al.* (2006)
GcvB	dppA	−29.40(80)	−31	−6	−125.2(1436)	−154	+110	−31	−14	Sharma *et al.* (2007)
GcvB	oppA	−25.10(263)	−3	45	−122.8(1837)	−156	+189	−8	+16	Sharma *et al.* (2007)

The number in parenthesis represents the quantity of predicted interactions involving the same ncRNA, overlapping with a 40 nt long region centered around the start codon and having a higher interaction energy than the functional hybrid. Positions in red indicate target sites predicted by RNAplex or RNAhybrid which do not overlap with the experimentally reported ones. For RNAplex a per nucleotide penalty of 0.3 /mol was used.

mouse embryonic stem cells (Kim *et al*., 2006; Velleca *et al*., 1994), and compared those results with the target predicted by RNAplex alone. We also assessed the specificity of both methods by recording the number of sequences that had a better interaction energy than the experimentally confirmed miR-134/Limk1 hybrid (Schratt *et al*., 2006).

For each 3′UTRs sequences which were downloaded from BIOMART (Durinck *et al*., 2005), we computed the minimal free energy of interaction (MFE). All sequences that had an MFE smaller than −15 kcal/mol were stored for subsequent inspection with RNAup (7503 sequences). Instead of using the whole 3′UTR sequence in RNAup we selected a 200 nt regions centered around the binding sites reported by RNAplex. Then each reported interaction was ranked based either on its RNAup or RNAplex interaction energy.

In case of our two steps method, where first putative targets are rapidly identified with RNAplex and further inspected with RNAup, Limk1 had a RNAup binding energy of −19.97 kcal/mol and was ranked among the 74 best targets (0.9%). In contrast, the same interaction was ranked 1057 when looking at the RNAplex energy (14.10%) (Fig. 3). Similarly using RNAhybrid instead of RNAplex would have resulted in 1445 hits. The 73 target mRNAs scoring higher than Limk1 are likely to contain additional true targets. 8 of the 73 targets were actually contained in a recent study of Miranda *et al*. (2006). For all of them miR-134 reduced the respective protein concentration by at least 45%.

4 CONCLUSION

The problem of folding more than one RNA strand can be treated at different levels of complexity. Because of the high-computational cost of many algorithms for RNA–RNA interaction prediction, target search may be best performed by a hierarchical search strategy, employing a series of filters that balance speed versus accuracy.

Here we have introduced the program RNAplex which reduces the time needed to localize putative hybridization sites, mainly by neglecting intramolecular interactions and by using a slighlty simplified energy model. Combined with, e.g. RNAup we can find high-confidence targets, with only a slight loss of sensitivity. As a consequence RNAplex is well suited for localizing putative ncRNAs interactions partners in large amount of genomic data.

Finally apart from the speed improvement, RNAplex in contrast to similar programs, can recover short, highly stable interactions between two RNAs, by introducing a per nucleotide penalty. We envision to improve this characteristic by integrating structural, position-dependent penalties into RNAplex. This would allow to take structural effect into account, improving the accuracy of RNAplex.

ACKNOWLEDGEMENTS

Funding: This work was supported by siemens, by the Wiener Wissenschafts-, Technologie- und Forschungsfonds and by the Austrian GEN-AU projects bioinformatics integration network and non-coding RNA.

Conflict of Interest: none declared.

REFERENCES

Akutsu,Y. (2000) Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Appl. Math*., **104**, 45–62.

Alkan,C. *et al*. (2006) RNA-RNA interaction prediction and antisense RNA target search. *J. Comput. Biol*., **13**, 267–282.

Altschul,S. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Ameres,S.L. *et al*. (2007) Molecular basis for target RNA recognition and cleavage by human RISC. *Cell*, **130**, 101–112.

Andronescu,M. *et al*. (2005) Secondary structure prediction of interacting RNA molecules. *J. Mol. Biol*., **345**, 987–1001.

Argaman,L. and Altuvia,S. (2000) fhlA repression by oxys RNA: kissing complex formation at two sites results in a stable antisense-target RNA complex. *J. Mol. Biol*., **300**, 1101–1112.

Bachellerie,J. *et al*. (2002) The expanding snoRNA world. *Biochimie*, **84**, 775–790.

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

Benne,R. (1992) RNA editing in trypanosomes. The us(e) of guide RNAs. *Mol. Biol. Rep*., **16**, 217–227.

Bernhart,S. *et al*. (2006) Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol. Biol*., **1**, 3–3.

Chen,S. *et al*. (2004) MicC, a second small-RNA regulator of Omp protein expression in Escherichia coli. *J. Bacteriol*., **186**, 6689–6697.

Durinck,S. *et al*. (2005) Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–3440.

Eguchi,Y. and Tomizawa,J. (1990) Complex formed by complementary RNA stem-loops and its stabilization by a protein: function of coie1 rom protein. *Cell*, **60**, 199–209.

Fire,A. *et al*. (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, **391**, 806–811.

Geissmann,T. and Touati,D. (2004) Hfq, a new chaperoning role: binding to messenger RNA determines access for small RNA regulator. *EMBO J*., **23**, 396–405.

Gerlach,W. and Giegerich,R. (2006) Guugle: a utility for fast exact matching under RNA complementary rules including g-u base pairing. *Bioinformatics*, **22**, 762–764.

Hofacker,I. *et al*. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem*., **125**, 167–188.

Kawamoto,H. *et al*. (2006) Base-pairing requirement for RNA silencing by a bacterial small RNA and acceleration of duplex formation by Hfq. *Mol. Microbiol*., **61**, 1013–1022.

Kertesz,M. *et al*. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet*., **39**, 1278–1284.

Kim,D.H. *et al*. (2006) Argonaute-1 directs siRNA-mediated transcriptional gene silencing in human cells. *Nat. Struct. Mol. Biol*., **13**, 793–797.

Kugel,J. and Goodrich,J. (2007) An RNA transcriptional regulator templates its own regulatory RNA. *Nat. Chem. Biol*., **3**, 89–90.

Lease,R.A. *et al*. (1998) Riboregulation in Escherichia coli: DsrA RNA acts by RNA:RNA interactions at multiple loci. *Proc. Natl Acad. Sci. USA*, **95**, 12456–12461.

Long,D. *et al*. (2007) Potent effect of target structure on microRNA function. *Nat. Struct. Mol. Biol*., **14**, 287–294.

Møller,T. *et al*. (2002) Spot 42 RNA mediates discoordinate expression of the E. coli galactose operon. *Genes Dev*., **16**, 1696–1706.

Mathews,D. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, **10**, 1178–1190.

Mathews,D. *et al*. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol*., **288**, 911–940.

Miranda,K.C. *et al*. (2006) A pattern-based method for the identification of microRNA binding sites and their corresponding heteroduplexes. *Cell*, **126**, 1203–1217.

Mückstein,U. *et al*. (2006) Thermodynamics of RNA–RNA binding. *Bioinformatics*, **22**, 1177–1182.
Mückstein,U. *et al*. (2008) *Translational control by RNA-RNA interaction*, *BIRD 2008, CCIS* (in press).
Pearson,W. and Lipman,D. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
Pervouchine,D.D. (2004) Iris: intermolecular RNA interaction search. *Genome Inform.*, **15**, 92–101.
Rasmussen,A.A. *et al*. (2005) Regulation of ompA mRNA stability: the role of a small regulatory RNA in growth phase-dependent control. *Mol. Microbiol.*, **58**, 1421–1429.
Rehmsmeier,M. *et al*. (2004) Fast and effective prediction of microRNA/target duplexes. *RNA*, **10**, 1507–1517.
Ren,J. *et al*. (2005) Hotknots: heuristic prediction of RNA secondary structures including pseudoknots. *RNA*, **11**, 1494–1504.
Schratt,G.M. *et al*. (2006) A brain-specific microRNA regulates dendritic spine development. *Nature*, **439**, 283–289.
Sethupathy,P. *et al*. (2006) TarBase: a comprehensive database of experimentally supported animal microRNA targets. *RNA*, **12**, 192–197.
Sharma,C.M. *et al*. (2007) A small RNA regulates multiple ABC transporter mRNAs by targeting C/A-rich elements inside and upstream of ribosome-binding sites. *Genes Dev.*, **21**, 2804–2817.
Turner,D. and Sugimoto,N. (1988) RNA structure prediction. *Annu. Rev. Biophys. Biophys. Chem.*, **17**, 167–192.
Urban,J.H. *et al*. (2007) A conserved small RNA promotes discoordinate expression of the glmUS operon mRNA to activate GlmS synthesis. *J. Mol. Biol.*, **373**, 521–528.
Velleca,M.A. *et al*. (1994) A novel synapse-associated noncoding RNA. *Mol. Cell Biol.*, **14**, 7095–7104.
Washietl,S. *et al*. (2007) Structured RNAs in the encode selected regions of the human genome. *Genome Res.*, **17**, 852–864.
Zorio,D. *et al*. (1997) Cloning of caenorhabditis u2af65: an alternatively spliced RNA containing a novel exon. *Mol. Cell Biol.*, **17**, 946–953.
Zuker,M. (2000) Calculating nucleic acid secondary structure. *Curr. Opin. Struct. Biol.*, **10**, 303–310.
Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.