BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 14 2011, pages 1934–1940
doi:10.1093/bioinformatics/btr281

Structural bioinformatics

Advance Access publication May 18, 2011

Fast accessibility-based prediction of RNA–RNA interactions

Hakim Tafer[1,*], Fabian Amman[2], Florian Eggenhofer[2], Peter F. Stadler[1,2,3,4,5] and Ivo L. Hofacker[2,*]

[1]Bioinformatics Group, Department of Computer Science, Interdisciplinary Center for Bioinformatics, University of Leipzig, D-04107 Leipzig, Germany, [2]Institute for Theoretical Chemistry, University of Vienna, A-1090 Vienna, Austria, [3]Max Planck Institute for Mathematics in the Sciences, [4]RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology, D-04103 Leipzig, Germany and [5]The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM-87501, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Currently, the best RNA–RNA interaction prediction tools are based on approaches that consider both the inter- and intramolecular interactions of hybridizing RNAs. While accurate, these methods are too slow and memory-hungry to be employed in genome-wide RNA target scans. Alternative methods neglecting intramolecular structures are fast enough for genome-wide applications, but are too inaccurate to be of much practical use.

Results: A new approach for RNA–RNA interaction was developed, with a prediction accuracy that is similar to that of algorithms that explicitly consider intramolecular structures, but running at least three orders of magnitude faster than `RNAup`. This is achieved by using a combination of precomputed accessibility profiles with an approximate energy model. This approach is implemented in the new version of `RNAplex`. The software also provides a variant using multiple sequences alignments as input, resulting in a further increase in specificity.

Availability: `RNAplex` is available at www.bioinf.uni-leipzig.de/Software/RNAplex.

Contact: htafer@bioinf.uni-leipzig.de; ivo@tbi.univie.ac.at

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

Received on September 9, 2010; revised on April 28, 2011; accepted on April 29, 2011

1 INTRODUCTION

The status of RNA in molecular biology has changed dramatically over the last decade. Instead of taking on a rather marginal role as messenger of genomic information, they are now considered as key regulatory elements in a wide spectrum of cellular processes. As of 2008, the number of known non-coding RNA sequences reached an overwhelming 29 million grouped into 1300 distinct families (Gardner *et al.*, 2009).

Non-coding RNAs (ncRNAs) frequently function by binding to other RNAs. For example, snoRNAs mediate pseudouridylation and methylation of rRNAs and snRNAs (Bachellerie *et al.*, 2002) and can influence the splicing of pre-mRNAs (Zorio *et al.*, 1997). ncRNAs are also involved in the editing of other RNA sequences (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 and Tomizawa, 1990). While siRNAs are often fully complementary to their targets, most other 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 formation of a two sites kissing complex (Argaman and Altuvia, 2000). Although there is statistical evidence that a plethora of ncRNAs interacts with other RNAs (The Athanasius F. Bompfünewerer RNA Consortium: *et al.*, 2007), targets remain unknown for most of them. The prediction of RNA–RNA interactions, therefore, has become an important field in computational biology.

RNA–RNA interactions are primarily governed by the same types of hydrogen bonds and stacking interactions as RNA secondary structure formation. The problem can, therefore, be tackled by similar algorithmic approaches and the same parametrization of the interaction energies. We may distinguish two distinct ways of addressing the RNA–RNA interaction problem. The most straightforward way consists in concatenating both sequences and subsequently folding them as a pseudo-single sequence. The precision of this kind of approach depends greatly on how the concatenation is handled. The crudest approaches use linker sequences to connect both RNA strands (Stark *et al.*, 2003). This can lead to erroneous structure prediction as the linker may interfere with the interacting sequences. Alternatively, a small modification of the folding algorithm keeps track of the concatenation point(s) and uses adjusted energy parameters for the loops in which the junctions occur (Andronescu *et al.*, 2003; Bernhart *et al.*, 2006b; Dimitrov and Zuker, 2004; Dirks *et al.*, 2007; Hofacker *et al.*, 1994). A combinatorially different model, known as RNA–RNA interaction problem (RIP), covers a larger set of possible structures (Alkan *et al.*, 2006; Chitsaz *et al.*, 2009; Huang *et al.*, 2010; Pervouchine, 2004).

The second type of approaches conceptually decomposes the RNA hybridization process into two stages: (i) the unfolding of the interacting regions of the two partners and (ii) the direct interaction of the exposed binding sites. In practice, one first computes the probability of being unpaired for each region (sequence interval) in both sequences. These probabilities are equivalent to the free energy necessary to expose the regions. In the second step, the interaction energy between each combinations of regions is evaluated (Busch *et al.*, 2008; Mückstein *et al.*, 2006, 2008). This approach was,

*To whom correspondence should be addressed.

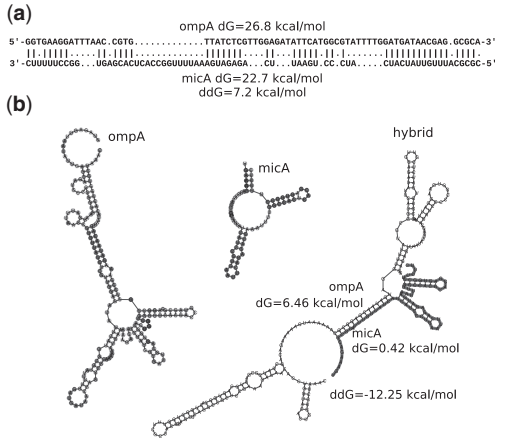

Fig. 1. Comparison of the ompA–micA hybrids predicted with and without considering intramolecular structures. (**a**) Hybrid structure predicted with RNAplex without considering the intramolecular structures of the RNA sequences. The hybrid extends over 67 and 69 nucleotides on ompA and micA, respectively, and has an hybridization energy of −42.3 kcal/mol. Still the energy needed to unfold both binding regions on ompA and micA amounts 22.7+26.8=49.5 kcal/mol, larger than the energy gained through binding. (**b**) ompA–micA interaction predicted by RNAup. OmpA–micA hybrid is shown on the right hand side, with the micA sequence represented by a bold line. Even though the hybrid is much smaller than the interaction in (**a**), it has a lower total interaction energy (ddG) of −12.25 kcal/mol, due to the fact that the interacting regions are less structured.

in particular, applied successfully to sRNA–mRNA interactions in bacteria.

While both types of algorithms proved useful in predicting the correct interaction structure of a ncRNA with its (known) target, they are computationally expensive, requiring at least $\mathcal{O}((n+m)^3)$ operations, where n and m are the size of the target and query sequences, respectively, and hence are impractical for genome-wide target predictions.

A drastic reduction in computational complexity can be achieved by omitting the computation of secondary structures within the monomers, as demonstrated by RNAhybrid (Rehmsmeier *et al.*, 2004), which runs in $\mathcal{O}(m \cdot n \cdot L^2)$ when restricting the maximum loop length to L. RNAplex, a conceptually very similar approach (Tafer and Hofacker, 2008), further reduces the time complexity to $\mathcal{O}(m \cdot n)$ by using a modified energy model. Neglecting the internal structure of the interacting sequences leads to a drastic decrease in specificity; however, see Figure 1. This issue is roughly addressed by RNAplex in that it mimics the effect of the competition between intra- and intermolecular interactions by adding a fixed per-nucleotide penalty (Tafer and Hofacker, 2008).

Currently, one therefore has to choose between precise but impractically slow methods or fast but imprecise methods for ncRNA target search, a situation that is quite unsatisfactory. In this contribution, we extend the RNAplex approach Tafer and Hofacker (2008) to tackle this problem. We mimic the effect of the competition between intra- and intermolecular interactions by adding a position-dependent per-nucleotide penalty instead of a fixed penalty. This penalty is derived from precomputed accessibility profiles produced by RNAplfold (Bernhart *et al.*, 2006a; Bompfünewerer *et al.*, 2008a) or RNAup (Mückstein *et al.*, 2008). More explicitly, these profiles contain the probabilities that any subsequence of arbitrary length is unpaired in thermodynamic equilibrium. These probabilities are converted to free energies that then enter as position-dependent penalties in the computation of the interaction energies, preserving RNAplex $\mathcal{O}(m \cdot n)$ run time. The main advantage is that the accessibility profiles can be precomputed and stored, making this approach particularly attractive for large-scale screening studies. In addition, we extended RNAplex so that it can also handle multiple alignment. This inclusion of comparative information into the target prediction process leads to a substantial increase in specificity.

2 METHODS

2.1 RNAplex novelties

The extension of RNAplex brings two novelties that increase its specificity. First, we introduce position-specific per-nucleotide penalties that approximate the effects of the competition between intra- and intermolecular interactions. Second, RNAplex is now able to compute the interactions between two alignments, allowing RNAplex to favor evolutionary conserved interactions. Similar to the single sequence version, the multiple sequences alignment version can also consider the accessibility of the targets.

2.2 Approximate opening energies

We first outline the design of RNAplex, which employs a two-steps approach. In the first step, the scanning phase, RNAplex identifies positions where putative interactions may end. For small interior loops (1×1, 2×1 and 2×2), as well as bulges of size 1, RNAplex still employs the original look-up tables provided by the Turner Energy Model. For larger interior loops and bulges, however, RNAplex uses a linear approximation of the size dependence of loop energies (Tafer and Hofacker, 2008). The resulting energy model is exact for small loops and slightly overestimates the loop energies of large interior, bulge loops and strongly asymmetric loops. A further advantage of the linear energy model is that RNAplex needs to store only the last four columns of the dynamic programming matrix during the scan phase. Once all high-scoring interactions are localized along the target sequence, RNAplex uses the standard energy model to recompute the energy and structure of the putative hybrids.

During the scan phase, in order to extend a hybrid by one nucleotide, we need to know the cost of freeing this nucleotide from all the intramolecular interactions it might be involved in. In thermodynamic equilibrium, this energy cost can be derived from the probability that the interacting stretch of nucleotides is unpaired. Since it is too expensive to compute this for all intervals, we seek a step-wise procedure. Consider an intermediary hybrid structure $\mathcal{S}_y^x$ between two sequences x and y that starts at base pair (x_i, y_j) and spans w_x nucleotides of sequence x and w_y nucleotides of sequence y. We need to determine the *conditional* probability ${}^{w_x}\mathrm{P}_u^x[i+w_x]$ that nucleotide x_{i+w_x} is not involved in any intramolecular interaction, *given* that its predecessors $i+w_x-1$ is unpaired, and the analogous quantity ${}^{w_y}\mathrm{P}_u^y[j-w_y]$. The subscript u emphasizes that the nucleotides x and y are supposed to be unpaired. Note that this is not the same as the problem of assessing the probability $P_u[i+w_x]$ that the individual nucleotides x_{i+w_x} is unpaired, because base pairing probabilities of adjacent nucleotides are highly correlated (Bompfünewerer *et al.*, 2008b).

The desired conditional probability can be written as:

$$ {}^{w_x}\mathrm{P}_u^x[i+w_x] = P_u^x([i+w_x]|[i,i+w_x-1]), \quad (1) $$

where the notation means that the interval $[i,i+w_x-1]$ is unpaired. An analogous expression holds for sequence y. Using the definition of the

conditional probability, we can write:

$$ {}^{w_x}\mathrm{P}_u^x[i+w_x] = \frac{P_u^x([i,i+w_x-1]\cup[i+w_x])}{P_u^x[i,i+w_x-1]} = \\ = \frac{P_u^x[i,i+w_x]}{P_u^x[i,i+w_x-1]} \tag{2} $$

Equation (2) tells us that the conditional probability ${}^{w_x}\mathrm{P}_u^x[i+w_x]$ depends only on the probabilities $P_u^x[i,i+w_x]$ and $P_u^x[i,i+w_x-1]$ that the corresponding intervals are unpaired. Conversely, the probability that an intervals is unpaired can be computed from the conditional probabilities and the probabilities that individual nucleotides are unpaired:

$$ P_u^x[i,i+w_x] = P_u^x[i]\cdot\prod_{j=1}^{w_x} {}^{j}\mathrm{P}_u^x[i+j] \tag{3} $$

A closer look at Equation (2) shows that the exact start position of the hybrid $\mathcal{S}_y^x$ has to be known in order to compute the desired conditional probability. Since RNAplex stores only a small number (four) of columns of the dynamic programming matrix, this cannot be done exactly. Instead we employ the approximation

$$ \frac{P_u^x[i,i+w_x]}{P_u^x[i,i+w_x-1]} \approx \frac{P_u^x[i+w_x-\delta+1,i+w_x]}{P_u^x[i+w_x-\delta+1,i+w_x-1]} \\ = {}^{\delta}\overline{\mathrm{P}}_u^x[i+w_x] \tag{4} $$

where δ represents the number of nucleotides considered prior to nucleotide x_{i+w_x} and ${}^{\delta}\overline{\mathrm{P}}_u^x[i+w_x]$ represents the conditional probability that x_{i+w_x} is unpaired for a given δ. This approximation is exact for $\delta = w_x$ and becomes worse with decreasing δ. This is a direct consequence of the fact that the state of the nucleotides in the interval $[i,i+w_x-\delta+1]$ is not taken into account for the computation of the conditional probability of nucleotide x_{i+w_x}.

Equation (3) can now be rewritten in the form

$$ P_u^x[i,i+w_x] \approx \\ {}^{\delta}\mathrm{P}_u^x[i,i+w_x] = P_u^x[i,i+\delta-1]\cdot\prod_{j=\delta}^{w_x} {}^{\delta}\overline{\mathrm{P}}_u^x[i+j] \tag{5} $$

The probability $P_u^x[i,i+w_x]$ of being unpaired is related to a corresponding opening energy

$$ \Delta G_u^x[i,i+w_x] = -\mathrm{RT}\ln P_u^x[i,i+w_x]. \tag{6} $$

The energy cost of adding one nucleotide to the hybrid therefore can be written as

$$ \Delta^{\delta}\overline{\mathrm{G}}_u^x[i+w_x] = -\mathrm{RT}\ln {}^{\delta}\overline{\mathrm{P}}_u^x[i+w_x] = \\ \Delta G_u^x[i+w_x-\delta+1,i+w_x] \\ -\Delta G_u^x[i+w_x-\delta+1,i+w_x-1]. \tag{7} $$

The opening energy of a region of size of w thus is given by

$$ \Delta^{\delta}\mathrm{G}_u^x[i,i+w_x] = -\mathrm{RT}\ln {}^{\delta}\mathrm{P}_u^x[i,i+w_x] = \\ \Delta G_u^x[i,i+\delta-1] + \sum_{j=\delta}^{w_x} \Delta^{\delta}\overline{\mathrm{G}}_u^x[i+j]. \tag{8} $$

Since RNAplex only stores the current four columns of the recursion matrix, we set $\delta = 4$ in practice.

2.3 Modified recursions of RNAplex

The energy $\Delta^4\overline{\mathrm{G}}_u^x[i]$ of freeing nucleotide x_i from all its intramolecular interactions can now easily be integrated into the dynamic programming recursion of RNAplex.

Let $C_{i,j}$ be the best interaction energy between the subsequences $x_1 \ldots x_i$ and $y_j \ldots y_m$. Similarly, $B_{i,j}^x$ and $B_{i,j}^y$ store the optimal interactions energy given that residue x_i or residue y_j, respectively, is part of a bulge; $I_{i,j}$ stores the optimal interaction energy given that x_i and y_j are in an interior loop.

The asymmetry penalty A models asymmetric extension of interior loops. $\mathcal{S}(i,j,i-1,j+1)$ represents the energy gained by stacking base pair (x_i,y_j) onto (x_{i+1},y_{j-1}). $\mathcal{M}(i,j,i-1,j+1)$ represents the mismatch energy of the unpaired nucleotide (x_{i-1},y_{j+1}) adjacent to the pair (x_i,y_j). The energy contribution of the small interior loops is represented by $\mathcal{I}$. Furthermore, we use the following abbreviations for the opening energies: $d_1^x = \Delta^4\overline{\mathrm{G}}_u^x[i]$, $d_2^x = d_1^x + \Delta^4\overline{\mathrm{G}}_u^x[i-1]$, $d_3^x = d_2^x + \Delta^4\overline{\mathrm{G}}_u^x[i-2]$; and $d_1^y = \Delta^4\overline{\mathrm{G}}_u^y[j]$, $d_2^y = d_1^y + \Delta^4\overline{\mathrm{G}}_u^y[j+1]$, $d_3^y = d_2^y + \Delta^4\overline{\mathrm{G}}_u^y[j+2]$. The full dynamic programming recursion then reads

$$ C_{i,j} = \min\begin{cases} C_{i-1,j+1} + \mathcal{S}(i,j;i-1,j+1) + d_1^x + d_1^y \\ C_{i-1,j+2} + \mathcal{S}(i,j;i-1,j+2) + P_{\mathrm{bulge}} + d_1^x + d_2^y \\ C_{i-2,j+1} + \mathcal{S}(i,j;i-2,j+1) + P_{\mathrm{bulge}} + d_2^x + d_1^y \\ C_{i-2,j+2} + \mathcal{I}(i,j;i-2,j+2) + d_2^x + d_2^y \\ C_{i-3,j+2} + \mathcal{I}(i,j;i-3,j+2) + d_3^x + d_2^y \\ C_{i-2,j+3} + \mathcal{I}(i,j;i-2,j+3) + d_2^x + d_3^y \\ C_{i-3,j+3} + \mathcal{I}(i,j;i-3,j+3) + d_3^x + d_3^y \\ I_{i-1,j+1} + \mathcal{M}(i,j;i-1,j+1) + d_1^x + d_1^y \\ B_{i-1,j+1}^x + d_1^x \\ B_{i-1,j+1}^y + d_1^y \end{cases} $$

$$ I_{i,j} = \min\begin{cases} C_{i-1,j+1} + \mathcal{M}(i-1,j+1;i,j) + \\ \qquad + g_{\mathrm{open}}^I + 2g_{\mathrm{ext}}^I + d_1^x + d_1^y \\ I_{i-1,j} + g_{\mathrm{ext}}^I + A + d_1^x \\ I_{i-1,j+1} + 2g_{\mathrm{ext}}^I + d_1^x + d_1^y \\ I_{i,j+1} + g_{\mathrm{ext}}^I + A + d_1^y \end{cases} $$

$$ B_{i,j}^x = \min\begin{cases} C_{i-1,j} + g_{\mathrm{open}}^B + g_{\mathrm{ext}}^B + d_1^x \\ B_{i-1,j}^x + g_{\mathrm{ext}}^B + d_1^x \end{cases} $$

$$ B_{i,j}^y = \min\begin{cases} C_{i,j+1} + g_{\mathrm{open}}^B + g_{\mathrm{ext}}^B + d_1^y \\ B_{i,j+1}^y + g_{\mathrm{ext}}^B + d_1^y \end{cases} $$

2.4 Hybrid structure and hybrid energy

The computation of the hybrid structure and interaction energy follows the strategy of RNAup. We assume that the binding region may contain mismatches and bulge loops. Thus, the most stable interaction between two segments (x_i,y_j) and (x_k,y_l) is obtained by minimizing over all possible interior loop closed by (x_p,y_q)

$$ C(x_i,y_j,x_k,y_l) = \min_{\substack{x_k<x_p<x_i \\ y_l>y_q>y_j}} C(x_i,y_j,x_p,y_q) + \\ I(x_p,y_q,x_k,y_l) + \Delta G_u^x[i,k] + \Delta G_u^y[j,l] \tag{9} $$

The overall most stable interaction is then obtained by minimizing over both duplex closing pairs (x_i,y_j) and (x_k,y_l):

$$ E_{\min} = \min_{\substack{x_1<x_k<x_i<x_n \\ y_1<y_j<y_l<y_m}} C(x_i,y_j,x_k,y_l) \tag{10} $$

where n and m are the length of sequences x and y, respectively. This leads to a theoretical run time of $\mathcal{O}(n^3\cdot m^3)$ and a memory footprint of $\mathcal{O}(n^2\cdot m^2)$.

Here we should note that one end of the hybrid, namely the base pair (x_i,y_j), was already found in the scanning phase of RNAplex. As a consequence, we only need to minimize over one closing pair instead of two. Equation (10) can thus be rewritten as:

$$ E_{\min} = \min_{\substack{x_1<x_k<x_i \\ y_j<y_l<y_m}} C(x_i,y_j,x_k,y_l) \tag{11} $$

Equations (10) and (11) show that the knowledge of base pair (x_i,y_j) allows to reduce memory and run time by a factor $n\cdot m$. Furthermore, the size of

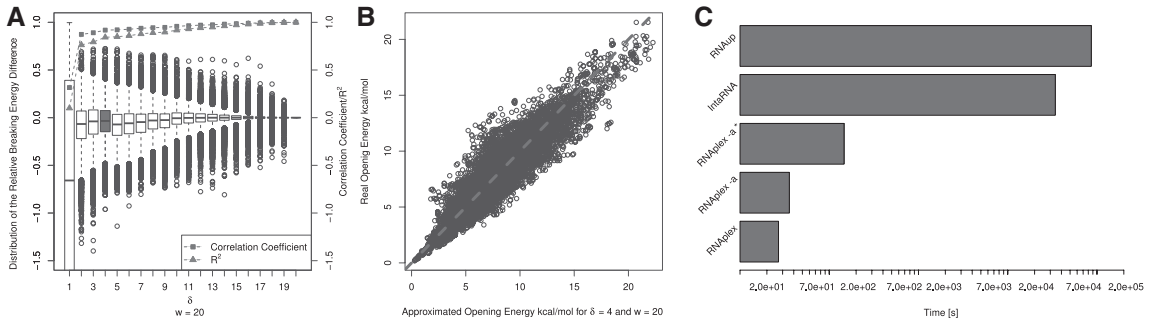

Fig. 2. (**A**) Boxplot representation of the distribution of the relative opening energy between our model and the standard energy model for different δ and a fixed target size of 20 nt. As expected, larger δ lead to smaller discrepancies. `RNAplex` uses $\delta = 4$. At this level of approximation, the Pearson's correlation coefficient between the approximated model and the real model reaches 0.92. (**B**) Scatterplot of the standard opening energies for 114 60 target sites of size 20 against the approximated opening energies as computed by `RNAplex`. (**C**) Bar plots representing the time necessary to complete the target search for 19 bacterial sRNAs in 100 random sequences of length 1200 nt for different RNA–RNA interaction tools. `RNAplex -c`, i.e. the old version of `RNAplex` is the fastest application with a completion time of 27 s. `RNAplex -a`, i.e. the new version of `RNAplex` considering accessibility, needs 36 s to achieve the same task. This grows to 120 s if one considers the time necessary to compute the accessibility profile. `RNAplex -a` is 1000 times faster than `IntaRNA` (Busch *et al.*, 2008) and 2422 times faster than `RNAup` (Mückstein *et al.*, 2008).

the interaction regions as well as the size of interior loops can be limited to arbitrary lengths ω and L, respectively, leading to a run time of $\mathcal{O}(\omega^2 \cdot L^2)$ and a memory usage of $\mathcal{O}(\omega^2)$, that is, the same complexity as `RNAduplex` or `RNAhybrid`.

2.5 Accuracy

We evaluated the performance of `RNAplex` at two levels. First, we looked at how well the opening energy derived by `RNAplex` from `RNAup` profiles matched the original `RNAup` values. Within the model of RNA secondary structures, this assess the quality of the approximations outlined in the previous section compared with the exact unpairing energies. Note that a comparison with experimentally measured opening energies is not possible since such measurements do not appear to be available in the published literature. The second test surveys how well `RNAplex` recovers the boundaries of known duplexes. This evaluates how well the different approximations made in `RNAplex` influence the quality of the predictions. The knowledge of the exact localization of RNA–RNA interactions is important, because ncRNAs may regulate their targets in different ways depending on the location of the binding sites.

In order to investigate the accuracy of the accessibility profiles, we used a set of 11 460 randomly generated sequences of length 400 nt for which the accessibility profiles was computed with `RNAup`. For each sequence, we then determined the difference of the `RNAup` opening energy and the `RNAplex` opening energy for the region located between nucleotides 181 and 200. Figure 2 shows the relative energy differences between both models as bar plots for different values of δ. The largest variations are seen for $\delta = 1$ with differences larger than 100%. R^2 (triangle) and the Pearson's correlation coefficient (square) reach their minimum there (0.09 and 0.37, respectively). Both coefficients then steadily improve with δ and reach their theoretical maximum of 1 for $\delta = w$. For $\delta < w$, our approximation slightly overestimates the opening energy. This can be seen for $\delta = 4$, the value used in `RNAplex` in the scatterplot in the middle of Figure 2. Half of the relative deviation are contained between +7% and −14%.

The accuracy of the energy model (interaction and opening energy) used in `RNAplex` was compared with that of `RNAup`, `biRNA` (Chitsaz *et al.*, 2009), and the old version of `RNAplex` (`RNAplex -c`) on a dataset of 17 known bacterial small RNA–mRNA interactions (Chitsaz *et al.*, 2009) (see Supplementary Material). In this dataset, both the opening energy of the interacting sequences and the hybridization energy affects the prediction.

`RNAplex -c` (old version) missed four interactions, while all `RNAplex -a` (with accessibility information) predictions overlapped with the corresponding experimentally determined interactions, as did the predictions of `RNAup` and `biRNA` (see Supplementary Table S2). These results emphasize the importance of accessibility for the correct prediction of RNA–RNA interactions. Furthermore, it confirms that the approximations used in `RNAplex` are sufficient to reach a level of accuracy similar to that of `RNAup` and `biRNA`.

The location of the predicted closing pairs was compared to the confirmed locations. For each prediction tool, the average over all 17 interactions of the sum of the magnitude of the deviation between the predicted and confirmed locations of the four closing nucleotides was computed. All three accessibility-based methods performed similarly with an average deviation of 16.76 for `RNAup`, 19.88 for `biRNA` and 20.60 for `RNAplex -a`, much smaller than the average deviation of `RNAplex -c` (59.76 nt) (see Supplementary Table S2).

It should be noted that `RNAup` and `RNAplex`, in contrast to `biRNA`, cannot handle interactions involving two or more interacting regions, such as the two kissing-hairpin complexes found in *OxyS-fhlA*. Still, in contrast to `RNAup`, `RNAplex` can return suboptimal predictions, without run time overhead, that can be used to identify disjoint interaction regions. For *OxyS-fhlA*, the confirmed binding regions are located at positions [22, 30] and [98, 104] on *OxyS* and [87, 95] and [39, 45] on *fhlA*, in accord with the two best suboptimals returned by `RNAplex` which are located on [23, 28] and [96, 100] on *OxyS* and [87, 92] and [41, 45] on *fhlA*.

2.6 Computational efficiency

The run time of the new version of `RNAplex` was compared with that of the old version (`RNAplex -c`, no accessibility), `RNAup` and `IntaRNA` (Busch *et al.*, 2008) on a dataset containing 19 *E.coli* sRNAs and 100 *E.coli* mRNAs (see Supplementary Material). For each gene, we defined the putative target region as the sequence interval from 200 nt upstream and 1000 nt downstream of the start codon.

`RNAplex` completed this task in 36 s, while `IntaRNA` and `RNAup` needed 34150 and 86487 s, respectively. The run time of `RNAplex` thus is reduced by a factor of 2400 and 950 compared with `RNAup` and `IntaRNA`, respectively. If we count the time needed to compute the accessibilities needed by `RNAplex`, the total run time reaches 120 s, still more than two orders of magnitude less than the other tools (Fig. 2).

We further compared the run time and the memory consumption of RNAup and IntaRNA against that of the new RNAplex, by generating a set of random target sequences of size 400, 800, 1600, 3200 and 6400 nt and query sequences of size 100, 200, 400 and 800 nt and searching for targets with all three tools. On this dataset, the new RNAplex is between 575 and 1600 times faster than IntaRNA and between 1500 and 65 400 times faster than RNAup. The memory consumption is also drastically reduced. RNAplex needs at least 17 and at most 1330 times less memory than IntaRNA, and 15–626 times less memory than RNAup (see Supplementary Table S1). Compared to the old version without accessibilities, the new RNAplex needs only four times more memory.

2.7 Conserved interactions

The absence of conserved target site in closely related species may indicate that the proposed interaction does not occur in nature. The presence of compensatory mutations between the sRNA and the target site, on the other hand, can lend further credibility to single sequence target predictions (Chen *et al.*, 2007). Alignments thus can improve the specificity of target search by focusing on evolutionary conserved interactions.

We, therefore, extended RNAplex to alignments. The approach follows the same idea as RNAalifold (Bernhart *et al.*, 2008; Hofacker *et al.*, 2002), where a thermodynamic energy minimization folding algorithm is coupled with a simple scoring model to assess structural evolutionary conservation. Base pairs are, therefore, restricted to pairs of positions in the alignments in which most or all sequences can form canonical pairs.

The evolutionary model used in RNAplex, while straightforward, performs well in predicting consensus secondary structure. Its simplicity allows it to be integrated into RNAplex without run time overhead (see Supplementary Material).

A potential weakness is the RNAalifold scoring model, which was trained and optimized for intramolecular interaction, instead for the intermolecular interactions to which it is applied here. More complex scoring schemes such as the one used in PETfold and PETcofold, where a maximum expected scoring approach combines the evolutionary probabilities of a consensus structure given an alignment with the thermodynamic probabilities of the associated structures in each sequence (Seemann *et al.*, 2008, 2010, 2011), perform slightly better than the RNAalifold scoring scheme. However, they can be incorporated only at the cost of a greatly increased run time, and thus are incompatible with the purpose of RNAplex.

Similar to the single sequence version, the alignment version of RNAplex only allows interior loops in the RNA–RNA hybrids. Like the single sequence, accessibility can be taken into account by averaging the position-dependent extension costs computed for the individual sequences in the alignment (see Supplementary Materials for a full description of the recursion).

2.8 Datasets

A complete description of all datasets used in this study can be found in the Supplementary Materials.

3 APPLICATION

As an application example, we consider the genome-wide prediction of sRNA targets in *E.coli*. As a reference set, we use the experimentally confirmed interactions published by Urban *et al.* (2007). We expect that, for a given sRNA, the number of predicted interactions with other (false positive) targets should decrease when accessibility of the target mRNA in included. Ideally, it should reach the low levels observed for RNAup (Mückstein *et al.*, 2008).

For each sRNA, the amount of false positives was estimated by counting genome wide the number of sRNA-target interactions that are more stable than the experimentally reported sRNA-target duplex. For each 4463 *E.coli* genes, a mRNA of length 1200 nt, including 200 nt upstream and 1000 nt downstream of the start codon were defined. Accessibility profiles were computed with RNAplfold, with a folding windows (option -W) of 240 nt and a maximal base pair distance of 160 (option -L). An interaction was reported if the corresponding sRNA–mRNA interaction energy is smaller than the experimentally confirmed interaction, and if it occurs in region encompassing 80 nt, 50 nt upstream and 30 nt downstream of the start codon.

The inclusion of the accessibility profiles in the new version of RNAplex leads to a substantial improvement as can be seen from Table 1. All native interaction sites are among the predictions, and the detailed target site localization is improved. Most importantly, the number of predictions with better interaction energies, i.e. the false positives, is reduced to a level similar to that of RNAup.

In order to better assess the number of false positives, the same method was applied on the dinucleotide-shuffled sRNAs and mRNAs. To this end, we compared the interaction energy of the non-shuffled, experimentally confirmed interactions, to the energy distribution of the shuffled sequences. Interestingly, in seven out of nine cases, the number of false positives is smaller (see Supplementary Material) in the shuffled case than in the non-shuffled one. This can be explained by the fact that in various bacteria, the region around the ribosomal entry site, which is also the preferred region of sRNA binding, is more accessible than the rest of the mRNA (see Supplementary Material). This in turn implies that compared with shuffled sequences, sRNAs have a greater chance to bind to the region around the start codon in non-shuffled mRNAs. Depending on the ncRNAs, one can expect between 7.5×10^{-7} false positives per nucleotide for micC and $1.5 \times 10{-}4$ false positives for gcvB (see Supplementary Material).

3.1 Multiple alignment

While RNAplex recovers all interactions, some of them like RyhB-sodB or GcvB-oppA are ranked lowly. A comparative version of RNAplex was designed (see Section 2) to reduce the number of false positives. Similar to consensus RNA folding, the quality of the input alignments is crucial to obtain meaningful results (Bernhart *et al.*, 2008).

The comparison of the performance of the single sequence with the comparative version of RNAplex was achieved by generating multiple sequences alignments clustalw (Larkin *et al.*, 2007) for the eight sRNAs from Table 1 and with MUSCLE (Edgar, 2004) for the 4463 *E.coli* mRNAs. The list of bacteria used for the alignment are found in the Supplementary Material.

In many cases, MUSCLE and clustalw were not able to satisfactorily align the sequences. This was caused e.g. by misannotations of the start codon as for the ompA gene in *E.coli* APEC O1, which was incorrectly annotated 70 nt upstream of the true start codon. In order to better handle these cases, we devised a method to produce multiple alignments of highly similar and strongly binding target sites (see Supplementary Materials).

Because highly conserved interactions are more credible than non-conserved interactions, ranking of interactions based on multiple sequences alignments should not only take the interaction energy into account, but also the number of organisms (in which a predicted interactions is detectable). This can be achieved by using *Z*-scores

Table 1. Summary of the predicted binding sites for the nine functional interactions reported by Urban *et al.* (2007)

sRNA	mRNA	Pos.lit.	$Pos_{RNAplex}$	ΔG RNAup	ΔG RNAplex	ΔG RNAplex -A	ΔG	*Z*-score	*Z*-score N°seq
RyhB	sodB	−7, +5	−4, +5	−10.50 (60)	−11.08 (50/87)	−9.31 (12)	65	57	2 (7)
DsrA	hns	+6, +21	+7, +19	−10.90 (17)	−12.74 (2/128)	−11.25 (10)	1	12	0 (0)
MicA	ompA	−21, −6	−21, −6	−13.46 (0)	−14.35 (1/67)	−14.04 (14)	0	11	0 (0)
MicC	ompC	−30, −15	−30, −15	−15.85 (1)	−16.24 (2/97)	−17.50 (9)	0	0	0 (0)
MicF	ompF	−8, +10	−16, +10	−17.00 (3)	−13.65 (8/34)	−18.28 (6)	0	0	0 (2)
Spot42	galK	−19, +14	−19, +21	−18.92 (0)	−13.02 (25/38)	−7.31 (9)	25	28	5 (12)
SgrS	ptsG	−28, −8	−28, +4	−17.17 (1)	−17.53 (0/170)	−11.17 (10)	5	4	0 (1)
GcvB	dppA	−31, −10	−31, −14	−16.90 (16)	−17.11 (8/80)	−13.15 (9)	14	14	7 (19)
GcvB	oppA	−4, 21	−8, 16	−11.64 (58)	−12.00 (36/263)	−14.43 (5)	27	26	14 (19)

The first and second columns show the name of interaction partners. Columns 3 and 4 give the predicted and experimentally reported binding regions, respectively. Columns 5 and 6 report the binding ΔG computed by RNAup and RNAplex, respectively. The numbers in parenthesis in the sixth column represent the number of interactions, located within a window of 80 nt centered around the start codon, with a lower interaction energy than the experimentally reported interaction for the predictions made by RNAplex with and without considering the opening energy, respectively. Column 7 gives the interaction energy for the multiple sequences interactions. The numbers in parenthesis in column 7 represent the number of sequences in the final alignments. Column 8 shows the rank of the interaction when looking only at the interaction energy. Column 9 shows the rank of the interactions based on the *Z*-score corrected for the number of sequences in the alignment. Finally, column 10 shows the rank of the interaction based on the *Z*-score, given that only interactions with a greater or equal number of sequences in the alignment are taken into account. The number in parenthesis in the last column represent the number of better scoring elements in the case of alignment when no accessibility information are taken into account.

as alternative ranking criterion. The *Z*-scores can be computed for all interactions having the same number of sequences in the alignments. This is important as highly conserved interactions tend to have a higher consensus interaction energy than interactions that are conserved in only few organisms (see Supplementary Figure S2).

In this way, extremely stable interactions can be compared without having to worry about the number of sequences in the alignments. The main drawback of this method is that highly conserved interactions with more than 10 sequences are rare, making the *Z*-score analysis unreliable. This is the case, for example, for the *micA-ompA* pair, which has the highest interaction energy among the interactions involving 14 species. In this case, the rank of MicA drops from 2 for the single sequence approach to 11 for the alignment approach.

Table 1 shows that the rank based on the interaction energy or the *Z*-score is similar to that of the single sequence energy ranking. However, when considering only interactions having a greater or equal number of sequences and a higher *Z*-score (column 10), the number of interactions that score better than the native one in the single sequence case (column 6) decreases significantly, with the greatest reduction being seen for *ryhB*. This is especially interesting because the *ryhB-sodB* is difficult to predict, probably due to its dependence upon *Hfq*, a protein known to facilitate sRNA–mRNA duplex formation (Sittka *et al.*, 2007). Similar to the single sequence case, the use of accessibility information in the case of multiple sequences alignments allows to improve the rank of the known interactions. This can be seen in the last column of Table 1.

It should be noted that some false positives turned out to be real interactions: for example, *iscS* and *acnB* score better than *sodB* as targets for *ryhB* and are true targets (Desnoyers *et al.*, 2009; Massé and Gottesman, 2002). Similar trends can be seen if the *Z*-score threshold is set to 0 and the number of sequences in the multiple alignment remains unchanged. If we look at the gene ontology of these targets in the case of *ryhB* (43 targets), we see that 35 are involved in catalytic activities ($P=0.006$), 9 are involved in iron–sulfur cluster binding ($P=0.007$), 39 are involved in binding ($P=0.01$). *ryhB* targets are also significantly overrepresented in the CO_2 fixation ($P=0.0001$) as well as citrate cycle cellular pathways ($P=0.0002$), in line with the gene ontology analysis. More examples can be found in the Supplementary Materials.

4 DISCUSSION

We presented a new version of RNAplex, a tool designed to rapidly and reliably predict RNA–RNA interactions. Compared with the previously published version, RNAplex now considers target site accessibility, by using accessibility profiles generated by RNAplfold to approximate the energy of removing a nucleotide from all intramolecular interactions. The introduction of position-specific, structure-dependent extension cost allows to greatly improve the specificity of RNAplex, bringing it close to that of RNAup, without modifying the linear run time of the original RNAplex.

Clearly, the main feature of RNAplex is its run time efficiency. On a dataset of 19 ncRNAs and 100 target mRNAs on length 1200, RNAplex runs 2400 faster than RNAup without noticeably loss of specificity, thus making ncRNAs target searches more affordable. In its present implementation, RNAplex can be used not only to predict ncRNA targets in small genomes, but can also be used to find miRNA targets and siRNA off-targets in large mammalian genomes and transcriptomes and it can be applied to microarray probes design. In contrast to RNAup or RNAhybrid, RNAplex can return suboptimal solutions efficiently on the fly without the need of recomputing the full recursion matrix.

The ability of RNAplex to perform comparative target search allows to discard poorly conserved interaction and to lend further credibility to interactions showing compensatory mutations. Based on a dataset of experimentally confirmed interactions, we show that RNAplex in its present form is an useful tool to predict new sRNA targets. We further show that suboptimal predictions from RNAplex may actually be real targets. Application of the comparative version

of RNAplex on larger genomes and other ncRNAs, e.g. miRNAs, is straightforward.

In order to make RNAplex more usable for the community, we plan to set up a web server especially designed to predict targets for sRNAs in bacteria. We further plan to use RNAplex to better understand the regulatory circuits found in *E. coli* (Shimoni *et al.*, 2007). Finally, a probe design method based on RNAplex is currently being developed.

Funding: European Union under the auspices of the FP-7 QUANTOMICS project (HT,PSD, in part); DFG priority program SPP1258 Sensory and Regulatory RNAs in prokaryotes (HT,PSD, in part); and Austrian GENAU project 'Regulatory Noncoding RNA' [ILH,FA].

Conflict of Interest: none declared.

REFERENCES

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

Andronescu,M. *et al.* (2003) RNAsoft: a suite of RNA secondary structure prediction and design software tools. *Nucleic Acids Res.*, **31**, 3416–3422.

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.

Backofen,R. *et al.*; The Athanasius F. Bompfünewerer RNA Consortium (2007). RNAs everywhere: genome-wide annotation of structured RNAs. *J. Exp. Zool. B Mol. Dev. Evol.*, **308B**, 1–25.

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.* (2006a) Local RNA base pairing probabilities in large sequences. *Bioinformatics*, **22**, 614–615.

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

Bernhart,S. *et al.* (2008) Rnaalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**, 474.

Bompfünewerer,A. *et al.* (2008) Variations on RNA folding and alignment: lessons from benasque. *J. Math. Biol.*, **56**, 129–144.

Busch,A. *et al.* (2008) IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, **24**, 2849–2856.

Chen,C. *et al*, (2007) Exploration of pairing constraints identifies a 9 base-pair core within box c/d snoRNA-rRNA duplexes. *J. Mol. Biol.*, **369**, 771–783.

Chitsaz,H. *et al.* (2009) *Algorithms in Bioinformatics*, Vol. 5724 of *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg.

Desnoyers,G. *et al.* (2009) Small RNA-induced differential degradation of the polycistronic mRNA iscrsua. *EMBO J.*, **28**, 1551–1561.

Dimitrov,R. and Zuker,M. (2004) Prediction of hybridization and melting for double-stranded nucleic acids. *Biophys. J.*, **87**, 215–226.

Dirks,R. *et al.* (2007) Thermodynamic analysis of interacting nucleic acid strands. *SIAM Rev.*, **49**, 65–88.

Edgar,R. (2004) Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

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.

Gardner,P. *et al.* (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res.*, **37**, 136–140.

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

Hofacker,I. *et al.* (2002) Secondary structure prediction for aligned rna sequences. *J. Mol. Biol.*, **319**, 1059–1066.

Huang,F. *et al.* (2010) Target prediction and a statistical sampling algorithm for RNA-RNA interaction. *Bioinformatics* , **26**, 175–181.

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

Larkin,M. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Massé,E. and Gottesman,S. (2002) A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli. *Proc. Natl Acad. Sci. USA*, **99**, 4620–4625.

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: Improved Computation of RNA-RNA Binding Thermodynamics. In Elloumi,M. *et al.* (eds) *Bioinformatics Research and Development*, Vol. 13, Springer, Berlin/Heidelberg, pp. 114–127.

Pervouchine,D. (2004) Iris: intermolecular rna interaction search. *Genome Inform.*, **15**, 92–101.

Rehmsmeier,M. *et al.* (2004) Fast and effective prediction of microRNA/target duplexes. *RNA*, **10**, 1507–1517.

Seemann,S. *et al.* (2008) Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Res.*, **36**, 6355–6362.

Seemann,S. *et al.* (2010) Hierarchical folding of multiple sequence alignments for the prediction of structures and RNA-RNA interactions. *Algorithms Mol. Biol.*, **5**, 22.

Seemann,S.E. *et al.* (2011) PETcofold: predicting conserved interactions and structures of two multiple alignments of RNA sequences. *Bioinformatics*, **27**, 211–219.

Shimoni,Y. *et al.* (2007) Regulation of gene expression by small non-coding RNAs: a quantitative view. *Mol. Syst. Biol.*, **3**, 138.

Sittka,A. *et al.* (2007) The RNA chaperone hfq is essential for the virulence of Salmonella typhirium. *Mol. Microbiol.*, **63**, 193–217.

Stark,A. *et al.* (2003) Identification of Drosophila MicroRNA targets. *PLoS Biol.*, **1**, e60.

Tafer,H. and Hofacker,I. (2008) Rnaplex: a fast tool for RNA-RNA interaction search. *Bioinformatics*, **24**, 2657–2663.

Urban,J. *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.

Zorio,D. *et al.* (1997) Cloning of caenorhabditis u2af65: an alternatively spliced RNA containing a novel exon. *Mol. Cell Biol.*, **17**, 946–953.