Fast Accessibility-Based Prediction of RNA-RNA Interactions

Hakim Tafer, Fabian Amman, Florian Eggenhofer, Peter F. Stadler, and Ivo L. Hofacker.

1 Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härterstrasse 16-18, D-04107 Leipzig, Germany
2 Inst. f. Theoretical Chemistry, University of Vienna, Währingerstrasse 17, A-1090 Vienna, Austria
3 Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany
4 RNomics Group, Fraunhofer Institut for Cell Therapy and Immunology, Perlikstraße 1,D-04103 Leipzig, Germany
5 The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, New Mexico, USA

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 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 pre-computed 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

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., 2007). ncRNAs are also involved in the editing of other RNA sequences (Benne, 1992), transcription and translation control (siRNA, miRNA, sRNA) (A. et al., 1998; Kugel and Goodrich, 2007; Banerjee and Slack, 2002) 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 E. 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. Bompfünnewerer 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 (Hofacker et al., 1994; Andronescu et al., 2003; Dimitrov and Zuker, 2004; Bernhart et al., 2006b; Dirks et al., 2007). A combinatorially different model, known as RNA-RNA Interaction Problem (RIP), covers a larger set of possible structures (Alkan et al., 2006; Pervouchine, 2004; Huang et al., 2010; Chitsaz et al., 2009).

* to whom correspondence should be addressed
The second type of approaches conceptually decomposes the RNA hybridization process into two stages: (1) the unfolding of the interacting regions of the two partners and (2) 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 combination of regions is evaluated (Mückstein et al., 2006, 2008; Busch et al., 2008). This approach was in particular applied successfully to rRNA-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 \( O((n + m)^3) \) 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 \( O(m \cdot n \cdot L^2) \) when restricting the maximum loop length to \( L \). RNAplex, a conceptually very similar approach (Tafer and Hofacker, 2008), further reduces the time complexity to \( O(m \cdot n) \) 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 RNAfold (Bernhart et al., 2006a; Bompfinewer 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 \( O(m \cdot n) \) 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-step approach. In the first step, the scanning phase, RNAplex identifies positions where putative interactions may end. For small interior loops \((1 \times 1, 2 \times 1, 2 \times 2)\), 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 4 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 \( S_y^x \) 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 \( u_y^x P^x_y[i + w_x, j + w_y] \) that nucleotide \( x_{i+w_x} \) is not involved in any intramolecular interaction, given that its predecessors \( x_{i+1}, \ldots, x_{i+w_x−1} \) is unpaired, and the analogous quantity \( u_y^x P^x_y[j−w_y, i−w_x] \). 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 $P_a[i + w_x]$ that the individual nucleotides $x_i, \ldots, x_{i+w_x}$ is unpaired, because base pairing probabilities of adjacent nucleotides are highly correlated (Bompeimüwezer et al., 2008b).

The desired conditional probability can be written as:

$$w_x P^u_a[i + w_x] = P^u_a ([i + w_x]), (1)$$

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:

$$w_x P^u_a[i + w_x] = \frac{P^u_a ([i, i + w_x - 1] \cup [i + w_x])}{P^u_a [i, i + w_x - 1]} (2)$$

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

$$P^u_a[i, i + w_x] = P^u_a [i] \prod_{j=1}^{w_x} P^u_a[i + j] (3)$$

A closer look at equation (2) shows that the exact start position of the hybrid $S^u_a$ 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

$$P^u_a[i, i + w_x] \approx P^u_a[i + w_x - \delta + 1, i + w_x - 1] (4)$$

where $\delta$ represents the number of nucleotides considered prior to nucleotide $x_i + w_x$ and $P^u_a[i + w_x]$ represents the conditional probability that $x_i + w_x$ is unpaired for a given $\delta$. This approximation is exact for $\delta = w_x$ and becomes worse with decreasing $\delta$. 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

$$P^u_a[i, i + w_x] \approx \delta P^u_a[i, i + w_x] = P^u_a[i, i + \delta - 1] \prod_{j=1}^{\delta} P^u_a[i + j] (5)$$

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

$$\Delta G^u_a[i, i + w_x] = -RT \ln P^u_a[i, i + w_x]. (6)$$

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

$$\Delta^a G^u_a[i + w_x] = -RT \ln \delta P^u_a[i + w_x] = \Delta G^u_a[i + w_x - 1, i + w_x] - \Delta G^u_a[i + w_x - \delta, i + w_x - 1]. (7)$$

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

$$\Delta G^u_a[i, i + \delta] = -RT \ln \delta P^u_a[i, i + \delta + \delta - 1] = \Delta G^u_a[i, i + \delta - 1] + \sum_{\delta} \Delta^a G^u_a[i + j]. (8)$$

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 $\Delta G^u_a[i]$ 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 between the sub-sequences $x_1 \ldots x_i$ and $y_1 \ldots y_j$. Similarly, $B_{i,j}$ and $B_{i,j}^o$ store the optimal interactions energy given that residue $x_i$ or residue $y_j$, resp., 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. $S(i, j, i - 1, j + 1)$ represents the energy gained by stacking base-pair $(x_i, y_j)$ onto $(x_{i-1}, y_{j-1})$. $M(i, j, i - 1, j + 1)$ 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 $I$. Furthermore, we use the following abbreviations for the opening energies:

$$d^u_1 = \Delta^a G^u_a[i], d^u_2 = d^u_1 + \Delta G^u_a[i - 1], d^u_3 = d^u_2 + \Delta G^u_a[i - 2], (9)$$

The full dynamic programming recursion then reads

$$C_{i,j} = \min \begin{cases} C_{i-1,j+1} + S(i,j;i - 1, j + 1) + d^u_1 + d^u_1 \\ C_{i-1,j+2} + S(i,j;i - 1, j + 2) + P_{i,j} + d^u_1 + d^u_2 \\ C_{i-2,j+3} + S(i,j;i - 1, j + 1) + d^u_1 + d^u_2 \\ C_{i-2,j+4} + S(i,j;i - 1, j + 2) + d^u_1 + d^u_2 \\ C_{i-3,j+5} + S(i,j;i - 1, j + 3) + d^u_1 + d^u_2 \\ C_{i-3,j+6} + S(i,j;i - 1, j + 4) + d^u_1 + d^u_3 \\ C_{i-4,j+7} + S(i,j;i - 1, j + 5) + d^u_1 + d^u_3 \\ B_{i+1,j+1} + M(i, j; i - 1, j + 1) + d^u_2 \\ B_{i+1,j+1} + d^u_2 \\ B_{i+1,j+2} + d^u_2 \\ B_{i+1,j+1} + d^u_3 \\ B_{i+1,j+1} + d^u_3 \end{cases}$$

$$I_{i,j} = \min \begin{cases} I_{i-1,j+1} + M(i-1,j+1; i, j) + d^u_1 + d^u_2 + \Delta G^u_a[i - 1] + d^u_3 + d^u_3 & \\ I_{i-1,j+2} + \Delta G^u_a[i - 1] + d^u_1 + d^u_2 + \Delta G^u_a[i - 1] + d^u_3 + d^u_3 \end{cases}$$

$$B_{i,j} = \min \begin{cases} B_{i,j} + \Delta G^u_a[i] + d^u_1 + d^u_2 + \Delta G^u_a[i] + d^u_3 \\ B_{i+1,j} + \Delta G^u_a[i] + d^u_1 + d^u_2 + \Delta G^u_a[i] + d^u_3 \end{cases}$$

2.4 Hybrid Structure and Hybrid Energy

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

$$C(x_k, y_k) = \min_{x_k < x_p < x_q} C(x_k, y_k; x_p, y_q) \quad (9)$$

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

$$E_{\text{min}} = \min_{x_k < x_p < x_q < x_m} C(x_k, y_k; x_p, y_q) \quad (10)$$

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

We should note that one end of the hybrid, namely the base-pair $(x_1, y_1)$, 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:

\[
E_{\text{min}} = \min_{x_i, y_j} \sum_{x_i, y_j < x_m} C(x_i, y_j)
\]  

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 the interaction regions as well as the size of interior loops can be limited to arbitrary lengths \(\omega\) and \(L\), respectively, leading to a run-time of \(O(\omega^2 L^7)\) and a memory usage of \(O(\omega^2)\), that is, the same complexity as 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 to 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 11460 randomly generated sequences of length 400nts 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 \(\delta\). The largest variations are seen for \(\delta = 1\) with differences larger than 100%. \(R^2\) (triangle) and the Pearson correlation coefficient (square) reach their minimum there (0.09 and 0.37, respectively). Both coefficients then steadily improve with \(\delta\) and reach their theoretical maximum of 1 for \(\delta = \infty\). For \(\delta < \infty\), 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 to 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 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 2). 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 nts) (see Supplementary Table 2).

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 OsyS-fhlA. Still, in contrast to RNAup, RNAplex can return suboptimal predictions, without runtime overhead, that can be used to identify disjoint interaction regions. For OsyS-fhlA, the confirmed binding regions are located at positions [22, 30] and [98, 104] on OsyS 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 OsyS and [87, 92] and [41, 45] on fhlA.

2.6 Computational efficiency

The runtime of the new version of RNAplex was compared to 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 nts upstream and 1000 nts downstream of the start codon.

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

We further compared the runtime 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's and querying sequences of size 100, 200, 400 and 800 nts 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 65400 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 to 626 times less memory than RNAup (see Supplementary Table 1). 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 runtime 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 runtime, 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.
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 is included. Ideally, it should reach the low levels observed for RNAup (Mückstein et al., 2008).

### 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 is included. Ideally, it should reach the low levels observed for RNAup (Mückstein et al., 2008).

We used the following methodology: for a given sRNA and the corresponding confirmed sRNA-mRNA interactions, we looked genome-wide at how many mRNAs binds with a lower energy to the sRNA than the reported sRNA-mRNA interaction, i.e. the number of false positive. For each 4463 e.Coli genes, a mRNA of length 1200 nts, including 200nts upstream and 1000nts downstream of the start codon were defined. Accessibility profiles were computed with RNAfold, 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 nts, 50 nts upstream and 30 nts 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 7 out of 9 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 to 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 $7.5 \times 10^{-1}$ 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 methods) to reduce the number of false positive. 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 8 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 Escherichia coli APEC 01, which was incorrectly annotated 70 nts 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...
### Table 1. Summary of the predicted binding sites for the 9 functional interactions reported by (Urban et al., 2007). The first and second columns show the name of interaction partners. Column 3 and 4 give the predicted and experimentally reported binding regions, respectively. Column 5 and 6 report the binding \( \Delta G \) 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 nts 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.

<table>
<thead>
<tr>
<th>sRNA</th>
<th>mRNA</th>
<th>Pos.lit.</th>
<th>PosRNAplex</th>
<th>( \Delta G ) RNAup</th>
<th>( \Delta G ) RNAplex</th>
<th>( \Delta G ) RNAplex -A</th>
<th>( \Delta G )</th>
<th>Z-score</th>
<th>Z-score N’seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>RyhB</td>
<td>sodB</td>
<td>-7.45</td>
<td>-4. +5</td>
<td>-10.50(60)</td>
<td>-11.08(50/87)</td>
<td>-9.31(12)</td>
<td>65</td>
<td>57</td>
<td>2 (7)</td>
</tr>
<tr>
<td>DsrA</td>
<td>has</td>
<td>+6,+21</td>
<td>+7, +19</td>
<td>-10.90(17)</td>
<td>-12.74(2/128)</td>
<td>-11.25(10)</td>
<td>1</td>
<td>12</td>
<td>0 (0)</td>
</tr>
<tr>
<td>MicA</td>
<td>ompA</td>
<td>-21, -6</td>
<td>-21, -6</td>
<td>-13.46(0)</td>
<td>-14.35(1/67)</td>
<td>-14.04(14)</td>
<td>0</td>
<td>11</td>
<td>0 (0)</td>
</tr>
<tr>
<td>MicC</td>
<td>ompC</td>
<td>-30, -15</td>
<td>-30, -15</td>
<td>-15.85(1)</td>
<td>-16.24(297)</td>
<td>-17.50(9)</td>
<td>0</td>
<td>0</td>
<td>0 (0)</td>
</tr>
<tr>
<td>MicF</td>
<td>ompF</td>
<td>-8, +10</td>
<td>-16, +10</td>
<td>-17.00(3)</td>
<td>-13.65(8/34)</td>
<td>-18.28(6)</td>
<td>0</td>
<td>0</td>
<td>0 (2)</td>
</tr>
<tr>
<td>Spot42</td>
<td>galK</td>
<td>-19, +14</td>
<td>-19, +21</td>
<td>-18.92(0)</td>
<td>-13.02(25/38)</td>
<td>-7.31(9)</td>
<td>25</td>
<td>28</td>
<td>5 (12)</td>
</tr>
<tr>
<td>SgrS</td>
<td>ptsG</td>
<td>-28, -8</td>
<td>-28, +4</td>
<td>-17.17(1)</td>
<td>-17.53(0/170)</td>
<td>-11.17(10)</td>
<td>5</td>
<td>4</td>
<td>0 (1)</td>
</tr>
<tr>
<td>GcvB</td>
<td>dppA</td>
<td>-31, -10</td>
<td>-31, -14</td>
<td>-16.90(16)</td>
<td>-17.11(8/80)</td>
<td>-13.15(9)</td>
<td>14</td>
<td>14</td>
<td>7 (19)</td>
</tr>
<tr>
<td>GcvB</td>
<td>oppA</td>
<td>-4, 21</td>
<td>-8, 16</td>
<td>-11.64(58)</td>
<td>-12.00(36/263)</td>
<td>-14.43(5)</td>
<td>27</td>
<td>26</td>
<td>14 (19)</td>
</tr>
</tbody>
</table>

4 DISCUSSION

We presented a new version of RNAplex, a tool designed to rapidly and reliably predict RNA-RNA interactions. Compared to the previously published version, RNAplex now considers target site accessibility, by using accessibility profiles generated by RNAfold 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 runtime 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 into account, but also the number of organisms (in which a predicted interaction is detectable). This can be achieved by using Z-scores 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 2).

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 acaN score better than sodB as targets for ryhB and are true targets (Massé and Gottesman, 2002; Desnoyers et al., 2009). 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 (\( p = 0.007 \)), and 39 are involved in binding (\( p = 0.01 \)). ryhB targets are also significantly overrepresented in the CO2 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.
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 compa-
rative 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 webserver especially designed to predict targets
for sRNAs in bacteria. We further plan to use RNAplex to bet-
tter 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 FP7 QUAN-
TOMICS project (in part), the Austrian GEN-AU project ‘Regu-
latory Noncoding RNA’ (in part), and the DFG priority program
SPP1258 Sensory and Regulatory RNAs in prokaryotes (in part).

REFERENCES

A., F., Xu, S., Montgomery, M., Kostas, S., Driver, S., and Mello,
C. (1998). Potent and specific genetic interference by double-
806–811.

Alkan, C., Karakoc, E., Nadeau, J., Sahinalp, S., and Zhang, K.
(2006). Rna-rna interaction prediction and antisense rna target

Andronescu, M., Aguirre-Hernandez, R., Condon, A., and Hoos,
H. H. (2003). RNAsoft: A suite of RNA secondary structure pre-
diction and design software tools. Nucleic Acids Res.,
31(13), 3416–3422.

rna: kissing complex formation at two sites results in a stable

Bachellerie, J., Cavaillé, J., and Hüttenhofer, A. (2002). The
expanding snoRNA world. Biochimie, 84(8), 775–790.

Banerjee, D. and Slack, F. (2002). Control of developmental timing
by small temporal rnas: a paradigm for rna-mediated regulation

Benne, R. (1992). Rna editing in trypanosomes, the use( ) of guide

pairing probabilities in large sequences. Bioinformatics, 22(5),
614–615.

Bernhart, S., Tafer, H., Mückstein, U., Flamm, C., Stadler, P.,
and Hofacker, I. (2006b). Partition function and base pairing

Bernhart, S., Hofacker, I., Will, S., Gruber, A., and Stadler,
rna alignments. BMC Bioinformatics, 9, 474–474.

Bompfünderer, A., Backofen, R., Bernhart, S., Hertel, J.,
Hofacker, I., Stadler, P., and Will, S. (2008a). Variations on

Bompfünderer, A., Backofen, R., Berhart, S., Hertel, J., Hofacker,
144.

cient prediction of bacterial sRNA targets incorporating target site
accessibility and seed regions. Bioinformatics, 24, 2849–2856.

of pairing constraints identifies a 9 base-pair core within box c/d

in Bioinformatics, volume 5724 of Lecture Notes in Computer

Small rna-induced differential degradation of the polycistronic
mrna iscrua. EMBO J, 28(11), 1551–1561.

melting for double-stranded nucleic acids. Biophys J, 87, 215–
226.

Dirks, R., Bois, J., Schaeffer, J., Winfree, E., and Pierce,
N. (2007). Thermodynamic analysis of interacting nucleic acid

accuracy and high throughput. Nucleic Acids Res, 32(5), 1792–
1797.

Eguchi, Y. and Tomizawa, J. (1990). Complex formed by comple-
timentary rna stem-loops and its stabilization by a protein: function

Gardner, P., Daub, J., Tate, J., Navrocki, E., Kolbe, D., Lindgreen,
S., Wilkinson, A., Finn, R., Griffiths-Jones, S., Eddy, S.,
Nucleic Acids Res, 37(Database issue), 136–140.

Hofacker, I., Fontana, W., Stadler, P., Bonhoeffer, S., Tacker,

1066.

Huang, F., Qin, J., Reidys, C., and Stadler, P. (2010). Target
prediction and a statistical sampling algorithm for RNA-RNA


Larkin, M., Blackshields, G., Brown, N., Chenna, R., McGettigan,
P., McWilliam, H., Valentin, F., Wallace, I., Wilm, A., Lopez,

Massé, E. and Gottesman, S. (2002). A small rna regulates the
expression of genes involved in iron metabolism in escherichia

Mückstein, U., Tafer, H., Hackermüller, J., Bernhart, S., Stadler,
Bioinformatics, 22(10), 1177–1182.

Mückstein, U., Tafer, H., Bernhart, S., Hernandez-Rosales, M.,
control by rna interaction. Submitted to BIRD 2008.


Rehmsmeier, M., Steffen, P., Hochsmann, M., and Giegerich,
duplexes. RNA, 10(10), 1507–1517.

ary and thermodynamic information for rna folding of multiple
alignments. *Nucleic Acids Res.*, **36**.