BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 24 2008, pages 2849–2856
doi:10.1093/bioinformatics/btn544

Sequence analysis

IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions

Anke Busch, Andreas S. Richter and Rolf Backofen*

Bioinformatics Group, Albert-Ludwigs-University Freiburg, Georges-Koehler-Allee 106, Freiburg D-79110, Germany

Received on June 16, 2008; revised on October 14, 2008; accepted on October 17, 2008

Advance Access publication October 21, 2008

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: During the last few years, several new small regulatory RNAs (sRNAs) have been discovered in bacteria. Most of them act as post-transcriptional regulators by base pairing to a target mRNA, causing translational repression or activation, or mRNA degradation. Numerous sRNAs have already been identified, but the number of experimentally verified targets is considerably lower. Consequently, computational target prediction is in great demand. Many existing target prediction programs neglect the accessibility of target sites and the existence of a seed, while other approaches are either specialized to certain types of RNAs or too slow for genome-wide searches.

Results: We introduce IntaRNA, a new general and fast approach to the prediction of RNA–RNA interactions incorporating accessibility of target sites as well as the existence of a user-definable seed. We successfully applied IntaRNA to the prediction of bacterial sRNA targets and determined the exact locations of the interactions with a higher accuracy than competing programs.

Availability: http://www.bioinf.uni-freiburg.de/Software/

Contact: IntaRNA@informatik.uni-freiburg.de

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

1 INTRODUCTION

Starting with the discovery of microRNAs (miRNAs) and the advent of genome-wide transcriptomics, it has become obvious that RNA plays a large variety of important, often regulatory, roles in living organism that extends far beyond being a mere intermediate in protein biosynthesis (Storz, 2002). Several of these non-protein coding RNAs (ncRNAs) regulate gene expression post-transcriptionally through base pairing to a target mRNA, like eukaryotic miRNAs and small interfering RNAs (siRNAs) (Bartel, 2004; Hannon, 2002; Zamore and Haley, 2005) as well as bacterial small regulatory RNAs (sRNAs) (Gottesman, 2005).

Typically, the size of bacterial sRNAs ranges from 50 nt to 250 nt (Vogel and Wagner, 2007). The post-transcriptional interaction with their target mRNA causes translational repression or activation, mRNA degradation or changes in mRNA stability (Storz *et al.*, 2004). They have been found to be crucial in the bacterial stress response and in bacterial virulence (Gottesman, 2005). Please note that we use the term sRNAs for small, regulatory RNAs in bacteria as done predominantly in the literature (e.g. Storz and Haas, 2007).

Since base pairing to a target mRNA is the main regulatory mechanism of sRNAs (Vogel and Wagner, 2007), the hybridization energy is a widely used criterion to predict RNA–RNA interactions (Rehmsmeier *et al.*, 2004; Tjaden *et al.*, 2006). For miRNAs and siRNAs, it was shown that the accessibility of the target sites has also an important influence (Ameres *et al.*, 2007; Kertesz *et al.*, 2007; Kretschmer-Kazemi Far and Sczakiel, 2003; Long *et al.*, 2007; Luo and Chang, 2004; Shao *et al.*, 2007). It can be assumed that the same holds for sRNAs. Furthermore, perfect Watson–Crick pairing of seven or eight consecutive bases (typically at positions 2–8) at the 5′ end of animal miRNAs (the *seed region*) is often sufficient for effective regulation (Bentwich, 2005; Brennecke *et al.*, 2005; Doench and Sharp, 2004). There is not much known about possible seed regions in bacterial sRNAs. A previous work about sRNA target prediction suggested that the interaction of sRNA and mRNA also starts with a stretch of bases that are unpaired in the sRNA and the mRNA and that form at least a minimal number of consecutive base pairs (Tjaden *et al.*, 2006). A very recent study in *Salmonella* confirms this assumption by showing that the conserved 5′-end of the RybB sRNA recognizes many *omp* mRNA targets by short seed pairings (Mika *et al.*, 2008).

In *Escherichia coli* (*E.coli*), there are >70 validated sRNAs, but only about 20 with a known cellular function since the experimental identification and verification of sRNA targets has been lagging behind (Vogel and Wagner, 2007). Thus, there is a high demand for computational target predictions for regulatory sRNAs. So far, there is no tool that integrates both accessibility of the target regions in the mRNA and in the sRNA and the existence of an arbitrary seed in a general approach.

This article presents IntaRNA, a new approach to the prediction of **int**e**r**acting **RNA**s (IntaRNA), where a combined energy score of the interaction is calculated as the sum of the free energy of hybridization and the free energy required for making the interaction sites accessible. In addition, the existence (but not the exact location) of a seed is enforced. The length of the seed can be set by the user. We present two variants: a complete approach, which is $O(n^2m^2)$ in time and $O(nm)$ in space, when restricting the size of internal loops and where n and m are the lengths of the interacting RNA sequences ($n > m$), and a heuristic simplification of the complete approach, which has a time complexity of $O(n\overline{m})$ and a space complexity of $O(nm)$, where $\overline{m} = \max\{m, L^3\}$ and L is the size of the sequence window in which both the target mRNA and the sRNA are folded.

*To whom correspondence should be addressed.

We successfully applied INTARNA to the prediction of sRNA targets. Additionally, INTARNA predicted the exact location of the RNA–RNA interactions.

Related work: to date, most existing target prediction approaches are based on one of the two following basic ideas. The first major group determines a common structure for ncRNA *and* targeted mRNA by concatenating the two RNA sequences and memorizing the linkage location. The new single sequence is folded by an usual RNA-folding algorithm, e.g. the algorithm of Zuker and Stiegler (1981), with slightly different parameters for the loop including the linkage location. The most prominent tools using this idea are PAIRFOLD (Andronescu *et al.*, 2005) and RNACOFOLD (Bernhart *et al.*, 2006b). They have a space complexity of $O((n+m)^2)$ and a time complexity of $O((n+m)^3)$ when restricting the size of interior loops. The main problem with these tools is that they can only predict interactions where the common structure is pseudoknot-free. In contrast, many interactions in living cells are located in loop regions and would represent a pseudoknot in the context of concatenated sequences.

The second major group of target prediction tools neglects intra-molecular binding in both RNA molecules. Algorithms based on this idea find the energetically most favorable hybridization of two RNA sequences. The most popular tools incorporating this idea are RNAHYBRID (Kruger and Rehmsmeier, 2006; Rehmsmeier *et al.*, 2004), RNADUPLEX and RNAPLEX (Tafer and Hofacker, 2008), and DINAMELT (Dimitrov and Zuker, 2004; Markham and Zuker, 2005). RNAHYBRID is primarily tailored for predicting potential miRNA binding sites in large target RNAs. This method uses a modification of the classical secondary structure prediction algorithm of Zuker and Stiegler (1981) that neglects multiloops. Furthermore, the loop size is restricted to a fixed value to reduce complexity. In principle, RNADUPLEX and RNAPLEX incorporate the same ideas as RNAHYBRID, but RNAPLEX uses a simplified energy scoring of loops and a length penalty to favor short stable interactions. By doing so, RNAPLEX performs 10–27 times faster than RNAHYBRID (Tafer and Hofacker, 2008).

There is a variety of additional tools that are specially designed to search for miRNA target sites [for a review see Bentwich (2005) and Yoon and De Micheli (2006)]. In contrast, there has been little investigation so far regarding the computational prediction of mRNA targets of bacterial sRNAs. RNAPLEX is also suitable for longer queries like sRNAs because it integrates a nucleotide penalty. It was recently applied to the prediction of sRNA targets (Tafer and Hofacker, 2008). Tjaden *et al.* (2006) developed a tool named TARGETRNA that predicts the targets of bacterial sRNAs (neglecting intra-molecular base pairs) and outputs them in a ranked list.

While most of the aforementioned approaches use the free energy of the hybridized duplex to predict the potential target site, in general, the free energy of the entire duplex is a poor predictor for that aim (Rajewsky, 2006). Several authors have shown that the secondary structure of the target mRNA (Ameres *et al.*, 2007; Kertesz *et al.*, 2007; Kretschmer-Kazemi Far and Sczakiel, 2003; Long *et al.*, 2007; Luo and Chang, 2004; Shao *et al.*, 2007) and the ncRNA (Koberle *et al.*, 2006) has a strong effect on target recognition. To the best of our knowledge, there are only two tools that incorporate the secondary structure of the mRNA (Kertesz *et al.*, 2007; Mückstein *et al.*, 2008). RNAUP (Mückstein *et al.*, 2008) calculates the thermodynamics of RNA–RNA interactions as the sum of two energy contributions: the energy needed to open the binding sites and make them accessible, and the hybridization energy. It has a space complexity of $O(n^2+nw^3)$ and a time complexity of $O(n^3+nw^5)$, when restricting the interior loop sizes to a fixed value and limiting the size of interaction to w. In contrast to our approach, RNAUP does not use any seed condition. PITA, developed by Kertesz *et al.* (2007) for miRNA target prediction, starts with a genome-wide search for initial seed regions and tries to extend these sites in one direction. This is typical for a large proportion of miRNAs, but is unlikely to hold for other ncRNAs.

Herein, we present the algorithmic details of INTARNA, our general approach to the prediction of RNA–RNA interactions, including target site accessibility and user-definable seeds.

2 METHODS

We are given two potentially interacting sequences S^1 and S^2 of lengths n and m, respectively. For every single RNA sequence S^k, a *target site* is a pair of positions $[x,y]$ that define an interval with x being the first and y being the last included position.

Hybridization energy: the first component determining the quality of an RNA–RNA interaction between target site $[i,k]$ in S^1 and target site $[j,l]$ in S^2 is the *hybridization energy* $E^{\text{hybrid}}(i,j,k,l)$. Its calculation is based on the energy model of RNAHYBRID. The energy parameters used are from Mathews *et al.* (1999). With $H(i,j)$, we denote the hybridization energy of the best interaction of subsequences $S^1_i \ldots S^1_n$ and $S^2_j \ldots S^2_m$, where the left-most positions of both subsequences i and j form a base pair. $H(i,j)$ can be calculated using a restricted variant of the algorithm of Zuker and Stiegler (1981) discarding multiloop structures. This algorithm has a time complexity of $O(nm)$ when restricting the internal loop length. Using the RNAHYBRID convention to number the first RNA $5' \rightarrow 3'$ and the second in the reverse direction, we get the following basic recursion for $H(i,j)$:

$$H(i,j) = \begin{cases} \min_{p,q} \left\{ E^{\text{loop}}(i,j,p,q) + H(p,q) \right\} & \text{if } S^1_i, S^2_j \text{ can pair} \\ \infty & \text{otherwise.} \end{cases} \quad (1)$$

Here, $E^{\text{loop}}(i,j,p,q)$ indicates the free energy of the loop including base pairs (i,j) and (p,q). We disregard dangling end energy contributions for the purpose of simplification here and in the following. Thus, matrix H is initialized with 0. The final hybridization energy is then calculated by choosing $\min_{i,j}\{H(i,j)\}$. The target site itself is calculated using a normal traceback.

Accessibility: the second component contributing to the quality of an RNA–RNA interaction is the *accessibility* of the target sites in each sequence, which is the energy required to make them single stranded. It is defined as the difference between the energy of the ensemble of all structures and the energy of the ensemble of structures, where the target site $[i,k]$ is single stranded. It is denoted by $ED(i,k)$ and calculated using a partition function approach (McCaskill, 1990). Let $\mathcal{S}$ be the set of all structures (called *ensemble*) that can be formed by a sequence S. Then

$$Z_{\mathcal{S}} = \sum_{Q \in \mathcal{S}} e^{-\frac{E(Q)}{RT}} \text{ and } E^{\text{ens}}(\mathcal{S}) = -RT \ln(Z_{\mathcal{S}})$$

where $Z_{\mathcal{S}}$ is the partition function of $\mathcal{S}$, $E(Q)$ is the free energy of sequence S folded into the secondary structure Q and $E^{\text{ens}}(\mathcal{S})$ denotes the ensemble energy of the set of structures $\mathcal{S}$. Let $\mathcal{S}^{\text{unpaired}}_{i,k}$ be the set of all structures of S that have nucleotides $S_i, S_{i+1}, \ldots, S_k$ unpaired. Then,

$$ED(i,k) = E^{\text{ens}}(\mathcal{S}^{\text{unpaired}}_{i,k}) - E^{\text{ens}}(\mathcal{S}),$$

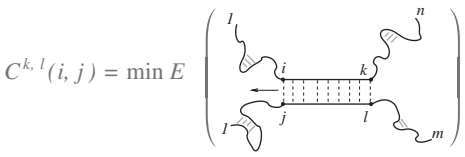

Fig. 1. Interpretation of the matrix $C^{k,l}(i,j)$.

which is greater or equal 0 by definition. The probability $PU(i,k)$ that the complete region between i and k is unpaired can be calculated by the equation $PU(i,k)=e^{-\frac{ED(i,k)}{RT}}$. $ED(i,k)$ can be obtained using RNAPLFOLD (parameter -u) (Bernhart *et al.*, 2006a; Bompfünewerer *et al.*, 2008) in $O(nL^2)$, where L represents the size of the locally folded subsequence.

Next, we combined both energy contributions. The *extended hybridization energy* of a specific interaction of two target sites $[i,k]$ and $[j,l]$ is now defined by summing up the *ED*-values and the hybridization energy. For calculating the *ED*-values, we must know the first and the last interacting base in both sequences. Hence, the basic recursion for calculating the extended hybridization energy requires a 4D array $C(i,j,k,l)$. Note that the basic assumption in $C(i,j,k,l)$ is that both (i,j) and (k,l) form a base pair. Thus, we have

$$C(i,j,k,l)=H(i,j,k,l)+ED(i,k)+ED(j,l)$$

where $H(i,j,k,l)$ is the 4D variant of the matrix calculated in Equation (1):

$$H(i,j,k,l)=\begin{cases}\min\limits_{p,q}\left\{E^{\text{loop}}(i,j,p,q)+H(p,q,k,l)\right\} & \text{if } S_i^1, S_j^2 \text{ can pair}\\ \infty & \text{otherwise.}\end{cases}$$

We achieve a complexity of $O(n^2m^2)$ time and $O(n^2m^2)$ space when limiting the size of the loops similar to RNAUP. When we limit the interaction length to w (as done in RNAUP), this approach has a complexity of $O(nmw^2)$ time and $O(nmw^2)$ space. In the following, we will show how to improve the time and space complexity without restricting the interaction size while integrating seed information in addition.

2.1 Reducing the space complexity

The space complexity can be improved by calculating all interactions for a common interaction start in one step. This leads to a 2D matrix $C^{k,l}(i,j)$, which is basically the slice of $C(i,j,k,l)$ for fixed k,l. The hybridization that starts at base pair (k,l), is elongated to the left and ends with base pair (i,j) (Fig. 1).

Considering the recursion for $C^{k,l}(i,j)$, note that the *ED*-values are already included. Since *ED*-values are not additive, we have to subtract the old *ED*-values before adding the new ones. The idea of the recursion for $C^{k,l}(i,j)$ is illustrated in Figure 2.

Formally, we get the following recursion:

$$C^{k,l}(i,j)=\begin{cases}\min\limits_{p,q}\begin{pmatrix}E^{\text{loop}}(i,j,p,q)+C^{k,l}(p,q)\\ -ED(p,k)-ED(q,l)\\ +ED(i,k)+ED(j,l)\end{pmatrix} & \text{if } S_i^1, S_j^2 \text{ can pair}\\ \infty & \text{otherwise.}\end{cases} \quad (2)$$

For the initial case we have:

$$C^{k,l}(k,l)=\begin{cases}ED(k,k)+ED(l,l) & \text{if } S_k^1, S_l^2 \text{ can pair}\\ \infty & \text{otherwise.}\end{cases} \quad (3)$$

Finally, we get a 2D matrix $C(i,j)$ that stores for all left-end base pairs (i,j) the best value found so far for all (k,l) with $i\le k$ and $j\le l$, i.e.

$$C(i,j)=\min_{k,l}\left\{C^{k,l}(i,j)\right\}. \quad (4)$$

The calculation of $C^{k,l}(i,j)$ for all (k,l) first and finding the minimal value $C(i,j)$ afterwards has still time and space complexity $O(n^2m^2)$.

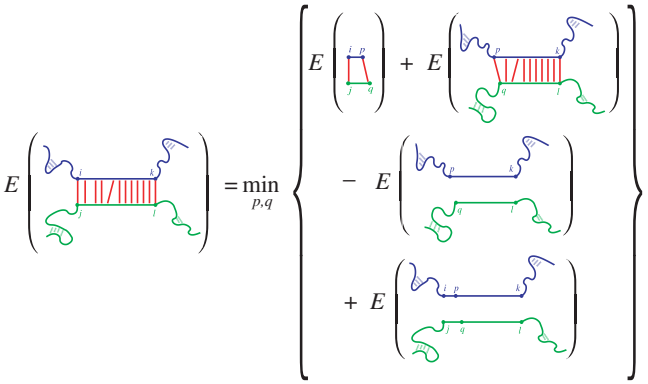

Fig. 2. Visualization of the recursion for $C^{k,l}(i,j)$. The hybridized part is shown in red, while the energy required to make the mRNA and the sRNA target site accessible (*ED*) is given in blue and green, respectively. Since *ED*-values are not additive, e.g. $ED(i,k)\neq ED(i,p)+ED(p,k)$, we need to substract $ED(p,k)$ and $ED(q,l)$, and add $ED(i,k)$ and $ED(j,l)$ to get the final result of $C^{k,l}(i,j)$.

Instead, we can update $C(i,j)$ successively after each evaluation of a right-end base pair (k,l) and reuse the matrix $C^{k,l}$ in order to reduce the space complexity. Thus

$$C(i,j)=\min\left\{C(i,j);\ C^{k,l}(i,j)\right\}. \quad (5)$$

This gives an $O(n^2m^2)$ time algorithm (applying the usual trick of restricted loops) with only $O(nm)$ space requirement.

2.2 Incorporation of seed features

According to the findings on seed regions presented in Section 1, we introduce *seed features* that define their properties:

- P: the number of bases perfectly paired in the seed region
- $b^{\max}$, $b_m^{\max}$ and $b_s^{\max}$: the maximal number of bases not hybridized in the seed region of both RNAs, the mRNA and the sRNA, respectively.

The seed features are a variable part of our algorithm and can be specified by the user. Although there is a preferred position for the seed (in the case of miRNA, it is the 5′-end), it has been shown that seeds could also be on other positions (Brennecke *et al.*, 2005). Hence, we require only the existence of a single seed sequence at any position. For this purpose, we introduce a function seed$(i,j,k,l;P)$, which stores the minimal free energy between $[i,k]$ and $[j,l]$ such that the interaction includes exactly P base pairs. If i, j, k, l and P are given, the numbers of unpaired bases in the mRNA and the sRNA are fixed to $k-i+1-P$ and $l-j+1-P$, respectively.

$$\text{seed}(i,j,k,l;P)=\begin{cases}\min\limits_{\substack{p,q \text{ with}\\ k-p+1\ge P-1\\ l-q+1\ge P-1}}\left\{\begin{matrix}E^{\text{loop}}(i,j,p,q)\\ +\text{seed}(p,q,k,l;P-1)\end{matrix}\right\} & P>2\\ E^{\text{loop}}(i,j,k,l) & P=2\\ \infty & \text{otherwise.}\end{cases} \quad (6)$$

The conditions $k-p+1\ge P-1$ and $l-q+1\ge P-1$ ensure that $P-1$ base pairs are possible in $[p,k]$ and $[q,l]$, respectively. Let $l_m=k-i+1$ and $l_s=l-j+1$ be the lengths of intervals $[i,k]$ and $[j,l]$. Then, seed$(i,j,k,l;P)$ is only valid if $l_m-P\le b_m^{\max}$, $l_s-P\le b_s^{\max}$ and $l_m+l_s-2P\le b^{\max}$. These three conditions assure compliance with the seed features.

While seed$(i,j,k,l;P)$ finds the minimal free energy for two fixed intervals $[i,k]$ and $[j,l]$, all valid intervals have to be analyzed to find the optimal seed region. This is done during the calculation of a second 2D array

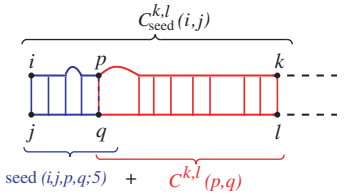

Fig. 3. $C^{k,l}_{\text{seed}}(i,j)$ matrix and its relation to the other matrices. Note that both $C^{k,l}(p,q)$ and seed$(i,j,p,q;5)$ consistently assume that (p,q) is a pair. Here, a seed of 5 bp and one unpaired base is shown.

$C^{k,l}_{\text{seed}}(i,j)$, which contains energy scores of interactions with a seed region. The interpretation of $C^{k,l}_{\text{seed}}(i,j)$ is given as in Figure 3. This leads to the following recursion:

$$C^{k,l}_{\text{seed}}(i,j)= \begin{cases} \min \begin{cases} \min\limits_{p,q} \begin{pmatrix} E^{\text{loop}}(i,j,p,q)+C^{k,l}_{\text{seed}}(p,q) \\ -ED(p,k)-ED(q,l) \\ +ED(i,k)+ED(j,l) \end{pmatrix} \\ \min\limits_{\substack{p,q \text{ with} \\ l_m \le b^{\max}_m+P \\ l_s \le b^{\max}_s+P \\ l_m+l_s \le b^{\max}+2P}} \begin{pmatrix} \text{seed}(i,j,p,q;P)+C^{k,l}(p,q) \\ -ED(p,k)-ED(q,l) \\ +ED(i,k)+ED(j,l) \end{pmatrix} \end{cases} & \text{if } S^1_i, S^2_j \text{ can pair} \\ \infty & \text{otherwise.} \end{cases} \quad (7)$$

where $l_m = p-i+1$ and $l_s = q-j+1$ now are the lengths of intervals $[i,p]$ and $[j,q]$, respectively. The first of the two inner minima addresses the case where a seed region was already found right of base pair (p,q). The second minimum refers to the case where a seed region was not found right of base pair (p,q), but between pairs (i,j) and (p,q).

Note that this extension of the algorithm does not increase its complexity. The final values are stored in $C(i,j)$ by replacing $C^{k,l}(i,j)$ with $C^{k,l}_{\text{seed}}(i,j)$ in Equation (5).

2.3 Reducing the space and time complexity

Although the IntaRNA algorithm presented above has a space complexity of $O(nm)$ like RNAhybrid, its time complexity of $O(n^2m^2)$ is still impractical for a genome-wide search. Hence, we want to achieve also the same time complexity as RNAhybrid. In the following, the idea is described for the case without seeds to simplify the presentation. However, seeds are integrated in the same way as described above.

Before introducing the version of IntaRNA with reduced space and time complexity, we summarize the full version of the algorithm to clarify the differences between them. For this purpose, Equations (2), (3) and (4) are combined into the following single equation:

$$C(i,j)= \begin{cases} \min \begin{cases} \min\limits_{p,q,k,l} \begin{pmatrix} E^{\text{loop}}(i,j,p,q)+C^{k,l}(p,q) \\ -ED(p,k)-ED(q,l) \\ +ED(i,k)+ED(j,l) \end{pmatrix} \\ \big(ED(i,i)+ED(j,j)\big) \end{cases} & \text{if } S^1_i, S^2_j \text{ can pair} \\ \infty & \text{otherwise.} \end{cases} \quad (8)$$

To reduce both time and space complexity, we use a heuristic simplification that is inspired by the sparsification technique. Here, the basic idea is that the matrix $C(i,j,k,l)$ is sparse in the sense that many entries will have the same values. This is due to the fact that many right hybridization ends will not be used in the next recursion steps. Our idea is to store the $C^{k,l}(i,j)$ values only for one starting point (k,l). Thus, we use a matrix $che(i,j)$ that stores for every (i,j) the right hybridization end (k,l), which yields the best extended hybridization energy until the left end (i,j) [see Equation (10)]. Given $che(i,j)$, we can directly use a 2D matrix $C'(i,j)$ and do not need $C^{k,l}(i,j)$ and $C(i,j)$ any longer. Denoting with $che_1(i,j)$ the first component k of the pair $(k,l)=che(i,j)$, and with $che_2(i,j)$ the second component l, we get the following recursion:

$$C'(i,j)= \begin{cases} \min \begin{cases} \min\limits_{p,q} \begin{pmatrix} E^{\text{loop}}(i,j,p,q)+C'(p,q) \\ -ED(p,che_1(p,q)) \\ -ED(q,che_2(p,q)) \\ +ED(i,che_1(p,q)) \\ +ED(j,che_2(p,q)) \end{pmatrix} & \text{(A)} \\ \big(ED(i,i)+ED(j,j)\big) & \text{(B)} \end{cases} & \text{if } S^1_i, S^2_j \text{ can pair} \\ \infty & \text{otherwise.} \end{cases} \quad (9)$$

After the calculation of $C'(i,j)$, the corresponding value in $che(i,j)$ has to be updated according to the following recursion:

$$che(i,j)= \begin{cases} che(p,q) & \text{if (A) is the minimum in Equation (9)} \\ (i,j) & \text{if (B) is the minimum in Equation (9)} \end{cases} \quad (10)$$

Equivalent simplifications are applied to the recursions that incorporate a seed region. The respective values are stored in $C'_{\text{seed}}(i,j)$. The final best hybridization score including a seed can be found by $\min_{i,j}\{C'_{\text{seed}}(i,j)\}$.

Compared with Equation (8), Equation (9) computes the minimum over only two instead of four variables. Since these two variables p and q are restricted by the maximal loop size, IntaRNA has a space complexity of $O(nm)$ and a time complexity of $O(nm+nL^3)=O(n\overline{m})$, where $\overline{m}=\max\{m,L^3\}$ and L is the size of the sequence window in which both mRNA and sRNA are folded. All *ED*-values are calculated in $O(nL^3)$ time by integrating RNAplfold into IntaRNA via the Vienna RNA library (Hofacker *et al.*, 1994).

Suboptimal hybridizations: IntaRNA can predict multiple potential target sites per sRNA. The computation of suboptimal hybridizations is implemented by multiple traceback. Since target sites at different locations within the mRNA are especially of interest, an interaction is accepted as suboptimal if it does not overlap with any other interaction predicted thus far. The desired number of suboptimal hybridizations are computed iteratively from the matrix C'_{seed}.

2.4 Functional analysis of sRNA target sites

In prokaryotes, the interaction between ribosome and mRNA is promoted by the Shine–Dalgarno (SD) sequence, a sequence motif typically 4–5 nt in length and located around 5–8 nt upstream of the start codon. The SD sequence is bound by a complementary motif of the 3′-tail of the 16S ribosomal RNA (rRNA). The translation is usually regulated by blocking access to this initiation site (Kozak, 2005).

The majority of bacterial sRNAs act as antisense regulators on *trans*-encoded mRNAs. Often, the base pairing occurs at the ribosome binding site (RBS), which leads to blockage of ribosome entry and thus to translation inhibition. In contrast, some sRNAs activate translation of their target mRNAs. Thereby, the sRNA binding results in melting of an inhibitory structure that sequesters the RBS (Vogel and Wagner, 2007).

Here, we study the consequences of sRNA binding to the target mRNA regarding the accessibility of the SD sequence, and thus the translational regulation. First, we predict the SD sequence location for every studied gene by simulating the hybridization between the mRNA and the single stranded 16S rRNA 3′-tail. The SD sequence is located by the position of the minimum free energy hybridization with a free energy below a significance threshold (Starmer *et al.*, 2006). Then, we calculate the probability that the SD sequence is unpaired before ($PU^{\text{nohybrid}}_{SD}$) and after ($PU^{\text{hybrid}}_{SD}$) the hybridization of the

Table 1. Prediction accuracy of IntaRNA compared with leading RNA–RNA interaction prediction methods on a set of experimentally verified interactions

		Sensitivity					PPV				
sRNA–target	Reference	IntaRNA	TargetRNA	RNAhybrid	RNAplex	RNAup	IntaRNA	TargetRNA	RNAhybrid	RNAplex	RNAup
DsrA-*rpoS*	Repoila *et al.* (2003)	0.808	0.808	0.000	0.808	0.808	0.778	0.778	0.000	0.778	0.778
GcvB-*argT*	Sharma *et al.* (2007)	0.950	1.000	1.000	0.000	0.900	0.950	0.625	0.160	0.000	0.947
GcvB-*dppA*	Sharma *et al.* (2007)	1.000	0.941	0.941	0.765	1.000	0.586	0.421	0.132	0.448	0.459
GcvB-*gltI*	Sharma *et al.* (2007)	0.000	–	0.875	1.000	0.000	0.000	–	0.210	0.857	0.000
GcvB-*livJ*	Sharma *et al.* (2007)	0.955	–	1.000	0.955	0.955	0.955	–	0.180	0.955	0.955
GcvB-*livK*	Sharma *et al.* (2007)	0.542	–	0.542	0.542	0.542	0.565	–	0.108	0.565	0.565
GcvB-*oppA*	Sharma *et al.* (2007)	1.000	1.000	1.000	1.000	1.000	0.957	0.957	0.200	0.957	0.957
GcvB-*STM4351*	Sharma *et al.* (2007)	0.760	0.000	0.000	0.000	0.880	0.905	0.000	0.000	0.000	0.957
IstR-*tisAB*	Vogel *et al.* (2004)	0.879	0.939	0.939	0.750	0.667	0.690	0.775	0.403	1.000	1.000
MicA-*ompA*	Udekwu *et al.* (2005)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.302	1.000	1.000
MicA-*lamB*	Bossi and Figueroa-Bossi (2007)	1.000	–	0.609	1.000	0.826	0.821	–	0.318	1.000	0.704
MicC-*ompC*	Chen *et al.* (2004)	1.000	0.636	1.000	0.000	0.727	0.537	0.286	0.333	0.000	0.410
MicF-*ompF*	Schmidt *et al.* (1995)	0.960	0.560	0.960	0.920	0.800	0.960	0.636	0.545	0.958	0.952
OxyS-*fhlA*	Argaman and Altuvia (2000)	0.500	–	0.938	0.563	0.375	1.000	–	0.288	0.750	1.000
RyhB-*sdhD*	Masse and Gottesman (2002)	0.588	0.882	0.794	0.824	0.794	1.000	0.909	0.403	1.000	0.794
RyhB-*sodB*	Geissmann and Touati (2004)	1.000	1.000	1.000	1.000	1.000	0.818	0.375	0.167	0.818	0.900
SgrS-*ptsG*	Kawamoto *et al.* (2006)	0.739	–	0.000	0.739	0.739	1.000	–	0.000	1.000	1.000
Spot42-*galK*	Møller *et al.* (2002)	0.409	0.545	0.523	0.432	0.523	0.643	0.558	0.280	0.655	0.523
Average		**0.783**	0.776	0.729	0.683	0.752	**0.787**	0.610	0.224	0.708	0.772

For every sRNA–target pair, sensitivity and PPV were calculated for the highest scoring interaction predicted. '-' means that no interaction was predicted. The best average result for each measure is highlighted in bold.

sRNA and the mRNA. The change in this probability, ΔPU_{SD}, is defined as

$$\Delta PU_{SD} = PU_{SD}^{\text{hybrid}} - PU_{SD}^{\text{nohybrid}} = e^{\frac{E^{\text{ens}}(\mathcal{S}_{k,l}^{\text{unpaired}}) - E^{\text{ens}}(\mathcal{S}_{i,j,k,l}^{\text{unpaired}})}{RT}} - e^{\frac{E^{\text{ens}}(\mathcal{S}) - E^{\text{ens}}(\mathcal{S}_{i,j}^{\text{unpaired}})}{RT}},$$

where the region between i and j is the location of the SD sequence, the region between k and l is the mRNA target site and $\mathcal{S}_{i,j,k,l}^{\text{unpaired}}$ is the set of all structures of the sequence S having $S_i \ldots S_j$ and $S_k \ldots S_l$ unpaired. $\Delta PU_{SD} > 0$ suggests translational activation by the sRNA–mRNA interaction, whereas $\Delta PU_{SD} < 0$ suggests translational repression. The higher the absolute value, the higher is the expected regulatory outcome. However, a special case arises if the mRNA target site overlaps with or is in close vicinity to the SD sequence. Then, the RBS is blocked and translation inactivation is expected. This measurement of single strandness has the advantage that it is based on base pair probabilities. Thus, it accounts for all possible secondary structures within the thermodynamic ensemble. The concept has been previously applied for searching binding motifs of RNA-binding proteins (Hiller *et al.*, 2006).

3 RESULTS

In order to assess the performance of the IntaRNA algorithm as presented in Section 2.3, we used the program to predict targets of bacterial regulatory sRNAs. The test set consisted of 10 biochemically mapped sRNA–mRNA interactions from *E.coli* and eight interactions from *Salmonella typhimurium* (denoted *Salmonella* in the following) that were previously published. For each sRNA, we predicted interactions for all genes of the respective genome. The genome sequences were downloaded from GenBank database of the National Center for Biotechnology Information (NCBI) (Benson *et al.*, 2008). Since the majority of the known sRNAs bind their target gene in close proximity to the RBS, we defined a subsequence of 150-nt upstream and 50-nt downstream of the first base of the start codon as the (putative) target region. We obtained 4294 target regions from the *E.coli* genome (GenBank accession number NC_000913) and 4425 target regions from the *Salmonella* genome (GenBank accession number NC_003197).

The seed features and other parameters used for the target prediction were chosen according to the known interactions from our test set. They included a seed of at least eight paired nucleotides length and no restriction on the interaction length. All interactions except OxyS-*fhlA* have a continuous hybridization pattern. Amongst these examples, the Spot42-*galK* interaction is the longest one with a length of 75 nt.

We compared the results with several state-of-the-art methods for the prediction of RNA–RNA interactions, namely TargetRNA, RNAhybrid, RNAplex and RNAup. Although RNAhybrid is primarily designed for the prediction of miRNA target sites, it has been used occasionally for prediction related to sRNAs (see, e.g. Sharma *et al.*, 2007; Urban and Vogel, 2008). Therefore, it has been included in our comparison for the sake of completeness using the default parameters. For TargetRNA, we used the web application (Tjaden, 2008) with default parameters, except that the search was focused on our target regions. RNAplex was used with a penalty of 0.3 kcal/mol per nucleotide as suggested by Tafer and Hofacker (2008). We used RNAup including the probability of unpaired regions in both RNAs (parameter -b) (Mückstein *et al.*, 2008) and set the maximal length of interaction to 80, which is slightly longer than the maximal interaction length in our dataset.

3.1 Accuracy of predicted sRNA–mRNA interactions

In a first experiment, we assessed whether IntaRNA is able to predict precisely the interaction between each sRNA and its mRNA target. Therefore, we computed the sensitivity and the positive predictive value (PPV) for each sRNA–target pair, where $\text{sensitivity} = \frac{\text{number of correctly predicted base pairings}}{\text{number of true base pairings}}$ and $\text{PPV} = \frac{\text{number of correctly predicted base pairings}}{\text{number of predicted base pairings}}$. These measures have been used in the past to compare different RNA secondary structure prediction methods (see e.g. Do *et al.*, 2006).

As shown in Table 1, IntaRNA outperforms existing methods in the accuracy of the predicted interactions. TargetRNA achieves

the second best sensitivity and the third best PPV, but reports only 12 out of 18 interactions due to its cutoff. The program RNAhybrid tends to maximize the length of hybridization, which leads to high sensitivity, but very low PPV. Thus, the program is more appropriate to predict interactions between short RNAs (like miRNAs) and long RNAs. To overcome this problem, RNAplex introduced a length penalty, which significantly increased its PPV compared with RNAhybrid. RNAup achieves the third best sensitivity and the second best PPV. Among all programs compared, it has an overall accuracy closest to IntaRNA. IntaRNA and RNAup are the only programs whose optimal solution locates every sRNA binding site correctly, except for the interaction GcvB-*gltI*. For this interaction, both IntaRNA and RNAup predict optimal hybridizations, which do not share a base pair with the experimentally verified location. However, IntaRNA, RNAhybrid and RNAplex also give suboptimal solutions, see Supplementary Table 1. When these predictions are additionally taken into account, the average sensitivity/PPV of RNAhybrid and RNAplex improve to 0.774/0.251 and 0.736/0.761, respectively. IntaRNA achieves, on average, both a sensitivity and a PPV greater than 0.8 when the first suboptimal prediction for GcvB-*gltI* is included.

To study the influence of seed features on IntaRNA's prediction quality, we repeated the experiments neglecting them (Supplementary Table 2). In this case, the averaged values of sensitivity and PPV are 0.699 and 0.728, respectively, which is below the accuracy of IntaRNA with seed features and RNAup. The difference to the latter, which uses a similar energy model, can be explained by the heuristic of IntaRNA.

Altogether, the results demonstrate that RNAup and IntaRNA, which both incorporate the accessibility of binding sites, perform better in the prediction of sRNA–mRNA interactions than the other programs neglecting the accessibility. Furthermore, the quality of IntaRNA's predictions is substantially improved when seed features are additionally taken into account.

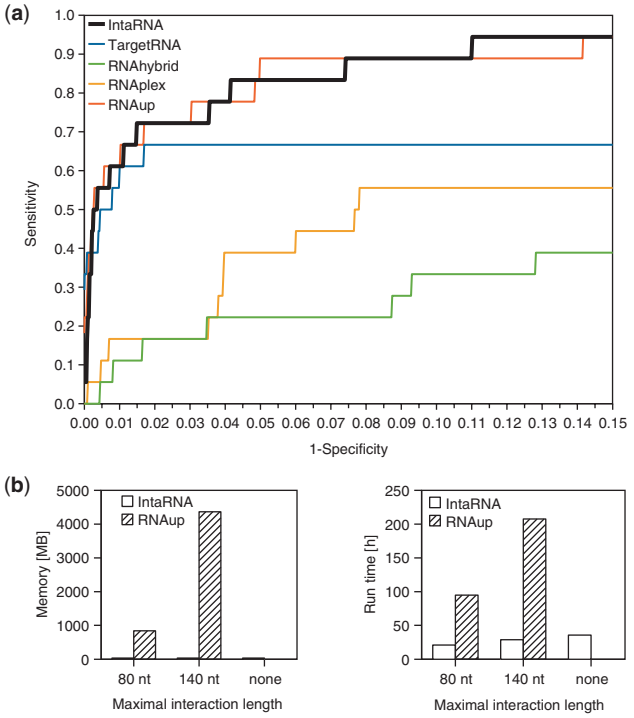

Fig. 4. Performance of IntaRNA. (**a**) Comparison of IntaRNA and other leading methods in the prediction of targets on our test set of 18 sRNAs with experimentally verified targets. The sensitivity (true positive rate) is shown as a function of the false positive rate (1 − specificity). For each prediction method, the target candidates for each sRNA were sorted by energy score. Each ROC curve was generated from the rate of true and false predictions, while varying the number of considered interactions per sRNA. (**b**) Comparison of resource requirements of IntaRNA (including computation of *ED*-values) and RNAup for a GcvB target search in *Salmonella*. Without restricting the interaction length, RNAup uses up the available complete memory and, as a consequence, crashes.

3.2 Performance on prediction of sRNA targets

In a second experiment, we compared IntaRNA and the existing methods with respect to the ability of finding sRNA targets. We applied every program to our test set and for each sRNA searched all target regions for potential target sites. The resulting list of target candidates for each sRNA was sorted by the computed energy score. All programs except TargetRNA and IntaRNA give an interaction for each putative target region. TargetRNA reports at most 100 putative interaction sites per sRNA. IntaRNA returns interactions that have both a seed with specified features and an energy score below 0.0 kcal/mol. For each method, we calculated the sensitivity and specificity. For our test set, there are 18 true interactions. Each of the nine sRNAs in *E.coli* may interact with any of the 4294 target regions, and each sRNA in *Salmonella* may interact with any of the 4425 target regions. Consequently, there are 47 496 potential interactions, of which 47 478 are considered non-interactions. A similar approach to evaluate the performance on prediction of sRNA targets has been used by Tjaden *et al.* (2006).

The ROC curves in Figure 4a illustrate the performance of different target prediction methods on our test set. We generated each ROC curve by calculating sensitivity and specificity while varying the number of computed interactions that were taken into account for each sRNA. The plot shows that IntaRNA and RNAup are the methods performing best on prediction of sRNA targets. Both RNAhybrid and RNAplex achieve a low sensitivity suggesting that these programs are suitable only to a limited degree for genome-wide sRNA target searches. Taking in consideration that TargetRNA limits the number of reported putative interaction sites, it achieves a fairly high sensitivity at a low false positive rate, although only an alignment-like algorithm based on base pairing potential is used. However, it can be assumed that the program will perform worse on interactions that show lower sequence complementarity, but underlie more complex duplex formation rules. The curves show that IntaRNA and RNAup have a similar performance on predicting sRNA targets and perform best among all studied programs. However, there is a clear difference in the practical applicability of both programs (Fig. 4b). On an Intel Xeon 5160 (3.0 GHz) with 7.8 GB available RAM, a GcvB target search in all *Salmonella* target regions allowing a maximal interaction length of 80 nt takes 21 h and requires 33 MB RAM with IntaRNA. The same search with RNAup needs 95 h and 840 MB RAM. An increase of the maximal interaction length to 140 nt raises IntaRNA's runtime to 29 h with unchanged memory usage, whereas RNAup now requires 207 h and 4.3 GB RAM. Without a restriction on the interaction length, IntaRNA takes 36 h and requires again 33 MB RAM. Since

RNAUP requires a restriction, we limited the interaction size to the length of the sRNA. This causes exhaustion of the complete memory and, as a consequence, a crash of RNAUP. The dramatic increase of RNAUP's resource requirements results from its higher asymptotic complexity and impairs its applicability on normal work stations with limited available memory.

Furthermore, it should be noted that the calculation of sensitivity and specificity is conservative, since we rely only on biochemically mapped interactions from literature. For instance, there are several unpublished experimentally verified targets among the top-ranked predictions of INTARNA (Jörg Vogel lab, personal communication) that have not been taken into consideration when evaluating the performance.

3.3 Prediction of the type of regulation

For all interactions of our test set, we also analyzed the regulatory outcome of the sRNA binding to its target mRNA. Many sRNAs regulate the translation of their target by changing the accessibility of the SD sequence where translation initiation occurs. Therefore, we studied the change ΔPU_{SD} in the probability that the SD sequence is unpaired as a consequence of sRNA–target interaction. We predicted SD sequence locations within a region of 35-nt up- and downstream of the first base of the start codon for each gene following the approach of Starmer *et al.* (2006). Then, we determined whether the sRNA binding site predicted by INTARNA is at or close to the SD sequence. In this case, the SD sequence is inaccessible for ribosome binding. Otherwise, we calculated ΔPU_{SD} for the gene. Supplementary Table 3 shows the results for all interactions of our test set.

In 11 out of 18 examples, our method successfully predicted the type of translational regulation by the sRNA. For three of the remaining seven interactions, the SD sequence could either not be located (GcvB-*STM4351*) or was located at an incorrect position (MicF-*ompF* and OxyS-*fhlA*). Another sRNA, IstR, blocks translation of its mRNA target *tisAB* by binding 100-nt upstream of the start codon without inducing structural changes at the RBS. Instead, IstR blocks a ribosome standby site that is essential for translation initiation (Unoson and Wagner, 2007). The remaining three interactions all involve the sRNA GcvB. Its targets *argT*, *livJ* and *gltI* are bound upstream of the ribosome binding site, and the inhibitory activity cannot be directly explained by competition with ribosome binding. At least for the last example, translational repression by a simple interference model or by masking a ribosome standby site is unlikely (Sharma *et al.*, 2007). Consequently, the regulation cannot be predicted by our model.

4 CONCLUSIONS

Although numerous regulatory ncRNAs have already been identified, the number of experimentally verified targets is much smaller. Consequently, computational target prediction is in great demand to restrain the list of putative targets.

In this article, we presented INTARNA, a new method for the prediction of interactions between two RNAs based on minimization of an extended hybridization energy. Our algorithm accounts for two important features that influence the strength of RNA–RNA interactions: the accessibility of the interaction sites and the existence of a user-specified seed. In contrast to previous methods for the prediction of RNA–RNA interactions, both features are integrated in a general approach for arbitrary RNAs. Although INTARNA was applied to predict targets for bacterial sRNAs in this work, the program can readily be used to find other RNA–RNA interactions as well.

The INTARNA target predictions for bacterial sRNA were compared to results of several state-of-the-art methods for the prediction of RNA–RNA interactions. INTARNA outperforms existing methods in the accuracy of the predicted interaction, and our method performs as well as the best existing program on finding putative sRNA targets, while the required CPU time and memory decrease drastically. Overall, the results show that our method is well suited both for general searches for putative target sites and the prediction of accurate RNA–RNA interactions. The comparison with RNAHYBRID, whose hybridization energy model is the basis of our more sophisticated extended hybridization energy, shows how the incorporation of the free energy required for making both interaction sites single stranded and the existence of a seed can improve the prediction quality. In addition, we were also able to successfully predict the regulatory effect on translation initiation for a number of sRNAs.

The interaction between the OxyS sRNA and the *fhlA* mRNA is the only one in our test set with a discontinuous hybridization pattern. In fact, the two RNAs form kissing hairpins at two sites (Argaman and Altuvia, 2000). We are aware of two approaches that can be used to predict interactions consisting of such independent substructures (Aksay *et al.*, 2007; Alkan *et al.*, 2006; Pervouchine, 2004). However, these algorithms are rather expensive with a time complexity of $O(n^3m^3)$. Further extensions of INTARNA could incorporate some basic ideas of those approaches to allow for prediction of more complex interactions with multiple sites.

Many bacterial sRNA genes and their mRNA interaction sites are conserved in closely related bacteria (see e.g. Delihas, 2003; Udekwu *et al.*, 2005). It can be assumed that the sRNA–target interaction mechanisms are conserved as well. Therefore, the performance on finding sRNA targets may be further improved by restricting the prediction to orthologous genes. A scoring scheme that accounts for interactions conserved between related species may be a promising extension of our combined energy score.

ACKNOWLEDGMENTS

We thank Sven Siebert for his initial work and fruitful discussions, Cynthia M. Sharma and Jörg Vogel for assistance with the biological data and Michael Beckstette for discussions on statistical analysis. We also thank the anonymous referees for their helpful comments.

Funding: German Federal Ministry of Education and Research (BMBF grant 0313921 FRISYS); German Research Foundation (DFG grant BA 2168/2-1 SPP 1258).

Conflict of Interest: none declared.

REFERENCES

Aksay,C. *et al*. (2007) taveRNA: a web suite for RNA algorithms and applications. *Nucleic Acids Res.*, **35**, W325–W329.

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

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.

Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Benson,D.A. *et al*. (2008) GenBank. *Nucleic Acids Res.*, **36**, D25–D30.

Bentwich,I. (2005) Prediction and validation of microRNAs and their targets. *FEBS Lett.*, **579**, 5904–5910.

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

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

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

Bossi,L. and Figueroa-Bossi,N. (2007) A small RNA downregulates LamB maltoporin in Salmonella. *Mol. Microbiol.*, **65**, 799–810.

Brennecke,J. *et al*. (2005) Principles of microRNA-target recognition. *PLoS Biol.*, **3**, e85.

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

Delihas,N. (2003) Annotation and evolutionary relationships of a small regulatory RNA gene micF and its target ompF in Yersinia species. *BMC Microbiol.*, **3**, 13.

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

Do,C.B. *et al*. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**, e90–e98.

Doench,J.G. and Sharp,P.A. (2004) Specificity of microRNA target selection in translational repression. *Genes Dev.*, **18**, 504–511.

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

Gottesman,S. (2005) Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends Genet.*, **21**, 399–404.

Hannon,G.J. (2002) RNA interference. *Nature*, **418**, 244–251.

Hiller,M. *et al*. (2006) Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Res.*, **34**, e117.

Hofacker,I.L. *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.

Koberle,C. *et al*. (2006) Selecting effective siRNAs based on guide RNA structure. *Nat. Protoc.*, **1**, 1832–1839.

Kozak,M. (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene*, **361**, 13–37.

Kretschmer-Kazemi Far,R. and Sczakiel,G. (2003) The activity of siRNA in mammalian cells is related to structural target accessibility: a comparison with antisense oligonucleotides. *Nucleic Acids Res.*, **31**, 4417–4424.

Kruger,J. and Rehmsmeier,M. (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res.*, **34**, W451–W454.

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

Luo,K.Q. and Chang,D.C. (2004) The gene-silencing efficiency of siRNA is strongly dependent on the local structure of mRNA at the targeted region. *Biochem. Biophys. Res. Commun.*, **318**, 303–310.

Markham,N.R. and Zuker,M. (2005) DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res.*, **33**, W577–W581.

Masse,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.

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

McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.

Mika,F. *et al*. (2008) Seed pairing and non-RBS target sites facilitate global *omp* mRNA regulation by a bacterial small RNA. (inpress).

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

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 of *Communications in Computer and Information Science*. Springer, Berlin, Heidelberg, pp. 114–127.

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

Rajewsky,N. (2006) microRNA target predictions in animals. *Nat. Genet.*, (Suppl. 38), S8–S13.

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

Repoila,F. *et al*. (2003) Small non-coding RNAs, co-ordinators of adaptation processes in Escherichia coli: the RpoS paradigm. *Mol. Microbiol.*, **48**, 855–861.

Schmidt,M. *et al*. (1995) Secondary structures of Escherichia coli antisense micF RNA, the 5′-end of the target ompF mRNA, and the RNA/RNA duplex. *Biochemistry*, **34**, 3621–3631.

Shao,Y. *et al*. (2007) Effect of target secondary structure on RNAi efficiency. *RNA*, **13**, 1631–1640.

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.

Starmer,J. *et al*. (2006) Predicting Shine-Dalgarno sequence locations exposes genome annotation errors. *PLoS Comput. Biol.*, **2**, e57.

Storz,G. (2002) An expanding universe of noncoding RNAs. *Science*, **296**, 1260–1263.

Storz,G. and Haas,D. (2007) A guide to small RNAs in microorganisms. *Curr. Opin. Microbiol.*, **10**, 93–95.

Storz,G. *et al*. (2004) Controlling mRNA stability and translation with small, noncoding RNAs. *Curr. Opin. Microbiol.*, **7**, 140–144.

Tafer,H. and Hofacker,I.L. (2008) RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics*. [Epub ahead of print; doi:10.1093/bioinformatics/btn193; April 23, 2008].

Tjaden,B. (2008) TargetRNA: a tool for predicting targets of small RNA action in bacteria. *Nucleic Acids Res.*, **36**, W109–W113.

Tjaden,B. *et al*. (2006) Target prediction for small, noncoding RNAs in bacteria. *Nucleic Acids Res.*, **34**, 2791–2802.

Udekwu,K.I. *et al*. (2005) Hfq-dependent regulation of OmpA synthesis is mediated by an antisense RNA. *Genes Dev.*, **19**, 2355–2366.

Unoson,C. and Wagner,E.G. (2007) Dealing with stable structures at ribosome binding sites: bacterial translation and ribosome standby. *RNA Biol.*, **4**, 113–117.

Urban,J.H. and Vogel,J. (2008) Two seemingly homologous noncoding RNAs act hierarchically to activate glmS mRNA translation. *PLoS Biol.*, **6**, e64.

Vogel,J. and Wagner,E.G.H. (2007) Target identification of small noncoding RNAs in bacteria. *Curr. Opin. Microbiol.*, **10**, 262–270.

Vogel,J. *et al*. (2004) The small RNA IstR inhibits synthesis of an SOS-induced toxic peptide. *Curr. Biol.*, **14**, 2271–2276.

Yoon,S. and De Micheli,G. (2006) Computational identification of microRNAs and their targets. *Birth Defects Res. C Embryo Today*, **78**, 118–128.

Zamore,P.D. and Haley,B. (2005) Ribo-gnome: the big world of small RNAs. *Science*, **309**, 1519–1524.

Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.