Abstract

Background: RNA has been recognized as a key player in cellular regulation in recent years. In many cases, non-coding RNAs exert their function by binding to other nucleic acids, as in the case of microRNAs and snoRNAs. The specificity of these interactions derives from the stability of inter-molecular base pairing. The accurate computational treatment of RNA-RNA binding therefore lies at the heart of target prediction algorithms.

Methods: The standard dynamic programming algorithms for computing secondary structures of linear single-stranded RNA molecules are extended to the co-folding of two interacting RNAs.

Results: We present a program, RNAcofold, that computes the hybridization energy and base pairing pattern of a pair of interacting RNA molecules. In contrast to earlier approaches, complex internal structures in both RNAs are fully taken into account. RNAcofold supports the calculation of the minimum energy structure and of a complete set of suboptimal structures in an energy band above the ground state. Furthermore, it provides an extension of McCaskill’s partition function algorithm to compute base pairing probabilities, realistic interaction energies, and equilibrium concentrations of duplex structures.

Availability: RNAcofold is distributed as part of the Vienna RNA Package, http://www.tbi.univie.ac.at/RNA/.

Contact: Stephan H. Bernhart – berni@tbi.univie.ac.at

Background

Over the last decade, our picture of RNA as a mere information carrier has changed dramatically. Since the discovery of microRNAs and siRNAs (see e.g. [1,2] for a recent reviews), small noncoding RNAs have been recognized as key regulators in gene expression. Both computational surveys, e.g. [3-7] and experimental data [8-11] now provide compelling evidence that non-protein-coding transcripts are a common phenomenon. Indeed, at least in higher eukaryotes, the complexity of the non-coding RNome appears to be comparable with the complexity of the proteome. This extensive inventory of non-coding
RNAs has been implicated in diverse mechanisms of gene regulation, see e.g. [12-16] for reviews.

Regulatory RNAs more often than not function by means of direct RNA-RNA binding. The specificity of these interactions is a direct consequence of complementary base pairing, allowing the same basic mechanisms to be used with very high specificity in large collections of target and effector RNAs. This mechanism underlies the post-transcriptional gene silencing pathways of microRNAs and siRNAs (reviewed e.g. in [17]), it is crucial for snoRNA-directed RNA editing [18], and it is used in the gRNA directed mRNA editing in kinetoplastids [19]. Furthermore, RNA-RNA interactions determine the specificity of directed mRNA editing in kinetoplastids [19]. Further-

The energy model for RNA secondary structures and recall algorithms for molecular biology, see e.g. [12-16] for reviews.

As pointed out by Dimitrov and Zuker [27], the concentration of the two interacting RNAs as well as the possibility to form homo-dimers plays an important role and cannot be neglected when quantitative predictions on RNA-RNA binding are required. In our implementation of RNAcofold we therefore follow their approach and explicitly compute the concentration dependencies of the equilibrium ensemble in a mixture of two partially hybridizing RNA species.

This contribution is organized as follows: We first review the energy model for RNA secondary structures and recall the minimum energy folding algorithm for simple linear RNA molecules. Then we discuss the modifications that are necessary to treat intermolecular base pairs in the partition function setting and describe the computation of base pairing probabilities. Then the equations for concentration dependencies are derived. Short sections summarize implementation, performance, as well as an application to real-world data.

**RNA secondary structures**

A secondary structure $S$ on a sequence $x$ of length $n$ is a set of base pairs $(i, j)$, $i < j$, such that

1. $0 (i, j) \in S$ implies that $(x_i, x_j)$ is either a Watson-Crick (GC or AU) or a wobble (GU) base pair.

2. Every sequence position $i$ that takes part in at most one base pair, i.e., $S$ is a matching in the graph of "legal" base pairs that can be formed within sequence $x$.

3. $(i, j) \in S$ implies $|i - j| \geq 4$, i.e., hairpin loops have at least three unpaired positions inside their closing pair.

4. If $(i, j) \in S$ and $(k, l) \in S$ with $i < k$, then either $i < j < k < l$ or $i < k < l < j$. This condition rules out knots and pseudoknots. Together with condition 1 it implies that $S$ is a circular matching [32,33].

The "loops" of $S$ are planar faces of the unique planar embedding of the secondary structure graph (whose edges are the base pairs in $S$ together with the backbone edges $(i, i+1)$, $i = 1 \ldots, n-1$). Equivalently, the loops are the elements of the unique minimum cycle basis of the secondary structure graph [34]. The external loop consists of all those nucleotides that are not enclosed by a base pair in $S$.
The standard energy model for RNA secondary structures associates an energy contribution to each loop \( L \) that depends on the loop type \( \text{type}(L) \) (hairpin loop, interior loop, bulge, stacked pair, or multi-branch loop) and the sequence of some or all of the nucleotides in the loop, \( x_{|L|} \):

\[
\alpha(L) = \alpha(\text{type}(L), x_{|L|}). \quad (1)
\]

The external loop does not contribute to the folding energy. The total energy of folding sequence \( x \) into a secondary structure \( S \) is then the sum over all loops of \( S \). Energy parameters are available for both RNA [35] and single stranded DNA [36].

Hairpin loops are uniquely determined by their closing pair \((i, j)\). The energy of a hairpin loop is tabulated in the form

\[
\mathcal{H}(i, j) = \mathcal{H}(x_i, x_{i+1}, \ell, x_{j-1}, x_j) \quad (2)
\]

where \( \ell \) is the length of the loop (expressed as the number of its unpaired nucleotides). Each interior loop is determined by the two base pairs enclosing it. Its energy is tabulated as

\[
I(i, j; k, l) = I(x_i, x_{i+1}, \ell_1, x_j, x_{j-1}, x_j) \quad (3)
\]

where \( \ell_1 \) is the length of the unpaired strand between \( i \) and \( k \) and \( \ell_2 \) is the length of the unpaired strand between \( l \) and \( j \). Symmetry of the energy model dictates \( I(i, j; k, l) = I(l, k; j, i) \) if \( \ell_1 = \ell_2 = 0 \) we have a (stabilizing) stacked pair, if only one of \( \ell_1 \) and \( \ell_2 \) vanish we have a bulge. For multiloops, finally we have an additive energy model of the form \( \mathcal{M} = a + b \times \beta + c \times \ell \) where \( \ell \) is the length of multiloop (again expressed as the number of unpaired nucleotides) and \( \beta \) is the number of branches, not counting the branch in which the closing pair of the loop resides.

So-called dangling end contributions arise from the stacking of unpaired bases to an adjacent base pair. We have to distinguish two types of dangling ends: (1) interior dangles, where the unpaired base \( i + 1 \) stacks onto \( i \) of the adjacent basepair \((i, j)\) and correspondingly \( j - 1 \) stacks onto \( j \) and (2) exterior dangles, where \( i - 1 \) stack onto \( i \) and \( j + 1 \) stacks on \( j \). The corresponding energy contributions are denoted by \( d_{i,j}^E \) and \( d_{i,j}^E \), respectively. Within the additive energy model, dangling end terms are interpreted as the contribution of 3' and 5' dangling nucleotides.

\[
d_{i,j}^E = d^S(k-1 \mid k, l) + d^S(k, l \mid l + 1)
\]

Here \( k \) separates the dangling nucleotide position from the adjacent base pair, \( d^S(k-1 \mid k, l) \) thus is the energy of the nucleotide at position \( k - 1 \) when interacting with following base pair \((k, l)\), while \( d^S(k, l \mid l + 1) \) scores the interaction of position \( l + 1 \) with the preceding pair \((k, l)\).

The Vienna RNA Package currently implements three different models for handling the dangling-end contributions: They can be (a) ignored, (b) taken into account for every combination of adjacent bases and base pairs, or (c) a more complex model can be used in which the unpaired base can stack with at most one base pair. In cases (a) and (b) one can absorb the dangling end contributions in the loop energies (with the exception of contributions in the external loop). Model (c) strictly speaking violates the secondary structure model in that an unpaired base \( x_j \) between two base pairs \((x_p, x_{i+1})\) and \((x_{i+1}, x_p)\) has three distinct states with different energies: \( x_i \) does not stack to its neighbors, \( x_i \) stacks to \( x_{i+1} \), or \( x_{i+1} \). The algorithm then minimizes over these possibilities. While model (c) is the default for computing minimum free energy structures in most implementations such as RNAfold and mfold, it is not tractable in a partition function approach in a consistent way unless different positions of the dangling ends are explicitly treated as different configurations.

### RNA secondary structure prediction

Because of the no-(pseudo)knot condition 3 above, every base pair \((i, j)\) subdivides a secondary structure into an interior and an exterior structure that do not interact with each other. This observation is the starting point of all dynamic programming approaches to RNA folding, see e.g. [32,33,37]. Including various classes of pseudoknots is feasible in dynamic programming approaches [38-40] at the expense of a dramatic increase in computational costs, which precludes the application of these approaches to large molecules such as most mRNAs.

In the course of the "normal" RNA folding algorithm for linear RNA molecules as implemented in the Vienna RNA Package [41,42], and in a similar way in Michael Zuker's mfold package [43-45] the following arrays are computed for \( i < j \):

\[
F_{ij} \text{ free energy of the optimal substructure on the subsequence } x[i, j];
\]

\[
C_{ij} \text{ free energy of the optimal substructure on the subsequence } x[i, j] \text{ subject to the constraint that } i \text{ and } j \text{ form a basepair.}
\]
The adaptation of the recursion to the folding of two RNAs $A$ and $B$ of length $n_1$ and $n_2$ into a dimeric structure is straightforward: the two molecules are concatenated to form a single sequence of length $n = n_1 + n_2$. It follows from the algorithmic considerations below that the order of the two parts is arbitrary.

A basic limitation of this approach arises from the no-pseudoknots condition: It restricts not only the intramolecular base pairs but also affects intermolecular pairs. Let $S^A$ and $S^B$ denote the intramolecular pairs in a cofolded structure $S$. These sets of base pairs define secondary structures on $A$ and $B$ respectively. Because of the no-pseudoknot condition on $S$, an intermolecular base pair in $S\setminus(S^A \cup S^B)$ can only connect nucleotides in the external loops of $A$ and $B$. This is a serious restriction for some applications, because it excludes among other pseudoknot-like structures also the so-called kissing hairpin complexes [46]. Taking such structures into account is equivalent to employing folding algorithms for structure models that include certain types of pseudoknots, such as the partition function approach by Dirks and Pierce [40]. Its high computational cost, however, precludes the analysis of large mRNAs. In an alternative model [29], no intramolecular interactions are allowed in the small partner $B$, thus allowing $B$ to form basepairs with all contiguous unpaired regions in $S^A$. From a biophysical point of view, however, it makes sense to consider exclusively hybridization in the exterior loop provided both partners are large structured RNAs. In this case, hybridization either stops early, i.e., at a kissing hairpin complex (in the case of very stable local structures) or it is thermodynamically controlled and runs into the ground state via a complete melting of the local structure. In the latter case, the no-pseudoknots condition is the same approximation that is also made when folding individual molecules. Note that this approximation does not imply that the process of hybridization could only start at external bases.
Let us now consider the algorithmic details of folding two concatenated RNA sequences. The missing backbone edge between the last nucleotide of the first molecule, position \( n_1 \) in the concatenated sequence, and the first nucleotide of the second molecule (now numbered \( n_1+1 \)) will be referred to as the cut \( c \). In each dimeric structure there is a unique loop \( L_c \) that contains the cut \( c \). If \( c \) lies in the external loop of a structure \( S \) then the two molecules \( A \) and \( B \) do not interact in this structure. Algorithmically, \( L_c \) is either a hairpin loop, interior loop, or multibranch loop. From an energetic point of view, however, \( L_c \) is an exterior loop, i.e., it does not contribute to the folding energy (relative to the random coil reference state). For example, an interior loop \( I \) (i, j, k, l) does not contribute to the energy if either \( i \leq n_1 < k \) or \( l \leq n_1 < j \). Naturally, dangling end contributions must not span the cut, either. Hairpin loops and interior loops (including the special cases of bulges and stacked pairs) can therefore be dealt with by a simple modification of the energy rules. In the case of the multiloop there is also no problem as long as one is only interested in energy minimization, since multiloops are always destabilizing and hence have strictly positive energy contribution. Such a modified MFE algorithm has been described already in [41].

For partition function calculations and the generation of suboptimal structures, however, we have to ensure that every secondary structure is counted exactly once. This requires one to explicitly keep track of loops that contain the cut \( c \). The cut \( c \) needs to be taken into account explicitly only in the recursion for the \( Z^P \) terms, where one has to distinguish between true hairpin and interior loops with closing pair (i, j) (upper alternatives in eq. (6)) and loops containing the cut \( c \) in their backbone (lower alternatives in eq. (6)). Explicitly, this means \( i \leq n_1 < j \) in the hairpin loop case, in the interior loop case, this either means \( i \leq n_1 < k \) or \( l \leq n_1 < j \). Since multiloops are decomposed into two components, it is sufficient to ensure during the construction of \( Z^M_A \) and \( Z^N_A \) that these components neither start nor end adjacent to the cut, see Fig. 1.

In their full form including dangling end terms, the forward recursions for the partition function of an interacting pair of RNAs become

\[
Z_q = Z_{i+1,j} + \sum_{i<k<j} Z^P_{ij} Z^M_{k+1,j} Z^N_{i+1,j}
\]

\[
Z^P_{ij} = \begin{cases} 
\mathcal{H}(i,j) & \text{if } i = n_1 + 1 \wedge j = n_1 \\
0 & \text{otherwise}
\end{cases}
\]

\[
Z^M_{ij} = \begin{cases} 
Z^M_{i+1,j} + Z^M_{i,j} + \sum_{i<k<j} Z^P_{ik} Z^M_{kj} Z^N_{k+1,j} & \text{if } i = n_1 + 1 \vee j = n_1 \\
0 & \text{otherwise}
\end{cases}
\]

Upper alternatives refer to regular loops, lower alternatives to the loop containing the cutpoint. For brevity we have used here abbreviations \( \mathcal{H} (i,j) = \exp(-\mathcal{H}(i,j)/RT) \), and equivalently \( I, d, a, b, c \), for the Boltzmann factors of the energy contributions. In the remainder of this presentation we will again suppress the dangling end terms for simplicity of presentation.

A second complication arises from the initiation energy \( \Phi_i \) that describes the entropy necessary to bring the two molecules into contact. This term, which is considered to be independent of sequence length and composition [47], has to be taken into account exactly once for every dimer structure if and only if the structure contains at least one base pair (i, j) that crosses the cut, i.e., \( i \leq n_1 < j \). The resulting bookkeeping problems fortunately can be avoided by introducing this term only after the dynamic programming tables have been filled. To this end we observe that

\[
Z_{i,j} = Z^M_{i,j}, \quad 1 \leq i, j \leq n_1
\]

\[
Z^N_{n_1+i,n_1+j} = Z^B_{i,j}, \quad 1 \leq i, j \leq n_2
\]

which are the corresponding quantities for the second interaction partner. Thus we can immediately compute the partition function \( Z^{AB} - Z^A Z^B \) that counts only the structures with intermolecular pairs, i.e., those that carry the additional initiation energy contribution. The total partition function including the initiation term is therefore

\[
Z^* = [Z_{n_1+i,n_1+j} - Z_{n_1+i,n_1+j} e^{-\Phi_i/RT}]
+ Z_{n_1+i,n_1+j}
\]

\[
(7)
\]
Base pairing probabilities
McCaskill’s algorithm [31] computes the base pairing probabilities from the partition functions of subsequences. Again, it seems easier to first perform the backtracking recursions on the "raw" partition functions that do not take into account the initiation contribution. This yields pairing probabilities $P_{kl}$ for an ensemble of structures that does not distinguish between true dimers and isolated structures for $A$ and $B$ and ignores the initiation energy. McCaskill’s backwards recursions are formally almost identical to the case of folding a single linear sequence. We only have to exclude multiloop contributions with the cut point $c$, which are computed without the initiation term, can now be corrected for this effect. To this end, we separately run the backward recursion starting from $Z_{1,n}$ and from $Z_{n+1,n_1+n_2}$ to obtain the base pairing probability matrices $P^A_{ij}$ and $P^B_{i',j'}$ for the isolated molecules. Note that equivalently we could compute $P^A_{ij}$ and $P^B_{i',j'}$ directly using the partition function version of RNAfold.

In solution, the probability of an intermolecular base pair is proportional to the (concentration dependent) probability that a dimer is formed at all. Thus, it makes sense to consider the conditional pair probabilities given that a dimer is formed, or not. The fraction of structures without intermolecular pairs in our partition function $Z$ (i.e. in the cofold model without initiation contributions) is $Z^A Z^B / Z$, and hence the fraction of true dimers is

$$p^* = 1 - \frac{Z^A Z^B}{Z}. \quad (9)$$

Now consider a base pair $(i, j)$. If $i \in A$ and $j \in B$, it must arise from the dimeric state. If $i, j \in A$ or $i, j \in B$, however, it arises from the dimeric state with probability $p^*$ and from the monomeric state with probability $1 - p^*$. Thus the conditional pairing probabilities in the dimeric complexes can be computed as

$$P_{ij} = \begin{cases} 
\frac{P_{ij} - (1 - p^*)P_{ij}^A}{p^*} & \text{if } i, j \in A \\
\frac{P_{ij} - (1 - p^*)P_{ij}^B}{p^*} & \text{if } i, j \in B \\
P_{ij} & \text{otherwise}
\end{cases} \quad (10)$$

The fraction of monomeric and dimeric structures, however, cannot be directly computed from the above model. As we shall see below, the solution of this problem requires that we explicitly take the concentrations of RNAs into account.

Concentration dependence of RNA-RNA hybridization
Consider a (dilute) solution of two nucleic acid sequences $A$ and $B$ with concentrations $a$ and $b$, respectively. Hybridization yields a distribution of five molecular species: the two monomers $A$ and $B$, the two homodimers $AA$ and $BB$, and the heterodimer $AB$. In principle, of course, more complex oligomers might also arise, we will, however, neglect them in our approach. We may argue that ternary and higher complexes are disfavored by additional destabilizing initiation entropies.

The presentation in this section closely follows a recent paper by Dimitrov [27], albeit we use here slightly different definitions of the partitions functions. The partition functions of the secondary structures of the monomeric states are $Z^A$ and $Z^B$, respectively, as introduced in the previous section. In contrast to [27], we include the unfolded states in these partition functions. The partition functions $Z^{AA}$, $Z^{BB}$, and $Z^{AB}$, which are the output of the RNAcofold algorithm (denoted $Z$ in the previous section), include those states in which each monomer forms base-pairs only within itself as well as the unfolded monomers. We can now define

$$Z^{AA} = Z^{AA} - (Z^A)^2,$$

$$Z^{BB} = Z^{BB} - (Z^B)^2,$$

$$Z^{AB} = Z^{AB} - (Z^A Z^B) \quad (11)$$

as the partition functions restricted to the true dimeric states, but neglecting the initiation energies $\Theta_i$. An additional symmetry correction is needed in the case of the homo-dimers: A structure of a homo-dimer is symmetric if for any base pair $(i, j)$ there exists a pair $(i', j')$, where $i'$ ($j'$) denotes the equivalent of position $i$ in the other copy.
of the molecule. Such symmetric structures have a two-fold rotational symmetry that reduces their conformation space by a factor of 2, resulting in an entropic penalty of $\Delta G_{\text{sym}} = RT \ln 2$. On the other hand, since the recursion for the partition functions eq. 6 assumes two distinguishable molecules $A$ and $B$, any asymmetric structures of a homodimer are in fact counted twice by the recursion. Leading to the same correction as for symmetric structures.

Since both the initiation energy $\Theta$ and the symmetry correction $\Delta G_{\text{sym}}$ are independent of the sequence length and composition, the thermodynamically correct partition functions for the three dimer species are given by

$$
\begin{align*}
Z_{AA}' &= Z_{AA} \exp(-\Theta / RT) / 2, \\
Z_{BB}' &= Z_{BB} \exp(-\Theta / RT) / 2, \\
Z_{AB}' &= Z_{AB} \exp(-\Theta / RT).
\end{align*}
$$

(12)

From the partition functions we get the free energies of the dimer species, such as $F_{AB} = -RT \ln Z_{AB}'$, and the free energy of binding $\Delta F = F_{AB} - F_{A} - F_{B}$. We assume that pressure and volume are constant and that the solution is sufficiently dilute so that excluded volume effects can be neglected. The many particle partition function for this system is therefore [27]

$$
Q = \frac{a! b!}{n_A! n_B! 2 n_{AA}! 2 n_{BB}! n_{AB}!} \times (Z^{AA})^{a_A} (Z^{BB})^{a_B} (Z^{AB})^{a_{AB}} (Z^{BBB})^{a_{BBB}}
$$

(13)

where $a = n_A + 2n_{AA} + n_{AB}$ is the total number of molecules of type $A$ put into the solution (equivalently for $b$); $n_A$, $n_B$, $n_{AA}$, $n_{BB}$, $n_{AB}$, $n_{BBB}$ are the particle numbers for the five different monomer and dimer species, $V$ is the volume and $n$ is the sum of the particle numbers. The system now minimizes the free energy $-kT \ln Q$, i.e., it maximizes $Q$, by choosing the particle numbers optimally.

As in [27], the dimer concentrations are therefore determined by the mass action equilibria:

$$
\begin{align*}
[AA] &= K_{AA}[A]^2, \\
[BB] &= K_{BB}[B]^2, \\
[AB] &= K_{AB}[A][B]
\end{align*}
$$

(14)

with

$$
K_{AA} = \frac{Z_{AA}'}{(Z_A^A)^2} = \frac{(Z_{AA} - (Z_A^A)^2) e^{-\Theta / RT} / 2}{(Z_A^A)^2}
$$

$$
K_{BB} = \frac{1}{2} e^{-\Theta / RT} \left( \frac{Z_{BB}^B}{(Z_B^B)^2} - 1 \right)
$$

$$
K_{AB} = e^{\Theta / RT} \left( \frac{Z_{AB}}{Z_A^A Z_B^B} - 1 \right)
$$

Concentrations in eq.(14) are in mol/l.

Note, however, that the equilibrium constants in eq.(15) are computed from a different microscopic model than in [27], which in particular also includes internal base pairs within the dimers.

Together with the constraints on particle numbers, eq.(14) forms a complete set of equations to determine $x = [A]$ and $y = [B]$ from $a$ and $b$ by solving the resulting quadratic equation in two variables:

$$
0 = f(x, y) = x + K_{AB} y + 2 K_{BB} x^2 - a
$$

$$
0 = g(x, y) = y + K_{AB} x y + 2 K_{BBB} y^2 - b
$$

(16)

The Jacobian

$$
J(x, y) = \begin{vmatrix}
\frac{\partial f / \partial x}{\partial g / \partial x} & \frac{\partial f / \partial y}{\partial g / \partial y} \\
\frac{\partial g / \partial x}{\partial g / \partial y} & \frac{\partial g / \partial y}{\partial g / \partial y}
\end{vmatrix}
$$

(17)

of this system is strictly positive and diagonally dominated, and hence invertible on $\mathbb{R}^+ \times \mathbb{R}^+$. Furthermore $f$ and $g$ are thrice continuously differentiable on $B = [0, a] \times [0, b]$ and we know (because of mass conservation and the finiteness of the equilibrium constants) that the solution $(\hat{x}, \hat{y})$ is contained in the interior of the rectangle $B$. Newton’s iteration method

$$
\begin{align*}
x' &= x + \frac{g(x, y) \partial_x f(x, y) - f(x, y) \partial_x g(x, y)}{\Delta} \\
y' &= y + \frac{f(x, y) \partial_y g(x, y) - g(x, y) \partial_y f(x, y)}{\Delta} \\
\Delta &= \partial_x f(x, y) \partial_y g(x, y) - \partial_y f(x, y) \partial_x g(x, y)
\end{align*}
$$

(18)

$= \det J$
thus converges (at least) quadratically [48, 5.4.2]. We use \((a, b)\) as initial values for the iteration.

**Implementation and performance**

The algorithm is implemented in ANSI C, and is distributed as part of the of the Vienna RNA package. The resource requirements of RNAcofold and RNAfold are theoretically the same: both require \(O(n^3)\) CPU time and \(O(n^2)\) memory. In practice, however, keeping track of the cut makes the evaluation of the loop energies much more expensive and increases the CPU time requirements by an order of magnitude: RNAcofold takes about 22 minutes to cofold an about 3000 nt mRNA with a 20 nt miRNA on an Intel Pentium 4 (3.2 GHz), while RNAfold takes about 3 minutes to fold the concatenated molecule.

The base pairing probabilities are represented as a *dot plot* in which squares with an area proportional to \(P_{ij}\) represent the raw pairing probabilities, see Fig. 2. The dot plot is provided as Postscript file which is structured in such a way that the raw data can be easily recovered explicitly. RNAcofold also computes a table of monomer and dimer concentrations dependent on a set of user supplied initial conditions. This feature can readily be used to investigate the concentration dependence of RNA-RNA hybridization, see Fig. 3 for an example.

Like RNAfold, RNAcofold can be used to compute DNA dimers by replacing the RNA parameter set by a suitable set of DNA parameters. At present, the computation of DNA-RNA heterodimers is not supported. This would not only require a complete set of DNA-RNA parameters (stacking energies are available [49], but we are not aware of a complete set of loop energies) but also further complicate the evaluation of the loop energy contributions since pure RNA and pure DNA loops will have to be distinguished from mixed RNA-DNA loops.

**Applications**

Intermolecular binding of RNA molecules is important in a broad spectrum of cases, ranging from mRNA accessibility to siRNA or miRNA binding, RNA probe design, or designing RNA openers [50]. An important question that arises repeatedly is to explain differences in RNA-RNA binding between seemingly very similar or even identical binding sites. As demonstrated e.g. in [22,29,51,52], different RNA secondary structure of the target molecule can have dramatic effects on binding affinities even if the sequence of the binding site is identical.

Since the comparison of base pairing patterns is a crucial step in such investigations we provide a tool for graphically comparing two dot plots, see Fig. 4. It is written in Perl-Tk and takes two dot plot files and, optionally, an alignment file as input. The differences between the two dot plots are displayed in color-code, the dot plot is zoomable and the identity and probability (difference) of a base pair is displayed when a box is clicked.

As a simple example for the applicability of RNAcofold, we re-evaluate here parts of a recent study by Doench and Sharp [53]. In this work, the influence of GU base pairs on the effectivity of translation attenuation by miRNAs is assayed by mutating binding sites and comparing attenuation effectiveness to wild type binding sites.

Introducing three GU base pairs into the mRNA/miRNA duplex did, with only minor changes to the binding energy, almost completely destroy the functionality of the binding site. While Doench and Sharp concluded that miRNA binding sites are not functional because of the GU base pairs, testing the dimer with RNAcofold shows that there is also a significant difference in the cofolding structure that might account for the activity difference without invoking sequence specificities: Because of the secondary structure of the target, the binding at the 5’ end of the miRNA is much weaker than in the wild type, Fig. 4.

**Limitations and future extensions**

We have described here an algorithm to compute the partition function of the secondary structure of RNA dimers and to model in detail the thermodynamics of a mixture of two RNA species. At present, RNAcofold implements the most sophisticated method for modeling the interac-
pattern, the binding energies (10 nmol/1 for both experiments. Despite the similar binding remain unpaired. We assume an mRNA concentration of fraction of the calculation shows that even if siRNA is added in excess, a large straight

In [54], the authors observed 10% expression for several mRNAs with identical target sites for an siRNA which are located in different secondary structures. In vari-
several mRNAs with identical target sites for an siRNA VR1 straight

Acknowledgements
This work has been funded, in part, by the Austrian GEN-AU projects bioinformatics integration network & bioinformatics integration network II, and the German DFG Bioinformatics Initiative BIZ-6/1-2.

References
1. Ambros V: The functions of animal microRNAs. Nature 2004, 431:350-355.
2. Kidner CA, Martienssen RA: The developmental role of microRNA in plants. Curr Opin Plant Biol 2005, 8:38-44.
3. Rivas E, Klein RJ, Jones TA, Eddy SR: Computational identification of noncoding RNAs in E. coli by comparative genomics. Curr Biol 2001, 11:1369-1373.
4. McCutcheon JP, Eddy SR: Computational identification of non-coding RNAs in Saccharomyces cerevisiae by comparative genomics. Nuc Acids Res 2003, 31:4119-4128.
5. Klein RJ, Misulovin Z, Eddy SR: Noncoding RNA genes identified in AT-rich hyperthermophiles. Proc Natl Acad Sci USA 2002, 99:7542-7547.
6. Washietl S, Hofacker IL, Lukasser M, Hüttenhofer A, F SP: Genomewide mapping of conserved RNA Secondary Structures Reveals Evidence for Thousands of functional Non-Coding RNAs in Human. Nature Biotech 2005, 23:1383-1390.
7. Missal K, Rose D, Stadler PF: Non-coding RNAs in Ciona intestinalis. Bioinformatics 2005, 21(S2)i77-i78. [ECCB 2005 Supplement].
8. Bertone P, Stoc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M: Global Identification of Human Transcribed Sequences with Genome Tiling Arrays. Science 2004, 306:2242-2246.
9. Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, Wheeler R, Venter J, Yamanaka M, Patel S, Brubaker S, Maniataki E, Sharma A, Mourelatos Z: Algorithms for Molecular Biology 2004, 1:169-172.

10. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, Tammana H, Gingeras TR: Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22 points to widespread regulation of noncoding. RNAs Cell 2004, 11:149-154.

11. Mathews DH, Sabina J, Zuker M, Turner DH: Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure. J Mol Biol 1999, 288:911-940.

12. Bartel DP, Chen CZ: Micromanagers of gene expression: the potentially wide-spread influence of metazoan microRNAs. Nature Genetics 2004, 36(3):396-400.

13. Hobert O: Common logic of transcription factor and microRNA action. Trends Biochem Sci 2004, 29:462-468.

14. Mattick JS: Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. Bioessays 2003, 25:930-939.

15. Mattick JS: RNA regulation: a new genetics? Nature Genetics 2004, 36:316-323.

16. Rehmsmeier M, Steffen P, Höchsmann M, Giegerich R: Fast Folding and Comparison of RNA Secondary Structures. Monat Chemie 1994, 125(2):167-188.

17. Nelson P, Kirikidou M, Sharma A, Maniataki E, Mourelatos Z: The micrRNA world: small is mighty. Trends Biochem Sci 2003, 28:534-540.

18. Zucke M: On finding all suboptimal foldings of an RNA molecule. Science 1989, 244:48-52.

19. Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, Wheeler R, Venter J, Yamanaka M, Patel S, Brubaker S, Maniataki E, Mourelatos Z: Algorithms for Molecular Biology 2004, 1:167-172.

20. Bompfünewerer A, Flamm C, Fried C, Fritzsch G, Hofacker IL, Lehner B, Leitner GDS, Gingeras TR: Transcriptional Maps of 10 Human Chromosomes at 5-Nucleotide Resolution. Science 2005, 308:1149-1154.

21. Mathews DH, Sabina J, Zuker M, Turner DH: Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure. J Mol Biol 1999, 288:911-940.

22. Nussinov R, Piecznik G, Griggs JR, Kleitman DJ: Pseudoknots in RNA secondary structure prediction. Nature 2004, 359:84-89.

23. Mathews DH, Sabina J, Zuker M, Turner DH: Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure. J Mol Biol 1999, 288:911-940.

24. Zhang L, Mathews DH, Turner DH: Vienna RNA secondary structure server. Nucl Acids Res 2003, 31:3429-3434.

25. Ding Y, Lawrence CE: Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond. Nucleic Acids Res 2001, 29:1034-1046.

26. Zucke M, Sankoff D: RNA secondary structures and their prediction. Bull Math Biol 1994, 46:591-621.

27. Zucke M: On finding all suboptimal foldings of an RNA molecule. Science 1989, 244:48-52.

28. Zucke M: RNA secondary structure prediction including pseudoknots. J Comput Chem 2002, 23:1664-1677.

29. Zucke M, Sankoff D: RNA secondary structures and their prediction. Bull Math Biol 1994, 46:591-621.

30. Ding Y, Lawrence CE: Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond. Nucleic Acids Res 2001, 29:1034-1046.

31. Zucke M, Sankoff D: RNA secondary structures and their prediction. Bull Math Biol 1994, 46:591-621.

32. Waterman MS: Secondary Structure of single-stranded nucleic acids. Adv Math Suppl Studies 1978, 1:167-212.

33. Stadler PF, Mückstein U, Tafer H, Hackermüller J, Bernhard SB, Stadler BM, Stadler PS, Hackermüller J, Bernhard SB, Stadler BM, Stadler PS: Vienna RNA secondary structure server. Nucleic Acids Res 2004, 32:5126-5133.