Partition Function and Base Pairing Probabilities of RNA Heterodimers

Stephan H Bernhart*[1], Hakim Tafer[1], Ulrike Mückstein[1], Christoph Flamm[2,1], Peter F Stadler[2,1,3], Ivo L Hofacker[1]

[1]Theoretical Biochemistry Group, Institute for Theoretical Chemistry, University of Vienna, Währingerstrasse 17, Vienna, Austria
[2]Bioinformatics Group, Department of Computer Science and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstrasse 16-18, D-04170 Leipzig, Germany
[3]The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, New Mexico

Email: Stephan H Bernhart*- berni@tbi.univie.ac.at;

*Corresponding author

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 in 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 interaction 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 important experimental techniques for changing the gene expression patterns including RNAi [20] and modifier RNAs [21–24].

RNA-RNA binding occurs by formation of stacked intermolecular base pairs, which of course compete with the propensity of both interacting partners to form intramolecular base pairs. These base pairing patterns, usually referred to as *secondary structures*, not only comprise the dominating part of the energetics of structure formation, they also appear as intermediates in the formation of the tertiary structure of RNAs [25], and they are in many cases well conserved in evolution. Consequently, secondary structures provide a convenient, and computationally tractable, approximation not only to RNA structure but also to the thermodynamics of RNA-RNA interaction.

From the computational point of view, this requires the extension of RNA folding algorithms to include intermolecular as well as intramolecular base pairs. Several approximations have been described in the literature: Rehmsmeier *et al.* [26] as well as Dimitrov and Zuker [27] introduced algorithms that consider exclusively intermolecular base pairs, leading to a drastic algorithmic simplification of the folding algorithms since multi-branch loops are by construction excluded in this case. Andronescu *et al.* [28], like the present contribution, consider all base pairs that can be formed in secondary structures in a concatenation of the two hybridizing molecules. This set in particular contains the complete structural ensemble of both partners in isolation. Mückstein *et al.* [29] recently consider an asymmetric model in which base pairing is unrestricted in a large target RNA, while the (short) interaction partner is restricted to intermolecular base pairs.

A consistent treatment of the thermodynamic aspects of RNA-RNA interactions requires that one takes into account the entire ensemble of suboptimal structures. This can be approximated by explicitly computing all structures in an energy band above the ground state. Corresponding algorithms are discussed in [30] for single RNAs and in [28] for two interacting RNAs. A more direct approach, that becomes much more efficient for larger molecules, is to directly compute the partition function of the entire ensemble along the lines of McCaskill's algorithm [31]. This is the main topic of the present contribution.

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

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.

1 Every sequence position i 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.

2 $(i, j) \in S$ implies $|i - j| \geq 4$, i.e., hairpin loops have at least three unpaired positions inside their closing pair.

3 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 con-

dition 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 $\mathsf{type}(L)$ (hairpin loop, interior loop, bulge, stacked pair, or multi-branch loop) and the sequence of some or all of the nucleotides in the loop, $x|_L$:

$$\varepsilon(L) = \varepsilon(\mathsf{type}(L), x|_L)\,. \tag{1}$$

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

$$\mathcal{H}(i, j) = \mathcal{H}(x_i, x_{i+1}, \ell, x_{j-1}, x_j) \tag{2}$$

where ℓ 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

$$\begin{aligned} &\mathcal{I}(i, j; k, l) = \\ &\mathcal{I}(x_i, x_{i+1}; \ell_1; x_{k-1}, x_k; x_l, x_{l+1}; \ell_2; x_{j-1}, x_j) \end{aligned} \tag{3}$$

where ℓ_1 is the length of the unpaired strand between k and p and ℓ_2 is the length of the unpaired strand between q and l. Symmetry of the energy model dictates $\mathcal{I}(i, j; k, l) = \mathcal{I}(l, k; j, i)$. If $\ell_1 = \ell_2 = 0$ we have a (stabilizing) stacked pair, if only one of ℓ_1 and ℓ_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 ℓ is the length of multiloop (again expressed as the number of unpaired nucleotides) and β 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_{ij} and $d^E_{i,j}$, respectively.

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_i between two base pairs (x_p, x_{i-1}) and (x_{i+1}, x_q) 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} free energy of the optimal substructure on the subsequence $x[i,j]$.

C_{ij} free energy of the optimal substructure on the subsequence $x[i,j]$ subject to the constraint that i and j form a basepair.

M_{ij} free energy of the optimal substructure on the subsequence $x[i,j]$ subject to the constraint that that $x[i,j]$ is part of a multiloop and has at least one component, i.e., a sub-sequence that is enclosed by a base pair.

M^1_{ij} free energy of the optimal substructure on the subsequence $x[i,j]$ subject to the constraint that that $x[i,j]$ is part of a multiloop and has exactly one component, which has the closing pair i,h for some h satisfying $i \leq h < j$.

The "conventional" energy minimization algorithm (for simplicity of presentation without dangling end contributions) for linear RNA molecules can be summarized in the following way, which corresponds to the recursions implemented in the `Vienna RNA Package`:

$$\begin{aligned}
F_{ij} &= \min\left\{F_{i+1,j},\ \min_{i<k\leq j} C_{ik} + F_{k+1,j}\right\} \\
C_{ij} &= \min\left\{\mathcal{H}(i,j),\ \min_{i<k<l<j} C_{kl} + \mathcal{I}(i,j;k,l),\right. \\
&\qquad \left.\min_{i<u<j} M_{i+1,u} + M^1_{u+1,j-1} + a\right\} \\
M_{ij} &= \min\left\{M^1_{ij}, \min_{i<u<j} C_{i,u} + M_{u+1,j} + b,\right. \\
&\qquad \left. M_{i+1,j} + c\right\} \\
M^1_{ij} &= \min\left\{C_{ij}, M^1_{i,j-1} + c\right\}
\end{aligned} \tag{4}$$

It is straightforward to translate these recursions into recursions for the partition function because they already provide a partition of the set of all secondary structures that can be formed by the sequence x. This unambiguity of the decomposition of the ensemble structure is not important for energy minimization, while it is crucial for enumeration and hence also for the computation of the partition function [31]. Let us write Z_{ij} for the partition function on $[x_i,x_j]$, Z^P_{ij} for the partition function constrained to structures with an (i,j) pair, and Z^M_{ij}, Z^{M1}_{ij} for the partition function versions of the multiloop terms M_{ij} and M^1_{ij}.

The adaptation of the recursion to 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.

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 no 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., is does not contribute to the folding energy (relative to the random coil reference state). For example, an interior loop $\mathcal{I}(i,j;k,l)$ 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.(5)) and loops containing the cut c in their backbone (lower alternatives in eq.(5)). Explicitly, this means $i \le n_1 < j$ in the hairpin loop case, in the interior loop case, this either means $i \le n_1 \le k$ or $l \le n_1 < j$. In the multiloop recursions we have to suppress those terms in which the cut is located between the multiloop components to avoid double-counting: These "exterior loop cases" are already taken care of by the alternative hairpin term.

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

$$
\begin{aligned}
Z_{ij} &= Z_{i+1,j} + \sum_{i<k\le j} Z^P_{i,k}\hat{d}^E_{ik} Z_{k+1,j} \\
Z^P_{ij} &= \begin{cases} \hat{\mathcal{H}}(i,j) \\ Z_{i+1,n_1} Z_{n_1+1,j-1}\hat{d}^I_{ij} \end{cases} \\
&\quad + \sum_{i<k<l<j} Z^P_{kl} \begin{cases} \hat{\mathcal{I}}(i,j;k,l) \\ \hat{d}^E_{k,l}\hat{d}^I_{i,j} \end{cases} \\
&\quad + \hat{d}^I_{i,j}\hat{a} \sum_{i<u<j} Z^M_{i+1,u} + Z^{M1}_{u+1,j-1} \\
Z^M_{ij} &= \begin{cases} Z^M_{i+1,j}\hat{c} \\ 0 \end{cases} + Z^{M1}_{i,j} + \sum_{\substack{i<k\le j \\ i,k,j\ne n_1}} Z^P_{i,k}\hat{d}^E_{ik} Z^M_{k+1,j}\hat{b} \\
Z^{M1}_{ij} &= \begin{cases} Z^P_{ij} + \begin{cases} Z^{M1}_{i,j-1}\hat{c} \\ 0 \end{cases} & \text{if} \quad i \ne n_1 \\ 0 & \text{if} \quad i = n_1 \end{cases}
\end{aligned} \tag{5}
$$

Upper alternatives refer to regular loops, lower alternatives to the loop containing the cutpoint. For brevity we have used here abbreviations $\hat{\mathcal{H}}(i,j) = \exp(-\mathcal{H}(i,j)/RT)$, etc., 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* Φ_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., $i \le n_1 < j$. The resulting book-keeping problems fortunately can be avoided by introducing this term only after the dynamic programming tables have been filled. To this end we observe that $Z_{i,j} = Z^A_{i,j}$, $1 \le i,j \le n_1$ are the partition functions for subsequences of the isolated A molecule, while $Z_{n_1+i,n_1+j} = Z^B_{i,j}$, $1 \le i,j \le n_2$ 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

$$
\begin{aligned}
Z^* = &\left[Z_{1,n_1+n_2} - Z_{1,n_1} Z_{n_1+1,n_1+n_2}\right] e^{-\Theta_I/RT} \\
&+ Z_{1,n_1} Z_{n_1+1,n_1+n_2}
\end{aligned} \tag{6}
$$

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 in which the cut-point u between components coincides with the cut point c. All other cases are

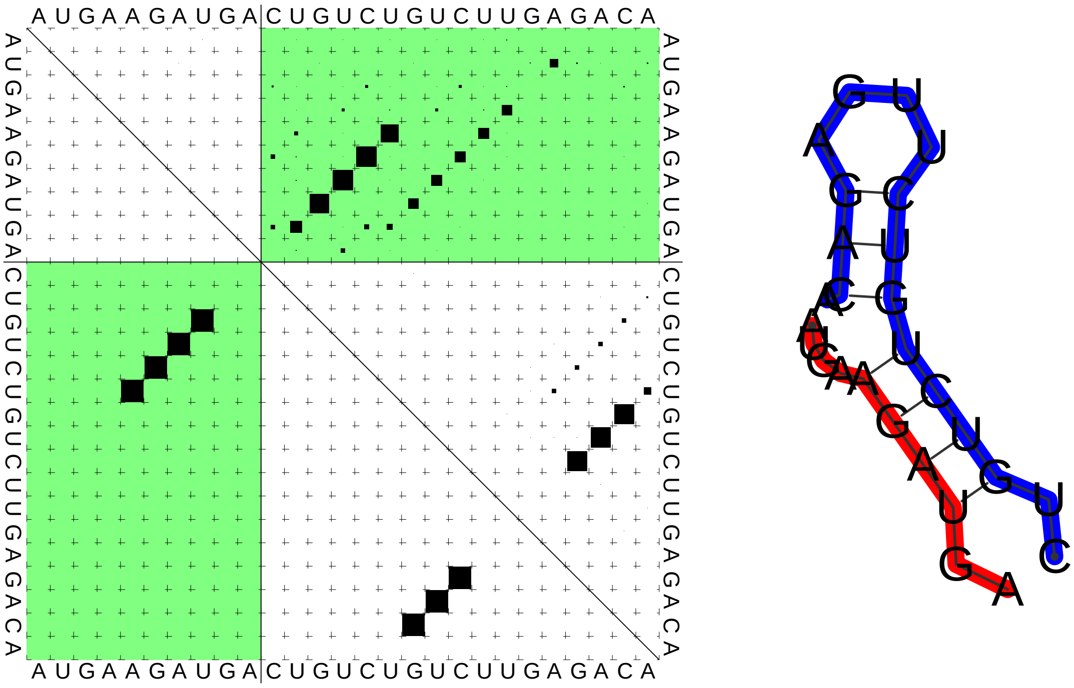


Figure 1: Dot plot (left) and mfe structure representation (right) of the cofolding structure of the two RNA molecules `AUGAAGAUGA` (red) and `CUGUCUGUCUUGAGACA` (blue).
Dot Plot: Upper right: Partition function. The area of the squares is proportional to the corresponding pair probabilities. Lower left: Minimum free energy structure. The two lines forming a cross indicate the cut point, intermolecular base pairs are depicted in the green upper right (partition function) and lower left (mfe) rectangle.

already taken care of in the forward recursion. Thus:

$$P_{kl} = \frac{Z_{1,k-1} Z^B_{k,l} Z^{l+1,n}}{Z_{1,n}} + \sum_{p<k;q>l} P_{pq} \frac{Z^B_{k,l}}{Z^B_{p,q}} \Big\{ \hat{\mathcal{I}}(p,q,k,l) + Z^M_{p+1,k-1} \hat{a} \hat{c}^{q-l-1} + Z^M_{l+1,q-1} \hat{a} \hat{c}^{k-p-1} + Z^M_{p+1,k-1} Z^M_{l+1,q-1} \hat{a} \Big\} \tag{7}$$

The "raw" values of P_{ij}, 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+1,n_1+n_2} to obtain the base pairing probability matrices P^A_{ij} and $P^B_{n_1+i,n_1+j}$ for the isolated molecules. Note that equivalently we could compute P^A_{ij} and P^B_{ij} directly using the partition function version of `RNAfold`. The fraction of structures without intermolecular pairs (in the cofold model without initiation contributions) is $Z^A Z^B / Z$, and hence the fraction of true dimers is

$$p^* = 1 - \frac{Z^A Z^B}{Z}\,. \tag{8}$$

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

$$P'_{ij} = \frac{1}{p^*} \begin{cases} P_{ij} - (1-p^*) P^A_{ij} & \text{if } i,j \in A \\ P_{ij} - (1-p^*) P^B_{ij} & \text{if } i,j \in B \\ P_{ij} & \text{otherwise} \end{cases} \tag{9}$$

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 homo-dimers 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

$$\begin{aligned} Z^{=}_{AA} &= Z^{AA} - (Z^A)^2\,, \\ Z^{=}_{BB} &= Z^{BB} - (Z^B)^2\,, \\ Z^{=}_{AB} &= Z^{AB} - (Z^A Z^B) \end{aligned} \tag{10}$$

as the partition functions restricted to the true homo- and hetero-dimers, but neglecting the initiation energies Θ_I. Since this term is assumed to be independent of the sequence length and sequence composition, the thermodynamically correct partition functions for the three dimer species are given by

$$\begin{aligned} Z'_{AA} &= (Z^{AA} - (Z^A)^2)\exp(-\Theta_I/RT)\,, \\ Z'_{BB} &= (Z^{BB} - (Z^B)^2)\exp(-\Theta_I/RT)\,, \\ Z'_{AB} &= (Z^{AB} - Z^A Z^B)\exp(-\Theta_I/RT)\,. \end{aligned} \tag{11}$$

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 grand-canonical ensemble for this system is therefore [27]

$$\begin{aligned} \mathcal{Q} = \quad & V^n \sum_{\substack{n_A+2n_{AA}+n_{AB}=a \\ n_B+2n_{BB}+n_{AB}=b}} \frac{a!b!}{n_A!n_B!n_{AA}!n_{BB}!n_{AB}!} \\ &\times (Z'^A)^{n_A}(Z'^{AA})^{n_{AA}}(Z'^{AB})^{n_{AB}}(Z'^{BB})^{n_{BB}}(Z'^B)^{n_B} \end{aligned} \tag{12}$$

which sums over all possibilities to arrange our A- and B-type molecules (of which we have a and b many, respectively) in our five molecular species and, within these, in all possible secondary structure states with Boltzmann weights. The system minimizes the grand-canonical free energy $-kT\ln\mathcal{Q}$, i.e., it maximizes $\mathcal{Q}$, by choosing the particle numbers n_A, n_B, n_{AA}, n_{BB}, n_{AB} optimally.

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

$$\begin{aligned} [AA] &= K_{AA}\,[A]^2 \\ [BB] &= K_{BB}\,[B]^2 \\ [AB] &= K_{AB}\,[A]\,[B] \end{aligned} \tag{13}$$

with

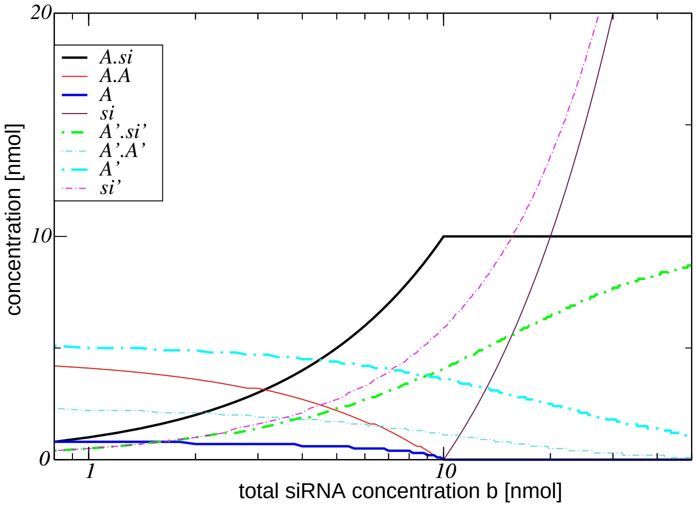
$$\begin{aligned} K_{AA} &= \frac{Z'^{AA}}{(Z^A)^2} = \frac{(Z^{AA} - (Z^A)^2)e^{-\Theta_I/RT}}{(Z_A)^2} \\ &= e^{-\Theta_I/RT}\left(\frac{Z^{AA}}{(Z^A)^2} - 1\right) \\ K_{BB} &= e^{-\Theta_I/RT}\left(\frac{Z^{BB}}{(Z^B)^2} - 1\right) \\ K_{AB} &= e^{-\Theta_I/RT}\left(\frac{Z^{AB}}{Z^A Z^B} - 1\right) \end{aligned} \tag{14}$$

Concentrations in Eq.(13) are in mol/l.

Note, however, that the equilibrium constants in Eq.(14) 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.(13) 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:

$$\begin{aligned} 0 &= f(x,y) := x + K_{AB}xy + 2K_{AA}x^2 - a \\ 0 &= g(x,y) := y + K_{AB}xy + 2K_{BB}y^2 - b \end{aligned} \tag{15}$$



Binding energies: $\Delta F(A) = -24.53\text{kcal/mol}$
$\Delta F(A') = -11.76\text{kcal/mol}$.

Figure 2: Example for the concentration dependency for two mRNA-siRNA binding experiments. In [48], Schubert *et al.* designed several mRNAs with identical target sites for an siRNA *si*, which are located in different secondary structures. In variant A, the *VR1 straight* mRNA, the binding site is unpaired, while in the mutant mRNA *VR1 HP5-11*, A', only 11 bases remain unpaired. We assume an mRNA concentration of $a = 10$ nmol/l for both experiments. Despite the similar binding pattern, the binding energies ($\Delta F = F^{AB} - F^A - F^B$) differ dramatically. In [48], the authors observed 10% expression for *VR1 straight*, and 30% expression for the HP5-11 mutant. Our calculation shows that even if siRNA is added in excess, a large fraction of the *VR1 HP5-11* mRNA remains unbound.

The Jacobian

$$\mathbf{J}(x,y) = \begin{pmatrix} \partial f/\partial x & \partial f/\partial y \\ \partial g/\partial x & \partial g/\partial y \end{pmatrix}$$
$$= \begin{pmatrix} 1 + K_{AB}y + 4K_{AA}x & K_{AB}x \\ K_{AB}y & 1 + K_{AB}x + 4K_{BB}y \end{pmatrix} \qquad (16)$$

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 $\mathbb{B} = [0,a] \times [0,b]$ 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 $\mathbb{B}$. Newton's iteration method

$$\begin{aligned} x' &= x + \frac{g(x,y)\partial_y f(x,y) - f(x,y)\partial_y g(x,y)}{\Delta} \\ y' &= y + \frac{f(x,y)\partial_x g(x,y) - g(x,y)\partial_x f(x,y)}{\Delta} \\ \Delta &= \partial_x f(x,y)\partial_y g(x,y) - \partial_y f(x,y)\partial_x g(x,y) \\ &= \det\mathbf{J} \end{aligned} \qquad (17)$$

thus converges (at least) quadratically [49, 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 $\mathcal{O}(n^3)$ CPU time and $\mathcal{O}(n^2)$ 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 3000nt mRNA with a 20nt 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 the raw pairing probabilities, see Fig. 1. 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. 2 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 [50], but we are not aware of a complete set

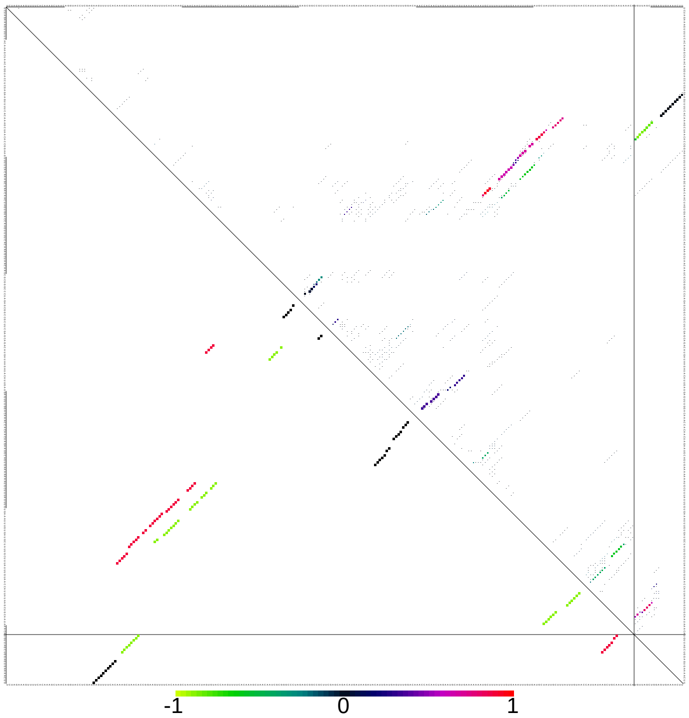


Figure 3: Difference dot Plot of native and mutated secondary structure of a 3GU mutation of the CXCR4 siRNA gene. The red part on the right hand side shows the base pairing probability of the 5' part of the micro RNA, which is 80% higher in the native structure. This is an alternative explanation for the missing function of the mutant. Because of the mutations, the stack a little to the left gets more stable, and the probability of binding of the 5' end of the siRNA is reduced significantly.
The color of the dots encodes the difference of the pair probabilities in the two molecules such positive (red) squares denote pairs more more probable in the second molecule (see color bar). The area of the dots is proportional to the larger of the two pair probabilities.

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 [51]. 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 [22,29,52,53], 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. 3. 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 [54]. 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 effectivity 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. 3.

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 interactions of two (large) RNAs. Because the no-pseudoknot condition is enforced to limit computational costs, our approach disregards certain interaction structures that are known to be important, including kissing hairpin complexes.

The second limitation, which is of potential importance in particular in histochemical applications, is the restriction to dimeric complexes. More complex oligomers are likely to form in reality. The generalization of the present approach to trimers or tetramers is complicated by the fact that for more than two molecules the results of the calculation are not independent of the order of the concatenation any more, so that for M-mers $(M-1)!$ permutations have to be considered separately. This also leads to bookkeeping problems since every secondary structure still has to be counted exactly once.

Acknowledgments

This work has been funded, in part, by the Austrian GEN-AU bioinformatics integration network sponsored by bm:bwk, grant number 200.067/3-VI/1/2002, 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**. *Nucl. 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: **Genome-wide 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, Wong B, Drenkow J, Yamanaka M, Patel S, Brubaker S, Tammana H, Helt G, Struhl K, Gingeras TR: **Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs**. *Cell* 2004, **116**:499–509.
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**. *Genome Res.* 2004, **14**:331–342.
11. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, Sementchenko V, Piccolboni A, Bekiranov S, Bailey DK, Ganesh M, Ghosh S, Bell I, Gerhard DS, Gingeras TR: **Transcriptional Maps of 10 Human Chromosomes at 5-Nucleotide Resolution**. *Science* 2005, **308**:1149–1154.
12. Bartel DP, Chen CZ: **Micromanagers of gene expression: the potentially wide-spread influence of metazoan microRNAs**. *Nature Genetics* 2004, **5**: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, **5**:316–323.
16. Bompfünewerer AF, Flamm C, Fried C, Fritzsch G, Hofacker IL, Lehmann J, Missal K, Mosig A, Müller B, Prohaska SJ, Stadler BMR, Stadler PF, Tanzer A, Washietl S, Witwer C: **Evolutionary Patterns of Non-Coding RNAs**. *Th. Biosci.* 2005, **123**:301–369.
17. Nelson P, Kiriakidou M, Sharma A, Maniataki E, Mourelatos Z: **The microRNA world: small is mighty**. *Trends Biochem. Sci.* 2003, **28**:534–540.
18. Gott JM, Emeson RB: **Functions and Mechanisms of RNA Editing**. *Annu. Rev. Genet.* 2000, **34**:499–531.
19. Stuart K, Allen TE, Heidmann S, Seiwert SD: **RNA editing in kinetoplastid protozoa**. *Microbiol. Mol. Biol. Rev.* 1997, **61**:105–120.
20. Elbashir W S Lendeckel, Tuschl T: **RNA interference is mediated by 21- and 22-nucleotide RNAs**. *Genes Dev.* 2001, **15**:188–200.
21. Childs JL, Disney MD, Turner DH: **Oligonucleotide directed misfolding of RNA inhibits *Candida albicans* group I intron splicing**. *Proc. Natl. Acad. Sci. USA* 2002, **99**:11091–11096.

22. Meisner NC, Hackermüller J, Uhl V, Aszódi A, Jaritz M, Auer M: **mRNA openers and closers: A methodology to modulate AU-rich element controlled mRNA stability by a molecular switch in mRNA conformation**. *Chembiochem.* 2004, **5**:1432–1447.

23. Nulf CJ, Corey D: **Intracellular inhibition of hepatitis C virus (HCV) internal ribosomal entry site (IRES)-dependent translation by peptide nucleic acids (PNAs) and locked nucleic acids (LNAs)**. *Nucl. Acids Res.* 2004, **32**:3792–3798.

24. Paulus M, Haslbeck M, Watzele M: **RNA stem-loop enhanced expression of previously non-expressible genes**. *Nucl. Acids Res.* 2004, **32**:9/e78. [Doi 10.1093/nar/gnh076].

25. Uhlenbeck OC: **A coat for all Sequences**. *Nature Struct. Biol.* 1998, **5**:174–176.

26. Rehmsmeier M, Steffen P, Höchsmann M, Giegerich R: **Fast and effective prediction of microRNA/target duplexes**. *RNA* 2004, **10**:1507–1517.

27. Dimitrov RA, Zuker M: **Prediction of Hybridization and Melting for Double-Stranded Nucleic Acids**. *Biophys. J.* 2004, **87**:215–226.

28. Andronescu M, Zhang ZC, Condon A: **Secondary Structure Prediction of Interacting RNA Molecules**. *J. Mol. Biol.* 2005, **345**:987–1001.

29. Mückstein U, Tafer H, Hackermüller J, Bernhard SB, Stadler PF, Hofacker IL: **Thermodynamics of RNA-RNA Binding**. In *German Conference on Bioinformatics 2005*, Volume P-71. Edited by Torda A, Kurtz S, Rarey M, Bonn: Gesellschaft f. Informatik 2005:3–13.

30. Wuchty S, Fontana W, Hofacker IL, Schuster P: **Complete Suboptimal Folding of RNA and the Stability of Secondary Structures**. *Biopolymers* 1999, **49**:145–165.

31. McCaskill JS: **The Equilibrium Partition Function and Base Pair Binding Probabilities for RNA Secondary Structure**. *Biopolymers* 1990, **29**:1105–1119.

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

33. Nussinov R, Piecznik G, Griggs JR, Kleitman DJ: **Algorithms for Loop Matching**. *SIAM J. Appl. Math.* 1978, **35**:68–82.

34. Leydold J, Stadler PF: **Minimal Cycle Basis of Outerplanar Graphs**. *Elec. J. Comb.* 1998, **5**:209–222 [R16: 14 p.]. [See `http://www.combinatorics.org/` R16 and Santa Fe Institute Preprint 98-01-011].

35. 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.

36. SantaLucia jr J: **A Unified View of Polymer, Dumbbell, and Oligonucleotide DNA Nearest-Neighbor Thermodynamics**. *Proc. Natl. Acad. Sci. USA* 1998, **95**:1460–1465.

37. Zuker M, Stiegler P: **Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information**. *Nucl. Acids Res.* 1981, **9**:133–148.

38. Rivas E, Eddy SR: **A dynamic programming algorithm for RNA structure prediction including pseudoknots**. *J. Mol. Biol* 1999, **85**(5):2053–2068.

39. Reeder J, Giegerich R: **Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics**. *BMC Bioinformatics* 2004, **5**.

40. Dirks RM, Pierce NA: **A parition function algorithm for nucleic acid secondary structure including pseudoknots**. *J. Comput. Chem.* 2003, **24**:1664–1677.

41. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P: **Fast Folding and Comparison of RNA Secondary Structures**. *Monatsh. Chemie* 1994, **125**(2):167–188.

42. Hofacker IL: **Vienna RNA secondary structure server**. *Nucl. Acids Res.* 2003, **31**:3429–3431.

43. Zuker M, Sankoff D: **RNA secondary structures and their prediction**. *Bull. Math. Biol.* 1984, **46**:591–621.

44. Zuker M: **On finding all suboptimal foldings of an RNA molecule**. *Science* 1989, **244**:48–52.

45. Zuker M: `Mfold` **web server for nucleic acid folding and hybridization prediction**. *Nucl. Acids Res.* 2003, **31**:3406–3415.

46. Weixlbaumer A, Werner A, Flamm C, Westhof E, Schroeder R: **Determination of thermodynamic parameters for HIV DIS type loop-loop kissing complexes**. *Nucl. Acids Res.* 2004, **32**:5126–5133.

47. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH: **Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure**. *Proc. Natl. Acad. Sci. USA* 2004, **101**:7287–6297.

48. Schubert S, Grunweller A, Erdmann V, Kurreck J: **Local RNA Target Structure Influences siRNA Efficacy: Systematic Analysis of Intentionally Designed Binding Regions.** *J. Mol. Biol.* 2005, **348**(4):883–93.

49. Schwarz HR: *Numerische Mathematik*. Stuttgart: B.G. Teubner 1986.

50. Wu P, Nakano Si, Sugimoto N: **Temperature dependence of thermodynamic properties for DNA/DNA and RNA/DNA duplex formation**. *Eur. J. Biochem.* 2002, **269**:2821–2830.

51. Hackermüller J, Meisner NC, Auer M, Jaritz M, Stadler PF: **The Effect of RNA Secondary Structures on RNA-Ligand Binding and the Modifier RNA Mechanism: A Quantitative Model**. *Gene* 2005, **345**:3–12.

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

53. Ding Y, Lawrence CE: **A statistical sampling algorithm for RNA secondary structure prediction**. *Nucl. Acids Res.* 2003, **31**:7180–7301.

54. Doench JG, Sharp PA: **Specificity of microRNA target selection in translational repression**. *Genes Devel.* 2004, **18**:504–511.