Target prediction and a statistical sampling algorithm for RNA–RNA interaction

Fenix W. D. Huang1, Jing Qin1, Christian M. Reidys1,2,* and Peter F. Stadler3–8

1Center for Combinatorics, LPMC-TJKLC, 2College of Life Science, Nankai University, Tianjin 300071, P. R. China, 3Bioinformatics Group, Department of Computer Science, 4Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelestrasse 16-18, D-04107 Leipzig, 5Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, 6 RNomics Group, Fraunhofer Institut for Cell Therapy and Immunology, Peršicstraße 1, D-04103 Leipzig, Germany, 7Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Vienna, Austria and 8The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM, USA

Received on September 9, 2009; revised on November 4, 2009; accepted on November 5, 2009
Advance Access publication November 13, 2009
Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: It has been proven that the accessibility of the target sites has a critical influence on RNA–RNA binding, in general and the specificity and efficiency of miRNAs and siRNAs, in particular. Recently, (kW) and (kW) space dynamic programming (DP) algorithms have become available that compute the partition function of RNA–RNA interaction complexes, thereby providing detailed insights into their thermodynamic properties.

Results: Modifications to the grammars underlying earlier approaches enables the calculation of interaction probabilities for any given interval on the target RNA. The computation of the ‘hybrid probabilities’ is complemented by a stochastic sampling algorithm that produces a Boltzmann weighted ensemble of RNA–RNA interaction structures. The sampling of k structures requires only negligible additional memory resources and runs in O(kW).

Availability: The algorithms described here are implemented in C as part of the rip package. The source code of rip can be downloaded from http://www.combinatorics.cn/cbpc/rip.html and http://www.bioinf.uni-leipzig.de/Software/rip.html.

Contact: duck@santafe.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

RNA–RNA binding is a major mode of action of various classes of non-coding RNAs and plays a crucial role in many regulatory processes in all living organisms. Examples include the regulation of translation in both prokaryotes (Narberhaus and Vogel, 2007) and eukaryotes (Banerjee and Slack, 2002; McMans and Sharp, 2002), the targeting of chemical modifications (Bachellerie et al., 2002), insertion editing (Benne, 1992) and transcriptional control (Kugel and Goodrich, 2007). Emerging evidence suggests, furthermore, that RNA–RNA interactions also play a role for the functionality of long mRNA-like ncRNAs (Hekimoglu and Ringrose, 2009). A common theme in many RNA classes, including miRNAs, snRNAs, gRNAs, snoRNAs and in particular many of the procaryotic small RNAs, is the formation of RNA–RNA interaction structures that are much more complex than simple complementary sense–antisense interactions. Thermodynamically, the binding of two RNA molecules A and B can be described by the binding energy \( \Delta G_{\text{bind}} = G_{\text{AB}} - G_A - G_B \), i.e. by the difference of the energy of structure formation \( G_{\text{AB}} \) of the AB complex and the folding energies \( G_A \) and \( G_B \) of the two individual RNAs A and B. Thus, the binding or hybridization energy has been widely used as a criterion to predict RNA–RNA interactions (Busch et al., 2008; Rehmsmeier et al., 2004; Tjaden et al., 2006).

The interaction between two RNAs is governed by the same physical principles that determine RNA folding: the formation of specific base pairing patterns whose energy is largely determined by base pair stacking and loop strains. Secondary structures, therefore, are an appropriate level of description to quantitatively understand the thermodynamics of RNA–RNA binding. Just as the general RNA folding problem with unrestricted pseudoknots (Akutsu, 2000), the RNA–RNA interaction problem (RIP) is Non-Polynomial (NP)-complete in its most general form (Alkan et al., 2006; Mneimneh, 2009). Polynomial-time algorithms can be derived, however, by restricting the space of allowed configurations in ways that are similar to pseudoknot folding algorithms (Rivas and Eddy, 1999). The simplest approach concatenates two (or more) interacting sequences and then employs the standard secondary structure folding algorithm with a slightly modified energy model that treats loops containing cut-points as external elements. The software tools RNAcofold (Bernhart et al., 2006; Hofacker et al., 1994), pairfold (Andronescu et al., 2005) and NUPACK (Dirks et al., 2007) subscribe to this strategy. The main problem of this approach is that it cannot predict important motifs such as kissing-hairpin loops. The paradigm of concatenation has also been generalized to the pseudoknot folding algorithm of Rivas and Eddy (1999). The resulting model, however, still does not generate all relevant interaction structures (Chitsaz et al., 2009; Qin and Reidys, 2007). An alternative line of thought,
implemented in RNAduplex and RNAhybrid (Rehmsmeier et al., 2004), is to neglect all internal base pairings in either strand, i.e. to compute the minimum free energy (MFE) secondary structure of hybridization of otherwise unstructured RNAs. RNAup (Mückstein et al., 2006, 2008) and interaRNA (Busch et al., 2008) restrict interactions to a single interval that remains unpaired in the secondary structure for each partner. As a special case, snoRNA/target complexes are treated more efficiently using a specialized tool (Tafer et al., 2009) due to the highly conserved interaction motif. Algorithmically, the approaches mentioned so far are close relatives of the RNA folding recursions given by Zuker and Sankoff (1984).

A different approach was taken independently by Pervouchine (2004) and Alkan et al. (2006), who proposed MFE folding algorithms for predicting the joint structure of two interacting RNA molecules. In this model, “joint structure” means that the intramolecular structures of each partner is pseudoknot free, the intermolecular binding pairs are non-crossing and there is no so-called “zig-zag” configuration (see below for details). The optimal joint structure can be computed in \(O(N^5)\) time and \(O(N^4)\) space by means of dynamic programming (DP). More recently, extensions to the partition function were proposed by Chitsaz et al. (2009b) \(\text{piRNA}\) and Huang et al. (2009) \(\text{rip2}\). In contrast with the RNA folding problem, where minimum energy folding and partition functions can be obtained by very similar algorithms, this is much more complicated for joint structures. The reason is that simple unambiguous grammars are known for RNA secondary structures (Dowell and Eddy, 2004), while the disambiguation of grammar underlying the Alkan-Pervouchine algorithm requires the introduction of a large number of additional non-terminals (which algorithmically translate into additional DP tables). Although the partition function of joint structures can be computed in \(O(N^5)\) time and \(O(N^4)\) space, the current implementations require very large computational resources. Salari et al. (2009) recently achieved a substantial speed-up making use of the observation that the external interactions mostly occur between pairs of unpaired regions of single structures. Chitsaz et al. (2009a), on the other hand, use tree-structured Markov random fields to approximate the joint probability distribution of multiple \((\geq 3)\) contact regions.

The binding energies provides a useful overall characterization of an RNA–RNA interaction. In many cases, however, the locations of the intermolecular base pairs and the detailed structure of the interaction complex is of crucial importance. Bacterial sRNAs, for example, may either up- or down-regulate mRNA translation depending on the structural changes induced by the interaction (Urban and Vogel, 2007). In particular, in RNA–RNA complexes with multiple interaction sites, i.e. in the class of structures for which the expensive computation of joint structures is necessary, one is interested in the probabilities of hybridization in individual regions and in the interdependencies of alternative conformations, see Fig. 1. The probabilities of the individual building blocks of the DP recursions of Huang et al. (2009), furthermore, do not lend themselves to direct biophysical interpretations (see Supplementary Material).

We therefore extend our previous framework in two directions: (i) A modification of the underlying grammar explicitly treats hybrids, i.e. maximal regions with exclusively intermolecular interactions. This allows us to investigate local aspects in much more detail. (ii) A stochastic backtracing algorithm, in analogy to similar approaches for RNA secondary structure prediction (Ding and Lawrence, 2003; Tacker et al., 1996), which can be used to produce representative structure and to generate samples from the thermodynamic properties. These samples can be useful to assess complex structural features for which it would be too tedious or expensive to design and implement dedicated exact backtracing algorithms.

2 THE HYBRID-PARTITION FUNCTION

2.1 Some basic facts

We briefly review some basic concepts and outline the notation introduced in Huang et al. (2009). Full details are given in the Supplementary Material. Given two RNA sequences \(R = (R_1)_R\) and \(S = (S_1)_S\) (e.g. an antisense RNA and its target or an mRNA and its sRNA regulator) with \(N\) and \(M\) vertices, we label the vertices such that \(R_1\) is the 5’ end of \(R\) and \(S_1\) denotes the 3’ end of \(S\). The arcs of \(R\) and \(S\) then represent the respective, intramolecular base pairs. An arc is called exterior if it is of the form \(R_iS_j\) and interior, otherwise.

Fig. 1. Examples of RNA–RNA interactions structures. The primary interaction region(s) are highlighted in grey in the experimentally supported structural models from the literature: (A) ompA-MucC (Udekwu et al., 2005); (B) sodB-OxyS (Gehrman and Touati, 2004); (C) FliA-ompA (Argaman and Almogia, 2000). Hybridization probabilities computed by rip2 are annotated by black boxes for regions with a probability larger than 10%. In many cases, the computational predictions identify additional hybridization regions that may further stabilize the interaction.
Next, we formally define joint structures (Alkan et al., 2006, Chitaev et al., 2009). A joint structure, $(R,S,\alpha)$ is a specific subgraph of $G$ having both incident vertices in $V$. The subgraph of a joint structure $J(R,S,\alpha)$ induced by a pair of subsequences $(R_i,S_j)$ and $(S_j,R_i)$ is denoted by $J_{i,j}(\alpha)$. In particular, $J(R,S,\alpha) = J_{i,j}(\alpha)$ and $J_{i,j}(\alpha) \subseteq C_{i,j}(\alpha)$ if and only if $J_{i,j}(\alpha)$ is a subgraph of $C_{i,j}(\alpha)$ induced by $(R_i,S_j)$.

Fig. 2. (A) A zigzag, generated by $R_3S_1$, $R_1S_2$ and $R_3S_1$. (B) We partition the joint structure $J_{3,1,7}$ in segments and tight structures.

Fig. 3. The four basic types of TS. (A) $\triangleright \triangleright \triangleright \triangleright$ $J_{i,j}(\alpha)$ and its reciprocal $\triangleright \triangleright \triangleright \triangleright$ $J_{j,i}(\alpha)$; (B) $\triangleright \triangleright \triangleright \triangleright$ $R_1R_j, S_iS_j$ and $S_jR_i \triangleright \triangleright \triangleright \triangleright$; (C) $\triangleright \triangleright \triangleright \triangleright$ $R_1R_j, S_iS_j \triangleright \triangleright \triangleright \triangleright$ $R_jR_i \triangleright \triangleright \triangleright \triangleright$ $S_iS_j$; and $R_jR_i \triangleright \triangleright \triangleright \triangleright$. If $R_jR_i \triangleright \triangleright \triangleright \triangleright$, and $R_jR_i \triangleright \triangleright \triangleright \triangleright$. Then TS equals $\triangleright \triangleright \triangleright \triangleright$.

(a) $\triangleright \triangleright \triangleright \triangleright$ $J_{i,j}(\alpha)$ and its reciprocal $\triangleright \triangleright \triangleright \triangleright$ $J_{j,i}(\alpha)$; (B) $\triangleright \triangleright \triangleright \triangleright$ $R_1R_j, S_iS_j$ and $S_jR_i \triangleright \triangleright \triangleright \triangleright$; (C) $\triangleright \triangleright \triangleright \triangleright$ $R_1R_j, S_iS_j \triangleright \triangleright \triangleright \triangleright$ $R_jR_i \triangleright \triangleright \triangleright \triangleright$ $S_iS_j$; and $R_jR_i \triangleright \triangleright \triangleright \triangleright$. If $R_jR_i \triangleright \triangleright \triangleright \triangleright$, and $R_jR_i \triangleright \triangleright \triangleright \triangleright$. Then TS equals $\triangleright \triangleright \triangleright \triangleright$.

Next, we formally define joint structures (Alkan et al., 2006, Chitaev et al., 2009). A joint structure, $(R,S,\alpha)$ is a specific subgraph of $G$ having both incident vertices in $V$. The subgraph of a joint structure $J(R,S,\alpha)$ induced by a pair of subsequences $(R_i,S_j)$ and $(S_j,R_i)$ is denoted by $J_{i,j}(\alpha)$. In particular, $J(R,S,\alpha) = J_{i,j}(\alpha)$ and $J_{i,j}(\alpha) \subseteq C_{i,j}(\alpha)$ if and only if $J_{i,j}(\alpha)$ is a subgraph of $C_{i,j}(\alpha)$ induced by $(R_i,S_j)$.

Fig. 3. The four basic types of TS. (A) $\triangleright \triangleright \triangleright \triangleright$ $J_{i,j}(\alpha)$ and its reciprocal $\triangleright \triangleright \triangleright \triangleright$ $J_{j,i}(\alpha)$; (B) $\triangleright \triangleright \triangleright \triangleright$ $R_1R_j, S_iS_j$ and $S_jR_i \triangleright \triangleright \triangleright \triangleright$; (C) $\triangleright \triangleright \triangleright \triangleright$ $R_1R_j, S_iS_j \triangleright \triangleright \triangleright \triangleright$ $R_jR_i \triangleright \triangleright \triangleright \triangleright$ $S_iS_j$; and $R_jR_i \triangleright \triangleright \triangleright \triangleright$. If $R_jR_i \triangleright \triangleright \triangleright \triangleright$, and $R_jR_i \triangleright \triangleright \triangleright \triangleright$. Then TS equals $\triangleright \triangleright \triangleright \triangleright$.

(a) $\triangleright \triangleright \triangleright \triangleright$ $J_{i,j}(\alpha)$ and its reciprocal $\triangleright \triangleright \triangleright \triangleright$ $J_{j,i}(\alpha)$; (B) $\triangleright \triangleright \triangleright \triangleright$ $R_1R_j, S_iS_j$ and $S_jR_i \triangleright \triangleright \triangleright \triangleright$; (C) $\triangleright \triangleright \triangleright \triangleright$ $R_1R_j, S_iS_j \triangleright \triangleright \triangleright \triangleright$ $R_jR_i \triangleright \triangleright \triangleright \triangleright$ $S_iS_j$; and $R_jR_i \triangleright \triangleright \triangleright \triangleright$. If $R_jR_i \triangleright \triangleright \triangleright \triangleright$, and $R_jR_i \triangleright \triangleright \triangleright \triangleright$. Then TS equals $\triangleright \triangleright \triangleright \triangleright$.

(a) $\triangleright \triangleright \triangleright \triangleright$ $J_{i,j}(\alpha)$ and its reciprocal $\triangleright \triangleright \triangleright \triangleright$ $J_{j,i}(\alpha)$; (B) $\triangleright \triangleright \triangleright \triangleright$ $R_1R_j, S_iS_j$ and $S_jR_i \triangleright \triangleright \triangleright \triangleright$; (C) $\triangleright \triangleright \triangleright \triangleright$ $R_1R_j, S_iS_j \triangleright \triangleright \triangleright \triangleright$ $R_jR_i \triangleright \triangleright \triangleright \triangleright$ $S_iS_j$; and $R_jR_i \triangleright \triangleright \triangleright \triangleright$. If $R_jR_i \triangleright \triangleright \triangleright \triangleright$, and $R_jR_i \triangleright \triangleright \triangleright \triangleright$. Then TS equals $\triangleright \triangleright \triangleright \triangleright$.

Next, we formally define joint structures (Alkan et al., 2006, Chitaev et al., 2009). A joint structure, $(R,S,\alpha)$ is a specific subgraph of $G$ having both incident vertices in $V$. The subgraph of a joint structure $J(R,S,\alpha)$ induced by a pair of subsequences $(R_i,S_j)$ and $(S_j,R_i)$ is denoted by $J_{i,j}(\alpha)$. In particular, $J(R,S,\alpha) = J_{i,j}(\alpha)$ and $J_{i,j}(\alpha) \subseteq C_{i,j}(\alpha)$ if and only if $J_{i,j}(\alpha)$ is a subgraph of $C_{i,j}(\alpha)$ induced by $(R_i,S_j)$.

Fig. 3. The four basic types of TS. (A) $\triangleright \triangleright \triangleright \triangleright$ $J_{i,j}(\alpha)$ and its reciprocal $\triangleright \triangleright \triangleright \triangleright$ $J_{j,i}(\alpha)$; (B) $\triangleright \triangleright \triangleright \triangleright$ $R_1R_j, S_iS_j$ and $S_jR_i \triangleright \triangleright \triangleright \triangleright$; (C) $\triangleright \triangleright \triangleright \triangleright$ $R_1R_j, S_iS_j \triangleright \triangleright \triangleright \triangleright$ $R_jR_i \triangleright \triangleright \triangleright \triangleright$ $S_iS_j$; and $R_jR_i \triangleright \triangleright \triangleright \triangleright$. If $R_jR_i \triangleright \triangleright \triangleright \triangleright$, and $R_jR_i \triangleright \triangleright \triangleright \triangleright$. Then TS equals $\triangleright \triangleright \triangleright \triangleright$.

(a) $\triangleright \triangleright \triangleright \triangleright$ $J_{i,j}(\alpha)$ and its reciprocal $\triangleright \triangleright \triangleright \triangleright$ $J_{j,i}(\alpha)$; (B) $\triangleright \triangleright \triangleright \triangleright$ $R_1R_j, S_iS_j$ and $S_jR_i \triangleright \triangleright \triangleright \triangleright$; (C) $\triangleright \triangleright \triangleright \triangleright$ $R_1R_j, S_iS_j \triangleright \triangleright \triangleright \triangleright$ $R_jR_i \triangleright \triangleright \triangleright \triangleright$ $S_iS_j$; and $R_jR_i \triangleright \triangleright \triangleright \triangleright$. If $R_jR_i \triangleright \triangleright \triangleright \triangleright$, and $R_jR_i \triangleright \triangleright \triangleright \triangleright$. Then TS equals $\triangleright \triangleright \triangleright \triangleright$.
is unambiguous. The two panels of Fig. 5 contrast the grammars of zipl (Huang et al., 2009) and the hybrid grammar of zipl introduced here. In zipl, hybrids were immediately decomposed into individual external base pairs and their associated interior loops, so that individual hybrids were not tractable in a straightforward manner.

Let us now have a closer look at the energy evaluation of J_{l,h}. Each decomposition step in Fig. 4 results in substructures whose energies are assumed to contribute additively and generalized loops that can be evaluated directly. There are the following two basic scenarios:

(I) Interior Arc removal: the first type of decomposition is derived from the decomposition of TS of Huang et al. (2009). Most of the decomposition operations in Procedure (b) displayed in Fig. 4 can be viewed as the "removal" of an arc (corresponding to the closing pair of a loop in secondary structure folding) followed by decomposition. Both, loop type as well as the subsequent decomposition steps depend on the newly exposed structural elements. Following the approach of Zuker and Sankoff (1984) for secondary structures, we treat the loop-decomposition problem by introducing additional matrices. Without loss of generality, we can assume that we open an interior base pair R_{l,h}.

The set of base pairs on R_{l,h} consists of all interior pairs R_{l,h} with |l-p|<|j| and all exterior pairs S_{h,k} with |l-p|>|j|. An interior arc is exposed on R_{l+1,j-1} if and only if it is not exposed by any interior arc in R_{l,j}. An exterior arc is exposed on R_{l+1,j-1} if and only if it is not a descendant of any interior arc in R_{l+1,j-1}.

Given R_{l,h}, the arcs exposed on R_{l+1,j-1} correspond to the base pairs immediately interior of R_{l,h}. Let us write E_{0,l,h}:=E_{l,h}(R_{l,h}), for this set of 'exposed base pairs' and its subsets of interior and exterior arcs. As in secondary structure folding, the loop type is determined by E_{0,l,h}=E_{l,h} as follows: E_{l,h}=0, hairpin loop; E_{l,h}=E_{l,h} and |E_{l,h}|=1, interior loop (including bulge and stack); E_{l,h}=E_{l,h} and |E_{l,h}|≥2, multi-branch loop; E_{l,h}=E_{l,h} kissing-hairpin loop; E_{l,h}≥2, general kissing loop.

This picture needs to be refined even further since the arc removal is coupled with further decomposition of the interval R_{l+1,j-1}. This prompts us to distinguish TS and DTS with different classes of exposed base pairs on one or both strands. It will be convenient, furthermore, to include information on the type of loop in which it was found. A TS J^{TDT}_{l,h,k} is of type E, if S_{l,h} is not enclosed in any base pair (T^{TDT}_{l,h,k}). Suppose J^{TDT}_{l,h,k} is located immediately interior to the closing pair S_{l,h}. Then, if the loop closed by S_{l,h} is a multi-loop, then J^{TDT}_{l,h,k} is of type M (T^{TDT}_{l,h,k}). If S_{l,h} is contained in a kissing loop, we distinguish the types F and K, depending on whether or not E_{l,h}≠0.

Analogously, there are in total four types of a hybrid J^{QDT}_{l,h,k}, i.e., J^{QDT}_{l,h,k} EE, J^{QDT}_{l,h,k} KE, J^{QDT}_{l,h,k} KE, J^{QDT}_{l,h,k} KK.

(II) Block decomposition: the second type of decomposition is the splitting of joins into "blocks". Here, the hybrid grammar differs from that of Huang et al. (2009) in two ways. First, we use the hybrid as a new block of the grammar, decomposing a hybrid by removing its exterior arcs in parallel simultaneously starting from the right. Second, we split a single structure into blocks via alternating decompositions of RTS and DTS as shown in the Procedure (a) of Fig. 4.

In order to guarantee the maximality hybrids, we observe that the RTS's R_{l+1,h,k}, R_{l,h+k}, R_{l+1,h+k} and R_{l,h+k} can appear in two scenarios, depending on whether or not there exists an exterior arc S_{h,k}, such that R_{l+1,h,k} and S_{h,k} are isolated segments. In case such an exterior arc exists, we say the RTS is of type (B) or (A), otherwise. Similarly, a DTS, R_{l,h+k}, R_{l,h+k} or R_{l+h,k} is of type (B) or (A) depending on whether R_{l,h,k} is an exterior arc. In Fig. 6A, we display the decomposition of J^{QDT}_{l,h,k} into hybrids and RTSs of type (A) and in Fig. 6B, we display the decomposition of J^{QDT}_{l,h,k} into secondary structure segments and DTS accordingly.

Suppose J^{QDT}_{l,h,k} is a DTS contained in a kissing loop, that is, we have either E_{l,h}≠0 or E_{l,h}≠0. Without loss of generality, we may assume E_{l,h}≠0. Then, at least one of the two 'blocks' belonging to E_{l,h} labeled by K or F, otherwise, see Fig. 6A.

2.3 Forward recursions

The computation of the partition function proceeds 'from the inside to the outside', see Equation (3). The recursions are initialized with the energies of individual external base pairs and empty secondary structures on subsequences of length up to 4. In order to differentiate multi- and kissing-loop contributions, we introduce the partition functions Q_{l,h} and Q_{l,h,k}. Here, Q_{l,h} denotes the partition function of secondary structures on R_{l,h} or S_{l,h} having at least one arc contained in a multi-loop. Similarly, Q_{l,h,k} denotes the partition function of secondary structures on R_{l,h,k} or S_{l,h,k} in which at least one arc is contained in a kissing loop. Let J^{TDT}_{l,h,k} be the set of substructures J^{TDT}_{l,h,k}⊂J_{l,h,k}ε, isolated from some joint structure J_{l,h,k}ε, such that J^{TDT}_{l,h,k} appears in TDS as an interaction structure of type ε (HT, RT, TV, C, Cε) with loop-subtypes Y_{ε}ε, ε∈(M,K,F) on the subintervals R_{l,h}, S_{l,h} and ε∈{A,B}. Let J^{QDT}_{l,h,k} denote the partition function of the set J^{QDT}_{l,h,k}. All recursions for Q^{QDT}_{l,h,k} represent a reformulation of the hybrid grammar specified in Fig. 4.

For instance, the recursion for Q^{QDT}_{l,h,k} displayed in Fig. 6A is given by:

\[ Q^{QDT}_{l,h,k} = \sum_{i,j} Q^{QDT}_{l+1,h+h,1+i} + Q^{QDT}_{l,h+k+i} + Q^{QDT}_{l,h+k+i} + Q^{QDT}_{l,h+k+i} + Q^{QDT}_{l,h+k+i} + Q^{QDT}_{l,h+k+i} + Q^{QDT}_{l,h+k+i} \]

where the corresponding recursion for Q^{QDT}_{l,h,k} is

\[ Q^{QDT}_{l,h,k} = \sum_{i,j} Q^{QDT}_{l,h,k+i} + Q^{QDT}_{l,h+k+i} + Q^{QDT}_{l,h+k+i} + Q^{QDT}_{l,h+k+i} \]

(2)
2.4 Hybrid probabilities

Since the probabilities of individual base pairs are not independent, it is not possible to compute the probabilities for particular hybrids directly from them. Hybrid probabilities thus cannot be obtained in a simple way from the backward recursions described by Huang et al. (2009).

Given two RNA sequences, our notion of probability is based on the ensemble of all possible joint interaction structures. Let \( \mathcal{Q} \) denote the partition function of all these joint structures that can be formed by two input RNA sequences. The probability of a fixed joint structure \( \mathcal{J}_{i,k} \) is given by

\[
\mathbb{P}(\mathcal{J}_{i,k}) = \frac{\mathcal{Q}(\mathcal{J}_{i,k})}{\mathcal{Q}}
\]

In difference to the computation of the hybrid-particle function 'from the inside to the outside' (IO), the computation of probabilities of specific substructures is obtained 'from the outside to the inside'. The same principle applies to the computation of base pairing computation of base pairing probabilities of secondary structures (McCaskill, 1990) and joint structures (Huang et al., 2009).

Let \( J_{i,k} \in \mathcal{J} \), with associated decomposition tree \( T(J) \) and let \( \mathcal{J}_{i,k} \subset J_{i,k} \) denote the set of all joint structures \( J \) such that \( \mathcal{J}_{i,k} \) is contained in the decomposition tree \( T(J) \). Then we have, by construction,

\[
\mathbb{P}(\mathcal{J}_{i,k}) = \sum_{J_{i,k} \in \mathcal{J}} \mathbb{P}(J_{i,k}).
\]

Following the (IO)-paradigm, the probability of a parent structure, \( \mathbb{P}(J_{i,k}) \), is computed prior to the calculation of \( \mathbb{P}(\mathcal{J}_{i,k}) \). The conditional probability \( \mathbb{P}(J_{i,k} | \mathcal{J}_{i,k}) \) equals \( Q(J_{i,k} | \mathcal{J}_{i,k}) / Q(\mathcal{J}_{i,k}) \), where \( Q(\mathcal{J}_{i,k}) \) is the partition function of \( \mathcal{J}_{i,k} \) and \( Q(J_{i,k} | \mathcal{J}_{i,k}) \) is the partition function of all those \( \mathcal{J}_{i,k} \) that have in addition \( J_{i,k} \) as a child in their parse trees. Consequently, \( \mathbb{P}(\mathcal{J}_{i,k}) \) can inductively be computed by summing over all probabilities \( \mathbb{P}(J_{i,k}) \), i.e.

\[
\mathbb{P}(\mathcal{J}_{i,k}) = \sum_{J_{i,k} \in \mathcal{J}} \mathbb{P}(J_{i,k}).
\]

Let \( \mathbb{P}(J_{i,k} | \mathcal{J}_{i,k}) \) denote the probability of the set of substructures \( J \) such that the specific hybrid substructure, \( \mathcal{J}_{i,k} \), appears in the decomposition tree \( T(J) \) and \( \mathcal{J}_{i,k} \in \mathcal{J} \). Since each joint structure \( \mathcal{J}_{i,k} \) is either one of the four types \( \mathcal{J}_{i,k} \), \( \mathcal{J}_{i,k} \), \( \mathcal{J}_{i,k} \), \( \mathcal{J}_{i,k} \), \( \mathcal{J}_{i,k} \) or \( \mathcal{J}_{i,k} \), we arrive at

\[
\mathbb{P}(J_{i,k} | \mathcal{J}_{i,k}) = \mathbb{P}(\mathcal{J}_{i,k} | \mathcal{J}_{i,k}) + \mathbb{P}(\mathcal{J}_{i,k} | \mathcal{J}_{i,k}) + \mathbb{P}(\mathcal{J}_{i,k} | \mathcal{J}_{i,k}).
\]

We remark, that, by construction, for \( [i,j] \neq [h,t] \), the hybrid probabilities \( \mathbb{P}(\mathcal{J}_{i,k} | \mathcal{J}_{i,k}) \) quantify disjoint classes of joint structures. This is a consequence of the maximality of hybrids, which implies that, for fixed interval \( [i,j] \), each \( [h,t] \) corresponds to a unique hybrid \( \mathcal{J}_{i,k} \). Based on the notion of hybrid probability, we can introduce

\[
\mathbb{P}(i,j) = \sum_{k \in [i,j]} \mathbb{P}(\mathcal{J}_{i,k} | \mathcal{J}_{i,k})
\]

which, according to the above, the probability of the target site \( [i,j] \) and furthermore

\[
\pi(x) = \sum_{j \in [i,j]} \mathbb{P}(\mathcal{J}_{i,k} | \mathcal{J}_{i,k})
\]

measuring, for each base \( i \) in \( R \) the probability that \( i \) is contained in a hybrid. A particular instructive observable is the interaction base pairing matrix, given by

\[
\pi_{i,k} = \sum_{j \in [i,j]} \sum_{p \in [j.p'] \in R} \rho_{p,j'}
\]

Fig. 7. Hybrid probability: the maximality of hybrids implies that—although the intervals \( [h,t] \) and \( [h,t] \) overlap—they belong to two distinct hybrids (gray).

2.5 Boltzmann sampling

A dynamic programming scheme for the computation of a partition function implies a corresponding stochastic backtracking procedure that can be used to sample from the associated distribution (Tucker et al., 1996). The usefulness of this approach for RNA secondary structures is discussed by Ding and Lawrence (2003). The same ideas can of course also produce representative samples from the Boltzmann equilibrium distribution of RNA interaction structures (Fig. 8).

The basic data structure of the algorithm is a stack \( A \) that stores tuples of the form \( (i,j,h,t,F) \) describing a pair of intervals \( [i,j] \) in \( R \) and \( [h,t] \) in \( S \) and the type \( F \) of the—not further specified—joint structure formed by the two intervals. The stack \( A \) is initialized with \( (1,1,1,1,1,7) \) where \( 7 \) denotes the unspecified type, guides the backtracking which is complete as soon as \( A \) is empty. A list \( L \) is used to collect the interior and exterior arcs and unpaired bases generated by the decompositions and eventually define the sampled interaction structure. In the first step, \( (1,1,1,1,1,7) \) is decomposed according to the grammar in Fig. 4 into either (i) a pair of secondary structures, or (ii) a RTS \((i,j,N,M,RTEE)\) with probabilities derived as explained above. Depending on the stochastic choice, we push either (i) \((1,1,1,1,1,7)\) or (ii) \((1,1,1,1,1,7)\) into the stack \( A \).

Given \( A \) and \( L \), we can associate a probability by considering the decomposition of the particular type of joint structure. For instance, suppose we have extracted \((i,j,h,t,DEYKB)\) from stack \( A \), see Fig. 6. Then, the probabilities for continuing with one of the five decompositions displayed in Fig. 6, for each position of the break points \( i \in [i,j] \) and \( h \in [h,t] \), is given by

\[
\rho_{i,h} = Q_{i,h}(KE) + Q_{i,h}(KE) + Q_{i,h}(KE) + Q_{i,h}(KE).
\]

179
We presented here a modified and improved unambiguous grammar routine. Applications, it is suitable for investigating the fine details rip2 hardware. While QDT rip2 the output produced by a stochastic backtracing facility. Fig. 9 gives an example of about a third. In the Supplementary Material, we contrast the computational efforts, in particular the memory consumption, for the RIP. Compared with the Supplementary Material. It comprises 9 4D-arrays Qmolecules. The complete set of recursions is compiled in Section 3 and show that hybrids (as opposed to TS, RTS or DTS) are rip1 180 2009), it reduces 20 4D-arrays QRT of various types, 20 4D-arrays QRT for RTS and 20 4D-arrays QRT for DTS. The implementation has been complemented by a stochastic backtracing facility. Fig. 9 gives an example of the output produced by rip2 (see also Supplementary Material, Fig. 4). Despite algorithmic improvements, rip2 still requires quite substantial computational resources for practical applications. rip2 is in practice limited to problem sizes of \( N_1 + N_2 \leq 350 \) on current hardware. While rip2 is still not an efficient tool for large-scale routine applications, it is suitable for investigating the fine details of particular interactions. Future work will thus focus on controlled approximations with the aim of a drastic reduction of both: CPU and memory consumption.

The major advantage of stochastic sampling is that it provides a generic and convenient means to estimate quantities that cannot be easily computed directly by backwards recursion (Ding and Lawrence, 2003). Both, the ompA-MlcA and sodB-RhyB complexes show a primary, highly likely, hybrid region and several additional less stable points of contact, see Fig. 10. In these examples, it
Target prediction and Boltzmann sampling for RIP

is of interest to investigate in detail how the putative interaction regions influence each other: is the binding cooperative so that the major hybrids in Fig. 10 are positively correlated, or do they constitute mutually exclusive contacts? Once a sufficiently large Boltzmann sample is obtained, we can easily compute, e.g. correlations $\rho_{pq}$ between indicator variables $P$ and $Q$ that measure the existence of external base pairs in two different hybrids. Fig. 10C provides examples, showing that there are strong correlations between hybridization regions. These multiple contacts can contribute substantially to the total interaction energy.

**ACKNOWLEDGEMENTS**

We want to thank Sven Findeiß for discussions. We are grateful to Sharon Selzo of the Modular and BICoC Benchmark Center, IBM and Kathy Treng of IBM Life Sciences Solutions Enablement. Their great support was vital for all computations presented here.

Funding: 973 Project of the Ministry of Science and Technology; PC3R7T Project of the Ministry of Education; National Science Foundation of China (to C.M.R. and his lab); Deutsche Forschungsgemeinschaft under the auspices of SPP-1258 'Small RNA'.

Forschungsgemeinschaft under the auspices of SPP-1258 'Small RNA' interaction. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 6, 682–688.

**REFERENCES**

Akutsu,T. (2000) Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Disc. Appl. Math.*, 104, 45–62.

Alkan,C. et al. (2000) fhlA et al. Chitsaz,H. et al. Bernhart,S. Benne,R. (1992) RNA editing in trypanosomes. the use of guide RNAs. Banerjee,D. and Slack,F. (2002) Control of developmental timing by small temporal RNAs. *Disc. Appl. Math.*, 119–129.

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

McManus,M.T. and Sharp,P.A. (2002) Gene silencing in mammalian cells using small interfering RNA. *Nat. Rev.*, 3, 737–747.

Mcmahon,S. (2009) On the approximation of optimal structures for RNA-RNA interaction. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 6, 682–688.

Mcmahon,S. et al. (2006) Thermodynamics of RNA-RNA binding. *Bioinformatics*, 22, 1177–1182.

Nickolay,U. et al. (2008) Transcriptional control by RNA-RNA interaction: improved computation of RNA-RNA binding thermodynamics. In Elloumi,M. et al. (eds) Bioinformatics Research and Development — *BIRD 2008*, Vol. 13 of Communication in Computer and Information Science. Springer, Berlin, pp. 314–327.

Niederhagen,F. and Vogel,J. (2007) Sensory and regulatory RNAs in prokaryotes: A new german research focus. *RNA Biol.*, 4, 160–164.

Pevzovitch,D. (2004) 2312: intermolecular RNA interaction search. Proc. Genome Inform., 15, 92–101.

Pevzovitch,D. (2004) 2312: intermolecular RNA interaction search. Proc. Genome Inform., 15, 92–101.

Rivas,E. and Eddy,S.R. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, 285, 2053–2068.

Sabatier,R. et al. (2009) Fast prediction of RNA-RNA interaction. In Proceedings of the 9th Workshop on Algorithms in Bioinformatics (WABI), Vol. 5724 of Lecture Notes in Computer Science. Springer, pp. 261–272.

Tacker,M. et al. (1996) Algorithm independent properties of RNA structure prediction. *Bioinformatics*, 12, 155–161.

Tafer,H. et al. (2009) Effizienzefficient target prediction for box H/ACA snoRNAs. *Bioinformatics*, University of Leipzig. Available at http://www.bioinf.uni-leipzig.de/Publications/PREPRINTS/09-025.pdf.

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

UdGorver,M. et al. (2009) High-performance translation of OmpA synthesis is mediated by an antisense RNA. *Genes Dev.*, 19, 2355–2366.

Urban,J.H. and Vogel,J. (2007) Translational control and target recognition by Escherichia coli small RNAs on sRNA. *Nucleic Acids Res.*, 35, 1018–1037.

Zaken,M. and Sankoff,D. (1984) RNA secondary structures and their prediction. *Bull. Math. Biol.*, 46, 591–621.