Structural bioinformatics

Partition function and base pairing probabilities for RNA–RNA interaction prediction

Fenix W. D. Huang¹, Jing Qin¹, Christian M. Reidys¹,²,⁶ and Peter F. Stadler³–⁷

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

Received on April 16, 2009; revised on June 21, 2009; accepted on August 2, 2009

ABSTRACT

Motivation: The RNA–RNA interaction problem (RIP) consists in finding the energetically optimal structure of two RNA molecules that bind to each other. The standard model allows secondary structures in both partners as well as additional base pairs between the two RNAs subject to certain restrictions that ensure that RIP is solvable by a polynomial time dynamic programming algorithm. RNA–RNA binding, like RNA folding, is typically not dominated by the ground state structure. Instead, a large ensemble of alternative structures contributes to the interaction thermodynamics.

Results: We present here an O(N²) time and O(N) dynamics programming algorithm for computing the full partition function for RIP which is based on the combinatorial notion of ‘tight structures’. Albeit equivalent to recent work by H. Chitsaz and collaborators, our approach in addition provides a full-fledged computation of the base pairing probabilities, which relies on the notion of a decomposition tree for joint structures. In practise, our implementation is efficient enough to investigate, for instance, the interactions of small bacterial RNAs and their target RNAs.

Availability: The program zrip is implemented in C. The source code is available for download from http://www.combinatorics.cn/cbpc/zrip.html and http://www.bioinf.uni-leipzig.de/Software/zrip.html.

Contact: duck@santafe.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

RNA–RNA interactions constitute one of the fundamental mechanisms of cellular regulation. In an important subclass, small RNAs specifically bind a larger (m)RNA target. Examples include the regulation of translation in both prokaryotes (Narberhaus and Vogel, 2007) and eukaryotes (Banerjee and Slack, 2002; McManus and Sharp, 2002); the targeting of chemical modifications (Bachellerie et al., 2002); and insertion editing (Benne, 1992) and transcriptional control (Kugel and Goodrich, 2007). The common theme in many RNA classes, including miRNAs, siRNAs, sRNAs, gRNAs and snoRNAs is the formation of RNA–RNA interaction structures that are more complex than simple sense–antisense interactions. The ability to predict the details of RNA–RNA interactions in terms of the thermodynamics of binding in its structural consequences is a necessary prerequisite for understanding RNA-based regulation mechanisms. The exact location of binding and the subsequent impact of the interaction on the structure of the target molecule can have profound biological consequences. In the case of sRNA–mRNA interactions, these details decide whether the sRNA is a positive or negative regulator of transcription depending on whether binding exposes or covers the Shine–Dalgarno sequence (Majdalani et al., 2002; Sharma et al., 2007). Similar effects have been observed while using artificially designed opener and closer RNAs that regulate the binding of the Holl protein to human mRNAs (Hackermüller et al., 2005; Meisner et al., 2004).

In its most general form, the RNA–RNA interaction problem (RIP) is NP-complete (Alkan et al., 2006; Meineche, 2007). The argument for this statement is based on an extension of the work of Akutsu (2000) for RNA folding with pseudoknots. 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 second major problem concerns the energy parameters since the standard loop types (hairpins, internal and multi-loops) are insufficient; for the additional types, such as kissing hairpins, experimental data are virtually absent. Tertiary interactions, furthermore, are likely to have a significant impact.

Several restricted versions of RNA–RNA interaction have been considered in the literature. The simplest approach concatenates the two interacting sequences, essentially employing a slightly modified secondary structure folding algorithm. The algorithms RNACOFOLD (Bernhart et al., 1996; Hofacker et al., 1994), pairfold (Andronescu et al., 2005) and SUFFPACK (Dirks et al., 2007) belong to this class. One major shortcoming 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 approach is to neglect all internal base pairings and to compute the minimum free energy (mfe) of the secondary structure for their hybridization under this constraint. For instance, RNAup (Rehmmsmeier et al., 2004) follow this paradigm. mfeRNA (Mückstein et al., 2006) and intaRNA (Busch et al., 2008) restrict interactions to a single interval that remains unpaired in the secondary structure for each partner. These models have proved particularly useful for bacterial sRNA–mRNA interactions. Due to the highly conserved interaction motif, sRNA-target interaction structures can be dealt with efficiently using specialized tools (Tafer et al., 2009).

Pervouchine (2004) and Alkan et al. (2006) independently derived and implemented mfe folding algorithms for predicting the joint secondary structure of two interacting RNA molecules with polynomial time complexity. In their model, a ‘joint structure’ means that the intramolecular structures of each molecule are pseudoknot-free, the intermolecular binding pairs are non-crossing and there exist no so-called ‘ziggzags’, see Figures 1A and 2A for examples of the ‘joint structures’. The optimal ‘joint structure’ can be computed in $O(N^3)$ time and $O(N^2)$ space by means of dynamic programming.

Recently, Chitsaz et al. (2009) presented ripRNA, a tool that uses dynamic programming algorithm to compute the partition function of ‘joint structures’, also in $O(N^3)$ time. The algorithmic cores of the forward recursions of ripRNA and zip were developed independently. Albeit differing in design details, they are equivalent. In addition, we identified here a basic data structure that forms the basis for computing additional important quantities such as the base pairing probability matrix and probabilities of hybrid formations (see Huang et al., 2009, for the latter). Further differences between the two approaches will be discussed in Section 5.

The key innovation for passing from the mfe folding of Alkan et al. (2006) to the partition function is an unambiguous grammar with which each interaction structure can be generated. Then, the computation of the partition function follows McCaskill’s approach for RNA secondary structures (McCaskill, 1990). The key idea is to identify a certain subclass of interaction structures that serve as building blocks in a recursive decomposition generalizing the loop decomposition of secondary structures. These are the ‘tight structures’, a generalization of the subsecondary structures enclosed by a unique closing pair. In the following two sections, we first derive a grammar that allows the unambiguous parsing of zigzag-free interaction structures, thus forming the basis for the computation of the partition function in $O(N^3)$ time and $O(N^2)$ memory, corresponding the mfe algorithm of Alkan et al. (2006). Then, we proceed by deriving the recursions for the base pairing probabilities, which are based on a conceptual reversing of the production rules. Indeed, one has to compute the pairing probabilities by explicitly ‘tracing back’ all contributing joint structures. The output of zip consists of the partition function, the base pairing probability matrix and the joint structure predicted by the maximal weighted matching (mwm) algorithm (in terms of the base pair probabilities) (Cary and Stormo, 1995; Gabow, 1973) and the most likely hybrid loops.

The soId-RhyB interaction structure (Geissmann and Touati, 2004) is a well-known paradigmatic example with a unique interaction region (Fig. 1). The zip software predicts this interaction region correctly. Results obtained with other algorithms deviate noticeably from the known structure (see Supplementary Fig. S2).

To date, only a handful of interaction structures are known that are more complex than those covered by intaRNA/ripRNA. The best known example is the repression of fhaA by OxyS RNA, which involves two widely separated kissing-hairpin loops (Argaman and Altuvia, 2000). In Figure 2, we display the natural interaction structure as well as the output of zip, which predicts two distinct interaction regions. The left one coincides exactly with the published structure, while the right one differs by only two base pairs. We shall return to the fhaA–OxyS prediction in more detail in Section 5.

2 JOINT STRUCTURES

Given two RNA sequences R and S (e.g. an antisense RNA and its target) with N and M vertices, we index the vertices such that $R_i$ is the 5’ end of $R$ and $S_j$ denotes the 3’ end of $S$. The edges of $R$ and $S$ represent the intramolecular base pairs. A prestructure $G(R, S, I)$ is a graph with the following properties:

1. $R$ and $S$ are secondary structures (each nucleotide being paired with at most one other nucleotide via hydrogen bonds, without internal pseudoknots);
2. I is a set of arcs of the form $R_i S_j$ without pseudoknots, i.e. if $R_i S_j \in I$, then $i < \ell_i, j < \ell_j$ holds.

An arc is called exterior, if it is of the form $R_S$, and interior, otherwise. Let $G$ be a graph and $V$ be a subset of $G$-vertices. The (induced) subgraph of $G$ induced by $V$ has vertex set $V$ and contains all $G$-edges having both incident vertices in $V$. In particular, we use $S_i(j, j)$ to denote the subgraph of the prestructure $G(R, S, I)$ induced by $S_i, S_i + 1, \ldots, S_j$, where $S_i(j,i)=S_j$ and $S_{i}(i-1)=\emptyset$. In absence of interactions, a prestructure is a pair of induced secondary structures on $R$ and $S$, which we will refer to as a pair of segments. A segment $S_i(j, j)$ is called maximal if there is no segment $S_i(j, j)$ strictly containing $S_i(j, j)$.

Fig. 1. (A) The natural structure of soId–RhyB (Geissmann and Touati, 2004). (B) The base pairing probability matrix (McCaskill, 1990) generated via zip. This matrix represents all potential base pairs of the soId–RhyB structure as squares, whose area is proportional to their respective probability. Inter- and intramolecular base pairs are depicted by the gray upper right rectangle and the two white triangles, respectively. (C) ‘Zoom’ into the most likely interaction region as predicted by zip. All base pairs of the hybrid are labeled by their probabilities.
The base pairing probability matrix (see the caption of Fig. 1 for notations).

Fig. 2. (A) The natural structure of fhlA-OxyS (Chitsaz et al., 2009). (B) The base pairing probability matrix (see the caption of Fig. 1 for notations). (C) ‘Zoom’ into the two distinct, most likely interaction regions, as predicted by xip.

Fig. 3. (A) Ancestors and parents: for the exterior arc RiSi, we have the following ancestor sets A0(RiSi) = {R|RiRiRjSi} and A0(RjSi) = {R|RjRiSiSj}. In particular, RiRj and SjSi are the R-parent and S-parent respectively. (B) Subsumed and equivalent arcs: RiRj subsumes SjSi and SjSi. Furthermore, RiRj is equivalent to SjSi.

Fig. 4. (A): A zigzag, generated by RjSj, RjRj and RjSj. (B): We partition the joint structure J1,2,3,2,3 in segments and tight structures.

An interior arc RiRj is an R-ancestor of the exterior arc RiSi if i < j < k. Analogously, SjSi is an S-ancestor of RiSi if i < j < k. The sets of R- and S-ancestors of RiSi are denoted by A0(RiSi) and A0(Si), respectively. We will also refer to RiSi as a descendant of RjRk and SiSk in this situation. The R- and S-ancestors of RiSi with minimum arc-length are referred to as R- and S-parents (Fig. 3A). Finally, we call RiRj and SjSi dependent if they have a common descendant and independent, otherwise. Suppose there is an exterior arc RjRj with ancestors RjRj and SjSi. Then RjRj is subsumed in SjSi, if for any RjSi ∈ Γ, i < k < j implies i < k′ < j′ (Fig. 3B). If RiSi is subsumed in SjSi and vice versa, we call these arcs equivalent. A zigzag, is a subgraph containing two dependent interior arcs RiRj and SjSi neither one subsuming the other (Fig. 4A).

Fig. 5. From left to right: tights of type □, △, ▽, and ▆.

A joint structure, J(R,S1), is a zigzag-free prestructure (Fig. 4B). Joint structures are exactly the configurations that are considered in the maximum matching approach of Perovshiche (2004), in the energy minimization algorithm of Alkan et al. (2006), and in the partition function approach of Chitsaz et al. (2009). The subgraph of a joint structure J(R,S1) induced by a pair of subsequences [R|RjRiRjSi] and [S|SiSjSi] is denoted by J1,2,i,j. In particular, J(R,S1) = J1,2,0,0. We say RjRj(SjSiRjSi) ∈ J1,2,i,j, if and only if RjRj(SjSiRjSi) is an edge of the graph J1,2,i,j. Furthermore, J1,2,i,j ⊂ J1,2,i,j if and only if J1,2,i,j is a subgraph of J1,2,i,j induced by (R1,...,Ri) and (Si,...,Sj).

We define next a tight structure (TS). Given a joint structure, J1,2,i,j, its tight J1,2,i,j is either a single exterior arc RjSiRjSi (in the case a′ = b′ and c′ = d′), or the minimal block centered around the leftmost and rightmost exterior arcs a,b,c,d, (possibly being equal) and an interior arc subsuming both, i.e., J1,2,i,j is tight in J1,2,i,j if it has either an arc RjRj or SjSi if a′ ≠ b′ or c′ ≠ d′.

More formally, let J1,2,i,j be contained in J1,2,i,j with rightmost and leftmost exterior arc RjSiRjSi and RjRj, and let M be the set of RjSiRjSi-ancestors in J1,2,i,j with maximal length. Then J1,2,i,j is tight in J1,2,i,j if

(1) for M = ∅: J1,2,i,j = [RjSiRjSi];

(2) for M = [RjRj | J1,2,i,j = J1,2,i,j, where c′ is the origin (left) of the S-ancestor of RjSiRjSi with maximal length (or if there is none).

The case M = {SjSi, SjSi} is analogous.

(3) for M = [RjRj, SjSi, SjSi]: suppose RjRj subsumes SjSi. Then J1,2,i,j = J1,2,i,j, where x1 is the origin of the S-ancestor of RjSiRjSi with maximal length (or if there is none). In particular, J1,2,i,j = J1,2,i,j, when RjRj is equivalent with SjSi. The case, where SjSi, RjRj is analogous.

In the following, a TS is denoted by J1,2,i,j. If J1,2,i,j is tight in J1,2,i,j, then we call J1,2,i,j its envelope. By construction, the notion of TS is depending on its envelope. There are only four basic types of TS (Fig. 5):

○: [RjSiRjSi] = J1,2,i,j and i = j = t;

◇: RjRj ∈ J1,2,i,j, and SjSi ∉ J1,2,i,j;

▽: [RjRj, SjSi] ∈ J1,2,i,j;

◆: SjSi ∈ J1,2,i,j and RjRj ∉ J1,2,i,j.

In the Supplementary Material we prove:

Proposition 2.1. Let J1,2,i,j be a joint structure. Then

(1) any exterior arc RjSi in J1,2,i,j is contained in a unique J1,2,i,j TS;

(2) J1,2,i,j decomposes into a unique collection of J1,2,i,j TS and maximal segments.

Given a TS J1,2,i,j, we introduce double tight structures as maximal substructure whose distinct leftmost and rightmost blocks are tight. By construction, therefore, each double tight structure (DTS) contains at least two tights.

More formally, given a TS J1,2,i,j, a double-tight structure, J1,2,i,j, in J1,2,i,j, where b ≤ k ≤ j and t ≤ l ≤ d ≤ e. Furthermore, two TS J1,2,i,j and J1,2,i,j in J1,2,i,j such that

J1,2,i,j = J1,2,0,0, where a = b ≤ j and t ≤ l ≤ d ≤ e.

(1)
Here, the disjoint union ∪ refers to both the vertex and arc sets of the joint structures (Fig. 6). The case of a DTS, $\mathcal{J}_{DTS}$, within a TS, $\mathcal{J}_{TS}$, is defined accordingly. By abuse of terminology, we simply use $\mathcal{J}^{\mathcal{J}_{DTS}}_{\mathcal{J}_{TS}}$ or $\mathcal{J}^{\mathcal{J}_{DTS}}_{\mathcal{J}_{TS}}$.

With the help of DTS, as illustrated in Figure 7 (Procedure (b)), we decompose a TS as follows: Let $\mathcal{J}_{TS}$ be a TS of type $\triangle$ and let $R_0$, $S_0$, and $R_0$, $S_0$ be the leftmost and rightmost exterior arcs in $\mathcal{J}_{TS}$, and $i + 1 \leq j \leq i - 1$. Then $\mathcal{J}_{TS}$ decomposes into

\[
\begin{align*}
R(i + 1, i + 1) & = \mathcal{J}^{\mathcal{J}_{DTS}}_{\mathcal{J}_{TS}}[R(i + 1, i - 1)], \\
R(i + 1, i - 1) & = \mathcal{J}^{\mathcal{J}_{DTS}}_{\mathcal{J}_{TS}}[R(i, i - 1)],
\end{align*}
\]

where $\mathcal{J}^{\mathcal{J}_{DTS}}_{\mathcal{J}_{TS}}[R(i + 1, i - 1)]$ denotes the unique TS in $\mathcal{J}_{TS}$ containing the leftmost arc $R_0$. Analogously, in case of a TS $\mathcal{J}_{TS}$ with leftmost and rightmost exterior arcs $R_0$, $S_0$, and $R_0$, $S_0$, and $i + 1 \leq r_1 \leq s_1 \leq i - 1$, $\mathcal{J}_{TS}$ can be decomposed in the form

\[
\begin{align*}
S(r + 1, r + 1) & = \mathcal{J}^{\mathcal{J}_{DTS}}_{\mathcal{J}_{TS}}[S(r + 1, r - 1)], \\
S(r + 1, r - 1) & = \mathcal{J}^{\mathcal{J}_{DTS}}_{\mathcal{J}_{TS}}[S(r, r - 1)],
\end{align*}
\]

where $\mathcal{J}^{\mathcal{J}_{DTS}}_{\mathcal{J}_{TS}}[S(r + 1, r - 1)]$ denotes the unique TS in $\mathcal{J}_{TS}$ containing the rightmost arc $S_0$.

For a TS, $\mathcal{J}_{TS}$, with $i + 1 \leq l \leq s \leq j - 1$ we analogously derive

\[
\begin{align*}
\mathcal{J}_{TS}(i + 1, j - 1) & = \mathcal{J}^{\mathcal{J}_{DTS}}_{\mathcal{J}_{TS}}[\mathcal{J}(i + 1, j - 1)],
\end{align*}
\]

where $\mathcal{J}^{\mathcal{J}_{DTS}}_{\mathcal{J}_{TS}}[\mathcal{J}(i + 1, j - 1)]$ denotes the unique TS in $\mathcal{J}_{TS}$ containing the arc $\mathcal{J}(i, j)$.

Proposition (2.1) and Equation (1–4) establish, for each joint structure, a unique decomposition into interior and exterior arcs.

3 THE PARTITION FUNCTION

3.1 Refined decomposition

The unique decomposition of TS would formally suffice to construct a partition function algorithm. Indeed, each decomposition step, such as Equation (1–4), corresponds to a multiplicative recursion relation for the partition function associated with the joint structures. However, this would result in an unwieldy expensive implementation. The reason is the multiple break points $a, b, c, d, \ldots$ each of which corresponds to a nested for-loop.

We therefore introduce a refined decomposition that reduces the number of break points. For this purpose, we call a joint structure right-tight if its rightmost block is a TS. We adopt the point of view of Algebraic Dynamic Programming (Giegerich and Meyer, 2002) and regard each decomposition rule as a production in a suitable grammar. Figure 7 summarizes two major steps in the decomposition, (I) the ‘arc-removal’ reducing TS and DTS. The scheme is complemented by the usual loop decomposition of secondary structures, and (II) the ‘block-decomposition’ splitting joint structures into blocks.

The details of the decomposition procedures are collected in the Supplementary Material, where we show that for each $\mathcal{J}_{n \times 1, 1 \times n}$ there exists a unique decomposition-tree (pne-tree), denoted by $\mathcal{J}^{\mathcal{J}_{DTS}}_{\mathcal{J}_{TS}}[\mathcal{J}(i, j)]$. This tree has root $\mathcal{J}_{n \times 1, 1 \times n}$ and all other vertices correspond to specific substructures of $\mathcal{J}_{n \times 1, 1 \times n}$ obtained by the successive application of the decomposition steps of Figure 7 and the loop decomposition of the secondary structures (Fig. 8).

3.2 Extended loop model

The standard energy model for RNA folding (Mathews et al., 1999), presented in the Supplementary Material, is consistent with the basic decomposition of secondary structures. In addition, joint structures give rise
to two further types of loops. Following Chitsaz et al. (2009), we call them hybrid and kissing-loop (Fig. 9).

- A hybrid is a maximal sequence of intermolecular interior loops formed by \( i \leq 2 \) exterior arcs \( R_iS_i \ldots R_nS_n \), where \( R_iS_i \) is nested within \( R_{i-1}S_i \) and the internal segments \( R_{i+1} \ldots R_{i+n-1} \) and \( S_{i-1} \ldots S_{i+n-1} \) consist of single-stranded nucleotides. That is, a hybrid is the maximal unbranched stem-loop formed by external arcs.

- A kissing-loop is either a pair \( (R_iR_j, R_{i+1} \ldots j-1) \), where the set of \( R_iR_j \)-children, \( S_iS_h \ldots i \leq h < j \), where \( i \leq h < j \) is empty, or a pair \( (S_iS_h, S_{i+1} \ldots j-1) \), where the set of \( S_iS_h \)-children \( R_iR_j \ldots i \leq j \) is empty.

Kissing-loops have been singled out for logical reasons and because some investigations into their thermodynamic properties have been reported in the literature (Gagos et al., 2009). For details of the parametrization employed in xips we refer to the Supplementary Material.

Let us now have a closer look at the energy evaluation of \( J_{\beta A} \). Each decomposition step in Figure 7 results in substructures whose energies we assume to contribute additively and generalized loops that need to be evaluated directly. There are the following two scenarios.

(I) Arc removal: most of the decomposition operations in Procedure (b) correspond 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 Stiegler (1981) 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 \( R_iR_j \).

The set of base pairs on \( R_iR_j \) consists of all interior pairs \( R_pR_q \) with \( i \leq p < q \leq j \) and all exterior pairs \( R_pS_i \) with \( i \leq p \leq j \). An interior arc is exposed on \( R_{i+1} \ldots j-1 \) if and only if it is not enclosed by any interior arc in \( R_iR_j \). An exterior arc is exposed on \( R_{i+1} \ldots j-1 \) if and only if it is not a descendant of any interior arc in \( R_{i+1} \ldots j-1 \). Given \( R_iR_j \), the arcs exposed on \( R_{i+1} \ldots j-1 \) corresponds to the base pairs immediately interior of \( R_iR_j \). Let us write \( E_xR_iR_j = E_{R_iR_j}^1 \cup E_{R_iR_j}^2 \) 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 \( E_xR_iR_j \) as follows:

- \( E_xR_iR_j = \emptyset \): hairpin loop; \( E_xR_iR_j = \{ \} \), interior loop (including bulge and stacks); \( E_xR_iR_j = E_{\text{full-loop}} \), multi-branch loop; \( E_xR_iR_j = E_{\text{kissing-hairpin}} \), kissing-hairpin loop; \( |E_xR_iR_j| \geq 2 \), general kissing-loop.

This picture needs to be refined even further since the arc removal is coupled with further decomposition of the interval \( R_{i+1}j-1 \). 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 the type of loop of in which it was found.

A TS \( f_{\text{TS},R_iR_j}^{\text{ex}} \) is of type \( R_iR_j \) if \( E_xR_iR_j \) is not enclosed in any base pair \( v_{\text{TS},R_iR_j}^{\text{ex}} \). Suppose \( f_{\text{TS},R_iR_j}^{\text{ex}} \) is immediately interior to the closing pair \( S_kS_{j} \), where \( p < h < k < q \). If the loop closed by \( S_kS_{j} \) is a multi-loop, then \( f_{\text{TS},R_iR_j}^{\text{ex}} \) is of type M in \( f_{\text{TS},R_iR_j}^{\text{ex}} \). If \( S_kS_{j} \) is contained in a kissing-loop, we distinguish the types \( F \) and \( K \), depending on whether or not \( E_xR_iR_j = \emptyset \). Figure 10 displays this decomposition for \( f_{\text{TS},R_iR_j}^{\text{ex}} \).

For a DTS \( f_{\text{DTS},R_iR_j}^{\text{ex}} \) (denoted by \( 'E' \) in Fig. 9), we need to determine the type of the exposed pairs of both \( R_{i+1}j-1 \) and \( S_{k}S_{j} \). Hence, each such structure will be indexed by two types. In total, we arrive at 18 distinct cases since some combinations cannot occur. For instance, a DTS cannot be external in both R and S; that is, type EE does not exist, where E means external.

(II) Block decomposition: the second type of decomposition is the splitting of joint structures into ‘blocks’, such as the decompositions of a right tight structure (RTS) in Procedure (a) and a DTS in Procedure (b) (Fig. 7). A RTS \( f_{\text{RTS},R_iR_j}^{\text{ex}} \) may appear in two ways, depending on whether or not there exists an exterior arc \( R_iS_k \) such that \( R_{i}R_{i+1} \ldots j-1 \). If such an exterior arc exists, \( f_{\text{RTS},R_iR_j}^{\text{ex}} \) is of type (R), otherwise it is of type (A). Analogously, (C) and (A) are defined for DTS. Figure 11 shows the decomposition of \( f_{\text{RTS},R_iR_j}^{\text{ex}} \).

Suppose \( f_{\text{RTS},R_iR_j}^{\text{ex}} \) is a DTS contained in a kissing-loop, that is, we have either \( E_xR_iR_j \neq \emptyset \) or \( E_{\text{ex}}R_iR_j \neq \emptyset \). Without loss of generality, we may assume \( E_{\text{ex}}R_iR_j \neq \emptyset \). Then, at least one of the two ‘blocks’ contains the exterior arc belonging to \( E_{\text{ex}}R_iR_j \) (i.e. direct bonds in the language of Chitsaz et al. (2009)) labeled by \( K \). Otherwise, the block is labeled \( F \) (Fig. 12). The situation is analogous if we decompose \( f_{\text{RTS},R_iR_j}^{\text{ex}} \) contained in a kissing-loop.

### 3.3 Recursions

The computation of the partition function is obtained ‘from the inside to the outside’, see Equations (5) and (6). The recursions are initialized with the energies of individual external base pairs and empty secondary structures on subsequences of length up to four. In order to differentiate multi- and
Analogously, the recursion for the DTS \( Q^{\text{TSM}}_{i,j} \) in Figure 10 reads:

\[
Q^{\text{TSM}}_{i,j} = \sum_{\xi} \left( Q^{\text{TSM}}_{i,j} e^{-2\alpha_1} + Q^{\text{TSM}}_{i,j} e^{-2\alpha_2} + Q^{\text{TSM}}_{i,j} e^{-2\alpha_3} + Q^{\text{TSM}}_{i,j} e^{-2\alpha_4} \right) \]

4 BASE PAIRING PROBABILITIES

Given two RNA sequences, our sample space is the ensemble of all zigzag-free joint interaction structures. Let \( Q^f \) denote the corresponding partition function. The probability of a joint structure \( I_{j,j',M} \) is then given by \( Q^f I_{j,j',M} = Q(I_{j,j',M}) / Q^f \).

4.1 Approach

While the computation of the partition function proceeds from smaller to larger subsequences, the computation of the substructure probabilities follows the order of the decomposition outlined in the previous section. That is, the longest range substructures are computed first, analogous to the tree constructed in Section 3.1.

4.2 Case study: secondary structures

In order to illustrate the logic of our backtracking procedure, we first consider the special case of secondary structures. Let \( P_{R_i} \) denote the base pairing binding probability of \( R_i \), i.e.,

\[
P_{R_i} = \sum_{R_j} Q_{ij} Q^{-1}
\]

where the sums are taken over all the partition functions of secondary structures \( W \) in \( R_i \) such that \( R_j \in R_i \). Let \( R_j \in R_i \) be the decomposition tree of a particular secondary structure \( W \) on \( R_i \) via Procedure (c), the key observation here is

\[
R_j \in R_i \text{ and } R_{ij} \in R_j
\]

4.3 Forward and backtracking

The decomposition is summarized in Procedure (c) (Fig. 13). In view of the fact that the \( z \) is not \( O(n^2) \) time complexity, we can differ here

\[
\text{Downloaded from https://academic.oup.com/bioinformatics/article-abstract/25/20/2646/193475 by guest on 21 January 2019}
\]
from the standard implementation of the RNA folding model. Inspection of Figure 13 shows that for a \( R^k(i,j) \)-parent, we have to distinguish the five cases displayed in the lower panel. Let \( R^k(i,j) \) be the set of segments \( R_l(i) \in T_{EC} \) containing at least one arc with an outer loop of type \( M \) and let \( R^{k}\{i,j\} \) be the set of all segments \( R_l(i) \in T_{EC} \). Furthermore, let \( \mathbb{P}_{R^k} \) and \( \mathbb{P}_{R^{k}\{i,j\}} \) be the corresponding probabilities. Note that it is possible, for (L1) and (L4) in Figure 13, that \( h \leftrightarrow i \) and \( j \leftrightarrow t \) holds. However, via further backtracking for \( R^k(i,j) \) and \( R^{k}\{i,j\} \) we can recursively calculate the binding probability.

Following the logic of Figure 13, we obtain

\[
\mathbb{P}_{R^k} = \sum_{(i,j) \in R^k} \mathbb{P}_{i,j} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E}) + \mathbb{P}_{R^k} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E}) + \mathbb{P}_{R^k} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E}) + \mathbb{P}_{R^k} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E}) + \mathbb{P}_{R^k} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E})
\]

where the lines correspond to the five loop types (L1-L5) in Figure 13. Analogously, the recursions for the base pairing probabilities \( \mathbb{P}_{R^k} \) and \( \mathbb{P}_{R^{k}\{i,j\}} \) are given by

\[
\mathbb{P}_{R^k} = \sum_{(i,j) \in R^k} \mathbb{P}_{i,j} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E}) + \mathbb{P}_{R^k} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E}) + \mathbb{P}_{R^k} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E}) + \mathbb{P}_{R^k} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E}) + \mathbb{P}_{R^k} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E})
\]

\[
\mathbb{P}_{R^{k}\{i,j\}} = \sum_{(i,j) \in R^{k}\{i,j\}} \mathbb{P}_{i,j} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E}) + \mathbb{P}_{R^{k}\{i,j\}} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E}) + \mathbb{P}_{R^{k}\{i,j\}} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E}) + \mathbb{P}_{R^{k}\{i,j\}} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E}) + \mathbb{P}_{R^{k}\{i,j\}} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E})
\]

4.3 Base pairing probabilities for joint structures

Set \( \Sigma_1 = \{ (R_{i,j} : i \in I) \} \). We apply the same strategy to the joint structures appearing in Figure 7. Let \( \mathbb{E} \) denote the partition function which sums over all the possible joint structures \( J_{i,j} : i \in I \). Then \( \mathbb{P}_{R_{i,j} : i \in I} = \sum_{(i,j) \in R^k} \mathbb{E} \). In order to compute \( \mathbb{P}_{R_{i,j} : i \in I} \), we classify \( \Sigma_1 \) according to the parent of \( R_{i,j} : i \in I \) in \( T \):

\[
\Sigma_1 = \{ (R_{i,j} : i \in I, R_{i,j} : i \in R^k) \}
\]

\[
\cup \{ (R_{i,j} : i \in I, R_{i,j} : i \in R^k) \}
\]

where the lines correspond to the five loop types (L1-L5) in Figure 13. Analogously, the recursions for the base pairing probabilities \( \mathbb{P}_{R_{i,j} : i \in I} \) and \( \mathbb{P}_{R_{i,j} : i \in I} \) are given by

\[
\mathbb{P}_{R_{i,j} : i \in I} = \sum_{(i,j) \in R^k} \mathbb{P}_{i,j} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E}) + \mathbb{P}_{R^k} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E}) + \mathbb{P}_{R^k} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E}) + \mathbb{P}_{R^k} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E}) + \mathbb{P}_{R^k} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E})
\]

\[
\mathbb{P}_{R_{i,j} : i \in I} = \sum_{(i,j) \in R^k} \mathbb{P}_{i,j} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E}) + \mathbb{P}_{R^k} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E}) + \mathbb{P}_{R^k} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E}) + \mathbb{P}_{R^k} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E}) + \mathbb{P}_{R^k} f_{i,j}^{\text{Dirks-Pierce}}(\mathbf{E})
\]

5 RESULTS AND DISCUSSION

In this contribution, we have introduced a framework in which both the partition function and the base pairing probabilities of zigzag-free RNA–RNA interactions can be derived in a natural way. Our approach is implemented in the software package rip by means of the Dirks–Pierce algorithm (Dirks and Pierce, 2003), whereas rip employs two distinct 2D-matrices inspired by Zucker’s recursion. For details, we refer to the Supplementary Material.

The notion of tights, which have a central role in our presentation, is also implicit in the work of Chitsaz et al. (2009). The focus on the underlying combinatorial aspects, however, leads us to highlight in particular the decomposition tree, which provides a natural framework in which to proceed beyond the algorithmic core of the partition function itself. Indeed, the decomposition tree facilitates the derivation of the base pairing probabilities. Related questions, such as that for the probability of complete hybrids (Huang et al., 2009) can be answered along the same lines. While the current implementation of rip (Chitsaz et al., 2009) concentrates on melting temperature in order to validate the partition function, rip focuses on a detailed analysis of the interaction structures themselves. To this end, we also compute the unweighted maximum expected accuracy structure, which is given as the maximum matching with weights given by the base pairing probabilities. On a more technical level, rip and rip differ in the decomposition of tights: rip utilizes a 4D gap-matrix by means of the Dirks–Pierce algorithm (Dirks and Pierce, 2003), whereas rip employs two distinct 2D-matrices inspired by Zucker’s recursion. For details, we refer to the Supplementary Material.

Back-tracing of the base pairing patterns that underlie the free energy of RNA–RNA binding is of great importance in detailed studies of ncRNA-mRNA interactions. The details of the binding sites have a crucial impact on the interpretation of the computational results and on the comparison of the computational prediction and experimental data. It was shown by Mückstein et al. (2008), for instance, that positive and negative regulation of bacterial mRNAs can be distinguished depending on whether the interaction structure contains the Shine–Dalgarno sequence in stable stem or exposed in an alternatively unpaired region.

Only a small number of interaction structures have been described so far that are more complex than those computable by rip. It is not clear, however, whether complex interactions are truly rare in nature, or whether multi-point contacts such as that of the fhlA–OxyS interaction structure (Argaman and Altuvia, 2000) are rarely observed experimentally because they are...
Table 1. The base pairing probabilities of the four alternative hybrids I, II, III and IV for \( fhl-A-OxyS \), predicted by \( r\!i\!p \) (Fig. 2)

<table>
<thead>
<tr>
<th>Hybrid</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>58, 47: 50.3%</td>
<td>21, 69: 56.5%</td>
<td>34, 55: 45.7%</td>
<td>83, 45: 17.3%</td>
</tr>
<tr>
<td></td>
<td>59, 46: 54.8%</td>
<td>20, 70: 59.5%</td>
<td>35, 54: 51.0%</td>
<td>82, 46: 18.9%</td>
</tr>
<tr>
<td></td>
<td>60, 45: 52.9%</td>
<td>19, 71: 29.8%</td>
<td>36, 53: 49.9%</td>
<td>81, 47: 18.5%</td>
</tr>
<tr>
<td></td>
<td>61, 44: 28.8%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Each entry represents the positions in \( fhl-A-OxyS \) and the base pairing probability. For instance, 58, 47: 50.3% is equivalent to \( P_{58,47} = 0.503 \).

Fig. 14. The natural structure of the \( fhl-A-OxyS \) interaction and the results predicted by several algorithms including \( RNA\!plex \).

The algorithmic approach taken here was motivated by a combinatorial analysis of zigzag-free interaction structures. From a mathematical point of view, our approach is centered around the notions of tight structures and their decomposition trees. A detailed mathematical analysis, in particular the derivation of the generating function and further enumeration results, will be discussed elsewhere.

In order to store the partition function and the base pairing probabilities of joint structures in \( r\!i\!p \), we employ 4D arrays. For the recursion of the partition function, \( Q^T \), we use 16 matrices, 24 matrices for \( Q^{RT} \), 18 matrices for \( Q^{RT} \) and 45 matrices for \( Q^T \), in the context of taking into account the loop energy. The complete set of partition function recursions and all details on the particular implementation of \( r\!i\!p \) can be found at http://www.combinatorics.cn/cbpc/rip.html. The space complexity of \( r\!i\!p \) is \( O(N^5) \). Summations in our recursion equations run over at most two independent indices. Therefore, the time complexity is \( r\!i\!p \) is \( O(N^5) \). In order to obtain the pairing probabilities, we trace back in the decomposition tree. Thus, we have the same space complexity and time complexity as for calculating the partition function.

ACKNOWLEDGEMENTS

We thank Bill Chen and Sven Findeiß for comments on the manuscript.

REFERENCES
