AMB ALGORITHMS FOR MOLECULAR BIOLOGY

RESEARCH Open Access

On the combinatorics of sparsification

Fenix WD Huang and Christian M Reidys*

Abstract

Background: We study the sparsification of dynamic programming based on folding algorithms of RNA structures. Sparsification is a method that improves significantly the computation of minimum free energy (mfe) RNA structures.

Results: We provide a quantitative analysis of the sparsification of a particular decomposition rule, Λ^*. This rule splits an interval of RNA secondary and pseudoknot structures of fixed topological genus. Key for quantifying sparsifications is the size of the so called candidate sets. Here we assume mfe-structures to be specifically distributed (see Assumption 1) within arbitrary and irreducible RNA secondary and pseudoknot structures of fixed topological genus. We then present a combinatorial framework which allows by means of probabilities of irreducible sub-structures to obtain the expectation of the Λ^*-candidate set w.r.t. a uniformly random input sequence. We compute these expectations for arc-based energy models via energy-filtered generating functions (GF) in case of RNA secondary structures as well as RNA pseudoknot structures. Furthermore, for RNA secondary structures we also analyze a simplified loop-based energy model. Our combinatorial analysis is then compared to the expected number of Λ^*-candidates obtained from the folding mfe-structures. In case of the mfe-folding of RNA secondary structures with a simplified loop-based energy model our results imply that sparsification provides a significant, constant improvement of 91% (theory) to be compared to an 96% (experimental, simplified arc-based model) reduction. However, we do not observe a linear factor improvement. Finally, in case of the "full" loop-energy model we can report a reduction of 98% (experiment).

Conclusions: Sparsification was initially attributed a linear factor improvement. This conclusion was based on the so called polymer-zeta property, which stems from interpreting polymer chains as self-avoiding walks. Subsequent findings however reveal that the $O(n)$ improvement is not correct. The combinatorial analysis presented here shows that, assuming a specific distribution (see Assumption 1), of mfe-structures within irreducible and arbitrary structures, the expected number of Λ^*-candidates is $\Theta(n^2)$. However, the constant reduction is quite significant, being in the range of 96%. We furthermore show an analogous result for the sparsification of the Λ^*-decomposition rule for RNA pseudoknotted structures of genus one. Finally we observe that the effect of sparsification is sensitive to the employed energy model.

Keywords: Sparsification, Generating function, Dynamic programming

Background

RNA structures, diagrams and genus filtration

An RNA sequence is a linear, oriented sequence of the nucleotides (bases) **A,U,G,C**. These sequences "fold" by establishing bonds between pairs of nucleotides. In this paper, we only consider the Watson-Crick base pair **A-U** or **G-C** and wobble base pairs **U-G**. The global conformation of an RNA molecule is determined by topological constraints encoded at the level of secondary structure, i.e., by the mutual arrangements of the base pairs [1].

Secondary structures can be interpreted as (partial) matchings in a graph of permissible base pairs [2]. They can be represented as diagrams, i.e. graphs over the vertices $1,\ldots,n$, drawn on a horizontal line with bonds (arcs) in the upper half-plane. The length of an arc (i,j) is denoted by $j-i$. Furthermore, we call two arc (i,j) and (r,s) (suppose $i < r$) cross if $i < r < j < s$ holds. In this representation one refers to a secondary structure without crossing arcs as a *simple* secondary structure and pseudoknot structure, otherwise, see Figure 1.

*Correspondence: duck@santafe.edu
Department of Mathematic and Computer science, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark

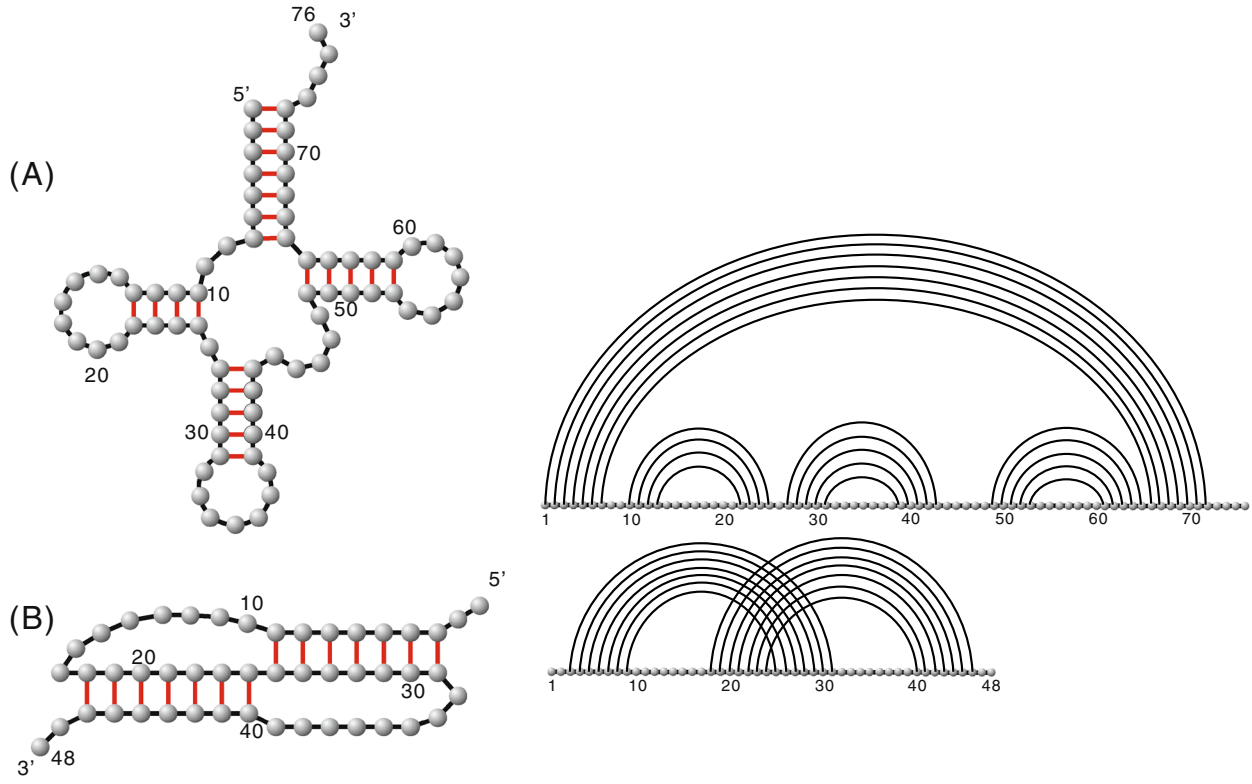

Figure 1 RNA structures as planar graphs and diagrams. (**A**) an RNA secondary structure and (**B**) an RNA pseudoknot structure.

A diagram is a labeled graph over the vertex set $[n] = \{1, \ldots, n\}$ in which each vertex has degree ≤ 3, represented by drawing its vertices in a horizontal line. The backbone of a diagram is the sequence of consecutive integers $(1, \ldots, n)$ together with the edges $\{\{i, i+1\} \mid 1 \leq i \leq n-1\}$. The arcs of a diagram, (i, j), where $i < j$, are drawn in the upper half-plane. We shall distinguish the backbone edge $\{i, i+1\}$ from the arc $(i, i+1)$, which we refer to as a 1-arc. A stack of length ℓ is a maximal sequence of "parallel" arcs, $((i, j), (i+1, j-1), \ldots, (i+(\ell-1), j-(\ell-1)))$ and is also referred to as a ℓ-stack, see Figure 2.

We shall consider diagrams as fatgraphs, $\mathbb{G}$, that is graphs G together with a collection of cyclic orderings, called fattenings. Each fatgraph $\mathbb{G}$ determines an oriented surface $F(\mathbb{G})$ [3,4] which is connected if G is and has some associated genus $g(G) \geq 0$ and number $r(G) \geq 1$ of boundary components. Clearly, $F(\mathbb{G})$ contains G as a deformation retract [5]. Fatgraphs were first applied to RNA secondary structures in [6,7].

A diagram $\mathbb{G}$ hence determines a unique surface $F(\mathbb{G})$ (with boundary). Filling the boundary components with discs we can pass from $F(\mathbb{G})$ to a surface without boundary. Euler characteristic, χ, and genus, g, of this surface is given by $\chi = v - e + r$ and $g = 1 - \frac{1}{2}\chi$, respectively, where v, e, r is the number of discs, ribbons and boundary components in $\mathbb{G}$, [5]. The genus of a diagram is that of its associated surface without boundary and a diagram of genus g is referred to as g-diagram.

A g-diagram without arcs of the form $(i, i+1)$ (1-arcs) is called a g-structure. A g-diagram that contains only vertices of degree three, i.e. does not contain any vertices not incident to arcs in the upper half-plane, is called a g-matching. A diagram is called irreducible, if and only if it cannot be split into two by cutting the backbone without cutting an arc, see Figure 2.

Folding algorithms

Folded configurations are energetically somewhat optimal. Here energy is obtained by adding contributions of loops [8] contained in RNA secondary and pseudoknot structures. Any RNA structure has a unique and disjoint decomposition into such loops which are really stems from the fatgraph [9,10] interpretation of such structures in which loops correspond to boundary components [11]. Additional constraints imply further properties, like for instance certain minimum arc-length conditions [12] and the nonexistence of isolated bonds. An mfe-RNA structure can be predicted in polynomials time by means of dynamic programming (DP) routines [12,13].

The most commonly used tools predicting simple RNA secondary structure `mfold` [13] and the `Vienna RNA Package` [14], require $O(n^2)$ space and $O(n^3)$ time. In the

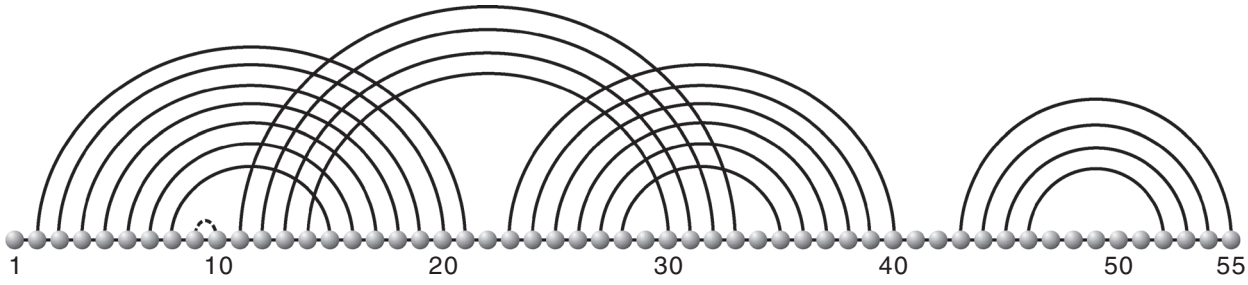

Figure 2 Diagram representation and irreducibility. A diagram over $\{1,\ldots,55\}$. The arcs $(1,21)$ and $(11,33)$ are crossing and the dashed arc $(9,10)$ is a 1-arc which is not allowed. This structure contains 4 stacks with length 7, 4, 6 and 4, from left to right respectively. Irreducibility relative also to a decomposition rule. The rule Λ^* splitting $S_{i,j}$ to $S_{i,k}$ and $S_{k+1,j}$, $S_{1,55}$ is not Λ^*-irreducible, while $S_{2,40}$ and $S_{43,55}$ are. However, for a specific decomposition rule Λ, which removes the outmost arc, $S_{43,55}$ is not Λ-irreducible while $S_{2,40}$ is.

following we omit "simple" and refer to secondary structures containing crossing arcs as pseudoknot structures.

Generalizing the matrices of the DP-routines of secondary structure folding [13,14] to gap-matrices [15], leads to a DP-folding of pseudoknotted structures [15] (`pknot-R&E`) with $O(n^4)$ space an $O(n^6)$ time complexity. The following references provide a certainly incomplete list of DP-approaches to RNA pseudoknot structure prediction using various structure classes characterized in terms of recursion equations and/or stochastic grammars: [9,15-26]. The most efficient algorithm for pseudoknot structures is [22] (`pknotsRG`) having $O(n^2)$ space and $O(n^4)$ time complexity. This algorithm however considers only a restricted class of pseudoknots.

Note that RNA secondary structures are exactly structures of topological genus zero [27]. The topological classification of RNA structures [10,11,28] has recently been translated into an efficient DP-algorithm [9]. Fixing the topological genus of RNA structures implies that there are only finitely many types, the so called irreducible shadows [11].

Sparsification

Let us have a closer look at sparsification and the results of [29-31]. Sparsification is a method tailored to speed up DP-algorithms predicting mfe-secondary structures [29,31]. The idea is to prune certain computation paths encountered in the DP-recursions, see Figure 3A. Let us consider the case of RNA secondary structure folding. Here sparsification reduces the DP-recursion paths to be based on so called candidates. A candidate is in this case an interval, for which the optimal solution cannot be written as a sum of optimal solutions of sub-intervals. This implies the structure over a candidate is an "irreducible" structures when tracing back from the optimal solution. Considering only these candidates gives the same optimal solution as considering all possible intervals. The crucial observation here is that if these irreducibles appear only at a low rate we have a significant reduction in time and space complexity.

Sparsification has been also applied in the context of RNA-RNA interaction structures [30] as well as RNA pseudoknot structures [32]. In difference to RNA secondary structures, however, not every decomposition rule in the DP-folding of RNA pseudoknot structures is amendable to sparsification. By construction, sparsification can only be applied for calculating mfe-energy structures. Since the computation of the partition function [20,33] needs to take into account *all* sub-structures, sparsification does not work.

Sparsification [29,31,32] can be described as follows: let $V = \{\nu_1, \nu_2, \ldots\}$ be a set whose elements ν_i are unions of pairwise disjoint intervals. Let furthermore L_ν denote an optimal solution (here optimal means to maximize the scores) of the DP-routine over ν. By assumption L_ν is recursively obtained. Suppose we are given a decomposition rule Λ_1, for which the optimal solution L_ν is $L_\nu = L_{\nu_1} + L_{\nu_2} + L_{\nu_3}$, where $\nu = \nu_1 \dot\cup \nu_2 \dot\cup \nu_3$. Then, under certain circumstances, the DP-routine may interpret L_ν either as $(L_{\nu_1} + L_{\nu_2}) + L_{\nu_3}$ or as $L_{\nu_1} + (L_{\nu_2} + L_{\nu_3})$, see Figure 3B. To be precise, this situation is encountered iff

- there exists an optimal solution $L_{\nu_1'}$ for a sub-structure over ν_1' where $\nu_1' = \nu_1 \dot\cup \nu_2$ via Λ_2 and L_ν is obtained from $L_{\nu_1'}$ and L_{ν_3} via Λ_1,
- there exists an optimal solution $L_{\nu_2'}$ for a sub-structure over ν_2' where $\nu_2' = \nu_2 \dot\cup \nu_3$ via Λ_3 and L_ν is obtained by L_{ν_1} and $L_{\nu_2'}$ via Λ_1.

Given a decomposition

$$L_\nu = \underbrace{\underbrace{L_{\nu_1} + L_{\nu_2}}_{\Lambda_2} + L_{\nu_3}}_{\Lambda_1},$$

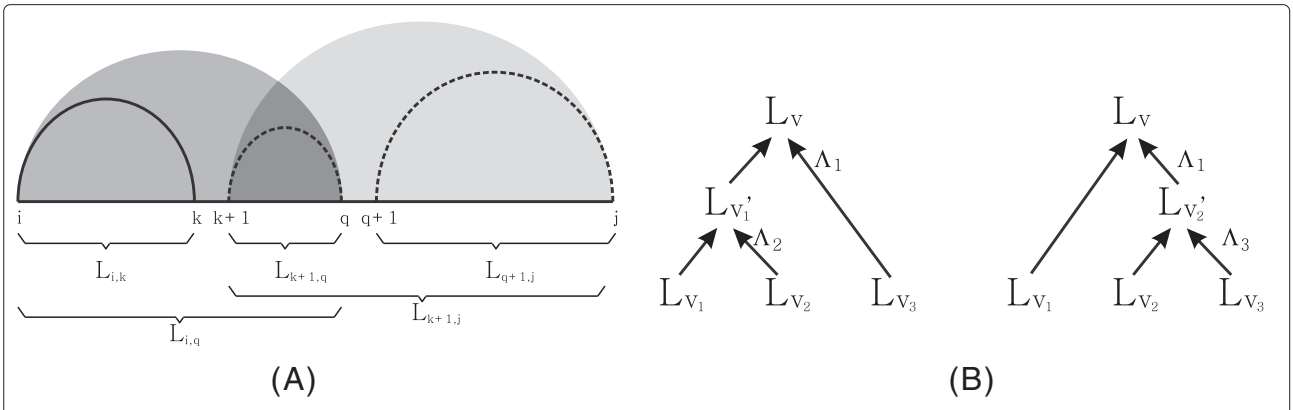

Figure 3 (A) Sparsification of secondary structure folding. Suppose the optimal solution $L_{i,j}$ is obtained from the optimal solutions $L_{i,k}$, $L_{k+1,q}$ and $L_{q+1,j}$. Based on the recursions of the secondary structures, $L_{i,k}$ and $L_{k+1,q}$ produce an optimal solution of $L_{i,q}$. Similarly, $L_{k+1,q}$ and $L_{q+1,j}$ produce an optimal solution of $L_{k+1,j}$. Now, in order to obtain an optimal solution of $L_{i,j}$ it is sufficient to consider either the grouping $L_{i,q}$ and $L_{q+1,j}$ or $L_{i,k}$ and $L_{k+1,j}$. **(B)** General idea of sparsification: L_v is alternatively realized via L_{v_1} and $L_{v_2'}$, or $L_{v_1'}$ and L_{v_3}. Thus it is sufficient to only consider one of the computation paths.

we call Λ_2 *s*-compatible to Λ_1 if there exists a decomposition rule Λ_3 such that

$$L_v = \underbrace{L_{v_1} + \underbrace{L_{v_2} + L_{v_3}}_{\Lambda_3}}_{\Lambda_1}.$$

Note that if Λ_2 is *s*-compatible to Λ_1 then Λ_3 is *s*-compatible to Λ_1. To summarize

Definition 1. (*s*-compatible) Suppose L_v is the optimal solution for S_v over v, $L_v = L_{v_1'} + L_{v_3}$ under decomposition rule Λ_1. $L_{v_1'}$ is obtained from two optimal solutions L_{v_1} and L_{v_2} under rule Λ_2. Then Λ_2 is called *s-compatible* to Λ_1 if there exist some rule Λ_3 such that $L_{v_2'} = L_{v_2} + L_{v_3}$ and $L_v = L_{v_1} + L_{v_2'}$.

Figure 3B depicts two such ways that realize the same optimal solution L_v. Sparsification prunes any such multiple computations of the same optimal value. Note that by symmetry, Λ_2 and Λ_3 are both *s*-compatible to Λ_1.

We next come to the important concept of candidates. The latter mark the essential computation paths for the DP-routine.

Definition 2. (Candidates) Suppose L_v is an optimal solution in a sense of maximizing. We call v is a Λ-*candidate* if for any $v_1 \subsetneq v$ obtained by Λ and $v = v_1 \dot{\cup} v_2$, we have

$$L_v > L_{v_1} + L_{v_2}$$

and we shall denote the set of Λ-candidates set by Q^{Λ}.

By construction a Λ-candidate v is a union of disjoint intervals such that its optimal solution L_v cannot be obtained via a Λ-splitting. This optimal solution allows to construct a non-unique arc-configuration (substructure) over v [13,14] and the above Λ-splitting consequently translates into a splitting of this sub-structure. This connects the notion of Λ-candidates with that of sub-structures and shows that a Λ-candidate implies a sub-structure that is Λ-irreducible.

Lemma 1. *[29,32] Suppose L_v is obtained by selecting the optimal solution from the decomposition rules $\Lambda_1, \Lambda_2, \ldots, \Lambda_n$. If Λ is s-compatible to all $\Lambda_i, \forall 1 \leq i \leq n$, then L_v can be obtained via Λ-candidates.*

In summary, as for the impact of sparsification, [29] claims that sparsification reduces the time complexity by a linear factor. This claim is based on the assumption that RNA molecules satisfy the *polymer-zeta property* [29]. Subsequent studies draw a slightly different picture [31] concluding that that sparsification requires $O(nZ)$ time, where n denotes the length of input sequence, and Z is a sparsity parameter satisfying $n \leq Z < n^2$. Recently, it has been shown in [34] that an asymptotic time complexity of a sparsified RNA folding algorithm using standard energy parameters remains $O(n^3)$ under a wide variety of condition.

Sparsification of RNA secondary structures

Here we recall some results of [29,31] on the sparsification of RNA secondary structures. Secondary structures satisfy a simple recursion which gives the optimal (maximum) solution over $[i,j]$ by $L_{i,j} = \max\{V_{i,j}, W_{i,j}\}$, where $V_{i,j}$ denotes the optimal solution in which (i,j) is a base pair, and $W_{i,j}$ denotes the optimal solution obtained by adding the optimal solutions of two subsequent intervals, respectively. Note that the optimal solution over a single vertex is denoted by $L_{i,i}$. We have the recursion equation

for $V_{i,j}$ and $W_{i,j}$:

$$(\Lambda_1) \quad V_{i,j} = L_{i+1,j-1} + w(i,j),$$

$$(\Lambda_2) \quad W_{i,j} = \max_{i<k<j}\{L_{i,k} + L_{k+1,j}\},$$

where $w(i,j)$ is the energy contribution of (i,j) forming a base pair, see Figure 4. In case two positions, i, j in the sequence are incompatible then we have $w(i,j) = -\infty$.

An interval $[i,j]$ is a Λ^*-candidate if the optimal solution over $[i,j]$ is given by $L_{i,j} = V_{i,j} > W_{i,j}$. Indeed, $[i,j]$ is a candidate iff $[i,j]$ is in the candidate set of Λ^*, and we denote the set Q^{Λ^*} by Q. Suppose the optimal solution $W_{i,j}$ is given by $W_{i,j} = L_{i,q} + L_{q+1,j}$ and suppose we have $L_{i,q} = L_{i,k} + L_{k+1,q}$. Then since $[i,q]$ is not a candidate, Lemma 1 shows that we can compute $W_{i,j} = L_{i,k} + L_{k+1,j}$, where $[i,k]$ is a candidate.

Sparsification on RNA pseudoknot structures

Sparsification can also be applied to the DP-algorithm folding RNA structures with pseudoknots [32]. In contrast to the decomposition rule Λ^* that spliced an interval into two subsequent intervals, we encounter in the grammar for pseudoknotted structures additional more complex decomposition rules [15]. As shown in [32] there exist some decomposition rules which are not s-compatible and which can accordingly not be sparsified at all, see Figure 5B. For instance, given a decomposition rule Λ in `pknot-R&E` subsequent decomposition rules which are s-compatible to Λ are referred to as split type of Λ [32].

In the following we will study RNA pseudoknot structures of fixed topological genus, see **RNA structures, diagrams and genus filtration** for details. An algorithm folding such pseudoknot structures, `gfold`, has been presented in [9]. The decomposition rules that appear in `gfold` are reminiscent to those of `pknot-R&E` but as they restrict the genus of sub-structures, the iteration of gap-matrices is severely restricted and the effect of sparsification of these decompositions is significantly smaller.

In the following, we restrict our analysis in pseudoknotted structures to only the decomposition rule Λ^*, which splices an interval into two subsequent intervals. Put differently, Λ^* cuts the backbone of an RNA pseudoknot structure of fixed genus g over one interval without cutting a bond.

Efficiency of sparsification

By construction, the fewer candidates the DP-routine encounters, the more efficient the sparsification. Thus it is of utmost importance to analyze the number of candidates. In the case of sparsification of RNA secondary structures we have one basic decomposition rule Λ^* acting on intervals, namely Λ^* splices an interval into two disjoint, subsequent intervals. The implied notion of a Λ^*-irreducible sub-structure is that of a sub-structure nested in a maximal arc, where maximal refers to the partial order of two arcs $(i,j) \leq (i',j')$ iff $i' \leq i \ \wedge j \leq j'$. This observation relates irreducibility to nesting of arcs and following this line of thought [29] identifies a specific property of polymer-chains introduced in [35,36] to be of relevance for the size of candidate sets:

Definition 3. (Polymer-zeta property) Let $\mathbb{P}(i,j)$ denote the probability of a structure over an interval $[i,j]$ under some decomposition rule Λ. Then we say Λ follows the polymer-zeta property if $\mathbb{P}(i,j) = b\ m^{-c}$ for some constant b, $c > 0$ and $m = j - i$.

Polymer-zeta comes from modeling the 2D-folding of a polymer chain as a self-avoiding walk (SAW) in a 2D lattice [37]. It implies that the probability of a base pair (i,j) depends only on the length of the arc, i.e. $\mathbb{P}(i,j) = \mathbb{P}(m)$, where $m = j - i$. In [29] stipulate that RNA molecules satisfy the polymer-zeta property and approximate $\mathbb{P}(i,j)$ by $\mathbb{P}(m) = bm^{-c}$ [29] using 50,000 mRNA sequences of an average length of 1992 nucleotides [38]. They find $b \approx 2.11$ and $c \approx 1.47$. The average probability $\mathbb{P}(m)$ is displayed in Figure 4, Page 865 [29] for increasing m. Furthermore, it is implied via Figure six, Page 867 [29] that the average number of candidates converges to a constant, implying that sparsification of DP-routine folding secondary structure takes $\Theta(n^2)$ time complexity.

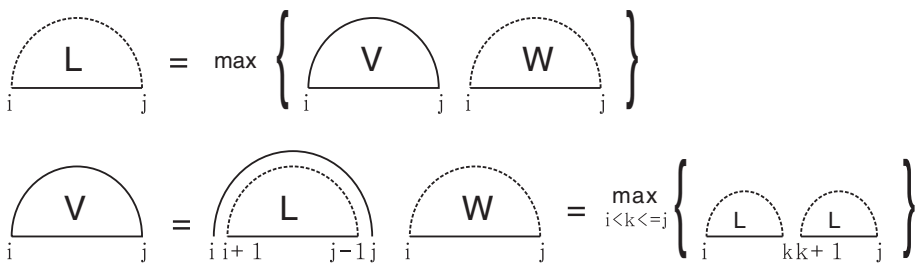

Figure 4 The recursion solving the optimal solution for secondary structures.

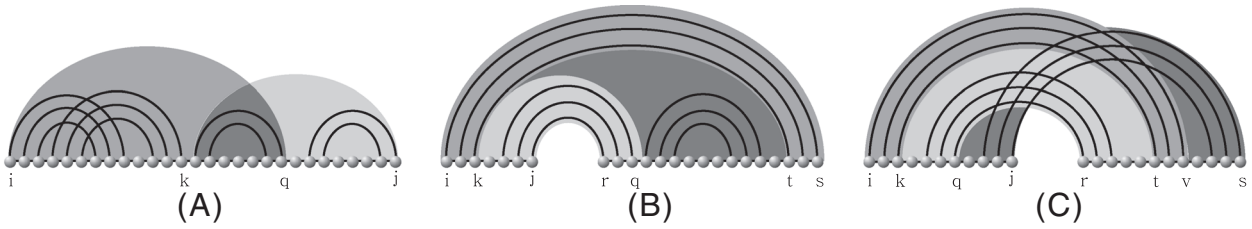

Figure 5 Decomposition rules for pseudoknot structures of fixed genus (decomposed into three colors). (**A**) three decompositions via the rule Λ^*, which is *s*-compatible to itself. (**B**) three decomposition rules $\Lambda_1, \Lambda_2, \Lambda_3$ where Λ_2, Λ_3 are *s*-compatible to Λ_1. (**C**) three decomposition rules $\Lambda_1, \Lambda_2, \Lambda_3$ where Λ_2, Λ_3 are not *s*-compatible to Λ_1.

These findings have been questioned by [34], where it has been observed that the time complexity of a sparsified RNA folding algorithm based on energy minimization remains $O(n^3)$ independently of the energy function used and the base composition of the RNA sequence. [34] argues that the significant effect of sparsification on the DP-routine is largely a finite-size effect. Namely, when the sequence length is below some threshold, the algorithm is dominated by the quadratic time factor. In this context, it may be worth pointing out that In [31] noticed that the improvement of a sparsified base-pairing maximization algorithm depends heavily on the base composition of the input. Backofen parameterizes explicitly the cardinality of candidate sets in [31].

Contribution

In this paper we study the sparsification of the decomposition rule Λ^* [31,32] for RNA secondary and RNA pseudoknot structures of fixed topological genus. Based on Assumption 1 below our paper provides a combinatorial framework for quantifying the effects of sparsification of the Λ^* rule.

We shall prove that the candidate set [29,31,32] is indeed small. We compute the probability of an interval being a candidate for two different energy models. For both models, this is facilitated via computing the generating function (GF) of structures and the generating function of irreducible structures. By studying the asymptotics of coefficients in these generating functions, we can compute the expected number of candidates of a uniformly random input sequence for large n. We show similar results for RNA pseudoknot structures of fixed topological genus. This provides new insights into the improvements of the sparsification of the concatenation-rule Λ^* in the presence of cross serial interactions. Our observations complement the detailed analysis of Backofen [31,32]. We show that although for pseudoknot structures of fixed topological genus [10,11] the effect of sparsification on the global time complexity is still unclear, the decomposition rule that splits an interval can be sped up significantly.

Methods

Suppose w is an energy function for RNA structures. Let $w_\delta(\sigma)$ denote the energy of an RNA structure σ over a sequence δ. The partition function of δ is given by

$$\mathbb{Q}(\delta) = \sum_{\sigma} e^{\frac{w_\delta(\sigma)}{RT}},$$

where R is the universal gas constant and T is the temperature. (Here we consider $w_\delta(\sigma)$ as a positive score.) The partition function induces a probability space in which the probability of a structure σ is

$$\mathbb{P}_\delta(\sigma) = \frac{e^{\frac{w_\delta(\sigma)}{RT}}}{\mathbb{Q}(\delta)}.$$

The concept of a partition function is close to that of a generating function. In case of $e^{w_\delta(\sigma)/RT} = 1$, i.e., each structure contributes equally regardless the underlying sequence and the partition function equals $[z^n]\,G(z)$, where G is the generating function and $[z^n]\,G$ is the coefficient of the term z^n.

Two important energy models are arc-based [39] and loop-based [8], respectively. The loop-based energy-filtration is different from the notion of "stickiness" [40]. The compatibility of two positions by folding random sequences is considered to be 6/16, reminiscent of the probability of two given positions to be compatible by Watson-Crick and Wobble base pairs rules.

Assumption 1. *Let*

$$W(\sigma) = \left(\frac{6}{16}\right)^{\ell} \eta^{w(\sigma)},$$

where $\eta > 1$ is a constant, $w(\sigma)$ is the energy value assigned to σ based on a given energy model and ℓ is the number of arcs contained in σ. Then the probability of a particular structure σ to be the mfe-structure of a uniformly random input sequence is

$$\mathbb{P}(\sigma) = \frac{W(\sigma)}{\sum_{\sigma'} W(\sigma')}. \quad (1)$$

Asymptotics

In this section we compute two generating functions and their singular expansions [11]. Let $\mathbf{c}_g(n)$ and $\mathbf{d}_g(n)$ denote the number of g-matchings and g-structures having n arcs and n vertices, respectively, with GF

$$\mathbf{C}_g(z) = \sum_{n=0}^{\infty} \mathbf{c}_g(n) z^n \qquad \mathbf{D}_g(z) = \sum_{n=0}^{\infty} \mathbf{d}_g(n) z^n.$$

The GF $\mathbf{C}_g(z)$ has been computed in the context of the virtual Euler characteristic of the moduli-space of curves in [41] and $\mathbf{D}_g(z)$ can be derived from $\mathbf{C}_g(z)$ by means of symbolic enumeration [11]. The GF of genus zero diagrams $\mathbf{C}_0(z)$ is well-known to be the GF of the Catalan numbers, i.e., the numbers of triangulations of a polygon with $(n+2)$ sides,

$$\mathbf{C}_0(z) = \frac{1-\sqrt{1-4z}}{2z}.$$

As for $g \geq 1$ we have the following situation [11]

Theorem 1. *Suppose $g \geq 1$. Then the following assertions hold*

(**a**) $\mathbf{D}_g(z)$ *is algebraic and*

$$\mathbf{D}_g(z) = \frac{1}{z^2-z+1}\, \mathbf{C}_g\left(\frac{z^2}{\left(z^2-z+1\right)^2}\right). \quad (2)$$

In particular, $z^2/(z^2-z+1)^2 = 1/4$ is the only dominant singularity of $D_g(z)$. we have for some constant a_g depending only on g and $\gamma \approx 2.618$:

$$[z^n]\, \mathbf{D}_g(z) \sim a_g\, n^{3(g-\frac{1}{2})} \gamma^n. \quad (3)$$

(**b**) *The bivariate GF of g-structures over n vertices, containing exactly m arcs, $\mathbf{E}_g(z,t)$, is given by*

$$\mathbf{E}_g(z,t) = \frac{1}{tz^2-z+1} \mathbf{D}_g\left(\frac{t\, z^2}{(t\, z^2-z+1)^2}\right). \quad (4)$$

Irreducible *g*-structures

In the context of Λ^*-candidates we observed that irreducible sub-structures are of key importance. It is accordingly of relevance to understand the combinatorics of these structures. To this end let $\mathbf{D}_g^*(z) = \sum_{n=0}^{\infty} \mathbf{d}_g^*(n) z^n$ denote the GF of irreducible g-structures.

Lemma 2. *For $g \geq 0$, the GF $\boldsymbol{D}_g^*(z)$ satisfies the recursion*

$$\mathbf{D}_0^*(z) = 1 - \frac{1}{\mathbf{D}_0(z)}$$

$$\mathbf{D}_g^*(z) = -\frac{(\mathbf{D}_0^*(z)-1)\mathbf{D}_g(z) + \sum_{g_1=1}^{g-1} \mathbf{D}_{g_1}^*(z)\mathbf{D}_{g-g_1}(z)}{\mathbf{D}_0(z)}.$$

For a proof of Lemma 2, see Section Proofs.

Theorem 2. *For $g \geq 1$ we have*

(**a**) *the GF of irreducible g-structures over n vertices is given by*

$$\mathbf{D}_g^*(z) = (z^2-z+1)\left(\frac{\mathbf{U}_g(u)}{(1-4u)^{3g-\frac{1}{2}}} + \frac{\mathbf{V}_g(u)}{(1-4u)^{3g-1}}\right), \quad (5)$$

where $u = \frac{z^2}{(z^2-z+1)^2}$, $\mathbf{U}_g(z)$ and $\mathbf{V}_g(z)$ are both polynomials with lowest degree at least $2g$, and $\mathbf{U}_g(1/4), \mathbf{V}_g(1/4) \neq 0$. In particular, for some constant $a_g^* > 0$ and $\gamma \approx 2.618$:

$$\mathbf{D}_g^*(n) \sim a_g^* n^{3\left(g-\frac{1}{2}\right)} \gamma^n. \quad (6)$$

(**b**) *the bivariate GF of irreducible g-structures over n vertices, containing exactly m arcs, $\mathbf{E}_g^*(z,t)$, is given by*

$$\mathbf{E}_g^*(z,t) = (tz^2-z+1)\left(\frac{\mathbf{U}_g(\nu)}{(1-4\nu)^{3g-\frac{1}{2}}} + \frac{\mathbf{V}_g(\nu)}{(1-4\nu)^{3g-1}}\right), \quad (7)$$

where $\nu = \frac{tz^2}{(tz^2-z+1)^2}$.

We shall postpone the proof of Theorem 2 to Section Proofs.

The main result

Nussinov-like energy model

In the following we mimic some form of mfe-g-structures: inspired by the Nussinov energy model [39] we consider the weight of a g-structure over n vertices $\sigma_{g,n}$ to be given by $w(\sigma_{g,n}) = c\ell$, where c is a constant contribution of a single arc and ℓ is the number of arcs in $\sigma_{g,n}$ [40]. Then by Assumption 1, we have the weight function $W(\sigma_{g,n}) = (6/16)^{\ell}\eta^{c\ell} = ((6/16)\eta^c)^{\ell}$. Note that the case $(6/16)\eta^c = 1$ corresponds to the uniform distribution, i.e. all g-structure have identical weight.

This approach requires to keep track of the number of arcs, i.e. we need to employ bivariate GF. In Theorem 1 **(b)** we computed this bivariate GF and in Theorem 2 **(b)** we derived from this bivariate GF $\mathbf{E}_g^*(z,t)$, the GF of irreducible g-structures over n vertices containing ℓ arcs.

The idea now is to substitute for the second indeterminant, t, some fixed $\tau = (6/16)\eta^c \in \mathbb{R}$. This substitution induces the formal power series

$$\mathbf{D}_{g,\tau}(z) = \mathbf{E}_g(z,\tau),$$

which we regard as being parameterized by τ. Obviously, setting $\tau = 1$ we recover $\mathbf{D}_g(z)$, i.e. we have $\mathbf{D}_g(z) =$

$\mathbf{D}_{g,1}(z) = \mathbf{E}_g(z,1)$. Note that for $\tau > 1/4$, the polynomial $\tau z^2 - z + 1$ has no real root. Thus we have for $\tau > 1/4$ the asymptotics

$$\mathbf{d}_{g,\tau}(n) \sim a_{g,\tau} n^{3\left(g-\frac{1}{2}\right)} \gamma_\tau^n \quad \text{and} \quad \mathbf{d}^*_{g,\tau}(n) \sim a^*_{g,\tau} n^{3\left(g-\frac{1}{2}\right)} \gamma_\tau^n, \tag{8}$$

with identical exponential growth rates as long as the supercritical paradigm [42] applies, i.e. as long as γ_τ, the real root of minimal modulus of

$$\left(\frac{\tau\, z^2}{(\tau\, z^2 - z + 1)^2}\right) = \frac{1}{4},$$

is smaller than any singularity of $\frac{1}{\tau z^2 - z + 1}$. In this situation τ affects the constant $a_{g,\tau}$ and the exponential growth rate γ_τ but *not* the sub-exponential factor $n^{3(g-\frac{1}{2})}$. The latter stems from the singular expansion of $\mathbf{C}_g(z)$. Analogously, we derive the τ-parameterized family of GF $\mathbf{D}^*_{g,\tau}(z) = \mathbf{E}^*_g(z,\tau)$. We set the contribution of a single arc $c = 1$ and the constant $\eta = e$, where e is the Euler number. Then we have the parameter $\tau = (6/16)e^1 \approx 1.0125$. By abuse of notation we will omit the subscript τ assuming $\tau = (6/16)e^1$.

The main result of this section is that the set of Λ^*-candidates is a small proportion of all entries. To put this size into context we note that the total number of entries considered for the Λ^*-decomposition rule is given by

$$\mathbb{M}(n) = \sum_{m=1}^{n} (n - m + 1).$$

Theorem 3. *Suppose an mfe-g-structure over an interval of length m is irreducible with probability* $\boldsymbol{d}^*_g(m)/\boldsymbol{d}_g(m)$, *then the expected number of candidates of g-structures for sequences of lengths n satisfies*

$$\mathbb{E}_g(n) = \Theta(n^2)$$

and furthermore, setting $\overline{\mathbb{E}}_g(n) = \mathbb{E}_g(n)/\mathbb{M}(n)$ *we have*

$$\overline{\mathbb{E}}_g(n) \sim \boldsymbol{d}^*_g(n)/\boldsymbol{d}_g(n) \sim b_g,$$

where $b_g > 0$ *is a constant.*

We provide an illustration of Theorem 3 in Figure 6.

Proof. We proof the theorem by quantifying the probability of $[i,j]$ being a Λ^*-candidate. In this case any (not necessarily unique) sub-structure, realizing the optimal solution $L_{i,j}$, is Λ^*-irreducible, and therefore an irreducible structure over $[i,j]$.

Let $m = (j - i + 1)$, by assumption, the probability that $[i,j]$ is a candidate conditional to the existence of a substructure over $[i,j]$ is given by

$$\mathbb{P}_*\left([i,j]\,|\,[i,j] \text{ is a candidate}\right) = \frac{\mathbf{d}^*_g(m)}{\mathbf{d}_g(m)}, \tag{9}$$

Note that $\mathbb{P}_*\left([i,j]\,|\,[i,j] \text{ is a candidate}\right)$ does not depend on the relative location of the interval but only on the interval-length. Let $\mathbb{P}_g(m) = \mathbf{d}^*_g(m)/\mathbf{d}_g(m)$, then according to Theorem 1,

$$(1-\epsilon) a_g m^{3\left(g-\frac{1}{2}\right)} \gamma^m \;\le\; \mathbf{d}_g(m) \;\le\; (1+\epsilon) a_g m^{3\left(g-\frac{1}{2}\right)} \gamma^m,$$

$$(1-\epsilon) a^*_g m^{3\left(g-\frac{1}{2}\right)} \gamma^m \;\le\; \mathbf{d}^*_g(m) \;\le\; (1+\epsilon) a^*_g m^{3\left(g-\frac{1}{2}\right)} \gamma^m,$$

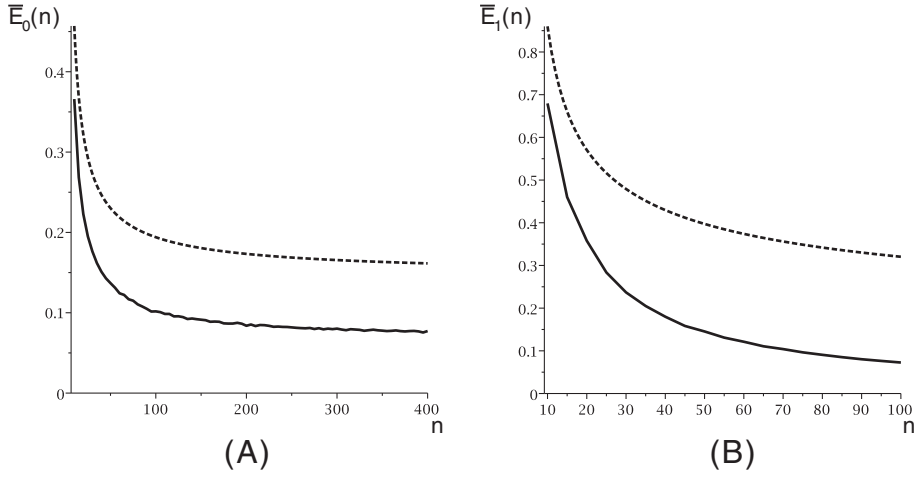

Figure 6 The expected number of candidates for secondary and 1-structures from an random input with a simplified arc-based energy model, $\overline{\mathbb{E}}_0(n)$ and $\overline{\mathbb{E}}_1(n)$: we compute the expected number of candidates obtained by folding 100 random sequences for secondary structures (A)(solid) and 1-structures (B)(solid). We also display the theoretical expectations implied by Theorem 3 (**A**)(dashed) and (**B**)(dashed).

for $m \geq m_0$ where $m_0 > 0$ and $0 < \epsilon < 1$ are constants. On the one hand

$$\mathbb{P}_g(m) = \frac{\mathbf{d}_g^*(m)}{\mathbf{d}_g(m)} \leq \frac{(1+\epsilon)a_g^* m^{3\left(g-\frac{1}{2}\right)}\gamma^m}{(1-\epsilon)a_g m^{3\left(g-\frac{1}{2}\right)}\gamma^m} = (1+\epsilon')\frac{a_g^*}{a_g} = (1+\epsilon')b_g, \tag{10}$$

where $b_g = a_g/a_g^* > 0$ is a constant. On the other hand, we have

$$\mathbb{P}_g(m) = \frac{\mathbf{d}_g^*(m)}{\mathbf{d}_g(m)} \geq \frac{(1-\epsilon)a_g^* m^{3\left(g-\frac{1}{2}\right)}\gamma^m}{(1+\epsilon)a_g m^{3\left(g-\frac{1}{2}\right)}\gamma^m} = (1-\epsilon'')\frac{a_g^*}{a_g} = (1-\epsilon'')b_g. \tag{11}$$

Setting $\epsilon = \max\{\epsilon', \epsilon''\}$, we can conclude that $\mathbb{P}_g(m) \sim \mathbf{d}_g^*(m)/\mathbf{d}_g(m)$, see Figure 7.

We next study the expected number of candidates over an interval of length m. To this end let

$$X_m = |\{[i,j]\,|[i,j] \text{ is a } \Lambda^*\text{-candidate of length } m\}|.$$

The expected cardinality of the set of Λ^*-candidates of length $m = (j-i+1)$ encountered in the DP-algorithm is given by

$$\mathbb{E}_g(X_m) \leq (n-(m-1))\,\mathbb{P}_g(m),$$

since there are $n-(m-1)$ starting points for such an interval $[i,j]$. Therefore, by linearity of expectation, for sufficiently large $m > m_0$, $\mathbb{P}_g(m) \leq (1+\epsilon)b_g$ with ϵ being a small constant. Thus we have

$$\mathbb{E}_g(n) = \mathbb{E}_g\left(\sum_m X_m\right) \leq \sum_{m=1}^{m_0}(n-m+1)\mathbb{P}_g(m) + (1+\epsilon)b_g \sum_{m=m_0}^{n}(n-m+1). \tag{12}$$

Consequently, the expected size of the Λ^*-candidate set is $\Theta(n^2)$. We proceed by comparing the expected number of candidates of a sequence with length n with $\mathbb{M}(n)$,

$$\begin{aligned}\frac{\mathbb{E}_g(n)}{\mathbb{M}(n)} &\leq \frac{\sum_{m=1}^{m_0}(n-m+1)\mathbb{P}_g(m) + (1+\epsilon)b_g\sum_{m=m_0}^{n}(n-m+1)}{\sum_{m=1}^{n}(n-m+1)} \\ &\leq (1+\epsilon)b_g + \frac{\sum_{m=1}^{m_0}(\mathbb{P}_g(m) - (1+\epsilon)b_g)(n-m+1)}{\sum_{m=1}^{n}(n-m+1)} \\ &\leq (1+\epsilon)b_g + \frac{k\cdot n}{n^2}.\end{aligned}$$

For sufficient large $n \geq n_0$, $\mathbb{E}_g(n)/\mathbb{M}(n) \leq (1+\epsilon')b_g$. Furthermore

$$\begin{aligned}\frac{\mathbb{E}_g(n)}{\mathbb{M}(n)} &\geq \frac{\sum_{m=1}^{m_0}(n-m+1)\mathbb{P}_g(m) + (1-\epsilon)b_g\sum_{m=m_0}^{n}(n-m+1)}{\sum_{m=1}^{n}(n-m+1)} \\ &\geq (1-\epsilon)b_g,\end{aligned}$$

from which we can conclude $\mathbb{E}_g(n)/\mathbb{M}(n) \sim \mathbf{d}_g^*(m)/\mathbf{d}_g(m) \sim b_g$ and the theorem is proved. □

Loop-based energy model

In this section we discuss the loop-based energy model of RNA secondary structure folding. To be precise we evoke here trivariate GFs $\mathbf{F}(z,t,v)$ and $\mathbf{F}^*(z,t,v)$ whose coefficients counting the numbers of secondary structures and irreducible secondary structures over n vertices having ℓ arcs and energy j, respectively. This becomes necessary since the loop-based model distinguishes between arcs

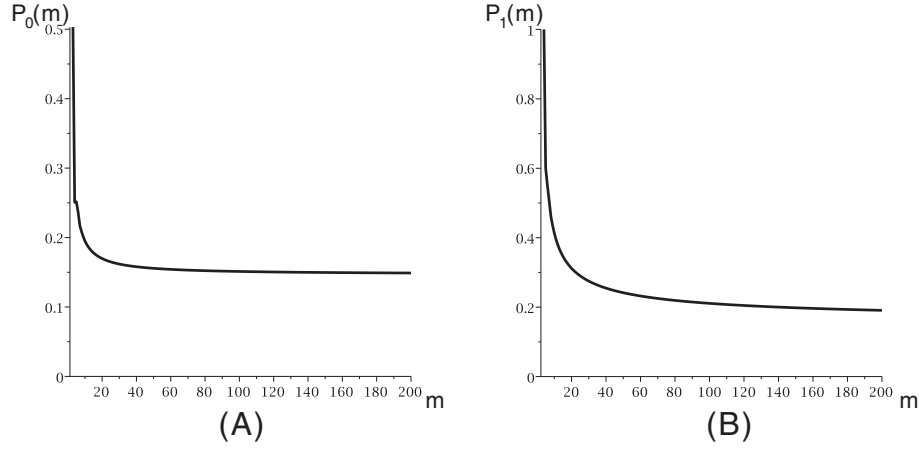

Figure 7 The probability distribution of $\mathbb{P}_0(m)$ (A) and $\mathbb{P}_1(m)$ (B) on a simplified arc-based energy model.

and energy. The "cancelation" effect or reparameterization of stickiness [40] to which we referred to before does not appear in this context. Thus we need both an arc- as well as an energy-filtration.

A further complication emerges. In difference to the GFs $\mathbf{E}_g(z,t)$ and $\mathbf{E}_g^*(z,t)$ the new GFs are not simply obtained by formally substituting $(tz^2/((tz^2-z+1)^2)$ into the power series $\mathbf{D}_g(z)$ and $\mathbf{D}_g^*(z)$ as bivariate terms. The more complicated energy model requires a specific recursion for irreducible secondary structures.

The energy model used in prediction of secondary structure is more complicated than the simple arc-based energy model. Loops which are formed by arcs as well as isolated vertices between the arcs are considered to give energy contribution. Loops are categorized as hairpin loops (no nested arcs), interior loops (including bulge loops and stacks) and multi-loops (more than two arcs nested), see Figure 8. An arbitrary secondary structure can be uniquely decomposed into a collection of mutually disjoint loops. A result of the particular energy parameters [8] is that the energy model prefers interior loops, in particular stacks (no isolated vertex between two parallel arc), and disfavors multi-loops. Based on this observation, we give a simplified energy model for a loop λ contained in secondary structure which only depends on the loop types by

- $w(\lambda) = 0.5$ if λ is a hairpin loop,
- $w(\lambda) = 1$ if λ is an interior loop,
- $w(\lambda) = -5$ if λ is a multi-loop,

where λ is a loop in a structure. The energy for a secondary structure σ accordingly is given by

$$w(\sigma) = \sum_{\lambda\in\sigma} w(\lambda). \tag{13}$$

Let $\mathbf{F}_0^*(z)$ and $\mathbf{F}_0(z)$ be the energy-filtered GFs obtained by setting $t = 6/16$ and $\nu = \eta = e$ in $\mathbf{F}^*(z,t,\nu)$ and $\mathbf{F}(z,t,\nu)$, where e is the Euler number. Then

$$\mathbf{f}_n = \sum_{\sigma}\left(\frac{6}{16}\right)^{\ell} e^{w(\sigma)} = \sum_{\sigma} W(\sigma),$$

$$\mathbf{f}_n^* = \sum_{\sigma'}\left(\frac{6}{16}\right)^{\ell'} e^{w(\sigma')} = \sum_{\sigma'} W(\sigma'),$$

where σ is an arbitary and σ' is an irreducible secondary structure. Along these lines, ℓ, ℓ' denote the number of arcs in σ and σ'. In other words, what happens here is that we find a suitable parameterization which brings us back to a simple univariate GF whose coefficients count the sum of weights of structures over n vertices.

Lemma 3. *The energy-filtered generating function of RNA secondary structures, $\mathbf{F}_0^*(z)$, satisfies the recursion*

$$\mathbf{F}_0^*(z) = \frac{6}{16}e^{0.5}z^2\frac{z}{1-z} + \frac{6}{16}e^1z^2\left(\frac{1}{1-z}\right)^2\mathbf{F}_0^*(z) + \frac{6}{16}e^{-5}z^2\frac{\left(\mathbf{F}_0^*(z)\frac{1}{1-z}\right)^2}{1-\boldsymbol{F}_0^*(z)\frac{1}{1-z}}\frac{1}{1-z}. \tag{14}$$

and $\mathbf{F}^*(z)$ is uniquely determined by the above equation. Furthermore

$$\mathbf{F}_0(z) = \frac{1}{1-z}\frac{1}{1-\mathbf{F}_0^*(z)\frac{1}{1-z}}. \tag{15}$$

Proof. We first consider the GF $\mathbf{F}_0^*(z)$ whose coefficient of z^n denotes the total weight of irreducible secondary structures over n vertices, where $(1,n)$ is an arc. Thus it gives a term $6/16z^2$. Isolated vertex lead to the term

$$z^p\sum_{i=0}^{\infty} z^i = z^p\frac{1}{1-z},$$

where p denotes the minimum number of isolated vertices to be inserted. Depending on the types of loops formed by (i,n), we have

- hairpin loops: $\frac{z}{1-z}$,
- interior loops: $\mathbf{F}_0^*(z)\left(\frac{1}{1-z}\right)^2$,
- multi-loops: there are at least two irreducible sub-structures, as well as isolated vertices, thus

$$\frac{1}{1-z}\sum_{i=2}^{\infty}\left(\mathbf{F}_0^*(z)\frac{1}{1-z}\right)^i = \frac{\left(\mathbf{F}_0^*(z)\frac{1}{1-z}\right)^2}{1-\mathbf{F}_0^*(z)\frac{1}{1-z}}\frac{1}{1-z}.$$

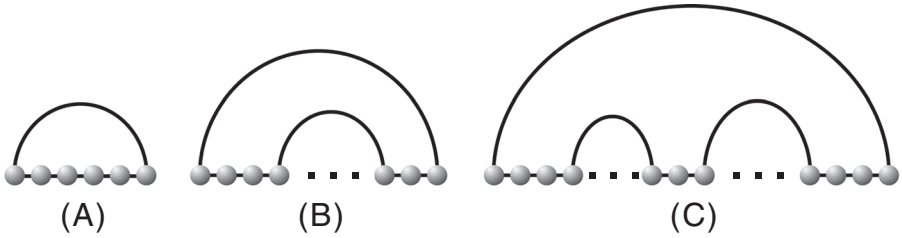

Figure 8 Diagram representation of loop types in secondary structures: (A) hairpin loop, (B) interior loop, (C) multi-loop.

Considering the contributions from the energy model we compute

$$\mathbf{F}_0^*(z) = \frac{6}{16}\left(e^{0.5}z^2\frac{z}{1-z} + e^1 z^2\left(\frac{1}{1-z}\right)^2\mathbf{F}_0^*(z) + e^{-5}z^2\frac{\left(\mathbf{F}_0^*(z)\frac{1}{1-z}\right)^2}{1-\mathbf{F}_0^*(z)\frac{1}{1-z}}\frac{1}{1-z}\right),$$

which establishes the recursion. The uniqueness of the solution as a power series follows from the fact that each coefficient can evidently be recursively computed.

An arbitrary secondary structure can be considered as a sequence of irreducible sub-structures with certain intervals of isolated vertices. Thus

$$\mathbf{F}_0(z) = \frac{1}{1-z}\sum_{i=0}^{\infty}\frac{1}{1-z}\mathbf{F}_0^*(z) = \frac{1}{1-z}\frac{1}{1-\mathbf{F}_0^*(z)\frac{1}{1-z}}.$$

□

Lemma 4. $\mathbf{F}_0^*(z)$ *and* $\mathbf{F}_0(z)$ *have the same singular expansion.*

$$\mathbf{f}_0^*(n) \sim \alpha n^{-\frac{3}{2}}\gamma^n, \quad \textit{and} \quad \mathbf{f}_0(n) \sim \beta n^{-\frac{3}{2}}\gamma^n, \tag{16}$$

where $\alpha \approx 0.24$ *and* $\beta \approx 2.88$ *are constants and* $\gamma \approx 2.1673$

Proof. Solving eq. 14 we obtain a unique solution for $\mathbf{F}_0^*(z)$ whose coefficient are all positive. Observing the dominant singularity of $\mathbf{F}_0^*(z)$ is $\rho \approx 0.4614$. $\mathbf{F}_0(z)$ is a function of $\mathbf{F}_0^*(z)$ and we examine the real root of minimal modulus of $1-\mathbf{F}_0^*(z)\frac{1}{1-z} = 0$ is bigger than ρ. Then by the supercritical paradigm [42] applying, $\mathbf{F}_0(z)$ and $\mathbf{F}_0^*(z)$ have identical exponential growth rates. Furthermore, $\mathbf{F}_0^*(z)$ and $\mathbf{F}_0(z)$ have the same sub-exponential factor $n^{-\frac{3}{2}}$, hence the lemma. □

Theorem 4. *Suppose an mfe-secondary structure over an interval of length m is irreducible with probability* $\mathbb{P}_0(m) = \frac{\mathbf{f}_0^*(m)}{\mathbf{f}_0(m)}$, *then the expected number of candidates from a random sequence of length n with a simplified loop-based energy model is*

$$\mathbb{E}_0(n) = \Theta(n^2)$$

and furthermore, setting $\overline{\mathbb{E}}_g(n) = \mathbb{E}_g(n)/\mathbb{M}(n)$, *we have*

$$\overline{\mathbb{E}}_0(n) \sim \mathbf{f}_0^*(n)/\mathbf{f}_0(n) \sim b,$$

where $b = \alpha/\beta \approx 0.08$.

Proof. By Lemma 4 we have $\mathbf{f}_0^*(m)/\mathbf{f}_0(m) \sim b$ where b is a constant. The proof is completely analogous to that of Theorem 3. □

We show the distribution of $\mathbb{P}_0(m)$ and $\overline{\mathbb{E}}_0(n)$ in Figure 9.

Conclusion

In this paper we quantify the effect of sparsification of the rule Λ^*. This rule splits intervals and separates concatenated sub-structures. The sparsification of Λ^* alone is claimed to provide a speed up of up to a linear factor of the DP-folding of RNA secondary structures [29].

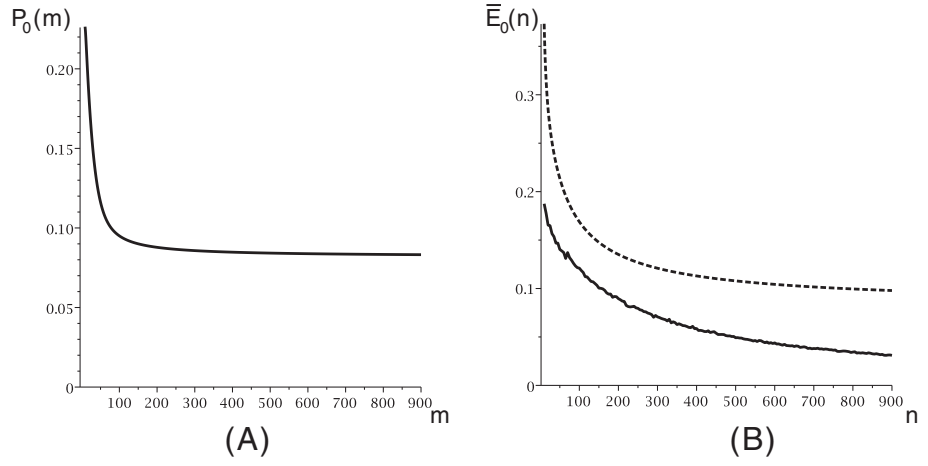

Figure 9 The distribution of $\mathbb{P}_0(m)$ (A) and $\overline{\mathbb{E}}_0(n)$ obtained by folding 100 random sequences on the loop-based model (B)(solid), as well as the theoretical expectation implied by Theorem 4 (B)(dashed).

A similar conclusion is drawn in [30] where the sparsification of RNA-RNA interaction structures is shown to experience also a linear reduction in time complexity. Both papers [29,30] base their conclusion on the validity of the polymer-zeta property. However, [34] comes to a different conclusion reporting a mere constant reduction in time complexity. While Λ^* is the key for the time complexity reduction of secondary structure folding, it is conceivable that for pseudoknot structures there may exist non-sparsifiable rules in which case the overall time complexity is not reduced.

In any case, the key is the set of candidates and we provide an analysis of Λ^*-candidates by combinatorial means. In general, the connection between candidates, i.e. unions of disjoint intervals and the combinatorics of structures is actually established by the algorithm itself via backtracking: at the end of the DP-algorithm a structure is being generated that realizes the previously computed energy as mfe-structure. This connects intervals and substructures.

So, does the condition $c > 1$ in polymer-zeta apply in the context of RNA structures? In fact this condition would follow *if* the intervals in question are distributed as in uniformly sampled structures. This however, is far from reasonable, due to the fact that the mfe-algorithm deliberately designs some mfe-structure over the given interval. What the algorithm produces is in fact antagonistic to uniform sampling. We here wish to acknowledge the help of one anonymous referee in clarifying this point.

Our results imply that polymer-zeta does not hold. Our framework critically depends on a specific distribution of mfe structures within irreducible and arbitrary structures, explicated in Assumption 1. We have cross-checked Assumption 1 with the number of candidates in DP-programs (using the same energy model), see Figure 7 and Figure 9. With this conclusion we are in accord with [31,34] but provide an entirely different approach.

The non validity of polymer-zeta has also been observed in the context of the limit distribution of the 5'-3' distances of RNA secondary structures [43]. Here it is observed that long arcs, to be precise arcs of lengths $O(n)$ *always* exist. This is of course a contradiction to the polymer-zeta property in case of $c > 1$.

The key to quantification of the expected number of candidates is the singularity analysis of a pair of energy-filtered GF, namely that of a class of structures and that of the subclass of all such structures that are irreducible. We show that for various energy models the singular expansions of both these functions are essentially *equal*–modulo some constant. This implies that the expected number of candidates is $\Theta(n^2)$ and all constants can explicitly be computed from a detailed singularity analysis. The good news is that depending on the energy model, a significant constant reduction, around 96% can be obtained. This is in accordance with data produced in [31] for the mfe-folding of random sequences. There a reduction by 98% is reported for sequences of length ≥ 500.

Our findings are of relevance for numerous results, that are formulated in terms of sizes of candidate sets [32]. These can now be quantified. It is certainly of interest to devise a full fledged analysis of the loop-based energy model. While these computations are far from easy our framework shows how to perform such an analysis.

Using the paradigm of gap-matrices Backofen has shown [32] that the sparsification of the DP-folding of RNA pseudoknot structures exhibits additional instances, where sparsification can be applied, see Figure 5B. Our results show that the expected number of candidates is $\Theta(n^2)$, where the constant reduction is around 90%. This is in fact very good new since the sequence length in the context of RNA pseudoknot structure folding is in the order of hundreds of nucleotides. So sparsification of further instances does have an significant impact on the time complexity of the folding.

Proofs

In this section, we prove Lemma 2 and Theorem 2.

Proof for Lemma 2: let $\mathbf{D}(z,u)$ and $\mathbf{D}^*(z,u)$ be the bivariate GF $\mathbf{D}(z,u) = \sum_{n\geq 0}\sum_{g=0}^{\lfloor \frac{n}{2}\rfloor} \mathbf{d}_g(n) z^n u^g$, and $\mathbf{D}^*(z,u) = \sum_{n\geq 1}\sum_{g=0}^{\lfloor \frac{n}{2}\rfloor} \mathbf{d}^*_g(n) z^n u^g$. Suppose a structure contains exactly j irreducible structures, then

$$\mathbf{D}(z,u) = \sum_{j\geq 0} \mathbf{R}(z,u)^j = \frac{1}{1-\mathbf{R}(z,u)} \tag{17}$$

and

$$\mathbf{D}^*_g(z) = [\,u^g]\,\mathbf{D}^*(z,u) = -[\,u^g]\,\frac{1}{\mathbf{D}(z,u)}, \quad g \geq 1, \tag{18}$$

as well as $\mathbf{D}^*_0(z) = 1 - [\,u^0]\,\frac{1}{\mathbf{D}(z,u)}$. Let $\mathbf{F}(z,u) = \sum_{n\geq 0}\sum_{g\geq 0} \mathbf{f}_g(n)\, z^n\, u^g = \frac{1}{\mathbf{D}(z,u)}$. Then $\mathbf{F}(z,u)\,\mathbf{D}(z,u) = 1$, whence for $g \geq 1$,

$$\sum_{g_1=0}^{g} \mathbf{F}_{g_1}(z)\mathbf{D}_{g-g_1}(z) = [\,u^g]\,\mathbf{F}(z,u)\mathbf{D}(z,u) = 0, \tag{19}$$

and $\mathbf{F}_0(z)\mathbf{D}_0(z) = 1$, where $\mathbf{F}_g(z) = \sum_{n\geq 0}\mathbf{f}_g(n)z^n = [\,u^g]\,\mathbf{F}(z,u) = [\,u^g]\,\frac{1}{\mathbf{D}(z,u)}$. Furthermore, we have $\mathbf{F}_0(z) = \frac{1}{\mathbf{D}_0(z)}$ and

$$\mathbf{F}_g(z) = -\frac{\sum_{g_1=0}^{g-1}\mathbf{F}_{g_1}(z)\mathbf{D}_{g-g_1}(z)}{\mathbf{D}_0(z)}, \quad g \geq 1, \tag{20}$$

which implies $\mathbf{D}_0^*(z) = 1 - \mathbf{F}_0(z) = 1 - \frac{1}{\mathbf{D}_0(z)}$ and

$$\begin{aligned}\mathbf{D}_g^*(z) &= -\mathbf{F}_g(z)\\ &= -\frac{(\mathbf{D}_0^*(z)-1)\mathbf{D}_g(z) + \sum_{g_1=1}^{g-1}\mathbf{D}_{g_1}^*(z)\mathbf{D}_{g-g_1}(z)}{\mathbf{D}_0(z)}.\end{aligned} \tag{21}$$

Proof for Theorem 2 Let $[\,n]_k$ denote the set of compositions of n having k parts, i.e. for $\sigma \in [\,n]_k$ we have $\sigma = (\sigma_1, \ldots, \sigma_k)$ and $\sum_{i=1}^k \sigma_i = n$.
Claim.

$$\begin{aligned}\mathbf{D}_{g+1}^*(z) = \frac{\mathbf{D}_{g+1}(z)}{\mathbf{D}_0(z)^2} + \sum_{j=0}^{g-1}\frac{(-1)^{g+2-j}}{\mathbf{D}_0(z)^{g+2-j}}\\ \times\left(\sum_{\sigma\in[g+1]_{g+1-j}}\prod_{i=1}^{g+1-j}\mathbf{D}_{\sigma_i}(z)\right).\end{aligned} \tag{22}$$

We shall prove the claim by induction on g. For $g = 1$ we have

$$\mathbf{D}_1^*(x) = \frac{\mathbf{D}_1(z)}{(\mathbf{D}_0(z))^2}, \tag{23}$$

whence eq. (22) holds for $g = 1$. By induction hypothesis, we may now assume that for $j \le g$, eq. (22) holds. According to Lemma 2, we have

$$\begin{aligned}\mathbf{D}_{g+1}^*(z) &= -\frac{(\mathbf{D}_0^*(z)-1)\mathbf{D}_{g+1}(z)+\sum_{g_1=1}^{g}\mathbf{D}_{g_1}^*(z)\mathbf{D}_{g+1-g_1}(z)}{\mathbf{D}_0(z)}\\ &= \frac{\mathbf{D}_{g+1}(z)}{\mathbf{D}_0(z)^2} - \sum_{g_1=1}^{g}\left(\frac{\mathbf{D}_{g_1}(z)}{\mathbf{D}_0(z)^3} + \sum_{j=0}^{g_1-2}\frac{(-1)^{g_1+1-j}}{\mathbf{D}_0(z)^{g_1+2-j}}\right.\\ &\quad\left.\times\left(\sum_{\sigma\in[g_1]_{g_1-j}}\prod_{i=1}^{g_1-j}\mathbf{D}_{\sigma_i}(z)\right)\right)\mathbf{D}_{g+1-g_1}(z).\end{aligned}$$

We next observe

$$\begin{aligned}&-\sum_{g_1=1}^{g}\frac{\mathbf{D}_{g_1}(z)}{\mathbf{D}_0(z)^3}\mathbf{D}_{g+1-g_1}(z)\\ &= \frac{(-1)^{g+2-(g-1)}}{\mathbf{D}_0(z)^{g+2-(g-1)}}\left(\sum_{\sigma'\in[g+1]_{g+1-(g-1)}}\prod_{i=1}^{g+1-(g-1)}\mathbf{D}_{\sigma_i'}(z)\right),\end{aligned} \tag{24}$$

and setting $h = g_1 - j$ we obtain,

$$\begin{aligned}&-\sum_{g_1=1}^{g}\sum_{j=0}^{g_1-2}\frac{(-1)^{g_1+1-j}}{\mathbf{D}_0(z)^{g_1+2-j}}\left(\sum_{\sigma\in[g_1]_{g_1-j}}\prod_{i=1}^{g_1-j}\mathbf{D}_{\sigma_i}(z)\right)\mathbf{D}_{g+1-g_1}(z)\\ &= \sum_{g_1=1}^{g}\sum_{h=2}^{g_1}\frac{(-1)^{h+2}}{\mathbf{D}_0(z)^{h+2}}\left(\sum_{\sigma\in[g_1]_h}\prod_{i=1}^{h}\mathbf{D}_{\sigma_i}(z)\right)\mathbf{D}_{g+1-g_1}(z)\\ &= \sum_{h=2}^{g}\frac{(-1)^{h+2}}{\mathbf{D}_0(z)^{h+2}}\left(\sum_{g_1=h}^{g}\left(\sum_{\sigma\in[g_1]_h}\prod_{i=1}^{h}\mathbf{D}_{\sigma_i}(z)\right)\mathbf{D}_{g+1-g_1}(z)\right)\\ &= \sum_{h=2}^{g}\frac{(-1)^{h+2}}{\mathbf{D}_0(z)^{h+2}}\left(\sum_{\sigma'\in[g+1]_{h+1}}\prod_{i=1}^{h+1}\mathbf{D}_{\sigma_i'}(z)\right)\end{aligned}$$

and setting $j = g - h$

$$= \sum_{j=0}^{g-2}\frac{(-1)^{g+2-j}}{\mathbf{D}_0(z)^{g+2-j}}\left(\sum_{\sigma'\in[g+1]_{g+1-j}}\prod_{i=1}^{g+1-j}\mathbf{D}_{\sigma_i'}(z)\right).$$

Consequently, the Claim holds for any $g \ge 1$.

For any $g \ge 1$, we have [11]

$$\begin{aligned}\mathbf{D}_g(z) &= \frac{1}{z^2-z+1}\frac{\mathbf{P}_g(u)}{(1-4u)^{3g-1/2}},\\ \mathbf{D}_0(z) &= \frac{1}{z^2-z+1}\frac{2}{(1+\sqrt{1-4u})},\end{aligned}$$

where $\mathbf{P}_g(u)$ is a polynomial with integral coefficients of degree at most $(3g-1)$, $\mathbf{P}_g(1/4) \ne 0$, $[\,u^{2g}]\,\mathbf{P}_g(u) \ne 0$ and $[\,u^h]\,\mathbf{P}_g(u) = 0$ for $0 \le h \le 2g-1$. Let $u = \frac{z^2}{(z^2-z+1)^2}$, the Claim provides in this context the following interpretation of $\mathbf{D}_g^*(z)$

$$\begin{aligned}\frac{1}{z^2-z+1}\mathbf{D}_g^*(z) &= \frac{\mathbf{P}_g(u)}{(1-4u)^{3g-1/2}}\left(\frac{1+\sqrt{1-4u}}{2}\right)^2\\ &\quad+\sum_{j=0}^{g-2}\left(-\frac{1+\sqrt{1-4u}}{2}\right)^{g+1-j}\\ &\quad\times\frac{\sum_{\sigma\in[g]_{g-j}}\prod_{i=1}^{g-j}\mathbf{P}_{\sigma_i}(u)}{(1-4u)^{3g-\frac{g-j}{2}}},\end{aligned} \tag{25}$$

and

$$\sum_{j=0}^{g-2}\left(-\frac{1+\sqrt{1-4u}}{2}\right)^{g+1-j}\frac{\sum_{\sigma\in[g]_{g-j}}\prod_{i=1}^{g-j}\mathbf{P}_{\sigma_i}(u)}{(1-4u)^{3g-\frac{g-j}{2}}}$$
$$=\sum_{j=0}^{g-2}\sum_{k=0}^{g+1-j}\left(-\frac{1}{2}\right)^{g+1-j}\binom{g+1-j}{k}\frac{\sum_{\sigma\in[g]_{g-j}}\prod_{i=1}^{g-j}\mathbf{P}_{\sigma_i}(u)}{(1-4u)^{3g-\frac{g-j+k}{2}}}$$
$$=\sum_{j=0}^{g-2}\sum_{s=g-j}^{2g+1-2j}\left(-\frac{1}{2}\right)^{g+1-j}\binom{g+1-j}{s-g+j}\frac{\sum_{\sigma\in[g]_{g-j}}\prod_{i=1}^{g-j}\mathbf{P}_{\sigma_i}(u)}{(1-4u)^{3g-\frac{s}{2}}}.$$

As $0\leq j\leq g-2$ and $g-j\leq s\leq 2g+1-2j$, we have $s\geq 2$. Consequently we arrive at

$$\frac{1}{z^2-z+1}\mathbf{D}_g^*(z)=\frac{\mathbf{U}_g(u)}{(1-4u)^{3g-1/2}}+\frac{\mathbf{V}_g(u)}{(1-4u)^{3g-1}},\tag{26}$$

where

$$\mathbf{U}_g(u)=\frac{\mathbf{P}_g(u)}{4}+\frac{\mathbf{P}_g(u)(1-4u)}{4}+\sum_{j=0}^{g-2}\sum_{\substack{g-j\leq s\leq 2g+1-2j\\ s\text{ is odd}}}$$
$$\times\left(-\frac{1}{2}\right)^{g+1-j}\binom{g+1-j}{s-g+j}\left(\sum_{\sigma\in[g]_{g-j}}\prod_{i=1}^{g-j}\mathbf{P}_{\sigma_i}(u)\right)$$
$$\times(1-4u)^{\frac{s-1}{2}},$$

and

$$\mathbf{V}_g(u)=\frac{\mathbf{P}_g(u)}{2}+\left(-\frac{1}{2}\right)^3\left(\sum_{\sigma\in[g]_2}\prod_{i=1}^{2}\mathbf{P}_{\sigma_i}(u)\right)+3\left(-\frac{1}{2}\right)^3$$
$$\times\left(\sum_{\sigma\in[g]_2}\prod_{i=1}^{2}\mathbf{P}_{\sigma_i}(u)\right)(1-4u)+\sum_{j=0}^{g-3}\sum_{\substack{g-j\leq s\leq 2g+1-2j\\ s\text{ is even}}}$$
$$\times\left(-\frac{1}{2}\right)^{g+1-j}\binom{g+1-j}{s-g+j}\left(\sum_{\sigma\in[g]_{g-j}}\prod_{i=1}^{g-j}\mathbf{P}_{\sigma_i}(u)\right)$$
$$\times(1-4u)^{\frac{s-2}{2}}.$$

We have for $\sigma\in[g]_k$, $k\geq 1$

$$[u^h]\left(\sum_{\sigma\in[g]_k}\prod_{i=1}^{k}\mathbf{P}_{\sigma_i}(u)\right)=\sum_{\sigma\in[g]_k}\prod_{i=1}^{k}[u^{h_i}]\,\mathbf{P}_{\sigma_i}(u),$$

where $\sum_{i=1}^{k}h_i=h$, $h_i\geq 0$. Then we obtain that

$$[u^h]\left(\sum_{\sigma\in[g]_k}\prod_{i=1}^{k}\mathbf{P}_{\sigma_i}(u)\right)=0,\quad 0\leq h\leq 2g-1.\tag{27}$$

Since $[u^{h_i}]\,\mathbf{P}_{\sigma_i}(u)=0$, $h_i\leq 2\sigma_i-1$, $[u^{2\sigma_i}]\,\mathbf{P}_{\sigma_i}(u)\neq 0$ and $\sum_{i=1}^{k}\sigma_i=g$. Thus for $0\leq h\leq 2g-1$,

$$[u^h]\,\mathbf{U}_g(u)=0\quad\text{and}\quad[u^h]\,\mathbf{V}_g(u)=0.\tag{28}$$

As shown in [11] we have

$$\mathbf{P}_g(1/4)=\frac{\Gamma\left(g-1/6\right)\Gamma\left(g+1/2\right)\Gamma\left(g+1/6\right)9^g4^{-g}}{6\pi^{3/2}\Gamma\left(g+1\right)}\tag{29}$$

and we obtain $\mathbf{U}_g(1/4)=\mathbf{P}_g(1/4)/4$. Furthermore,

$$\mathbf{V}_g(1/4)=\frac{\mathbf{P}_g(1/4)}{2}+\left(-\frac{1}{2}\right)^3\left(\sum_{\sigma\in[g]_2}\prod_{i=1}^{2}\mathbf{P}_{\sigma_i}(1/4)\right)$$
$$=\frac{1}{8}\left(4\mathbf{P}_g(1/4)-\sum_{j=1}^{g-1}\mathbf{P}_j(1/4)\mathbf{P}_{g-j}(1/4)\right)\neq 0.$$

We can recruit the computation of [11] in order to observe $4\mathbf{P}_g(1/4)-\sum_{j=1}^{g-1}\mathbf{P}_j(1/4)\mathbf{P}_{g-j}(1/4)\neq 0$. In order to compute the bivariate GF, $\mathbf{E}_g^*(z,t)$, we only need to replace in eq. (22) $\mathbf{D}_g(z)$ by $\mathbf{E}_g(z,t)$ and the proof is completely analogous.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
FWDH and CMR contributed equally to research and manuscript. Both authors read and approved the final manuscript.

Acknowledgements
We want to thank Thomas J.X. Li for discussions and comments. We want to thank an anonymous referee for pointing out an incorrect assumption of first version of this paper.

Received: 31 December 2011 Accepted: 11 October 2012
Published: 22 October 2012

References

1. Bailor MH, Sun X, Al-Hashimi HM: **Topology Links RNA Secondary Structure with Global Conformation, Dynamics, and Adaptation.** *Science* 2010, **327:**202–206.
2. Tabaska JE, Cary RB, Gabow HN, Stormo GD: **An RNA folding method capable of identifying pseudoknots and base triples.** *Bioinformatics* 1998, **14:**691–699.
3. Loebl M, Moffatt I: **The chromatic polynomial of fatgraphs and its categorification.** *Adv. Math.* 2008, **217:**1558–1587.
4. Penner RC, Knudsen M, Wiuf C, Andersen JE: **Fatgraph models of proteins.** *Comm Pure Appl Math* 2010, **63:**1249–1297.
5. Massey WS: *Algebraic Topology: An Introduction*. New York: Springer-Veriag; 1967.
6. Penner RC, Waterman MS: **Spaces of RNA secondary structures.** *Adv. Math* 1993, **101:**31–49.
7. Penner RC: **Cell decomposition and compactification of Riemann's moduli space in decorated Teichmü, ller theory.** In *Woods Hole Mathematics-perspectives in math and physics*. Edited by Tongring N, Penner RC: Singapore; World Scientific 2004:263–301. [ArXiv: math. GT/0306190].
8. Mathews D, Sabina J, Zuker M, Turner D: **Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure.** *J. Mol. Biol* 1999, **288:**911–940.

9. Reidys CM, Huang FWD, Andersen JE, Penner RC, Stadler PF, Nebel ME: **Topology and prediction of RNA pseudoknots.** *Bioinformatics* 2011, **27:**1076–1085.
10. Bon M, Vernizzi G, Orland H, Zee A: **Topological Classification of RNA Structures.** *J Mol Biol* 2008, **379:**900–911.
11. Andersen JE, Penner RC, Reidys CM, Waterman MS: **Topological classification and enumeration of RNA structrues by genus.** *J. Math. Biol* 2011 doi:10.1007/s00285-012-0594-x [Prepreint].
12. Smith T, Waterman M: **RNA secondary structure.** *Math. Biol* 1978, **42:**31–49.
13. Zuker M: **On finding all suboptimal foldings of an RNA molecule.** *Science* 1989, **244:**48–52.
14. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P: **Fast Folding and Comparison of RNA Secondary Structures.** *Monatsh Chem* 1994, **125:**167–188.
15. Rivas E, Eddy SR: **A Dynamic Programming Algorithm for RNA Structure Prediction Including Pseudoknots.** *J Mol Biol* 1999, **285:**2053–2068.
16. Uemura Y A Hasegawa, Kobayashi S, Yokomori T: **Tree adjoining grammars for RNA structure prediction.** *Theor Comp Sci* 1999, **210:**277–303.
17. Akutsu T: **Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots.** *Discr Appl Math* 2000, **104:**45–62.
18. Lyngsø RB, Pedersen CN: **RNA pseudoknot prediction in energy-based models.** *J Comp Biol* 2000, **7:**409–427.
19. Cai L, Malmberg RL, Wu Y: **Stochastic modeling of RNA pseudoknotted structures: a grammatical approach.** *Bioinformatics* 2003, **19 S1:** i66–i73.
20. Dirks RM, Pierce NA: **A partition function algorithm for nucleic acid secondary structure including pseudoknots.** *J Comput Chem* 2003, **24:**1664–1677.
21. Deogun JS, Donis R, Komina O, Ma F: **RNA secondary structure prediction with simple pseudoknots.** In *Proceedings of the second conference on Asia-Pacific bioinformatics (APBC 2004),* Australian Computer Society; 2004:239–246.
22. Reeder J, Giegerich R: **Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics.** *BMC Bioinformatics* 2004, **5:**104.
23. Li H, Zhu D: **A New Pseudoknots Folding Algorithm for RNA Structure Prediction.** In *COCOON 2005, Volume 3595*. Edited by Wang L. Berlin: Springer; 2005:94–103.
24. Matsui H, Sato K, Sakakibara Y: **Pair stochastic tree adjoining grammars for aligning and predicting pseudoknot RNA structures.** *Bioinformatics* 2005, **21:**2611–2617.
25. Kato Y, Seki H, Kasami T: **RNA Pseudoknotted Structure Prediction Using Stochastic Multiple Context-Free Grammar.** *IPSJ Digital Courier* 2006, **2:**655–664.
26. Chen HL, Condon A, Jabbari H: **An $O(n^5)$ Algorithm for MFE Prediction of Kissing Hairpins and 4-Chains in Nucleic Acids.** *J Comp Biol* 2009, **16:**803–815.
27. Waterman MS: **Secondary structure of single-stranded nucleic acids.** *Adv Math (Suppl Studies)* 1978, **1:**167–212.
28. Orland H, Zee A: **RNA folding and large N matrix theory.** *Nuclear Physics B* 2002, **620:**456–476.
29. Wexler Y, Zilberstein C, Ziv-Ukelson M: **A study of accessible motifs and RNA complexity.** *J Comput Biol* 2007, **14**(6):856–872.
30. Salari R, Möhl M, Will S, Sahinalp C, Backofen R: **Time and space efficient RNA-RNA interaction prediction via sparse folding.** *Proc of RECOMB* 2010, **6044:**473–490.
31. Backofen R, Tsur D, Zakov S, Ziv-Ukelson M: **Sparse RNA folding: Time and space efficient algorithms.** *J Disc Algor* 2011, **9**(1):12–31.
32. Möhl M, Salari R, Will S, Backofen R, Sahinalp SC: **Sparsification of RNA structure prediction including pseudoknots.** *Algorithms Mol Biol* 2010, **5:**39.
33. McCaskill JS: **The equilibrium partition function and base pair binding probabilities for RNA secondary structure.** *Biopolymers* 1990, **29:**1105–1119.
34. Dimitrieva S, Bucher P: **Practicality and time complexity of a sparsified RNA folding algorithm.** *J Bioinfo Comput Biol* 2012, **10**(2):1241007.
35. Kafri Y, Mukamel D, Peliti L: **Why is the DNA Denaturation Transition First Order?** *Phys Rev Lett* 2000, **85:**4988–4991.
36. Kabakcioglu A, Stella AL: **A scale-free network hidden in the collapsing polymer.** *Phys Rev E* 2005, **72:**055102.
37. Vanderzande C: *Lattic models of polymers*. New York: Cambridge University Press; 1998.
38. **NCBI database.** [http://www.ncbi.nlm.nih.gov/guide/dna-rna/#downloads_]
39. Nussinov R, Piecznik G, Griggs JR, Kleitman DJ: **Algorithms for Loop Matching.** *SIAM J Appl Math* 1978, **35:**68–82.
40. Nebel ME: **Investigation of the Bernoulli model for RNA secondary structures.** *Bull math biol* 2003, **66**(5):925–964.
41. Zagier D: **On the distribution of the number of cycles of elements in symmetric groups.** *Nieuw Arch Wisk IV* 1995, **13:**489–495.
42. Flajolet P, Sedgewick R: *Analytic Combinatorics*. New York: Cambridge University Press; 2009.
43. Han HSW, Reidys CM: **The 5′-3′ distance of RNA secondary structures.** *J Comput Biol* 2012, **19**(7):867–878.

doi:10.1186/1748-7188-7-28
Cite this article as: Huang and Reidys: On the combinatorics of sparsification. *Algorithms for Molecular Biology* 2012 **7**:28.