Bull Math Biol (2011) 73: 754–776
DOI 10.1007/s11538-010-9607-3

ORIGINAL ARTICLE

Parametric Analysis of RNA Branching Configurations

Valerie Hower · Christine E. Heitsch

Received: 22 April 2009 / Accepted: 4 November 2010 / Published online: 5 January 2011

Abstract Motivated by recent work in parametric sequence alignment, we study the parameter space for scoring RNA folds and construct an RNA polytope. A vertex of this polytope corresponds to RNA secondary structures with common branching. We use this polytope and its normal fan to study the effect of varying three parameters in the free energy model that are not determined experimentally. Our results indicate that variation of these specific parameters does not have a dramatic effect on the structures predicted by the free energy model. We additionally map a collection of known RNA secondary structures to the RNA polytope.

Keywords RNA secondary structure · Plane tree · Free energy · Thermodynamic model · Parametric analysis

1 Introduction

Determining the structure of RNA molecules remains a fundamental scientific challenge, since current methods cannot always identify the "correct" fold from the large number of possible configurations. A common method for predicting the secondary structure of a single RNA molecule, termed the *thermodynamic model*, involves free energy minimization (Mathews and Turner 2006; Zuker 2000; Zuker et al. 1999). Extensions to this approach, such as suboptimal structure prediction and partition function calculations, still depend on the parameters from the thermodynamic model to score possible secondary structures. The free energy of a secondary structure is

V. Hower (✉)
Department of Mathematics, University of California, Berkeley, 970 Evans Hall #3840, Berkeley, CA 94720, USA
e-mail: vhower@math.berkeley.edu

C.E. Heitsch
School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, USA

calculated by scoring substructures according to a set of parameters—most of which are determined experimentally (see SantaLucia and Turner 1997 for a review). A dynamic programming algorithm, used in software packages like mfold (Mathews et al. 1999; Zuker 2003), computes the minimal free energy as well as the optimal secondary structure(s) (Mathews et al. 2004).

In this work, we address variation in the parameter space for scoring secondary structures, focusing on three parameters from the multibranch loop energy function that are not based on measurement. Specifically, we address the following questions. What is the geometry of the parameter space for scoring RNA folds and how does this geometry relate back to the biology? How sensitive is the thermodynamic model to variation of the ad-hoc multibranch loop parameters? We answer these questions using geometric combinatorics. We find that variation of the multibranch loop parameters has a smaller effect than the change in the parameter space coming from improved measurement. Moreover, regardless of the choice of multibranch loop parameters used in the current version of the thermodynamic model, the minimal energy structures have a low degree of branching.

Our results are achieved by applying techniques from geometric combinatorics to give a parametric analysis of RNA folding. We construct an RNA polytope whose vertices correspond to sets of secondary structures with common branching. Its normal fan subdivides the parameter space so that the parameters lying in the same cone give the same minimal free energy structures. These approaches have been used recently in parametric sequence alignment (Dewey et al. 2006; Dewey and Woods 2005; Pachter and Sturmfels 2004a) and for more general hidden Markov models (Beerenwinkel et al. 2005; Pachter and Sturmfels 2004b). There is also earlier polyhedral work on parametric sequence alignment (Gusfield et al. 1992; Waterman et al. 1992) and related work on secondary structure comparison (Wang and Zhao 2003) and sequence/structure alignment (Lenhof et al. 1998). We additionally make comparisons with biological structures, and this work supports our theoretical results.

2 Biological Motivations and Implications

Under the thermodynamic model, the folding of an RNA sequence is predicted to maximize the stabilizing base pairs while minimizing the energetic cost of loop structures. These optimizations depend on the specific parameter values used to score the favorability of RNA secondary structures under the thermodynamic model. Hence, we focus here on a parametric analysis of RNA folding.

We investigate a combinatorial model of RNA folding to gain insight into the trade-offs among the different types of loop structures—in particular the dependence on the thermodynamic parameters which can be analyzed parametrically using geometric combinatorics. Since this is one of the first such analyses, we limited our thermodynamic model to the loop energies to keep the set of parameters to a reasonable size for these initial results. In Sect. 4, we give a complete geometric characterization of our simplified thermodynamic parameter space and the associated RNA polytope.

As illustrated in Fig. 3, we see that our RNA polytope is a tetrahedron whose boundary corresponds to (sets of) trees optimal for different possible loop parameters. As we discuss, the vertices correspond to different trade-offs in terms of the thermodynamic penalties/rewards for the number of hairpin loops, of bulge/internal loops, and of helices in the external loop. We note this analysis holds for arbitrary size structures and generic thermodynamic parameters in our combinatorial model of RNA folding.

We then investigate some specific variations in the parameter space, by considering four different types of combinatorial RNA sequences. This allows us to focus particularly on the three parameters which govern the ad hoc multibranch loop energy function. As described elsewhere, for instance in Mathews et al. (1999), most of the 10,000+ parameters in the current thermodynamic model are derived from experimental results measuring the stability of different loop structures. A notable exception to this is the affine energy function which is used to score the entropic effects for multibranch loops. This function, chosen primarily for computational expediency, is known to be a very low-order approximation to the complicated thermodynamics of branching loops. Although experimentalists are now beginning to measure the thermodynamics of branching loops (Diamond et al. 2001; Mathews and Turner 2002), the three function parameters currently in use were chosen through a knowledge-based approach. Thus, we investigate the effect that varying said parameters has on determining the optimal structures in our combinatorial model.

As in previous work (Bakhtin and Heitsch 2009), we find qualitative differences when the current thermodynamic parameters (version 3.0) are compared against the previous ones (version 2.3). In particular, when applied to our combinatorial RNA sequences, the version 2.3 parameters favor a high degree of branching in the external loop while the current ones favor a much lower degree of branching overall and show more dependence on the base composition of the sequence. Hence, we see that the behavior of the current thermodynamic parameters is more biologically realistic, although it is also likely to be more sensitive to changes in the ad hoc multibranch function parameters.

Finally, we compare the results of our parametric analysis with the branching of a set of known RNA secondary structures. When interpreted as plane trees, a substantial fraction of the known RNA secondary structures lie on the boundary of the appropriate RNA polytope, and so would be minimal for some choice of parameters. Since they are distributed among different facets, we conclude (as expected) that there are important aspects of RNA folding which our combinatorial model does not capture.

We can still, though, consider the implications for RNA folding gained from the parametric analysis of our combinatorial model. Trees on the boundary of our polytope are less than half the size of trees found in the interior on average. This suggests as we discuss that a simpler thermodynamic model may be sufficient for smaller RNA molecules. It is already known that even the full thermodynamic model is not adequate to accurately predict large RNA secondary structures. We find it intriguing that more than 80% of the known secondary structures are close to the two polytope vertices which would be optimal for our combinatorial RNA sequences under the current thermodynamic model. This suggests that there are essential biological characteristics, namely the thermodynamically favored branching configurations, which are being captured by our combinatorial model of RNA folding.

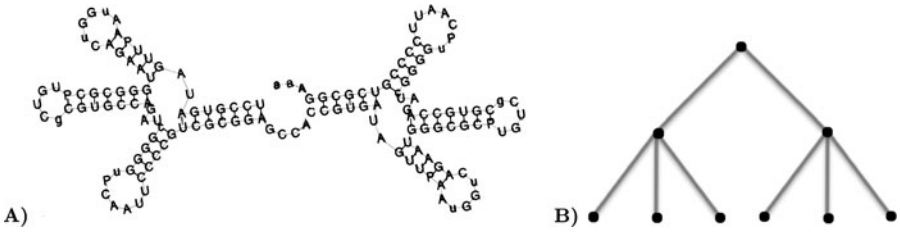

Fig. 1 Secondary structures as rooted plane trees

3 Background

3.1 Plane Trees and RNA Folding

We use a simplified model of RNA folding in which a secondary structure S is represented by a rooted plane tree $T = T(S)$. Single-stranded RNA sequences fold into molecular structures. One step in this folding process is the formation of Watson–Crick and also G-U base pairs. The set of (nested) base pairs determines the secondary structure of an RNA sequence. As illustrated in Fig. 1, a secondary structure has two basic types of substructures—runs of stacked base pairs which are called *helices* and the single-stranded regions known as *loops*. Every component of a secondary structure is given an associated free energy score by the thermodynamic model. To a first approximation, the score of a loop is determined its *degree*—the number of base pairs contained in the loop, or equivalently the number of helices meeting the loop. There are different energy functions for hairpin loops, which have degree 1, bulge/internal loops with degree 2, multibranch loops with degree greater than 2, and the exterior loop, which includes the unpaired bases not contained in any hairpin loop, bulge/internal loop, or multibranch loop. Suppose L is multibranch loop. The current free energy model uses the following formula for the energy of L:

$$E(L) = a + bn_1 + cn_2 + q, \tag{1}$$

where n_1 is the number of single-stranded bases in L, n_2 is the number of helices in L, q is the sum of the single-base stacking energies (also called "dangling energies" Zuker et al. 1999) in L, and a, b, c are the parameters for offset, free base, and helix penalties, respectively (Zuker et al. 1999). Equation (1) is not based on experimental measurement, but rather it is used in order to facilitate faster computations. In this work, our analysis is primarily focused on the three parameters a, b, and c from this function since they are not experimentally determined. Our results are obtained by considering rooted plane trees as a simplified model of RNA folding.

Plane trees have been used to enumerate possible RNA secondary structures (Schmitt and Waterman 1994) and also to compare them (Le et al. 1989; Shapiro and Zhang 1990) for some time now. The interaction between combinatorics and RNA folding has continued to develop over the last 20 years, including using trees as more abstract representations of RNA folding, for instance in Gan et al. (2003) and related work as well as in Bakhtin and Heitsch (2008, 2009), Heitsch (2010). A *rooted plane tree* (also called *plane tree* or *ordered tree* (Dershowitz and Zaks 1980;

Stanley 1999)) is a tree with a specified root vertex and such that the subtrees of any given vertex are ordered. This ordering comes from the $5' \to 3'$ linear arrangement of the RNA sequence. Plane trees with n edges are one of the many combinatorial objects counted by the Catalan numbers

$$C_n = \frac{1}{n+1}\binom{2n}{n}. \tag{2}$$

To obtain T, we assign the root vertex to the exterior loop of S and the non-root vertices of T correspond to the remaining loops in S. Two vertices in T share an edge when their loops in S are connected by helices. As an example, we give a secondary structure in Fig.1A together with its associated plane tree in Fig. 1B. Technically, a secondary structure S must be free of pseudoknots in order to construct T. While pseudoknots do occur in secondary structures, the thermodynamic model cannot predict them and moreover one can create a nested, pseudoknot-free structure from a given fold in several ways—some of which are in Smit et al. (2008) and our approach is described the Materials and Methods section. Given a plane tree T with n edges, we write r for the degree of the root vertex and for $0 \le k \le n$, d_k is the number of nonroot vertices with k children. Thus, d_k gives the number of non-root vertices in T with degree $k+1$, and this is the number of loops in S with $k+1$ branches. To assign an energy to a plane tree, we assign weights to the vertices, based on the down degree of the vertex. In terms of secondary structures, we are assigning the same energy to each type of loop in the fold. This is a simplification of the scoring for the thermodynamic model, in which the energy of a structure is the sum of the energies of the loops. For example, to assign an energy to a multibranch loop vertex of degree k, we use the energy function (1) for a multibranch loop where the number of free bases is a multiple of the number of helices. The parameters b and c in (1) are incorporated into one parameter, and (1) simplifies to $c_2 + a_2(k+1)$. If T is a plane tree with n edges, the free energy of T is written as

$$\begin{aligned} E(T) &= a_3 r + a_0 d_0 + a_1 d_1 + \sum_{k=2}^{n} \big[c_2 + a_2(k+1)\big] d_k \\ &= a_3 r + a_0 d_0 + a_1 d_1 + (c_2 + 2a_2) \sum_{k=2}^{n} d_k + a_2 \sum_{k=2}^{n} (k-1) d_k \\ &= a_3 r + a_0 d_0 + a_1 d_1 + (c_2 + 2a_2)(n - d_0 - d_1) + a_2(d_0 - r) \\ &= (c_2 + 2a_2)n + (a_3 - a_2)r + (a_0 - c_2 - a_2)d_0 + (a_1 - c_2 - 2a_2)d_1, \end{aligned}$$

where we have used the relations

$$\sum_{k=2}^{n} d_k = n - d_0 - d_1 \quad \text{and} \quad \sum_{k=2}^{n} (k-1) d_k = d_0 - r$$

that hold for all plane trees (Stanley 1999). We refer the reader to Sect. 4.4 for further discussion the energy of a plane tree. To minimize free energy, we must minimize $E(T)$ over the space of all plane trees. Since this space is infinite, we will typically think of n as being fixed but arbitrary and minimize the free energy function over the finite space of plane trees with n edges. For a given set of parameters a_0, a_1, a_2, a_3, c_2,

this is equivalent to minimizing the following inner product

$$E'(T) = (\theta_2, \theta_3, \theta_4) \cdot (r, d_0, d_1) \tag{3}$$

where

$$\theta_2 = a_3 - a_2, \qquad \theta_3 = a_0 - c_2 - a_2, \qquad \theta_4 = a_1 - c_2 - 2a_2.$$

3.2 Geometric Combinatorics

In this section, we present some basic definitions in geometric combinatorics. We refer the reader to Grünbaum (2003), Ziegler (1995) for a more detailed treatment. A set $U \subset \mathbb{R}^d$ is *convex* if for any two points $x, y \in U$, the line segment connecting x and y is contained in U, that is $\{\alpha x + (1-\alpha)y \mid 0 \le \alpha \le 1\} \subset U$. For any subset U of $\mathbb{R}^d$, the *convex hull* of U, written $\operatorname{conv} U$, is the intersection of all convex sets that contain U. A *lattice polytope* $\Delta \subset \mathbb{R}^d$ is the convex hull of a finite collection of lattice points: $\Delta = \operatorname{conv}\mathcal{A}$, where $\mathcal{A} = \{y_1, y_2, y_3, \ldots, y_r\} \subset \mathbb{Z}^d$.

Any lattice polytope Δ is characterized by a finite collection of defining inequalities

$$\{c_i \cdot x \ge b_i\}_{i \in I} \quad \text{where } c_i \in \mathbb{Z}^d,\ x \in \Delta,\ \text{and } b_i \in \mathbb{Z}. \tag{4}$$

A *face* F of Δ is a subset defined by setting some of the defining inequalities to equality, i.e.,

$$F = \left\{ x \in \Delta \;\middle|\; \begin{array}{l} c_{i_1} \cdot x = b_{i_1} \\ c_{i_2} \cdot x = b_{i_2} \\ \quad \vdots \\ c_{i_k} \cdot x = b_{i_k} \end{array} \right\},$$

and the dimension of F is the dimension of its affine span. The *vertices* of Δ are the 0-dimensional faces while the *facets* have dimension $\dim \Delta - 1$. The *boundary* of Δ, written $\partial\Delta$ is the union of all faces of Δ of dimensions $0, 1, 2, \ldots, \dim \Delta - 1$.

A *convex polyhedral cone* σ is the positive hull of a finite collection of lattice points in $\mathbb{Z}^d$: $\sigma = \{t_1 z_1 + t_2 z_2 + \cdots + t_s z_s \mid t_i \ge 0, z_i \in \mathbb{Z}^d\}$, and we write $\sigma = \langle z_1, z_2, \ldots, z_s \rangle$. Associated to each lattice polytope Δ is its normal fan $\mathcal{N}(\Delta)$ that will give us the set of parameter values which makes a particular face of Δ optimal. Geometrically, $\mathcal{N}(\Delta)$ is a collection of cones that are in one-to-one correspondence with faces F of Δ:

$$\sigma_F = \left\{ v \in \mathbb{R}^d \mid u \cdot v \le x \cdot v \quad \forall u \in F,\ \forall x \in \Delta \right\}. \tag{5}$$

Note that $\dim \sigma_F = \dim \Delta - \dim F$. As an example of the above concepts, we give a 2-dimensional polytope Δ in Fig. 2A and its normal fan $\mathcal{N}(\Delta)$ in Fig. 2B. The four vertices of Δ correspond to the four 2-dimensional cones in $\mathcal{N}(\Delta)$, and the four facets of Δ correspond to the four rays of $\mathcal{N}(\Delta)$.

In terms of minimization, (5) states that the points in F are minimizers of the dot product for vectors in σ_F, among all points in Δ. Readers familiar with linear programming will recognize that the polytope Δ is the feasible region of a linear program with constraints coming from the rays (1-dimensional cones) of $\mathcal{N}(\Delta)$. Taking

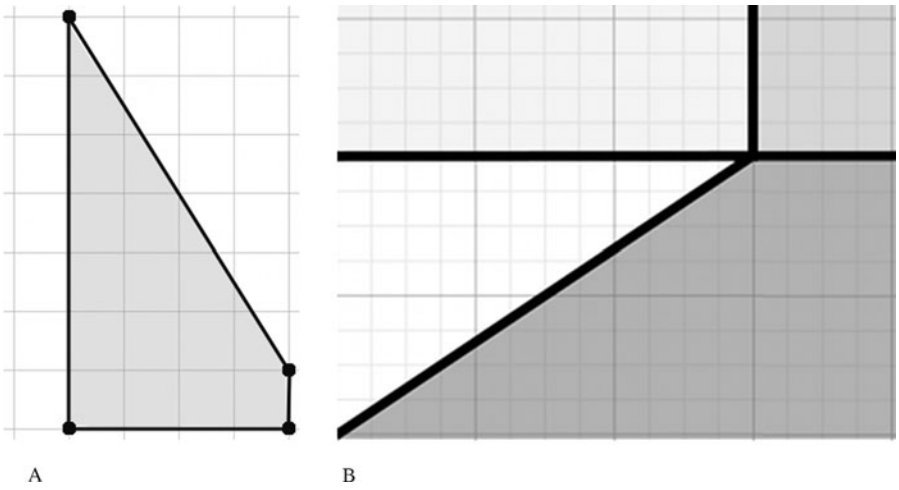

Fig. 2 A 2-dimensional polytope Δ (**A**) and its normal fan $\mathcal{N}(\Delta)$ (**B**)

the inner product with any vector in $\mathbb{R}^d$ gives the objective function for a linear program over this feasible region. The correspondence between faces of Δ and cones in $\mathcal{N}(\Delta)$ in (5) says vectors in the face F solve any linear program whose objective function lies in the cone σ_F. Our analysis in this work involves linear programming over all possible count vectors (r, d_0, d_1).

4 Results

4.1 Plane Trees that Minimize Energy

Fixing $n \geq 5$, the possible count vectors (r, d_0, d_1) of plane trees are classified by the second author (Heitsch 2010) and fall into one of four classes, as listed in Table 1 with $r, d_0, d_1 \geq 0$ in all cases. Since r, d_0, d_1 must all be integers, the vertices in Table 1B or Table 1D differ depending on whether or not n is even or odd. We want to minimize the linear energy function over this point set (which includes count vectors from all four cases), and hence we let P_n be the convex hull of the union of the four polytopes listed in Table 1. Regardless of our choice of energy parameters, a minimum energy plane tree with n edges will occur at a vertex of P_n. The following proposition describes the vertices of P_n.

Proposition 4.1 *Define* Ψ_n *as follows*:

$$\Psi_n := \begin{cases} \operatorname{conv}\{(1, \frac{n+1}{2}, 0), (1, n-1, 0), (1, 1, n-1), (n, n, 0)\} \\ \quad n \text{ odd} \\ \operatorname{conv}\{(1, \frac{n+2}{2}, 0), (1, \frac{n}{2}, 1), (2, \frac{n+2}{2}, 0), (1, n-1, 0), (1, 1, n-1), (n, n, 0)\} \\ \quad n \text{ even} \end{cases}$$

Then $\Psi_n = P_n$ *for* $n \geq 5$.

Table 1 Sets of inequalities and corresponding vertices for plane trees

	Set of inequalities	Vertices for n even	Vertices for n odd
(A)	$r = 1$ $d_0 = 1$ $d_1 = n - 1$	$\{(1, 1, n-1)\}$	$\{(1, 1, n-1)\}$
(B)	$r = 1$ $2 \le d_0 \le n$ $n - 2d_0 + 1 \le d_1$ $d_1 \le n - d_0 - 1$	$\left\{ \begin{array}{l} (1, 2, n-3), (1, \frac{n+2}{2}, 0), \\ (1, \frac{n}{2}, 1), (1, n-1, 0) \end{array} \right\}$	$\left\{ \begin{array}{l} (1, 2, n-3), (1, \frac{n+1}{2}, 0), \\ (1, n-1, 0) \end{array} \right\}$
(C)	$r = d_0$ $2 \le d_0 \le n$ $d_1 = n - d_0$	$\{(2, 2, n-2), (n, n, 0)\}$	$\{(2, 2, n-2), (n, n, 0)\}$
(D)	$2 \le r$ $r \le 2d_0 - n + d_1$ $3 \le d_0 \le n - 1$ $n - 2d_0 + 2 \le d_1$ $d_1 \le n - d_0 - 1$	$\left\{ \begin{array}{l} (2, n-1, 0), (n-2, n-1, 0), \\ (2, 3, n-4), (2, \frac{n+2}{2}, 0) \end{array} \right\}$	$\left\{ \begin{array}{l} (2, n-1, 0), (n-2, n-1, 0), \\ (2, \frac{n+3}{2}, 0), (3, \frac{n+3}{2}, 0), \\ (2, 3, n-4), (2, \frac{n+1}{2}, 1) \end{array} \right\}$

Proof Clearly, $\Psi_n \subset P_n$, and hence we'll show each lattice point of P_n in Table 1 is contained in Ψ_n. The normal fan of Ψ_n has rays

$$\begin{array}{ll} \big\{(-1, 2, 1), (1, 0, 0), (1, 1-n, 2-n), (0, 0, 1)\big\} & n \text{ odd} \\ \big\{(-1, 2, 1), (1, 0, 0), (1, 1-n, 2-n), (0, 0, 1), (0, 1, 1)\big\} & n \text{ even} \end{array}$$

Moreover, for each lattice point $t = (r, d_0, d_1)$ in Table 1, one can verify that t satisfies the defining inequalities of Ψ_n:

$$\begin{aligned} (r, d_0, d_1) \cdot (-1, 2, 1) &\ge n \\ (r, d_0, d_1) \cdot (1, 0, 0) &\ge 1 \\ (r, d_0, d_1) \cdot (1, 1-n, 2-n) &\ge 2n - n^2 \\ (r, d_0, d_1) \cdot (0, 0, 1) &\ge 0 \end{aligned}$$

and for n even we additionally have

$$(r, d_0, d_1) \cdot (0, 1, 1) \ge \frac{n+2}{2}.$$

This gives $P_n \subset \Psi_n$, and we have equality. □

In the sequel, we will primarily focus on the rational tetrahedron

$$\Delta_n := \operatorname{conv}\left\{ \left(1, \frac{n+1}{2}, 0\right), (1, n-1, 0), (1, 1, n-1), (n, n, 0) \right\}$$

regardless of whether n is even or odd. There are many reasons for this. First, asymptotically, there is no difference between P_n and Δ_n for n even. The normal fan $\mathcal{N}(P_n)$ is obtained from $\mathcal{N}(\Delta_n)$ by adding a single ray and subdividing the full dimensional cone $\sigma = \langle (1,0,0), (0,0,1), (-1,2,1) \rangle$ corresponding to the vertex $(1, \frac{n+1}{2}, 0)$. Thus, when n is even, the parameters giving $(1,1,n-1)$, $(1,n-1,0)$, or $(n,n,0)$ the minimal energy are the same regardless of whether we use the subdivision of $\mathbb{R}^3$ determined by $\mathcal{N}(P_n)$ or that determined by $\mathcal{N}(\Delta_n)$. Moreover, the parameters in σ will yield $(1, \frac{n+2}{2}, 0)$, $(1, \frac{n}{2}, 1)$, or $(2, \frac{n+2}{2}, 0)$ as minimal, and the trees corresponding to these three count vectors are all similar, as discussed in Proposition 4.3.

4.2 Lattice Points in ∂P_n

Suppose S is a secondary structure whose plane tree has count vector (r, d_0, d_1). If $(r, d_0, d_1) \in \operatorname{int} P_n$ then there is no choice of parameters that can make S have minimal free energy. Conversely, if $(r, d_0, d_1) \in \operatorname{int} F$ for some face F of P_n, then any parameter vector in the cone $\sigma_F \subset \mathcal{N}(P_n)$ yields S with minimal energy. We thus want to determine the count vectors lying on ∂P_n.

All four sets of inequalities in Table 1 intersect ∂P_n. Let Q_A, Q_B, Q_C, and Q_D be the polyhedra described in Table 1(A), (B), (C), and (D), respectively. Then $Q_A, Q_B, Q_C \subset \partial P_n$ and

$$\begin{aligned} Q_A &= \big\{(1,1,n-1)\big\} \\ (Q_A \cup Q_B) \cap \mathbb{Z}^3 &= \operatorname{conv}\left\{(1,n-1,0), (1,1,n-1), \left(1, \frac{n+1}{2}, 0\right)\right\} \cap \mathbb{Z}^3 \\ (Q_A \cup Q_C) \cap \mathbb{Z}^3 &= \operatorname{conv}\big\{(1,1,n-1), (n,n,0)\big\} \cap \mathbb{Z}^3. \end{aligned}$$

Since Q_D is 3-dimensional, it cannot be contained in the boundary of P_n. We do, however, have

$$(Q_D \cap \partial P_n) \cap \mathbb{Z}^3 = (\operatorname{int} E_1 \cup \operatorname{int} F_1 \cup \operatorname{int} F_2) \cap \mathbb{Z}^3, \tag{6}$$

where $E_1 = \operatorname{conv}\{(n,n,0), (1, \frac{n+1}{2}, 0)\}$, $F_1 = \operatorname{conv}\{(n,n,0), (1, \frac{n+1}{2}, 0), (1,1,n-1)\}$, and $F_2 = \operatorname{conv}\{(n,n,0), (1, \frac{n+1}{2}, 0), (1,n-1,0)\}$. Equation (6) follows from counting lattice points in the objects on the left and right-hand sides of the equation using the same technique as in Proposition 4.2. The plane trees defined in Table 1D that lie on ∂P_n satisfy $d_1 = 0$ or $r = 2d_0 - n + d_1$. Their associated secondary structures either have no bulges/internal loops or have a maximal number of helices in the exterior loop.

Next, we count the number of lattice points in the interior of each face of P_n. For an edge of the form $E = \operatorname{conv}\{(x_1, y_1, z_1), (x_2, y_2, z_2)\}$, we use the formula

$$\#\big(\operatorname{int} E \cap \mathbb{Z}^3\big) = \gcd\big(|x_1 - x_2|, |y_1 - y_2|, |z_1 - z_2|\big) - 1$$

and obtain the following counts. The edges $\operatorname{conv}\{(n,n,0), (1, \frac{n+1}{2}, 0)\}$ and $\operatorname{conv}\{(1,1,n-1), (1, \frac{n+1}{2}, 0)\}$ each have $\frac{1}{2}(n-3)$ lattice points in their interiors.

A total of $\frac{1}{2}(n-5)$ lattice points are in the interior of $\mathrm{conv}\{(1,n-1,0),(1,\frac{n+1}{2},0)\}$. The interior of $\mathrm{conv}\{(n,n,0),(1,1,n-1)\}$ contains $n-2$ lattice points, and there are no interior lattice points for the edges $\mathrm{conv}\{(n,n,0),(1,n-1,0)\}$ and $\mathrm{conv}\{(1,1,n-1),(1,n-1,0)\}$.

To determine the number of lattice points in a facet F of P_n, we use Pick's theorem (Iseri 2008)

$$\#\big(\mathrm{int}\,F\cap\mathbb{Z}^3\big)=\mathrm{Area}(F)-\frac{1}{2}\big[\#\big(\partial F\cap\mathbb{Z}^3\big)\big]+1,$$

where the area of F is normalized with respect to the 2-dimensional sublattice containing F. We illustrate Pick's theorem with the following proposition.

Proposition 4.2 *There are no interior lattice points in the facet*

$$F=\mathrm{conv}\big\{(1,1,n-1),(1,n-1,0),(n,n,0)\big\}.$$

Proof The triangle F lies on the hyperplane $-X+(n-1)Y+(n-2)Z=n^2-2n$ in $\mathbb{R}^3$, and thus we normalize the area of F by dividing by $\sqrt{(-1)^2+(n-1)^2+(n-2)^2}=\sqrt{2(n^2-3n+3)}$. Before normalization, the area of F is

$$\begin{aligned}&\frac{1}{2}\sqrt{\begin{vmatrix}1&1&n\\1&n-1&n\\1&1&1\end{vmatrix}^2+\begin{vmatrix}1&n-1&n\\n-1&0&0\\1&1&1\end{vmatrix}^2+\begin{vmatrix}n-1&0&0\\1&1&n\\1&1&1\end{vmatrix}^2}\\&=\frac{1}{2}\sqrt{2n^4-10n^3+20n-18n+6}\\&=\frac{1}{2}(n-1)\sqrt{2\big(n^2-3n+3\big)}.\end{aligned}$$

Moreover, using the counts above for the interior lattice points in the edges of F, we have

$$\#\big(\partial F\cap\mathbb{Z}^3\big)=(n-2)+0+0+3=n+1.$$

Applying Pick's theorem yields

$$\begin{aligned}\#\big(\mathrm{int}\,F\cap\mathbb{Z}^3\big)&=\frac{1}{2}(n-1)-\frac{1}{2}(n+1)+1\\&=0.\end{aligned}$$

□

For the other three facets of P_n, each contains $\frac{1}{4}(n-3)^2$ interior lattice points. In total, this gives $\frac{1}{4}(3n^2-8n+13)$ lattice points on ∂P_n, all of which correspond to plane trees.

4.3 Biological Meaning of P_n and $\mathcal{N}(P_n)$

4.3.1 The Vertices of P_n

The vertices of P_n represent the secondary structures with the maximum number of helices in a loop—so-called "maximal degree of branching"—and the fewest helices in a loop—or "minimal degree of branching"—as described below.

If T is a plane tree represented as a vertex of P_n, then T has n edges and $n+1$ vertices. If in addition, T has count vector $(n, n, 0)$ then the degree of the root vertex is n and the $n+1$ vertices are the root together with the n leaves (vertices with 0 children). Thus, a secondary structure corresponding to T has no internal loops, bulges, or multibranch loops and the exterior loop has n helices.

If T has count vector $(1, 1, n-1)$, the root vertex has degree 1, there is one leaf, and $n-1$ vertices of degree 2 (1 child). Thus, T is a straight line, and a secondary structure corresponding to T has no multibranch loops and the exterior loop has one helix.

If T has count vector $(1, n-1, 0)$, the $n+1$ vertices are the root (with degree 1), $n-1$ leaves, and one vertex of degree n. Secondary structures corresponding to T have no internal loops or bulges and one multibranch loop with n helices. In addition, the exterior loop has one helix.

The remaining vertices—$(1, \frac{n+2}{2}, 0)$ for n odd and $(1, \frac{n}{2}, 1)$, $(2, \frac{n+2}{2}, 0)$, or $(1, \frac{n+2}{2}, 0)$ for n even—are dealt with in the following proposition.

Proposition 4.3

(i) *For n odd, any plane tree with count vector* $(1, \frac{n+1}{2}, 0)$ *satisfies* $d_2 = \frac{n-1}{2}$ *and* $d_i = 0$ *for* $i > 2$.
(ii) *For n even, any plane tree with count vector* $(1, \frac{n}{2}, 1)$ *or* $(2, \frac{n+2}{2}, 0)$ *satisfies* $d_2 = \frac{n-2}{2}$ *and* $d_i = 0$ *for* $i > 2$.
(iii) *For n even, any plane tree with count vector* $(1, \frac{n+2}{2}, 0)$ *satisfies* $d_2 = \frac{n-4}{2}$, $d_3 = 1$, *and* $d_i = 0$ *for* $i > 3$.

Proof For (i), suppose n is odd and T is a plane tree with n edges, $r = 1$, $d_0 = \frac{n+1}{2}$, and $d_1 = 0$. Then T has $\frac{n+1}{2} + 1$ vertices of degree 1, and the remaining $n + 1 - (\frac{n+1}{2} + 1) = \frac{n-1}{2}$ vertices have degree at least 3. Thus,

$$\begin{aligned}\sum_{v \in V} \deg v &= \frac{1}{2}(n+1) + 1 + \sum_{\deg v \geq 3} \deg v \\ &\geq \frac{1}{2}(n+1) + 1 + \frac{3}{2}(n-1) \\ &= 2n.\end{aligned}$$

However, since $\sum_{v \in V} \deg v = 2|E|$, we must have equality. Thus, all other vertices must have degree 3 (2 children).

The proof of (ii) is nearly identical to that of (i).

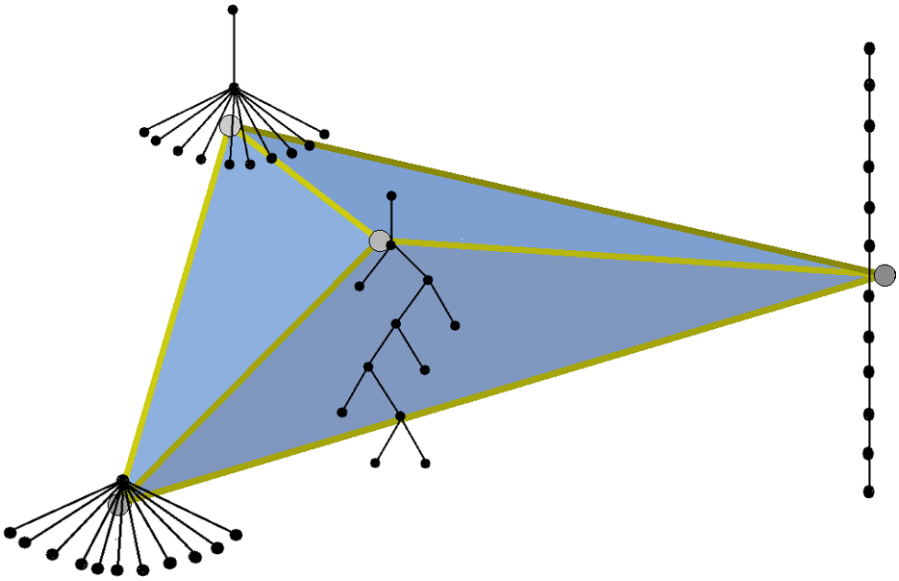

Fig. 3 The RNA polytope P_n

For (iii), a plane tree with n edges, $r = 1$, $d_0 = \frac{n+2}{2}$ and $d_1 = 0$ has $\frac{n+4}{2}$ vertices of degree 1 and zero vertices of degree 2. Such a tree cannot have all other vertices of degree 3 as this would yield a graph with an odd number of odd vertices. Thus, there is a vertex v_0 with degree p with $p \geq 4$ even. This gives

$$\begin{aligned}\sum_{v \in V} \deg v &= \frac{1}{2}(n+4) + p + \sum_{\substack{\deg v \geq 3 \\ v \neq v_0}} \deg v \\ &\geq \frac{1}{2}(n+4) + 4 + \frac{3}{2}(n-4) \\ &= 2n\end{aligned}$$

As before this inequality must be an equality, and hence $p = 4$ and all other vertices have degree 3. □

Thus, for n odd, the count vector $(1, \frac{n+1}{2}, 0)$ corresponds to secondary structures with no interior loops/bulges, all multibranch loops have 3 helices, and the exterior loop has one helix. When n is even, a secondary structure with n helices and all three of these properties is not possible. We instead have three cases, each with exactly one of the properties relaxed: a structure corresponding to $(1, \frac{n}{2}, 1)$ has one interior loop/bulge, the count vector $(1, \frac{n+2}{2}, 0)$ arises from structures having one multibranch loop with 4 helices (all other multibranch loops have 3 helices), and the exterior loop of a structure corresponding to $(2, \frac{n+2}{2}, 0)$ has 2 helices. For n odd, plane trees representative of those described in this section are shown in Fig. 3.

Remark The map from plane trees to count vectors is generically many-to-one. Three of the 4 vertices, however, correspond to exactly one tree: $(n, n, 0)$, $(1, 1, n-1)$, $(1, n-1, 0)$. The trees with count vector $(1, \frac{n+1}{2}, 0)$ are in one-to-one correspondence with full binary trees with $n-1$ edges (by removing the root vertex). There are $C_{\frac{n-1}{2}}$ such trees (Deutsch 2004), where $C_{\frac{n-1}{2}}$ is the $\frac{n-1}{2}$th Catalan number defined in (2).

4.3.2 The Rays in $\mathcal{N}(P_n)$

The energy function E' in (3) scores a secondary structure with n helices based on the number of helices in the exterior loop, the number hairpin loops, and the number of bulges/internal loops. The normal fan $\mathcal{N}(P_n)$ of P_n subdivides the $(\theta_2, \theta_3, \theta_4)$ parameter space. Each vector in $(x, y, z) \in \mathbb{R}[\theta_2] \times \mathbb{R}[\theta_3] \times \mathbb{R}[\theta_4]$ corresponds to a scoring function in which x gives the weight of a helix in the external loop, y gives the weight of a hairpin loop, and z gives the weight of a bulge/internal loop.

The fan $\mathcal{N}(P_n)$ consists of cones generated by elements in the power set

$$\mathcal{P}\big(\{(1,0,0), (0,0,1), (-1,2,1), (1,1-n,2-n)\}\big).$$

Thus, a parameter vector $v \in \mathbb{R}[\theta_2] \times \mathbb{R}[\theta_3] \times \mathbb{R}[\theta_4]$ has the form $c_1 y_1 + c_2 y_2 + c_3 y_3$ with $c_1, c_2, c_3 \geq 0$ and $y_1, y_2, y_3 \in \{(1,0,0), (0,0,1), (-1,2,1), (1,1-n,2-n)\}$. A generic vector in $\mathbb{R}^3$ lies in the interior of one of the 3-dimensional cones in $\mathcal{N}(P_n)$, and hence we give a brief interpretation of the parameter vectors with $c_i \neq 0$ for $i = 1, 2, 3$.

Scoring vectors in the interior of the cone $\langle (0,0,1), (1,0,0), (1,1-n,2-n) \rangle$ penalize for hairpin loops and can independently penalize or reward for helices in the exterior loop and bulges/internal loops. If $v \in \text{int}\langle (0,0,1), (1,0,0), (-1,2,1) \rangle$ then v gives a penalty for both hairpin loops and interior loops/bulges. Helices in the exterior loop can be beneficial or harmful with this scoring vector, and v can equally penalize helices in the exterior loop, hairpin loops, and internal loops/bulges. Scoring vectors in the interior of one of the two remaining cones can reward or penalize all three quantities. These are not independent, however. For instance, if $v \in \text{int}\langle (1,1-n,2-n), (0,0,1), (-1,2,1) \rangle$ and hairpin loops are disadvantageous under v's scoring scheme then helices in the exterior loop are beneficial. If $w \in \text{int}\langle (1,0,0), (1,1-n,2-n), (-1,2,1) \rangle$ and w rewards hairpin loops, then w rewards bulges/internal loops. Similarly, if w penalizes for bulges/internal loops then w penalizes for hairpin loops. Also, scoring vectors in the interior of the cone $\langle (1,1-n,2-n), (0,0,1), (-1,2,1) \rangle$ can equally reward hairpin loops, internal loops/bulges, and helices in the exterior loop.

4.4 Variation in the Parameter Space

In this section, we add additional information to the parameters $\{\theta_2, \theta_3, \theta_4\}$ in order to study the effect of varying the multibranch loop parameters in the thermodynamic model of RNA folding. We obtain free energy parameters for plane trees using one

Table 2 Energy parameters for plane trees

Sequence	Turner 3.0 values			Turner 2.3 values		
	a_3	a_0	a_1	a_3	a_0	a_1
$X=A, Y=G, Z=C$	−1.9	4.1	2.3	−1.9	3.5	3.0
$X=A, Y=C, Z=G$	−1.6	4.5	2.3	−1.6	3.8	3.0
$X=C, Y=A, Z=U$	−0.4	5.0	3.7	−0.4	4.3	4.0
$X=C, Y=U, Z=A$	−0.6	4.9	3.7	−0.6	4.2	4.0

of the four combinatorial sequences having the form

$$X^4\left(Y^6X^4Z^6X^4\right)^k \quad \text{where } k \geq 1 \text{ and } \begin{cases} X = A \text{ and } \{Y, Z\} = \{C, G\} \\ X = C \text{ and } \{Y, Z\} = \{A, U\}. \end{cases}$$

In these sequences, the segments of the form Y^6 pair with the Z^6 segments while the X nucleotides remain unpaired, and moreover all the loops of a given type have the same free energy. We do not include the possibilities $X = U$ and $\{Y, Z\} = \{C, G\}$ or $X = G$ and $\{Y, Z\} = \{A, U\}$ because we want to prevent the $G - U$ pairing. For a given sequence, we use both the current (version 3.0) (Mathews et al. 2004) and previous (version 2.3) (Walter and Turner 1994) thermodynamic parameters, determined by the Turner lab. Parameters in the thermodynamic model are calculated by measuring the change in free energy coming from the formation of a given motif in the structure at a fixed temperature (typically 37°C). See SantaLucia and Turner (1997) for a review of common methods used including optical melting and Burkardm et al. (2001), Chen et al. (2009) as examples of the experimental methods. The parameters a_3, a_0, and a_1 are based on this type of measurement for the combinatorial sequences and are listed in Table 2. The parameters a_2 and c_2 come from the multi-branch loop scoring function in (1), where the parameters a, b, and c in this function are not determined experimentally. If L is a multibranch loop with n_1 single-stranded bases and n_2 helices and L appears in a secondary structure for one of our 4 combinatorial sequences, then we have $n_1 = 4n_2$. Additionally, for each helix in L, the single-base stacking energy is a_3. Thus, free energy of L in (1) becomes $E(L) = a + 4bn_2 + cn_2 + a_3n_2$, and the parameters a_2 and c_2 in the free energy function E' in (3) can be written as $a_2 = 4b + c + a_3$ and $c_2 = a$. Table 3 illustrates three types of variation: variation of specific nucleotides in combinatorial sequence, variation of the version of Turner's energy parameters, and variation of a, b, c parameters. The effect of varying the multibranch loop parameters a, b, c is more or less the same for each sequence and energy table: two different count vectors can be minimal depending on the value of $a + 12b + 3c$. Technically, a third vertex of P_n has minimal energy in some cases when $b = c = 0$. However, if the offset and helix penalties are both zero, the multibranch energy function will have no penalties for the number of single-stranded bases and the number of stems in a loop. This does not agree with the free energy model.

Varying the sequence alone, we obtain differences in the cut-off values for $a + 12b + 3c$. On the whole, however, nucleotide variation in the combinatorial sequence does not give qualitative differences in the minimal energy plane trees.

Table 3 Restrictions on a, b, c parameters from full-dimensional cones in $\mathcal{N}(P_n)$

Vertex	Rays in $\mathcal{N}(P_n)$	Energy version	Restrictions on a, b, c	Sequence: $[X, Y, Z] =$
$(1, 1, n-1)$	$(1, 1-n, 2-n)$ $(1, 0, 0)$ $(-1, 2, 1)$	3.0 (2.3)	$a + 12b + 3c \geq$	N/A (N/A) $[A, G, C]$ 4.9 (N/A) $[A, C, G]$ 3.6 (N/A) $[C, A, U]$ 4.3 (N/A) $[C, U, A]$
$(1, \frac{n+1}{2}, 0)$	$(0, 0, 1)$ $(1, 0, 0)$ $(-1, 2, 1)$	3.0 (2.3)	$a + 12b + 3c \leq$	6.0 (5.4) $[A, G, C]$ 4.9 (5.4) $[A, C, G]$ 3.6 (4.7) $[C, A, U]$ 4.3 (4.8) $[C, U, A]$
$(n, n, 0)$	$(1, 1-n, 2-n)$ $(0, 0, 1)$ $(-1, 2, 1)$	3.0 (2.3)	$a + 12b + 3c \geq$	6.0 (5.4) $[A, G, C]$ N/A (5.4) $[A, C, G]$ N/A (4.7) $[C, A, U]$ N/A (4.8) $[C, U, A]$
$(1, n-1, 0)$	$(0, 0, 1)$ $(1, 0, 0)$ $(1, 1-n, 2-n)$	3.0 (2.3)	$b = c = 0, a =$	6.0 (5.4) $[A, G, C]$ N/A (5.4) $[A, C, G]$ N/A (4.7) $[C, A, U]$ N/A (4.8) $[C, U, A]$

We do see (in 3 of the 4 sequences) qualitative differences in the minimal energy trees when we compare version 3.0 parameters to version 2.3 parameters. For instance, when $a + 12b + 3c$ is large, 3 of the 4 sequences give the "straight line" tree with count vector $(1, 1, n-1)$ minimal with version 3.0 parameters. Using version 2.3, all four sequences result in the maximal degree of branching, with count vector $(n, n, 0)$ having minimal energy. This difference in minimal energy trees is not too surprising because the change from version 2.3 to version 3.0 was based on more accurate experimental measurement. The secondary predicted structures have indeed changed. It is worth noting that if we use the actual penalties for offset, free base, and helix from versions 2.3 and 3.0 of the Turner energies, we obtain

$$a + 12b + 3c = \begin{cases} 4.6 & \text{v3.0} \\ 9.7 & \text{v2.3.} \end{cases}$$

Thus, all four combinatorial sequences yield $(n, n, 0)$ with minimal energy for version 2.3. Moreover, as 9.7 is a fair amount greater than the cut off for all four sequences, slight variation in these parameters will not change the predicted structure. For version 3.0, the two combinatorial sequences with unpaired poly-A segments have $(1, \frac{n+1}{2}, 0)$ being minimal while $(1, 1, n-1)$ is minimal for the other two sequences. Also, 4.6 is much closer to the cut off values for the sequences. Small changes in these parameters could change which trees have minimal energy.

Table 4 Trees in RNA STRAND collection by size

Category	Range of n	# of trees	Average length	Median length	Average n	Median n
Small	5–12	50	244	220	9	9
Medium	13–40	40	676	512	22	19
Large	41–136	20	2104	1831	82	76

4.5 RNA STRAND Database Analysis

4.5.1 Overall Shape of Data

Our initial collection of secondary structures contains 145 structures from 137 distinct RNA sequences, as described in Materials and Methods. The sequences range from 19 to 4216 nucleotides. We exclude structures for which the number of helices is less than 5 from further analysis. This reason for this is that not all the vertices of P_n listed in Proposition 4.1 are valid and distinct when $n \leq 4$. We have 110 structures with $n \geq 5$ (from 103 sequences) having average (median) length of 739 (367) and $n = 27$ (13). We break these into classes, based on the number of helices, as depicted in Table 4.

While our collection contains more small and medium trees as compared to large trees, this reflects the frequency in the RNA STRAND database. For instance, according to an analysis done by RNA STRAND, the average (median) number of helices over the entire database is 28 (8). This count does, however, include the sequences with fewer than 5 helices and includes a less restrictive definition of bulges/internal loops and helices: internal loops/bulges can have any number of unpaired bases and helices can have any number of base pairs. Our large trees come from 16S ribosomal RNA and 23S ribosomal RNA sequences and have a minimum sequence length of 954 nucleotides. In the RNA STRAND database, only 20% of the 4,666 structures contain at least 954 nucleotides.

4.5.2 Location of Count Vectors on Polytope

It is of great importance to know when biologically correct secondary structures can be predicted by the free energy model. With our simplified energy function E' in (3), we ask if the biologically correct structures can be minimal for some choice of parameters. As mentioned in Sect. 4.2, this translates into determining when the corresponding count vectors lie on the boundary of P_n.

Seventy-one out of 110 count vectors lie on the boundary of P_n: 49 lying on the interior of a facet, 18 lying on the interior of an edge, and 4 occurring as vertices. For a generic choice of parameter values, the count vectors lying on a vertex will be predicted minimizers. Thus, our simple combinatorial model is not sufficient to capture the complexity of RNA folding, since the structures on the boundary are distributed in faces of different dimensions. Moreover, the results below illustrate the connection between the complexity of the folding model and the size of structures which it can handle.

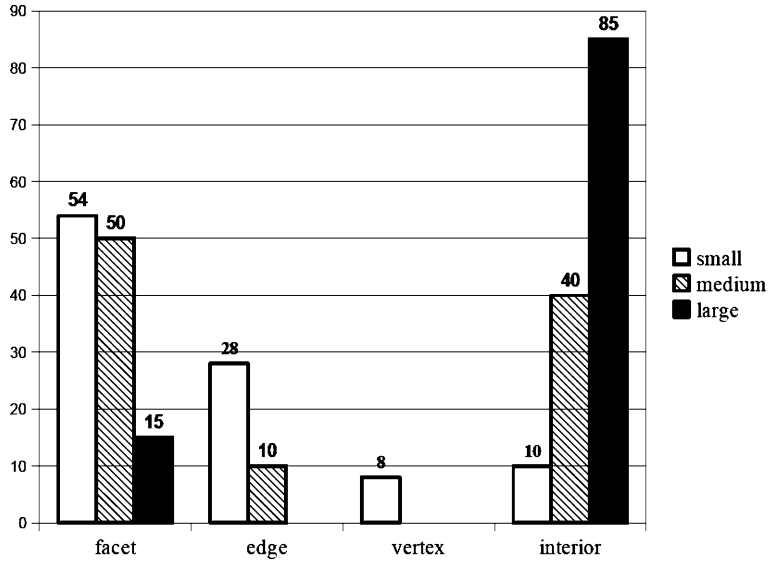

Fig. 4 Location of count vectors on P_n for small, medium, and large trees (in percentage)

The average number of edges for plane trees on the boundary of P_n is 17 and is 45 for plane trees in the interior of P_n. Of those contained in the interior of a facet, 28 are minimal for parameters in $\langle(-1, 2, 1)\rangle$, 7 are minimal for parameters in $\langle(0, 0, 1)\rangle$, and 14 are minimal for parameter values in $\langle(1, 0, 0)\rangle$. Of those contained in the interior of an edge, 8 are minimal for parameters in $\langle(-1, 2, 1), (1, 0, 0)\rangle$, 9 are minimal for parameters in $\langle(-1, 2, 1, (1, 1 - n, 2 - n)\rangle$, and 1 is minimal for parameters in $\langle(1, 0, 0), (0, 0, 1)\rangle$. The 4 count vectors that are vertices of P_n satisfy $n = 5$ or 6 and consist of the set $\{(5, 5, 0), (6, 6, 0), (1, 4, 0), (1, 1, 4)\}$. Figure 4 shows the location of the count vectors for small, medium, and large trees, given in terms of the percentage of trees in each category.

4.5.3 Closest Vertex to Count Vectors

In order to determine which of the 4 vertices of P_n is closest to a given count vector, we map the tetrahedron

$$\text{conv}\left\{(1, n, n, 0), (1, 1, 1, n - 1), (1, 1, n - 1, 0), \left(1, 1, \frac{n+1}{2}, 0\right)\right\}$$

onto the standard tetrahedron with vertices $\{(1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0), (0, 0, 0, 1)\}$. This is accomplished with the following matrix:

$$\begin{bmatrix} -\frac{1}{n-1} & \frac{1}{n-1} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{n-1} \\ -\frac{n}{n-3} & -\frac{1}{n-3} & \frac{2}{n-3} & \frac{1}{n-3} \\ \frac{2n(n-2)}{(n-1)(n-3)} & \frac{2}{(n-1)(n-3)} & -\frac{2}{n-3} & -\frac{2(n-2)}{(n-1)(n-3)} \end{bmatrix} \tag{7}$$

which has determinant $\frac{2}{(n-1)^2(n-3)}$. For a given n, any count vector (r, d_0, d_1) can be written as a sum

$$a_1(n, n, 0) + a_2(1, 1, n-1) + a_3(1, n-1, 0) + a_4\left(1, \frac{n+1}{2}, 0\right)$$

with $0 \le a_i \le 1$ and $a_1 + a_2 + a_3 + a_4 = 1$. After applying the linear transformation (7), the lattice point $(1, r, d_0, d_1)$ will have coordinates (a_1, a_2, a_3, a_4). The coordinate a_i gives a measure of the "closeness" to vertex i. For a given RNA structure, the largest of the a_i gives the vertex closest to its count vector. Moreover, if $t = \max\{a_1, a_2, a_3, a_4\}$ then $0.25 \le t \le 1$.

Fifty-two of the 110 structures are closest to $(1, \frac{n+1}{2}, 0)$, 38 are closest to $(1, 1, n-1)$, 10 are closest to $(n, n, 0)$, and 6 are closest to $(1, n-1, 0)$. Additionally, we have 2 that are closest to both $(1, 1, n-1)$ and $(n, n, 0)$ and 2 that are closest to both $(1, 1, n-1)$ and $(1, \frac{n+1}{2}, 0)$. The average values of (a_1, a_2, a_3, a_4) over the 110 structures are $(0.181, 0.357, 0.138, 0.332)$ which shows that as a whole, the count vectors are closest to $(1, 1, n-1)$ and $(1, \frac{n+1}{2}, 0)$.

We say a count vector is 'close' to vertex i if $a_i > 0.625$. The value 0.625 is halfway in between the smallest and largest possible values of a_i. With this definition, 22% of the small trees are close to vertices, 5% of the medium trees are close to vertices, and no large trees are close to vertices. Thirteen trees in total are close to vertices, of which 8 are close to $(1, 1, n-1)$, 2 are close to $(1, \frac{n+1}{2}, 0)$, 2 are close to $(n, n, 0)$, and 1 is close to $(1, n-1, 0)$. All thirteen of these lattice points lie on the boundary of P_n and hence correspond to minimal energy trees for some choice of parameter values.

5 Discussion and Conclusions

We have used a simple scoring scheme for scoring RNA folds: energy is assigned to a secondary structure based solely on the total number of helices, the number of helices in the exterior loop, and the numbers of hairpin loops and bulges/internal loops. Fixing the total number of helices, the extremal folds are those with the maximal and minimal degrees of branching. When a generic parameter vector is chosen, precisely one of those will have minimal energy. For more specific choices of parameters (biologically realistic or not), the number of minimal count vectors is on the order of the square of the total number of helices. While this seems large, the total number of count vectors that cannot be minimal for any choice of parameters is on the order

of the cube of the total number of helices. Thus, when this total is large, we would not expect such a scoring scheme to accurately predict the correct structures. This is supported by our RNA STRAND analysis in which 85% of the count vectors from known structures with a high number of helices cannot be minimal for any choice of parameters. None of these structures are "close" to the extremal folds. This is not unexpected, however, since even the highly detailed free energy model is not accurate for large RNA molecules (Doshi et al. 2004).

On the other hand, when the total number of helices is small, only 10% of the known structures cannot be minimal for our scoring scheme. While the scoring function used in this work is too simplistic to implement in a prediction software, our results suggest that for small RNA molecules, the full free energy model is not necessary for accurate predictions. We are not the first to make this observation, for Dowell and Eddy (2004) analyzed some simple probabilistic RNA folding models—one with as few 21 free parameters—whose accuracies are comparable to mfold's. In their study, the sequences used for testing came from ribonuclease P RNA, transfer mRNA, and signal recognition particle RNA sequences, all of which yield small to medium trees by our classification. While 21 parameters is far too many for parametric analysis using polyhedral geometry, perhaps a simple model incorporating some thermodynamics and some probabilistic parameters can accurately predict the folding of small RNA molecules.

We compared the variation of multibranch loop parameters to two other types of variation in the parameter space. Fixing the combinatorial sequence and energy version, two possible count vectors can be minimal by varying the multibranch loops parameters. If we use the most recent (accurate) energy version, we find that for 3 of the 4 sequences, these two count vectors include $(1, 1, n-1)$ and $(1, \frac{n+1}{2}, 0)$. Interestingly, these two vertices are closest to the known structures in our RNA STRAND collection. Moreover, regardless of the choices of multi-branch loop parameters in the current version of the thermodynamic model, predicted structures have a low degree of branching—both in the exterior loop and in the multibranch loops. Out of the three possible variations, the most significant changes come from varying the energy version, as the possible predicted structure for version 2.3 have a high degree of branching. Even though the penalties for off-set, free base and helix in the multibranch loop energy calculation are chosen without specific measurement, they do not appear to have a dramatic effect on the predicted structures. One would hope that the parameters determined experimentally are what truly govern the predicted structures, and our findings support this possibility.

6 Future Work

The parametric analysis of RNA branching configurations given here addressed variation in the parameter space for scoring RNA secondary structures. In particular, we focused on three parameters from the multibranch loop energy function which are not based on experimental measurement. We finish by discussing several ways in which this analysis can be extended beyond the scoring scheme for RNA branching investigated here.

An immediate extension of this work is to investigate the impact of varying these multibranch loop parameters on the folding of specific RNA sequences. We expect that this could be achieved either through polytope propagation or an incremental approach. For a given RNA sequence, it is possible to find the optimal, that is the "minimum free energy," secondary structure under the nearest neighbor thermodynamic model using dynamic programming. The polytope propagation method (Pachter and Sturmfels 2004a) would compute the dynamic programming optimization over a semiring of polytopes rather than the numerical parameters. The output is an "folding polytope" for the particular RNA sequence input. As with our results here, the vertices of the polytope correspond to the folds which are optimal under some choice of the multi-branch loop scoring parameters. In the incremental approach (Dewey et al. 2006) the folding polytope is computed by running the dynamic programming optimization for different scoring parameters. Different optimal folds yield different vertices of the polytope. In this way, the folding polytopes for different RNA sequences could be computed and then compared.

Further extensions of this work include investigating the folding polytopes' face complexity as a function of characteristics of the input sequences such as length or G-C content. Our initial results indicate a dependence on the length of the sequence but are less clear in terms of its base composition. Another interesting direction would be exploring different approximations to the complicated biophysics of branching loop thermodynamics. For instance, the Jacobson–Stockmeyer theory derives an energy function with a logarithmic dependence on the number of unpaired bases in the branching loop. Additionally, there are now prediction algorithms which take into account the coaxial stacking of adjacent helices in the multi-branch loops.

Finally, there are a range of other questions to be considered in the parametric analysis of RNA folding. One possibility would be to use parametric methods to perform sensitivity analyses of RNA folding. The decomposition of the parameter space, such as the one analyzed in Gusfield et al. (1992), can be used to determine how sensitive a particular alignment is to changes in the parameters. A related question would take the relatively few known secondary structures and compute the parameters which are optimal for those folds. More generally, there are the questions of face complexity and polyhedral algorithms. While sequence alignment and RNA folding share certain broad characteristics, the details in the RNA folding case are considerably more complicated. Hence, we expect that considerable future work is needed to address the complexity and computability of parametric RNA folding.

7 Materials and Methods

7.1 Selection of Secondary Structures from RNA STRAND Database

The RNA STRAND database (Andronescu et al. 2008) was searched by type of RNA (for example, 16S ribosomal RNA, cis-regulatory element, or group I intron). Each type of RNA was sorted by molecule length, and structures were selected from a variety of organisms to be representative of the different lengths appearing in the database for that type of RNA. Visual inspection of the secondary structures was important in

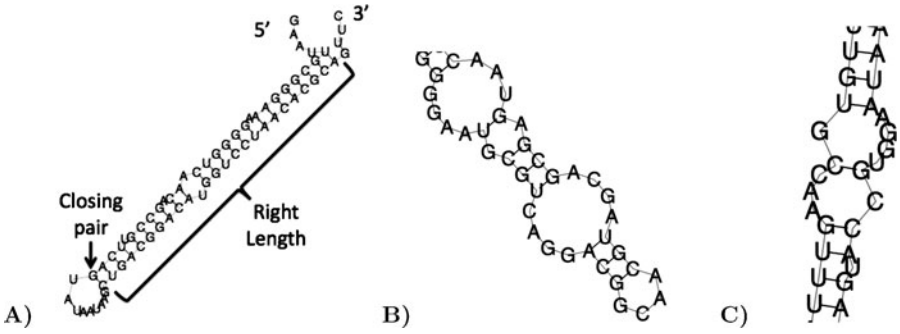

Fig. 5 Helices and internal loops: **A**, **B**, and **C** are fragments from structures in the RNA STRAND database

the selection of the structures for our collection. It allowed for the inclusion of similar length structures with different types of branching. It also prevented our collection from containing nearly identical structures formed by two different RNA molecules of the same type. Finally, visual inspection kept our collection from having a plethora of structures with only one or two helices; these structures are overrepresented in the RNA STRAND database.

7.2 Removal of Pseudoknots from .ct Files

In order to obtain a plane tree from a give secondary structure, pseudoknots were removed. A perl script read the .ct file and stored the closing pairs of all helices, where the helices are defined is Sect. 7.3. Each pair (i, j) and (i', j') of closing pairs was tested to see if $i < i' < j < j'$. If true, the pairs (i, j) and (i', j') were printed to a file. Next, for each pair (i, j) and (i', j') in the output file, one of the associated helices was removed according to the following rubric. If some closing pair (i, j) appears multiple times, its helix was removed under the assumption that it formed a pseudoknot. If both (i, j) and (i', j') were not listed with any other closing pairs, the shorter of the 2 corresponding helices was removed. In the event that the two helices had the same number of paired bases, two versions of the .ct file were saved—one with the first helix removed and one with the second helix removed.

7.3 Calculation of n, r, d_0, d_1 from .ct Files

After all the pseudoknots were removed from the .ct files of secondary structures in our collection, a perl script calculated $n, r, d_0,$ and d_1. In our simplified model of RNA folding, all helices have the same energy independent of the number of base pairs in the helix. Similarly, all bulges/internal loops have the same energy regardless of the number of free bases in the loop. Because of this, very small bulges/internal loops and very short helices were ignored. Bulges and interior loops were required to have at least 3 unpaired bases. No restrictions were placed on the number of free bases in a hairpin loop, which was important so as to maintain the graph structure (edges connecting two vertices).

Each helix with choice of closing pair has a "left length" and "right length" of the helix. The *left length* of a helix is the number of bases in the portion of the sequence that terminates at one of the closing bases. The *right length* of a helix is the number of bases in the portion of the sequence that originates at one of the closing bases. The closing pair of a helix as well as its right length are depicted in Fig. 5A. For this structure, the helix with closing pair G–C has left length 28 and right length 25. For our analyses, a helix was defined to have both the left and right length 3 or greater. Thus, the piece of secondary structure shown in Fig. 5B has two helices—one with left and right length 5 and one with left and right length 3—and one hairpin loop. Similarly, with our definitions, the fragment depicted in Fig. 5C has only 1 interior loop that contains the base pairs G–C and U–A. The single C–G base pair is not considered a helix, and since each of the internal loops containing the C–G pair have more than 3 unpaired bases, the C–G base pair is not considered a part of either helix.

Acknowledgements V.H. and C.E.H. were both supported by the NIH grant 1R01GM083621-01 (P.I. Heitsch). C.E.H. also acknowledges funding from a Career Award at the Scientific Interface (CASI) from the Burroughs Wellcome Fund (BWF). Additionally, V.H. would like to thank Justin Filoseta for the remarkable computer support at the Georgia Institute of Technology.

References

Andronescu, M., Bereg, V., Hoos, H., & Condon, A. (2008). RNA STRAND: The RNA secondary structure and statistical analysis database. *BMC Bioinformatics*, *9*(1), 340.

Bakhtin, Y., & Heitsch, C. E. (2008). Large deviations for random trees. *J. Stat. Phys.*, *132*(3), 551–560.

Bakhtin, Y., & Heitsch, C. E. (2009). Large deviations for random trees and the branching of RNA secondary structures. *Bull. Math. Biol.*, *71*(1), 84–106.

Beerenwinkel, N., Dewey, C. N., & Woods, K. M. (2005). Parametric inference of recombination in HIV genomes. Preprint available at arXiv:q-bio/0512019v1.

Burkardm, M. E., Xia, T., & Turner, D. H. (2001). Thermodynamics of RNA internal loops with a Guanosine-Guanosine pair adjacent to another noncanonical pair. *Biochemistry*, *40*(8), 2478–2483.

Chen, G., Kennedy, S. D., & Turner, D. H. (2009). A CA^{+} pair adjacent to a sheared GA or AA pair stabilizes size-symmetric RNA internal loops. *Biochemistry*, *48*(24), 5738–5752.

Dershowitz, N., & Zaks, S. (1980). Enumerations of ordered trees. *Discrete Math.*, *31*(1), 9–28.

Deutsch, E. (2004). Ordered trees with prescribed root degrees, node degrees, and branch lengths. *Discrete Math.*, *282*(1–3), 89–94.

Dewey, C. N., & Woods, K. (2005). Parametric sequence alignment. In B. Sturmfels & L. Pachter (Eds.), *Algebraic statistics for computational biology* (pp. 193–205). New York: Cambridge University Press.

Dewey, C. N., Huggins, P. M., Woods, K., Sturmfels, B., & Pachter, L. (2006). Parametric alignment of drosophila genomes. *PLoS Comput. Biol.*, *2*(6), 606–614.

Diamond, J. M., Turner, D. H., & Mathews, D. H. (2001). Thermodynamics of three-way multibranch loops in RNA. *Biochemistry*, *40*, 6971–6981.

Doshi, K. J., Cannone, J. J., Cobaugh, C. W., & Gutell, R. R. (2004). Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*, *5*(1), 105.

Dowell, R., & Eddy, S. (2004). Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*, *5*(1), 71.

Gan, H. H., Pasquali, S., & Schlick, T. (2003). Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *Nucleic Acids Res.*, *31*(11), 2926–2943.

Grünbaum, B. (2003). In *Graduate texts in mathematics: Vol. 221*. *Convex polytopes* (2nd ed.). New York: Springer. Prepared and with a preface by Volker Kaibel, Victor Klee and Günter M. Ziegler.

Gusfield, D., Balasubrama, K., & Naor, D. (1992). Parametric optimization of sequence alignment. *Algorithmica*, *12*(4–5), 312–326.

Heitsch, C. E. (2010). Combinatorial insights into RNA secondary structures. In preparation.

Iseri, H. (2008). An exploration of Pick's theorem in space. *Math. Mag.*, *81*(2), 106–115.

Le, S.-Y., Nussinov, R., & Maizel, J. V. (1989). Tree graphs of RNA secondary structures and their comparisons. *Comput. Biomed. Res.*, *22*(5), 461–473.

Lenhof, H.-P., Reinert, K., & Vingron, M. (1998). A polyhedral approach to RNA sequence structure alignment. *J. Comput. Biol.*, *5*, 517–530.

Mathews, D. H., & Turner, D. H. (2002). Experimentally derived nearest-neighbor parameters for the stability of RNA three- and four-way multibranch loops. *Biochemistry*, *41*, 869–990.

Mathews, D. H., & Turner, D. H. (2006). Prediction of RNA secondary structure by free energy minimization. *Curr. Opin. Struct. Biol.*, *16*(3), 270–278.

Mathews, D. H., Sabina, J., Zuker, M., & Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, *288*(5), 911–940.

Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., & Turner, D. H. (2004). Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. USA*, *101*(19), 7287–7292.

Pachter, L., & Sturmfels, B. (2004a). Parametric inference for biological sequence analysis. *Proc. Natl. Acad. Soc.*, *101*(46), 16138–16143.

Pachter, L., & Sturmfels, B. (2004b). Tropical geometry of statistical models. *Proc. Natl. Acad. Soc.*, *101*(46), 16132–16137.

SantaLucia, J., & Turner, D. H. (1997). Measuring the thermodynamics of RNA secondary structure formation. *Biopolymers*, *44*, 309–319.

Schmitt, W. R., & Waterman, M. S. (1994). Linear trees and RNA secondary structure. *Discrete Appl. Math.*, *51*(3), 317–323.

Shapiro, B. A., & Zhang, K. (1990). Comparing multiple RNA secondary structures using tree comparisons. *Comput. Appl. Biosci.*, *6*(4), 309–318.

Smit, S., Rother, K., Heringa, J., & Knight, R. (2008). From knotted to nested RNA structures: a variety of computational methods for pseudoknot removal. *RNA*, *14*(3), 410–416.

Stanley, R. P. (1999). *Cambridge studies in advanced mathematics: Vol. 62*. *Enumerative combinatorics: Vol. 2*. Cambridge: Cambridge University Press.

Walter, A. E., & Turner, D. H. (1994). Sequence dependence of stability for coaxial stacking of RNA helixes with Watson-Crick base paired interfaces. *Biochemistry*, *33*(42), 12715–12719.

Wang, L., & Zhao, J. (2003). Parametric alignment of ordered trees. *Bioinformatics*, *19*(17), 2237–2245.

Waterman, M. S., Eggert, M., & Lander, E. (1992). Parametric sequence comparisons. *Proc. Natl. Acad. Sci. USA*, *89*(12), 6090–6093.

Ziegler, G. M. (1995). *Graduate texts in mathematics: Vol. 152*. *Lectures on polytopes*. New York: Springer.

Zuker, M. (2000). Calculating nucleic acid secondary structure. *Curr. Opin. Struct. Biol.*, *10*(3), 303–310.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, *31*(13), 3406–3415.

Zuker, M., Mathews, D., & Turner, D. (1999). Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In J. Barciszewski & B. Clark (Eds.), *NATO ASI series*. *RNA biochemistry and biotechnology* (pp. 11–43). Amsterdam: Kluwer Academic.