SIAM J. APPL. ALGEBRA GEOMETRY
Vol. 2, No. 3, pp. 444–461
© 2018 Society for Industrial and Applied Mathematics

On the Structure of RNA Branching Polytopes*

Fidel Barrera-Cruz†, Christine Heitsch†, and Svetlana Poznanović‡

Abstract. The prevalent method for RNA secondary structure prediction for a single sequence is free energy minimization based on the nearest neighbor thermodynamic model (NNTM). One of the least well developed parts of the model is the energy function assigned to the multibranch loops. Parametric analysis can be performed to elucidate the dependence of the prediction on the branching parameters used in the NNTM. Since the objective function is linear, this boils down to analyzing the normal fans of the *branching polytopes*. Here we show that because of the way the multibranch loops are scored under the NNTM, certain branching patterns are possible for all sequences. We do this by characterizing the dominant parts of the parameter space obtained by looking at the relevant section of the normal fan; therefore, we conclude that the structures that are normally found in nature are obtained for a relatively small set of parameters.

Key words. RNA secondary structure, polytope, multibranch loops, parametric optimization

AMS subject classifications. 92D20, 52B99

DOI. 10.1137/17M1144076

1. Introduction.

Ribonucleic acid (RNA) is one of the three major biological macromolecules, along with DNA and proteins, that are essential for all known forms of life. Several decades of research in bioinformatics have resulted in a wide range of computational methods for predicting RNA secondary structures. The standard approach for single-sequence secondary structure prediction is free energy minimization [11], which uses a nearest-neighbor thermodynamic model (NNTM) with several hundred, mostly experimentally determined, parameters [18]. However, when these minimum free energy (MFE) predictions are compared to structures derived from information-theoretic means, the current gold standard, the average accuracy for longer ribosomal RNA sequences is only 40% [4]. Hence, it is critical to understand which aspects of RNA base pairing are not captured well by the NNTM.

We focus here on the part of the energy function which governs the branching of an RNA secondary structure, which is known to be a weakness of the current model [21]. For computational reasons, the entropic cost of the branching loops is modeled as an affine function with three parameters. A very natural question to ask is, *How does the optimal secondary structure depend on the branching loop parameters?* Methods from geometric combinatorics, specifically polytopes (which we term *branching polytopes*) and their normal fans, can be used

*Received by the editors August 18, 2017; accepted for publication (in revised form) June 20, 2018; published electronically September 18, 2018.
http://www.siam.org/journals/siaga/2-3/M114407.html
Funding: The second author was partially supported by a BWF CASI. The third author was partially supported by NSF DMS-1312817.
†School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 USA (fidelbc@math.gatech.edu, heitsch@math.gatech.edu).
‡Department of Mathematical Sciences, Clemson University, Clemson, SC 29634 USA (spoznan@clemson.edu).

Downloaded 05/21/24 to 3.239.95.36 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

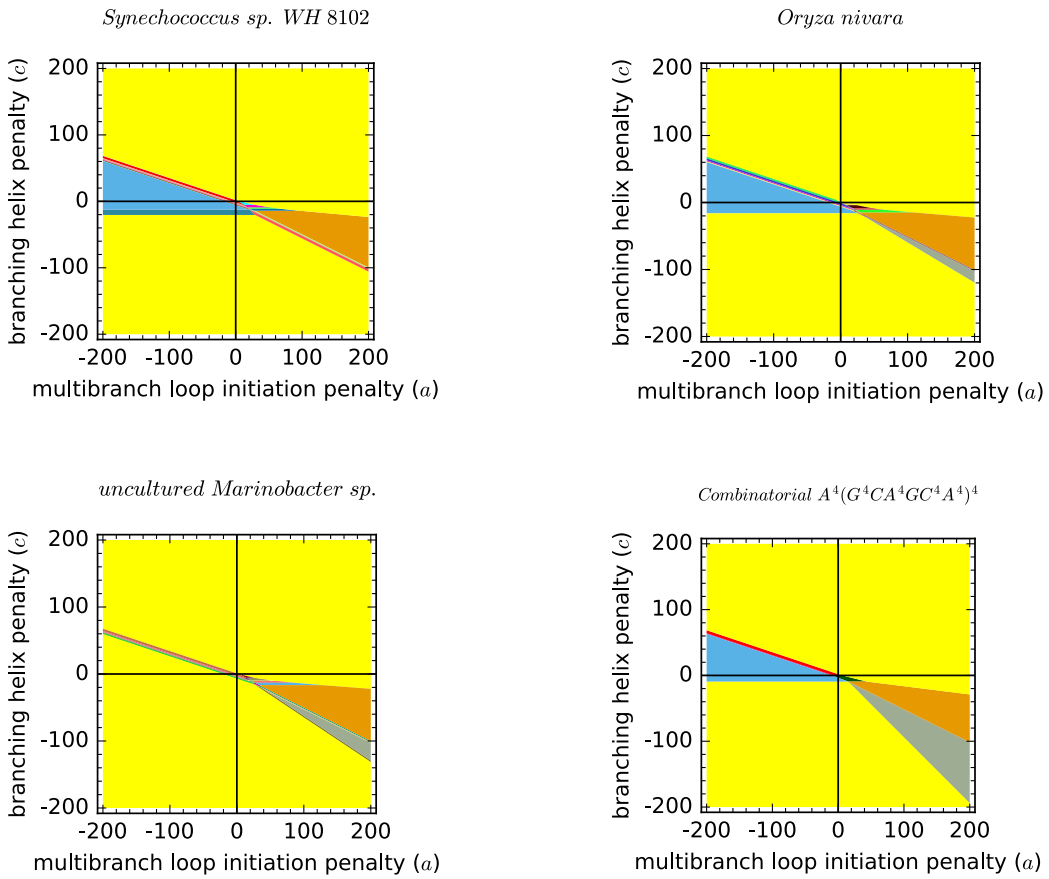


Figure 1. *The $\mathcal{R}_0$ slice of $\mathcal{N}(\mathcal{P})$ for three tRNA sequences and one combinatorial sequence.*

to perform a full parametric analysis of the branching part of the NNTM. The computational framework, and proof-of-principle results, which give the first complete parametric analysis of the branching part of the NNTM for real RNA sequences, were presented in [5].

The branching polytopes depend on the RNA sequence and have hundreds of vertices and facets even for sequences of fewer than 100 nucleotides, which makes it challenging to compare them in a biologically meaningful way. However, comparisons of the normal fans of the branching polytopes that we computed for RNA, random, and combinatorial sequences (see section 2 for details) revealed a lot of similarity. For example, Figure 1 shows slices of the normal fans for three tRNA and one combinatorial sequence, where one can observe a similarity in the position of the dominant, unbounded regions. In this paper we make this statement precise. Namely, we prove that under some mild conditions (Hypotheses 3.6 and 3.12, which are true for all RNA and random sequences we analyzed), the branching polytopes share certain dominant features. We do this by studying the dominant regions of the normal fans of the branching polytopes; more precisely, we study the intersection of the normal fan with certain hyperplanes because that is appropriate for this optimization problem. In particular,

Downloaded 05/21/24 to 3.239.95.36 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 05/21/24 to 3.239.95.36 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

we completely characterize the structures that are optimal for these dominant regions of the parametric space. Since these structures are not biologically realistic, our results imply that the biologically relevant secondary structures are attained for a relatively small portion of the parameter space.

Parametric optimization, where the optimization problem is solved as a function of one or multiple parameters, has arisen as an important problem for other biological models in the past. Most notably, the parametric sequence alignment dates back to [7, 20]. Tools for parametric alignment of ordered trees which could be used for comparison of RNA structures were developed in [19]. Graphical models are a unifying framework for many problems in biological sequence analysis [14], and the polytope propagation algorithm to construct the Newton polytope of an observation from a graphical model is given in [13]. An actual complete parametric alignment of two *Drosophila* genomes was computed in [3], where it was established that polytope propagation is outperformed by the beneath-beyond approach for incremental convex hull construction [6]. That is why the software we use for computing the branching polytopes [5] uses the beneath-beyond approach as well. Our results build nicely on a simplified model of RNA folding defined and analyzed in [8], where some of the same types of extremes were observed. However, this is the first parametric analysis of the NNTM for real RNA sequences, in which the energies of all the motifs are included in the same way as they are computed in the free energy minimization used to predict secondary structures.

The paper is organized as follows. In section 2 we give the preliminaries. We give a definition of a secondary structure and explain the part of the NNTM used to score the branching loops. Then we define the branching polytopes. In section 3 we characterize the dominant cones of the normal fan of the branching polytopes, where we specifically focus on the trade-off between the entropic cost of forming a branching loop and the stability of a helix branching off. We end with conclusions and a discussion in the last section.

2. Preliminaries.

Secondary structure. RNA is a chain of four nucleotides, abbreviated A, C, G, and U (instead of T), which form the familiar Watson–Crick pairings, similar to DNA. Traditionally, RNA has been thought of as an important nucleic acid that plays a role in the transcription of the genetic code stored in DNA and its translation into proteins. However, in the last few decades, it has been discovered that RNA also performs other critical biological functions, including gene splicing, editing, and regulation. Knowing the structure of noncoding RNA molecules is critical to understanding and manipulating their cellular functions, and that is why the prediction of the RNA structure has been an important problem in computational biology in the recent past.

The structure of RNA is understood hierarchically. Unlike DNA, most of RNA is found in nature as a self-complementary single strand which folds onto itself by formation of intra-sequence base pairings. The nucleotide chain itself is thought of as the *primary* structure, the intra-sequence base pairs form the *secondary* structure, while the *tertiary* structure includes more complex interactions, including pseudoknots and base triples. Determining the tertiary structure has been challenging, both experimentally and computationally, and hence a lot of attention has been devoted to the prediction of the secondary structure.

The secondary structure consists of runs of stacked base pairs (helices) separated by single-

stranded regions (loops). Mathematically, it is a partial noncrossing matching, which means that when the RNA sequence is written on a circle and the base pairings are drawn as straight lines, there are no crossing lines. Biologically, the noncrossing property means that the pseudoknots are not considered to be part of the secondary structure, and they are not predicted by the NNTM which we analyze in this paper. The *exterior loop* is the loop that contains the two ends of the RNA strand. In this paper we focus on the so called *multibranch loops*, i.e., the loops that have at least three helices meeting them. The exterior loop is not considered to be a multibranch loop, regardless of how many helices are incident with it. Nonexterior loops which are incident with a single helix are called *hairpin loops*. A nonexterior loop incident with two helices is said to be *interior* if it contains two strands of unpaired nucleotides, or a *bulge* if it contains only one such strand.

There are three types of base pairs allowed in a secondary structure: the Watson–Crick pairs A-U C-G and the wobble pair G-U. Even with these restrictions, there are multiple secondary structures for a given sequence (see Figure 2 for an example of two possible structures for a tRNA from *Synechococcus sp. WH* 8102[1] generated using [22]); in fact the number of secondary structures grows exponentially with sequence length [16].

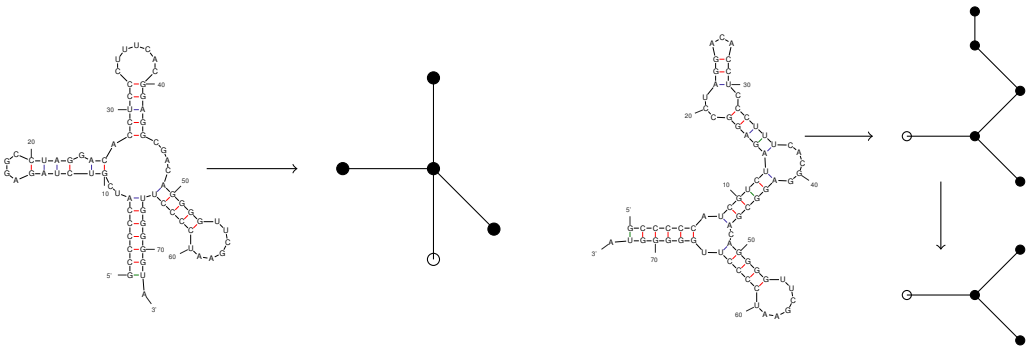


Figure 2. *Two of many possible structures for the same tRNA sequence with the corresponding tree representations.*

Representation of secondary structure. Since in this paper we are focusing on the multibranch loops, it is convenient to think of the reduced rooted plane tree representation of a secondary structure, in which the nonbranching interior loops have been smoothed away. In such a representation, the root is the exterior loop, the internal nodes other than the root are the branching loops, and the leaves represent the hairpin loops. Figure 2 shows the tree representations of two structures. The white nodes are the roots. For the second structure, an intermediary tree is shown in which each loop is represented by a node. Since we are interested only in multibranch loops, it is sufficient to consider the smaller tree, in which some of the edges of the intermediary tree have been contracted and the nodes correspond to the exterior loop, multibranch loops, or hairpin loops.

Another common way to represent secondary structures is via the so-called dot-bracket notation. The base pairs correspond to matching sets of parentheses, while the single-stranded nucleotides are represented by dots on a line. An example is given in the next subsection.

[1] GCCCCCAUCGUCUAGAGGCCUAGGACACCUCCCUUUCACGGAGGCGACAGGGGUUCGAAUCCCCUUGGGGGUA.

Downloaded 05/21/24 to 3.239.95.36 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 05/21/24 to 3.239.95.36 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Signatures. The energy function used for MFE prediction has evolved substantially over the years [10, 12, 23], with a significant increase in the number of parameters. It is a sum of the energy of the helices and the loops. Some of the contributions of the various kinds of substructures have been determined experimentally, others have been extrapolated from experimental data, and some have been learnt computationally. In particular, in the Turner99 version [18] of the NNTM, the initiation term assigned to a multibranch loop is

$$\Delta G_{\text{initiation}} = a + b \cdot (\#\text{unpaired nucleotides}) + c \cdot (\#\text{branching helices}),$$

where the values

$$a = 3.4, \quad b = 0, \quad c = 0.4 \tag{1}$$

have been computationally determined in [10]. The free energy of a secondary structure S is calculated as

$$\Delta G_S = a \cdot x + b \cdot y + c \cdot z + w, \tag{2}$$

where x is the number of multibranch loops in S, y is the total number of single-stranded nucleotides in the multibranch loops in S, z is the total number of helices incident with the multibranch loops in S (a helix is counted twice if it is incident with two multibranch loops), and w is the total sum of the energy contributions of the helices and the other kinds of loops in S. The value of w depends on the type of stacked base pairs forming the helices, initiation penalties, the type of bulges and internal and hairpin loops present, and unpaired nucleotides dangling on neighboring base pairs [18]. Condensing the energy contribution of these motifs into one term allows us to focus on the analysis of the effect of the multibranch loop parameters. In light of formula (2), for a given secondary structure, we define its *signature* to be (x, y, z, w), where x, y, z, w are as described above.

The signature is a map from the set of all secondary structures to $\mathbb{R}^4$, but even when the underlying RNA sequence is fixed, this map is not one-to-one. The following two structures (given in the dot-bracket notation introduced in the previous subsection) for a tRNA of *Oryza nivara*[2] have the same signature $(5, 18, 17, 19)$:

$$(\cdot\,\cdot(\cdot\,\cdot\,\cdot)\cdot((\cdot(\cdot\,\cdot\,\cdot\,\cdot)\cdot(\cdot(\cdot(\cdot\,\cdot\,\cdot)\cdot(\cdot\,\cdot\,\cdot\,\cdot)(\cdot\,\cdot\,\cdot))\cdot(\cdot\,\cdot\,\cdot\,\cdot)\cdot)\cdot(\cdot\,\cdot\,\cdot))\cdot(\cdot\,\cdot\,\cdot\,\cdot)\cdot\,\cdot\,\cdot\,\cdot)\cdot\,\cdot)\cdot$$
$$(\cdot\,\cdot(\cdot\,\cdot\,\cdot)\cdot((\cdot(\cdot\,\cdot\,\cdot\,\cdot)\cdot\,\cdot\,\cdot(\cdot(\cdot\,\cdot\,\cdot)\cdot(\cdot\,\cdot\,\cdot\,\cdot)(\cdot\,\cdot\,\cdot))\cdot)\cdot((\cdot\,\cdot\,\cdot\,\cdot)(\cdot\,\cdot\,\cdot)\cdot\,\cdot(\cdot\,\cdot\,\cdot\,\cdot)\cdot\,\cdot)\cdot)\cdot\,\cdot)\cdot$$

Branching polytope. For a fixed sequence s over the alphabet $\{\text{A}, \text{C}, \text{G}, \text{U}\}$, its *branching polytope* $\mathcal{P}$ is the convex hull of the set of signatures $\mathcal{S}$ that correspond to secondary structures over s. We are interested in its vertices because they minimize the linear functional

$$f_{a,b,c,d}(x, y, z, w) = a \cdot x + b \cdot y + c \cdot z + d \cdot w.$$

In particular, the signature of the MFE structure for s is a vertex of $\mathcal{P}$ as the optimum for $f_{a,b,c,1}$ over $\mathcal{S}$ for the values of a, b, c given in (1). Computing $\operatorname{argmin}_{\mathcal{S}} f_{a,b,c,1}$ corresponds to

[2] GGGGAUAUAGCUCAGUUGGUAGAGCUCCGCUCUUGCAAGGCGGAUGUCAGCGGUUCGAGUCCGCUUAUCUCCA.

Downloaded 05/21/24 to 3.239.95.36 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

calculating the signature of the MFE structure under an NNTM with modified multibranch parameters. While, as we said, the signature does not determine the structure completely, it contains information about its branching.

The normal fan of a polytope is the collection of its normal cones. For a face F of the polytope, its corresponding normal cone is the set of all vectors $\mathbf{B}$ such that any point $\mathbf{y} \in F$ minimizes the product $\mathbf{Bx}$. The full dimensional cones of the normal fan $\mathcal{N}(\mathcal{P})$ of $\mathcal{P}$ correspond to the vertices of $\mathcal{P}$. For a vertex (x, y, z, w), we will denote by $\text{cone}(x, y, z, w)$ the cone of parameters (a, b, c, d) in $\mathcal{N}(\mathcal{P})$ such that $(x, y, z, w) = \text{argmin}_{\mathcal{S}} f_{a,b,c,d}$. In particular, since in the NNTM we have $d = 1$, we are interested in the $d = 1$ slice of $\mathcal{N}(\mathcal{P})$, which is a polyhedral subdivision of $\mathbb{R}^3$, and the vertices (x, y, z, w) for which $\text{cone}(x, y, z, w) \cap \{(a, b, c, d) : d = 1\} \neq \emptyset$.

In order to understand the trade-off between the cost a of closing a multibranch loop and the cost c of starting a branching helix, we will consider the regions in

$$\mathcal{R}_{b_0} := \mathcal{N}(\mathcal{P}) \cap \{(a, b, c, d) \colon d = 1, b = b_0\}.$$

Figure 1 illustrates $\mathcal{R}_0$ for several sequences: tRNA from *Synechococcus sp. WH* 8102,[3] *Oryza nivara*,[4] *uncultured Marinobacter sp.*,[5] and a combinatorial sequence, generated with `CoCalc` [15]. The bounded regions, which are roughly around the origin, are not visible, and instead one can notice dominant unbounded regions. While the precise boundaries of the unbounded regions differ between the figures, all of them have two regions, colored yellow, which dominate the first and third quadrant. These regions are separated by a sequence of *unbounded stripes* and possibly an *unbounded wedge*, colored blue, in the second quadrant, and a fan of unbounded wedges (colored orange and grey) and unbounded stripes in the fourth quadrant. In the next section we characterize the vertices of $\mathcal{P}$ which correspond to these unbounded regions.

Sequences analyzed. We have computed the branching polytopes for many more biological and randomized sequences [1]. The real RNA sequences were obtained from the Comparative RNA Web Site [2]. Our sample of tRNA contains 50 sequences ranging from 72 to 79 nucleotides in length. The sequences were chosen so that their accuracy (as given by `gtfold` [17]) and GC content were distributed as uniformly as possible over the interval $[0, 1]$. Two other sets of real sequences were used: a set of 99 5S sequences (length 117–133) and a set of five RNase P sequences (length 208–360). The computing time (60 days for the longest RNase P in the set) currently prevents us from compute polytopes of the order of 1000 nucleotides. We have also looked at the branching polytopes for two sets of randomized sequences. The first set of 45 random sequences (five of each length from 72 to 80) was generated to have uniform individual nucleotide frequencies. Finally, for the second random set, using `uShuffle` [9] we produced five sequences with the same dinucleotide frequencies as the tRNA in the sample.

3. Characterization of the unbounded regions in $\mathcal{R}_{b_0}$.

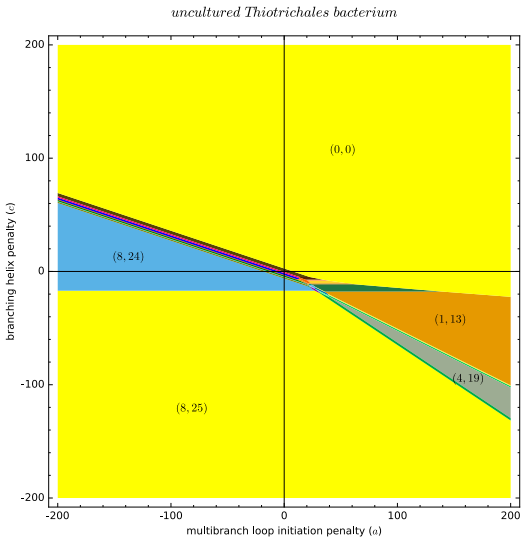
In this section we characterize the vertices of $\mathcal{P}$ which correspond to the unbounded regions in $\mathcal{R}_{b_0}$. See Figure 3 for an outline

[3] See footnote 1.

[4] See footnote 2.

[5] GGUCUGUAGCUCAGGUGGUUAGAGCGCACCCCUGAUAAGGGUGAGGUCGGUGGUUCAAGUCCACCCAGACCCACCAG.

of the section. Each of the unbounded regions in the illustrated $\mathcal{R}_0$ slice of an *uncultured Thiotrichales bacterium* tRNA[6] is labeled by the coordinates (x, z) from the vertex (x, y, z, w) which corresponds to that region. While the specific labels differ among different slices $\mathcal{R}_0$ and among different sequences, the reader can see that the location of the dominant regions is similar to that in Figure 1. We start by describing the regions that dominate the first and third quadrants of a slice $\mathcal{R}_0$. We then move onto the second quadrant and finish with the fourth one. The table in Figure 3 summarizes which results in this section describe each of these four quadrants. The unbounded regions in the table are listed consecutively as they appear when one walks in a counterclockwise direction. The labels of the stripes in the figure are omitted for clarity.



Unbounded regions	Type	Related results
(0,0)	Wedge	Corollary 3.3
(1,3) (2,6) (3,9) (4,12) (5,15) (6,18) (7,21)	Stripes extending northwest	Proposition 3.14
(8,24)	Wedge	Proposition 3.10
(8,25)	Wedge	Corollary 3.7
(6,22) (4,19) (3,17) (2,15)	Stripes and wedges extending southeast	Theorem 3.19
(1,13)	Wedge	Proposition 3.16

Figure 3. *An outline of the main results.*

Let s be a fixed sequence of length n over the alphabet $\{\text{A}, \text{C}, \text{G}, \text{U}\}$, let $\mathcal{S}$ be its set of branching signatures, let $\mathcal{P}$ be the associated branching polytope, and let $\mathcal{V}$ be the set of vertices of $\mathcal{P}$. Let $\mathbf{v} = (x, y, z, w) \in \mathcal{S}$. Let $\alpha = (a, b, c, d) \in \mathbb{R}^4$. Then $\mathbf{v}$ is optimal for α if, for all $\mathbf{v}' = (x', y', z', w') \in \mathcal{S}$, $\alpha \cdot (\mathbf{v} - \mathbf{v}') \leq 0$.

Let $x_{max} = \max\{x : (x, y, z, w) \in \mathcal{S}\}$. Similarly, we define x_{min}, y_{min}, y_{max}, z_{min}, and z_{max}.

Proposition 3.1. *There exists $(x, y, z, w) \in \mathcal{V}$ such that $x = x_{max}$.*

Proof. This holds since $x_{max} - x' \geq 0$ for all $\mathbf{v}' = (x', y', z', w') \in \mathcal{S}$. Thus, as $a \to -\infty$ for fixed $b, c, d \in \mathbb{R}$, then $\alpha \cdot (\mathbf{v} - \mathbf{v}') \leq 0$ for $\mathbf{v} = (x_{max}, y, z, w) \in \mathcal{S}$. Hence, $\mathbf{v}$ is a vertex of $\mathcal{P}$ for some choice of y, z, w. ∎

There may be more than one signature in $\mathcal{S}$ with $x = x_{max}$. In this case, optimality is determined by the relationship among the other three parameters. Clearly, comparable results hold for y_{max}, z_{max}, and w_{max}. However, since d is a dummy variable in the optimization, we

[6] CCAUAGCUCAGCUGGGAGAGCACCUGCUUUGCAAGCAGGGGGUCGGCGGUUCGACCCCGCCUGGCUCCACCAG.

Downloaded 05/21/24 to 3.239.95.36 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 05/21/24 to 3.239.95.36 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

will henceforth consider $d = 1$ fixed, which is the only case of interest. Also, dual definitions and results hold for x_{min}, y_{min}, z_{min}, and w_{min}. The minimum value of 0 is achieved for x, y, and z simultaneously in a structure with no branch points, and the empty structure is one such possible structure for any sequence. Let

$$w_0 = \min\{w : (0, 0, 0, w) \in \mathcal{S}\}.$$

Proposition 3.2. *For each $b \in \mathbb{R}$, there exist $a, c \in \mathbb{R}$ such that $(0, 0, 0, w_0)$ is optimal for $(a, b, c, 1)$.*

Proof. Let $\mathbf{v_0} = (0, 0, 0, w_0)$ and $\alpha = (a, b, c, 1) \in \mathbb{R}^4$. By construction, $\alpha \cdot \mathbf{v_0} \leq \alpha \cdot (0, 0, 0, w)$ for every other $(0, 0, 0, w) \in \mathcal{S}$. Hence, we consider $\mathbf{v} = (x, y, z, w) \in \mathcal{S}$ with $x > 0$.

Suppose $b \geq 0$. Let $a \geq 0$, $c \geq 0$ with $a + 3c \geq w_0 - w_{min}$, where $w_{min} = \min\{w : (x, y, z, w) \in \mathcal{S}\}$ as discussed above. Then since the parameters are all nonnegative and $x \geq 1$, $y \geq 0$, $z \geq 3$, $w \geq w_{min}$,

$$\alpha \cdot \mathbf{v_0} = w_0 \leq a + 3c + w_{min} \leq \alpha \cdot \mathbf{v}.$$

For $b < 0$, again let $a \geq 0$, $c \geq 0$ but with $a + 3c \geq w_0 - w_{min} - bn$. Then

$$\alpha \cdot \mathbf{v_0} = w_0 \leq a + bn + 3c + w_{min} \leq \alpha \cdot \mathbf{v}$$

since $y \leq n$ and $by \geq bn$. ∎

Recall that $\mathcal{R}_{b_0}$ denotes the intersection of $\mathcal{N}(\mathcal{P})$ with the hyperplanes $d = 1$ and $b = b_0$. We show that many of the characteristics visible in Figure 1 hold in general. To begin, the regions of $\mathcal{R}_{b_0}$ have two basic types: unbounded and bounded.

Let R be a region in $\mathcal{R}_{b_0}$ corresponding to $(x, y, z, w) \in \mathcal{V}$ and containing the point $(a_0, b_0, c_0, 1)$. We already know of some general bounds on R as a consequence of Proposition 3.1; if $x < x_{max}$, then R is bounded due west (along the ray $(a_0 - t, b_0, c_0, 1)$ for $t \geq 0$), with analogous conclusions for $0 < x$ and due east, for $0 < z$ and due north (along the ray $(a_0, b_0, c_0 + t, 1)$ for $t \geq 0$), and for $z < z_{max}$ and due south.

Those regions which are unbounded in at least one direction divide into two subtypes which we call "wedges" and "stripes." Recall that a convex polyhedron is the convex sum of its vertices plus the conical sum of the direction vectors of its infinite edges. For an unbounded 2-dimensional polyhedron R, we say that R is a "stripe" if its infinite edges have the same direction, and it is a "wedge" if its infinite edges have two different directions.

Let $\text{cone}(x, y, z, w)$ denote the cone in the normal fan of $\mathcal{P}$ associated to the vertex $(x, y, z, w) \in \mathcal{V}$. As a consequence of the proof of Proposition 3.2, we have the following.

Corollary 3.3. *For each $b_0 \in \mathbb{R}$, $\text{cone}(0, 0, 0, w_0) \cap \mathcal{R}_{b_0}$ is an unbounded wedge.*

Although redundant, we retain "unbounded" as an adjective to emphasize this as the primary characteristic, with the specific geometric subtype as a secondary characteristic.

Notice that the constant w_0 doesn't depend on the slice $\mathcal{R}_{b_0}$. Moreover, as can be seen from the choice of parameters in the proof of Proposition 3.2, $\text{cone}(0, 0, 0, w_0)$ dominates the northeast quadrant of the (a, c) plane. In Figure 3, this is the region labeled $(0, 0)$. Moving north ($c \to \infty$) or east ($a \to \infty$) from any point in $\mathcal{R}_{b_0}$ eventually leads to $\text{cone}(0, 0, 0, w_0)$.

Hence, there are no unbounded northeast rays outside of $\text{cone}(0,0,0,w_0)$. We will prove dual statements about wedges in the southwest quadrant and southwest rays based on the hypothesis that the maximums for x and z occur simultaneously.

For a particular $x_0 \in \mathbb{R}$, let $z_{max}(x_0)$ be the maximum number of branches for signatures with the given number of branch points, that is,

$$z_{max}(x_0) = \max\{z : (x_0, y, z, w) \in \mathcal{S}\}.$$

There are obvious analogous definitions exchanging x and z, etc., and dual ones for minimization.

Proposition 3.4. *For each $b_0 \in \mathbb{R}$, there exist $y, w \in \mathbb{R}$ such that $\text{cone}(x_{max}, y, z_{max}(x_{max}), w)$ is an unbounded wedge in $\mathcal{R}_{b_0}$.*

Proof. Let $x = x_{max}$, $z = z_{max}(x_{max})$, and y, w be such that $\mathbf{v} = (x, y, z, w) \in \mathcal{S}$ and $b_0 y + w$ is the least possible for the given b_0. Let $\alpha = (a_0, b_0, c_0, 1) \in \mathcal{R}_{b_0}$ and $\mathbf{v}' = (x', y', z', w') \in \mathcal{S}$. We may assume either that $x' = x$ and $3x \le z' < z$ or that $x > x' \ge 0$. We will use the fact that $w \ge w_{min}$, $n \ge y \ge 0$, and $z \le z_{max}$.

Suppose $b_0 \ge 0$. Then $w_{min} - w - b_0 y \le 0$. Let a_0, c_0 be such that

$$a_0 \le 0,\ c_0 \le w_{min} - w - b_0 y,\ a_0 + c_0(z - z_{max}) \le w_{min} - w - b_0 y.$$

If $x' = x$ and $z' \le z - 1$, then using the upper bound for c_0,

$$\begin{aligned}(a_0, b_0, c_0, 1) \cdot (x', y', z', w') &\ge (a_0, b_0, c_0, 1) \cdot (x, 0, z - 1, w_{min}) \\ &\ge (a_0, b_0, c_0, 1) \cdot (x, y, z, w).\end{aligned}$$

If $x' \le x - 1$, then from the choice of a_0 and c_0 and the fact that $c_0 \le 0$, we have

$$\begin{aligned}(a_0, b_0, c_0, 1) \cdot (x', y', z', w') &\ge (a_0, b_0, c_0, 1) \cdot (x - 1, 0, z_{max}, w_{min}) \\ &\ge (a_0, b_0, c_0, 1) \cdot (x, y, z, w).\end{aligned}$$

So, $\alpha \cdot (\mathbf{v} - \mathbf{v}') \le 0$. Now suppose $b_0 < 0$. Then $w_{min} - w + b_0(n - y) \le 0$. Let a_0, c_0 be such that

$$a_0 \le 0,\ c_0 \le w_{min} - w + b_0(n - y),\ a_0 + c_0(z - z_{max}) \le w_{min} - w + b_0(n - y).$$

Now $\alpha \cdot (\mathbf{v} - \mathbf{v}') \le 0$ follows from the choice of c_0 by comparison with $\alpha \cdot (x, n, z - 1, w_{min})$ if $x' = x$ and $z' \le z - 1$ and from the choice of a_0 and c_0 by comparison with $\alpha \cdot (x - 1, n, z_{max}, w_{min})$ if $x > x' \ge 0$. ∎

Using the same argument in which the roles of x and z are swapped, we can conclude that a part of the southwest quadrant belongs to another unbounded wedge.

Proposition 3.5. *For each $b_0 \in \mathbb{R}$, there exist $y, w \in \mathbb{R}$ such that $\text{cone}(x_{max}(z_{max}), y, z_{max}, w)$ is an unbounded wedge in $\mathcal{R}_{b_0}$.*

The wedges from Propositions 3.4 and 3.5 can coincide, which is indeed the case in all of the examples we have seen. Namely, all of the sequences described in section 2 have the following property.

Downloaded 05/21/24 to 3.239.95.36 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Hypothesis 3.6. We have $z_{max} = z_{max}(x_{max})$ or, equivalently, $x_{max} = x_{max}(z_{max})$.

In contrast to the situation with $(0, 0, 0, w_0)$, however, even under Hypothesis 3.6, there may be different optimal signatures $(x_{max}, y, z_{max}, w) \in \mathcal{V}$ depending on the choice of b. To summarize, we have the following dual of Corollary 3.3.

Corollary 3.7. *If Hypothesis 3.6 holds, then for each $b_0 \in \mathbb{R}$ there exist $y_m, w_m \in \mathbb{R}$ such that* $\text{cone}(x_{max}, y_m, z_{max}, w_m) \cap \mathcal{R}_{b_0}$ *is an unbounded wedge.*

While the existence of such a wedge in each $\mathcal{R}_{b_0}$ follows from Propositions 3.4 and 3.5, the proofs additionally describe its geometry. Namely, for $b_0 \geq 0$, the nonempty $\text{cone}(x_{max}, y_m, z_{max}, w_m) \cap \mathcal{R}_{b_0}$ contains the region

$$a_0, c_0 \leq w_{min} - w_m - b_0 y_m,$$

and for $b_0 < 0$ it contains the region

$$a_0, c_0 \leq w_{min} - w_m + b_0(n - y_m).$$

Thus, if it exists, $\text{cone}(x_{max}, y_m, z_{max}, w_m)$ dominates the southwest quadrant of $\mathcal{R}_{b_0}$.

We have already shown that among the edges of the unbounded regions of $\mathcal{R}_{b_0}$ there are no unbounded northeast rays. Hypothesis 3.6 implies the dual statement about southwest rays. As a consequence, the infinite edges of the unbounded regions of $\mathcal{R}_{b_0}$ conform to a particular geometry.

Theorem 3.8. *Hypothesis 3.6 holds if and only if no infinite edge of an unbounded region of $\mathcal{R}_{b_0}$ has a positive slope.*

Proof. The proof of Proposition 3.2 implies that the wedge $\text{cone}(0, 0, 0, w_0) \cap \mathcal{R}_{b_0}$ contains a ray in direction 0 to $\pi/2$ for any point in the region. This means that no unbounded region can have an infinite edge sloped in the northeast direction.

If Hypothesis 3.6 holds, by Corollary 3.7, there exist $y, w \in \mathbb{R}$ such that $\text{cone}(x_{max}, y, z_{max}, w) \cap \mathcal{R}_{b_0}$ is an unbounded wedge. Moreover, as we observed above, the proofs of Propositions 3.4 and 3.5 indicate that this wedge contains a ray in direction π to $3\pi/2$ for any point in the region. This means that no unbounded region can have an infinite edge sloped in the southwest direction. On, the other hand, if $z_{max} \neq z_{max}(x_{max})$, then the southwest quadrant of $\mathcal{R}_{b_0}$ contains unbounded portions of at least two wedges (possibly more), namely, $\text{cone}(x_{max}, y', z_{max}(x_{max}), w')$ and $\text{cone}(x_{max}(z_{max}), y'', z_{max}, w'')$, for some $y', w', y'', w'' \in \mathbb{R}$, and hence these wedges have an unbounded edge sloped in the southwest direction. ∎

In general, the slopes of finite boundary edges are also negative. However, horizontal and vertical edges are seen and, though more rare, positive slopes for a bounded region of $\mathcal{R}_{b_0}$ have been observed. See [1] for numerical results about the bounded regions.

For the sequence in Figure 3, $x_{max} = 8$ and $z_{max} = 25$. Although all slices of $\mathcal{R}_{b_0}$ contain a slice corresponding to these values, the other two coordinates y_m, w_m of the signature may be different. So, while we use the shortened signatures in the figure for clarity, in the proofs we work with the full signatures. We next describe the unbounded regions that we see as we traverse $\mathcal{R}_{b_0}$ counterclockwise from $(0, 0, 0, w_0)$ around to $(x_{max}, y_m, z_{max}, w_m)$. We start by proving that crossing a region boundary in the (a, c) plane in either a horizontal or a vertical

Downloaded 05/21/24 to 3.239.95.36 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

direction implies a strict change in the number of branching points or of branches, respectively, for the associated signatures.

Proposition 3.9. *Suppose* $(x,y,z,w),(x',y',z',w') \in \mathcal{V}$ *are optimal for* $(a,b_0,c,1)$ *and* $(a',b_0,c',1)$ *respectively, where these parameters lie in the interior of two distinct regions of* $\mathcal{R}_{b_0}$*. If* $a=a'$ *but* $c<c'$*, then* $z>z'$*. Similarly, if* $c=c'$ *but* $a<a'$*, then* $x>x'$*.*

Proof. By assumption, $\alpha\cdot(\mathbf{v}-\mathbf{v}')<0$ and $\alpha'\cdot(\mathbf{v}'-\mathbf{v})<0$ for $\mathbf{v}=(x,y,z,w)$, $\mathbf{v}'=(x',y',z',w')$, $\alpha=(a,b_0,c,1)$, and $\alpha'=(a',b_0,c',1)$. Hence, $(a-a')(x-x')+(c-c')(z-z')<0$. ∎

In comparison to Proposition 3.1, Proposition 3.9 says that the number of branch points in the corresponding optimal signatures increases monotonically whenever a region boundary is crossed as $a\to-\infty$ for fixed c in $\mathcal{R}_{b_0}$. Analogous statements can be made for the total number of branches, and for minimization.

The proof also shows that a particular combination of x and z can be associated with at most one region of $\mathcal{R}_{b_0}$with a nonempty interior. The consequence of Theorem 3.8 is that as we traverse the unbounded regions of $\mathcal{R}_{b_0}$ counterclockwise from $(0,0,0,w_0)$ around to $(x_{max},y_m,z_{max},w_m)$, we see a correlated increase in both x and z. We see a similar increase with a clockwise traversal. The difference is that counterclockwise, the number of branches is minimized, whereas clockwise it is maximized (subject to some other conditions). The cones which correspond to x_{max} and intersect $\mathcal{R}_{b_0}$ all produce unbounded regions in $\mathcal{R}_{b_0}$. In particular, we have the following wedge.

Proposition 3.10. *For each* $b_0\in\mathbb{R}$*, there exist* $y,w\in\mathbb{R}$ *such that* $\mathrm{cone}(x_{max},y,z_{min}(x_{max}),w)$ *is an unbounded wedge in* $\mathcal{R}_{b_0}$*.*

Proof. Let $x=x_{max}$, $z=z_{min}(x_{max})$, and y,w be such that $\mathbf{v}=(x,y,z,w)\in\mathcal{S}$ and b_0y+w is the least possible for the given b_0. Let $\alpha=(a_0,b_0,c_0,1)\in\mathcal{R}_{b_0}$ and $\mathbf{v}'=(x',y',z',w')\in\mathcal{S}$. Recall that n is the length of sequence s, and w_{min} is minimal over all signatures for s. Hence, $w\geq w_{min}$ and $n\geq y\geq 0$.

We claim (x,y,z,w) is optimal for parameters $\alpha=(a_0,b_0,c_0,1)\in\mathcal{R}_{b_0}$, where a_0 and c_0 satisfy the constraints below. By the choice of y and w, we may assume that for $\mathbf{v}'=(x',y',z',w')\in\mathcal{S}$ either $x'=x$ and $z'\geq z+1$ or $x>x'\geq 0$.

If $b_0\geq 0$, $w_{min}-w-b_0n\leq 0$. Let a_0,c_0 be such that

$$a_0\leq 0,\quad c_0\geq b_0n+w-w_{min},\quad a_0+c_0z\leq w_{min}-w-b_0n.$$

If $x'=x$ and $z'\geq z+1$, $\alpha\cdot(\mathbf{v}-\mathbf{v}')\leq 0$ because of the upper bound for c_0 by comparison with $(x,0,z+1,w_{min})$. If $x>x'\geq 0$, then $\alpha\cdot(\mathbf{v}-\mathbf{v}')\leq 0$ by comparison with $(x-1,0,0,w_{min})$ due to the choice of a_0 and c_0 and the fact that $c_0\leq 0$.

Now suppose $b_0<0$. Then $w_{min}-w+b_0(n-y)\leq 0$. Let a_0,c_0 be such that

$$a_0\leq 0,\quad c_0\geq b_0(y-n)+w-w_{min},\quad a_0+c_0z\leq w_{min}-w+b_0(n-y).$$

In this case, $\alpha\cdot(\mathbf{v}-\mathbf{v}')\leq 0$ follows from the choice of c_0 by comparison with $\alpha\cdot(x,n,z+1,w_{min})$ if $x=x'$ and $z'\geq z+1$ and from the choice of a_0 and c_0 by comparison with $\alpha\cdot(x-1,n,0,w_{min})$ if $x>x'\geq 0$. ∎

Downloaded 05/21/24 to 3.239.95.36 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 05/21/24 to 3.239.95.36 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

By the choice of parameters, we see that $\text{cone}(x_{max}, y, z_{min}(x_{max}), w)$ always intersects the northwest quadrant of $\mathcal{R}_{b_0}$. From the chosen a_0 and c_0, we know that it is unbounded to the west along the line $(a_0 - t, b_0, c_0, 1)$ and, since $z = z_{min}(x_{max}) > 0$ for $x_{max} > 0$, to the northwest along the line $(a_0 - t, b_0, c_0 + t/z, 1)$ for $t \geq 0$.

Moreover, one can readily see that if $z_{min}(x_{max}) \neq z_{max}(x_{max})$ and if $\text{cone}(x_{max}, y, z, w)$ contains $(a_0, b_0, c_0, 1)$ for $z_{min}(x_{max}) < z < z_{max}(x_{max})$, then it contains the whole ray $(a_0 - t, b_0, c_0, 1)$ for $t \geq 0$. Therefore $\text{cone}(x_{max}, y, z, w)$ is an unbounded stripe between the wedges $\text{cone}(x_{max}, y, z_{min}(x_{max}), w)$ and $\text{cone}(x_{max}, y, z_{max}(x_{max}), w)$.

We now show that, for a given number of branch points, having the minimal number of branches possible is necessary for signatures which correspond to regions that are unbounded to the northwest. Dually, for a given number of branch points, having the maximal number of branch points possible is necessary for the corresponding region to be unbounded to the northwest.

Proposition 3.11. *Let $(x, y, z, w) \in \mathcal{V}$ such that $R = \text{cone}(x, y, z, w) \cap \mathcal{R}_{b_0} \neq \emptyset$. If $x < x_{max}(z)$ or $z > z_{min}(x)$, then R is bounded to the northwest.*

Proof. Suppose $\mathbf{v} = (x, y, z, w)$ is optimal for $\alpha = (a_0, b_0, c_0, 1)$, where $z > z_{min}(x)$. By definition, there exist y', w' such that $\mathbf{v}' = (x, y', z', w') \in \mathcal{S}$ for $z' = z_{min}(x)$.

Let $m > 0$ and consider $\alpha' = (a_0 - t, b_0, c_0 + mt, 1)$ for $t > 0$. Then

$$\alpha'(\mathbf{v} - \mathbf{v}') = b_0(y - y') + (c_0 + mt)(z - z') + (w - w') = \alpha(\mathbf{v} - \mathbf{v}') + mt(z - z') > 0$$

for $mt > 0$ sufficiently large since $\alpha(\mathbf{v} - \mathbf{v}')$ is fixed and $z - z' > 0$. Hence, R cannot contain the ray $l = (a_0 - t, b_0, c_0 + mt, 1)$ for $t \geq 0$. We get the same contradiction if $x < x_{max}(z)$, and we consider a point $\mathbf{v}'' = (x'', y'', z, w'') \in \mathcal{S}$ with $x'' = x_{max}(z)$. ∎

We know that $z_{min}(x) \geq 3x$ since a branch point must have at least three branches by definition. Hence, a minimally branched structure resembles a binary tree in the sense that each branch point has exactly two children. We note, however, that this says nothing about nonbranching vertices and also that the root does not count as a branch point for our purposes since its energy function has no entropic penalty.

As far as we have seen, this lower bound on the total number of branches is always achieved by some signature having the maximum number of branch points, and it again has interesting geometric implications.

Hypothesis 3.12. We have $z_{min}(x_{max}) = 3x_{max}$.

Corollary 3.13. *If Hypothesis 3.12 holds, then for every $0 < x < x_{max}$, $z_{min}(x) = 3x$.*

Proof. Suppose S is a structure with signature $(x, y, 3x, w)$ for some $0 < x \leq x_{max}$, $y, w \in \mathbb{R}$; then in the rooted tree representation of S (Figure 2), all of its branching nodes have exactly two children. Take one of the branching nodes whose children are leaves (i.e., not branching nodes themselves) and unpair all the base pairs in the branches that meet at that node. The resulting structure has $x - 1$ branching points and $3x - 3$ branches. ∎

In this case, when the region corresponds to a signature having the minimal number of branches possible, this is sufficient for the region to be unbounded to the northwest.

Proposition 3.14. *Suppose* $(x, y, 3x, w) \in \mathcal{V}$ *with* $R = \text{cone}(x, y, 3x, w) \cap \mathcal{R}_{b_0} \neq \emptyset$. *Then* R *is unbounded to the northwest.*

Proof. Suppose $\mathbf{v} = (x, y, 3x, w)$ is optimal for $\alpha = (a_0, b_0, c_0, 1)$. Let $\alpha' = (a_0 - t, b_0, c_0 + \frac{t}{3}, 1)$ for $t > 0$ and $\mathbf{v}' = (x', y', z', w') \in \mathcal{V}$. Since $\alpha(\mathbf{v} - \mathbf{v}') \leq 0$, then $\alpha'(\mathbf{v} - \mathbf{v}') \leq 0$ follows from

$$-tx' + \frac{t}{3}z' \geq -tx + \frac{t}{3}3x,$$

which is equivalent to $z' \geq 3x'$. ∎

We begin our characterization of the east half of $\mathcal{R}_{b_0}$ with a result analogous to Proposition 3.10; the proof is a straightforward dualization.

Proposition 3.15. *For each* $b_0 \in \mathbb{R}$, *there exist* $y, w \in \mathbb{R}$ *such that* $\text{cone}(x_{min}(z_{max}), y, z_{max}, w)$ *is an unbounded wedge in* $\mathcal{R}_{b_0}$.

By duality, $\text{cone}(x_{min}(z_{max}), y, z_{max}, w)$ is partly in the southeast quadrant. In the examples of RNA sequences we have seen, $x_{min}(z_{max}) = x_{max}$, and this wedge coincides with $\text{cone}(x_{max}, y_m, z_{max}, w_m)$. However, this is not true for all sequences. For example, for the sequence $(\text{GACAAA})^6$, $z_{max} = 6$, $x_{max} = 2$, but $x_{min}(6) = 1$. Regardless, if the sequence is long enough so that $x_{max} > 1$, the southeast quadrant is guaranteed to contain at least one more unbounded wedge that we have not mentioned so far but show next.

Proposition 3.16. *If* $x_{max} \geq 1$, *then for each* $b_0 \in \mathbb{R}$, *there exist* $y, w \in \mathbb{R}$ *such that* $\text{cone}(1, y, z_{max}(1), w)$ *is an unbounded wedge in* $\mathcal{R}_{b_0}$.

Proof. Let $z_1 = z_{max}(1)$ and y, w be such that $\mathbf{v} = (1, y, z_1, w) \in \mathcal{S}$ and $b_0 y + w$ is the least possible. Let $\alpha = (a_0, b_0, c_0, 1) \in \mathcal{R}_{b_0}$ and $\mathbf{v}' = (x', y', z', w') \in \mathcal{S}$. By the choice of $b_0 y + w$, we may assume either that $x' = 1$ and $z' \leq z_1 - 1$ or that $x' \geq 2$.

Suppose $b_0 \geq 0$. Then, as in previous proofs, $w_{min} - w - b_0 y \leq 0$. Let a_0, c_0 be such that

$$a_0 \geq 0, \quad c_0 \leq w_{min} - w - b_0 y, \quad a_0 + c_0(z_{max} - z_1) \geq b_0 y + w - w_{min}.$$

In this case, $\alpha \cdot (\mathbf{v} - \mathbf{v}') \leq 0$ follows from the choice of c_0 by comparison with $\alpha \cdot (1, 0, z_1 - 1, w_{min})$ if $x' = 1$ and $z' \leq z_1 - 1$ and from the choice of a_0 and c_0 by comparison with $\alpha \cdot (2, 0, z_{max}, w_{min})$ if $x' \geq 2$.

Now suppose $b_0 < 0$. Again we have $w_{min} - w + b_0(n - y) \leq 0$. Let a_0, c_0 be such that

$$a_0 \geq 0, \quad c_0 \leq w_{min} - w + b_0(n - y), \quad a_0 + c_0(z_{max} - z_1) \geq b_0(y - n) + w - w_{min}.$$

Now $\alpha \cdot (\mathbf{v} - \mathbf{v}') \leq 0$ follows from the choice of c_0 by comparison with $\alpha \cdot (1, n, z_1 - 1, w_{min})$ if $x' = 1$ and $z' \leq z_1 - 1$ and from the choice of a_0 and c_0 by comparison with $\alpha \cdot (2, n, z_{max}, w_{min})$ if $x' \geq 2$. ∎

A result analogous to Proposition 3.11 also holds; the proof is a straightforward dualization.

Proposition 3.17. *Let* $(x, y, z, w) \in \mathcal{V}$ *such that* $R = \text{cone}(x, y, z, w) \cap \mathcal{R}_{b_0} \neq \emptyset$. *If* $x > x_{min}(z)$ *or* $z < z_{max}(x)$, *then* R *is bounded to the southeast.*

Downloaded 05/21/24 to 3.239.95.36 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Hence, only the regions corresponding to signatures in the set

$$\mathcal{S}_0 := \{(x, y, z, w) \in \mathcal{S} \colon x = x_{min}(z), z = z_{max}(x)\}$$

are candidates for regions in $\mathcal{R}_{b_0}$ which are unbounded to the southeast. In fact, the set of candidate unbounded regions can be reduced even further to those that satisfy

$$\text{for every} \quad (x', y', z', w') \in \mathcal{S}_0, \quad x' < x \iff z' < z, \tag{3}$$

or, equivalently, to those that satisfy

$$\text{for every} \quad (x', y', z', w') \in \mathcal{S}_0, \quad x' > x \iff z' > z, \tag{4}$$

as explained in the following proposition.

Proposition 3.18. *The regions which are not in*

$$\mathcal{S}_1 := \{(x, y, z, w) \in \mathcal{S}_0 \colon (x, y, z, w) \textit{ satisfies } (3)\} = \{(x, y, z, w) \in \mathcal{S}_0 \colon (x, y, z, w) \textit{ satisfies } (4)\}$$

are bounded to the southeast.

Proof. Using the same reasoning as in the proof of Proposition 3.11, $(x, y, z, w) \in \mathcal{S}_0$ is bounded unless

$$z_{max}(x') < z \quad \text{for all} \quad 1 \le x' < x, \tag{5}$$

$$x_{min}(z') > x \quad \text{for all} \quad z < z' \le z_{max}. \tag{6}$$

In fact, (5) and (6) are equivalent. To exclude the signatures from $\mathcal{S}_0$ that do not satisfy (5) and (6), it is sufficient to check against points from $\mathcal{S}_0$. Namely, we claim that $(x, y, z, w) \in \mathcal{S}_0$ satisfies (5) and (6) if and only if it satisfies (3).

It is clear that (5) and (6) imply (3). To see the converse, suppose $(x, y, z, w) \in \mathcal{S}_0$ satisfies (3) and suppose x_1 is such that $1 \le x_1 < x$ but $z_1 = z_{max}(x_1) \ge z$. Let

$$x_2 = x_{min}(z_1), z_2 = z_{max}(x_2), x_3 = x_{min}(z_2), z_3 = x_{min}(x_3), \ldots.$$

Then

$$x > x_1 \ge x_2 \ge x_3 \ge \cdots,$$
$$z \le z_1 \le z_2 \le z_3 \le \cdots$$

are two bounded sequences and, therefore, eventually stabilize. Suppose for all sufficiently large n, we have $x_n = x^*$ and $z_n = z^*$. Then there is a signature $(x^*, y^*, z^*, w^*) \in \mathcal{S}_0$ such that $x^* < x$ but $z^* \ge z$. This is a contradiction. Therefore, we conclude that (3) implies (5).

Notice that for two points $(x, y, z, w), (x', y', z', w') \in \mathcal{S}_0$, we have $x = x'$ if and only if $z = z'$. Hence, a signature $(x, y, z, w) \in \mathcal{S}_0$ satisfies (3) if and only if it satisfies (4). ∎

The next result completely characterizes the regions in $\mathcal{S}_1$ which are unbounded to the southeast.

Downloaded 05/21/24 to 3.239.95.36 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 05/21/24 to 3.239.95.36 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Theorem 3.19. *Suppose* $(x,y,z,w) \in \mathcal{S}_1$, $x > 1$, $z < z_{max}$, *is such that* $R = \text{cone}(x,y,z,w) \cap \mathcal{R}_{b_0} \neq \emptyset$. *Then* R *is bounded to the southeast if and only if there exist* $(x',y',z',w'), (x'',y'',z'',w'') \in \mathcal{S}_0$ *with* $x' < x < x''$ *(equivalently,* $z' < z < z''$*) such that*

$$\frac{x-x'}{z-z'} > \frac{x-x''}{z-z''}. \tag{7}$$

Proof. Suppose R is unbounded to the southeast and contains the ray $(t,0,-mt,0), t \geq 0$, for some $m > 0$. Let $\mathbf{v_0} = (x_0,y_0,z_0,w_0)$. Then $(1,0,-m,0)\cdot(\mathbf{v}-\mathbf{v_0}) \leq 0$ for every $\mathbf{v_0} \in \mathcal{S}$, which implies that $\frac{x-x'}{z-z'} \leq m$ for every $(x',y',z',w') \in \mathcal{S}$ with $z' < z$ and $m \leq \frac{x-x''}{z-z''}$ for every $(x'',y'',z'',w'') \in \mathcal{S}$ with $z'' > z$.

Conversely, suppose there are no two points $(x',y',z',w'), (x'',y'',z'',w'') \in \mathcal{S}_0$ with $x' < x < x''$ (equivalently, $z' < z < z''$) such that (7) holds. Let $m > 0$ be such that

$$\max\left\{\frac{x-x'}{z-z'} : (x',y',z',w') \in \mathcal{S}_0, z' < z\right\} \leq m \leq \min\left\{\frac{x-x''}{z-z''} : (x'',y'',z'',w'') \in \mathcal{S}_0, z'' > z\right\}. \tag{8}$$

The parameter m can be chosen to be positive because all the fractions in the sets in (8) are positive. We claim that the region R contains the ray $(t,0,-mt,0)$ and hence is unbounded to the southeast. Suppose this is not the case. Then there is $(x_0,y_0,z_0,w_0) \in \mathcal{S}$ such that

$$(x-x_0) - m(z-z_0) > 0.$$

We first consider the case $z_0 < z$. Then $m > 0$ implies $x > x_0$. Similarly, as in the proof of Proposition 3.18, we can take

$$x_1 = x_{min}(z_0), z_1 = z_{max}(x_1), x_2 = x_{min}(z_1), z_2 = z_{max}(x_2), \ldots,$$

which yields two bounded monotone sequences,

$$x > x_0 \geq x_1 \geq x_2 \geq \cdots$$
$$z_0 \leq z_1 \leq z_2 \leq \cdots,$$

that eventually stabilize. Suppose $x_n = x^*$, $z_n = z^*$ for all $n \geq n_0$ for some $n_0 \in \mathbb{N}$. Then there is a signature $(x^*,y^*,z^*,w^*) \in \mathcal{S}_0$ for some y^*, w^*. Since $(x,y,z,w) \in \mathcal{S}_1$, $x^* < x$ implies $z^* < z$ and, consequently, $z_n \leq z$ for all $n \in \mathbb{N}$. Moreover, by construction,

$$0 < m < \frac{x-x_0}{z-z_0} \leq \frac{x-x_1}{z-z_1} \leq \cdots \leq \frac{x-x^*}{z-z^*},$$

which contradicts (8). The case when $z_0 > z$ leads to a similar contradiction, while $z_0 = z$ is clearly impossible. ∎

The proof of Theorem 3.19 also gives a criterion for determining whether for $(x,y,z,w) \in \mathcal{S}_1$ the unbounded region $R = \text{cone}(x,y,z,w) \cap \mathcal{R}_{b_0}$ is an unbounded stripe. Namely, R is a stripe if and only if

$$\max\left\{\frac{x-x'}{z-z'} : (x',y',z',w') \in \mathcal{S}_0, z' < z\right\} = \min\left\{\frac{x-x''}{z-z''} : (x'',y'',z'',w'') \in \mathcal{S}_0, z'' > z\right\}.$$

Another property that we have observed for the RNA sequences described in section 2 is that

$$z_{max}(x) \leq z_{max}(x-1)+2 \quad \text{for } 2 \leq z \leq z_{max}. \tag{9}$$

Suppose (9) is satisfied. Then $z_{max} \leq z_{max}(1)+2(x-1)$ for $1 \geq x \geq x_{max}$. Let $(x,y,z,w) \in \mathcal{S}_1$ be such that $R = \text{cone}(x,y,z,w) \cap \mathcal{R}_{b_0} \neq \emptyset$ and $z = z_{max}(x) = z_{max}(1)+2(x-1)$. Then for $(x',y',z',w') \in \mathcal{S}_0$, since $z' \leq z_{max}(1)+2(x'-1)$, we have $z-z' \geq 2(x-x')$, which implies

$$\max\left\{\frac{x-x'}{z-z'} : (x',y',z',w') \in \mathcal{S}_0, z' < z\right\} \leq \frac{1}{2} \leq \min\left\{\frac{x-x''}{z-z''} : (x'',y'',z'',w'') \in \mathcal{S}_0, z'' > z\right\},$$

and therefore by Theorem 3.19, R is unbounded. This explains the arithmetic progression with common difference 2 in the z coordinates that we see in Figure 3 when we traverse the unbounded regions clockwise starting from $(0,0,0,w_0)$.

4. Conclusion and discussion. This paper addresses the question of parametric secondary structure prediction under the NNTM. The parameters of interest that are allowed to vary are those used to score the branching loops. Instead of performing a complete parametric RNA folding for a single sequence and describing which convex regions of the parameter space would yield the same structure, we focus on distinguishing the features of the parametric optimization that are common to all sequences.

We have shown that for each sequence, under certain conditions which are empirically true for naturally occurring RNA sequences, the structures with a minimal and maximal number of branches for a given number of branching points are optimal for a vast portion of the parameter space when the parameter b, which penalizes single-stranded nucleotides in the multibranch loops, is kept constant. This was done via a complete characterization of the unbounded regions in the $\mathcal{R}_{b_0}$ section of the normal fan of the branching polytope. While not all maximal branching structures correspond to unbounded regions, we have completely characterized those that do and have shown that this really depends on the combinatorics of the possible pairings for the sequence, not on the energy of the other motifs in those structures.

As a consequence of our descriptions of the vertices that correspond to the unbounded regions in $\mathcal{R}_{b_0}$, we can conclude that the secondary structures that are biologically reasonable have signatures that correspond to bounded regions, where the optimization is less stable under the change of branching parameters. To improve the average prediction accuracy, therefore, one would need to consider structures that are approximately correct. The accuracy, stability, and robustness are analyzed in [1].

We note that some of the results, such as Proposition 3.1, even though stated in the language of the branching polytopes, hold for general polytopes. However, the main results are affected by the fact that the signatures, especially the first and third coordinates, encode a tree structure and are, therefore, subject to certain relations such as $z \geq 3x$.

Some of our results depend on hypotheses which we have observed to be true for branching polytopes of RNA sequences. We believe that these assumptions need not be true for all sequences over the four letter alphabet, but that the counterexamples would be pathological. For example, for the sequence

$$\text{ACCCCGACCCUUUUCCCAGCCCCA} \tag{10}$$

Downloaded 05/21/24 to 3.239.95.36 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

we have $z_{max}(2) \geq 6$, but $z_{max}(1) = 3$ and hence does not satisfy (9). Notice that this sequence is very short and palindromic. However, if the rooted tree corresponding to the structure with signature (x, y, z, w) has depth more than one, there are two branching points separated by a stem, and breaking the base pairs in that stem produces a structure with $x - 1$ branching points and $z - 2$ branches. Therefore, any counterexample to (9) would have to satisfy the property that for some $x > 1$, all structures with x branching points and $z_{max}(x)$ branches do not have a path between the branching points that does not involve the root. The number of such structures is limited because of the possibility of alternative configurations, so we believe that this is true for RNA sequences. However, a detailed discussion of this (and certainly a proof) would need to involve a different approach, so we do not attempt to give one here.

In another case, for the precise condition in Theorem 3.19, we had to introduce the set $\mathcal{S}_1$, which is determined by the technical condition (3). However, for the branching polytopes that we have computed, we observe that

$$x_1 < x_2 \implies z_{max}(x_1) \leq z_{max}(x_2)$$

and that there is a structure with x branching points and z branches for every $x_{min} \leq x \leq x_{max}$, $z_{min}(x) \leq z \leq z_{max}(x)$. These conditions together imply (3), which means that in practice, $\mathcal{S}_1 = \mathcal{S}_0$. We expect that a counterexample would also be a pathological sequence. Therefore, the following questions are natural to ask: *Can our assumptions be mathematically justified? Is there a reasonable probability distribution of sequences under which Hypotheses 3.6 and 3.12 hold?*

Acknowledgment. We would like to thank Matthew Ielusic for helping us with the colors in Figures 1 and 3.

REFERENCES

[1] F. Barrera-Cruz, C. Heitsch, A. Kirkpatrick, M. Lelusic, and S. Poznanović, *Statistics on RNA branching polytopes: Accuracy, stability, robustness, and other characteristics*, in preparation.

[2] J. J. Cannone, S. Subramanian, M. N. Schnare, J. R. Collett, L. M. D'Souza, Y. Du, B. Feng, N. Lin, L. V. Madabusi, K. M. Müller, N. Pande, Z. Shang, N. Yu, and R. R. Gutell, *The comparative RNA web (CRW) site: An online database of comparative sequence and structure information for ribosomal, intron, and other RNAs*, BMC Bioinformatics, 3 (2002), 2.

[3] C. N. Dewey, P. M. Huggins, K. Woods, B. Sturmfels, and L. Pachter, *Parametric alignment of Drosophila genomes*, PLoS Comput. Biol., 2 (2006), e73.

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

[5] E. Drellich, A. Gainer-Dewar, H. A. Harrington, Q. He, C. Heitsch, and S. Poznanović, *Geometric combinatorics and computational molecular biology: Branching polytopes for RNA sequences*, in Algebraic and Geometric Methods in Discrete Mathematics, Contemp. Math. 685, AMS, 2017, pp. 137–154.

[6] B. Grünbaum, V. Klee, M. A. Perles, and G. C. Shephard, *Convex Polytopes*, Pure Appl. Math. 16, Interscience Publishers John Wiley & Sons, Inc., 1967.

[7] D. Gusfield, K. Balasubramanian, and D. Naor, *Parametric optimization of sequence alignment*, Algorithmica, 12 (1994), 312.

[8] V. Hower and C. E. Heitsch, *Parametric analysis of RNA branching configurations*, Bull. Math. Biol., 73 (2011), pp. 754–776.

Downloaded 05/21/24 to 3.239.95.36 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

[9] M. Jiang, J. Anderson, J. Gillespie, and M. Mayne, *uShuffle: A useful tool for shuffling biological sequences while preserving the k-let counts*, BMC Bioinformatics, 9 (2008), 192.
[10] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner, *Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure*, J. Molecular Biol., 288 (1999), pp. 911–940.
[11] D. H. Mathews and D. H. Turner, *Prediction of RNA secondary structure by free energy minimization*, Current Opinion in Structural Biology, 16 (2006), pp. 270–278.
[12] R. Nussinov, G. Pieczenik, J. R. Griggs, and D. J. Kleitman, *Algorithms for loop matchings*, SIAM J. Appl. Math., 35 (1978), pp. 68–82, https://doi.org/10.1137/0135006.
[13] L. Pachter and B. Sturmfels, *Parametric inference for biological sequence analysis*, Proc. Natl. Acad. Sci. USA, 101 (2004), pp. 16138–16143.
[14] L. Pachter and B. Sturmfels, eds., *Algebraic Statistics for Computational Biology*, Cambridge University Press, 2005.
[15] SageMath, Inc., *CoCalc: Collaborative Computation in the Cloud*, https://cocalc.com/, 2017.
[16] P. Stein and M. Waterman, *On some new sequences generalizing the Catalan and Motzkin numbers*, Discrete Math., 26 (1979), pp. 261–272.
[17] M. S. Swenson, J. Anderson, A. Ash, P. Gaurav, Z. Sükösd, D. A. Bader, S. C. Harvey, and C. E. Heitsch, *GTfold: Enabling parallel RNA secondary structure prediction on multi-core desktops*, BMC Research Notes, 5 (2012), 341.
[18] D. H. Turner and D. H. Mathews, *NNDB: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure*, Nucleic Acids Research, 38 (2009), pp. D280–D282.
[19] L. Wang and J. Zhao, *Parametric alignment of ordered trees*, Bioinformatics, 19 (2003), pp. 2237–2245.
[20] M. Waterman, M. Eggert, and E. Lander, *Parametric sequence comparisons*, Proc. Natl. Acad. Sci. USA, 89 (1992), pp. 6090–6093.
[21] J. Zuber, H. Sun, X. Zhang, I. McFadyen, and D. H. Mathews, *A sensitivity analysis of RNA folding nearest neighbor parameters identifies a subset of free energy parameters with the greatest impact on RNA secondary structure prediction*, Nucleic Acids Research, 45 (2017), pp. 6168–6176.
[22] M. Zuker, *Mfold web server for nucleic acid folding and hybridization prediction*, Nucleic Acids Research, 31 (2003), pp. 3406–3415.
[23] M. Zuker and P. Stiegler, *Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information*, Nucleic Acids Research, 9 (1981), pp. 133–148.

Downloaded 05/21/24 to 3.239.95.36 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy