Branching Polytopes for RNA Sequences

Fidel Barrera-Cruz

Joint work with: Christine Heitsch (GATech), and
Svetlana Poznanović (Clemson)

2017 SIAM Annual Meeting
July 12, 2017

Outline

Introduction

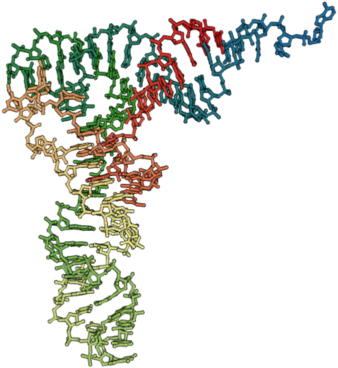
Nearest Neighbor Thermodynamic Model

Geometry: Polytopes and normal fans

Results

Introduction

RNA

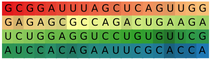


We can think of RNA as a sequence of nucleic acids (adenine (A), cytosine (C), guanine (G), and uracil (U)).

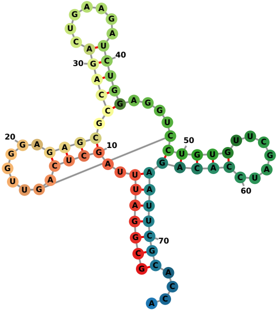
RNA is part of several biological processes such as gene expression and protein encoding.

1D, 2D and 3D structures

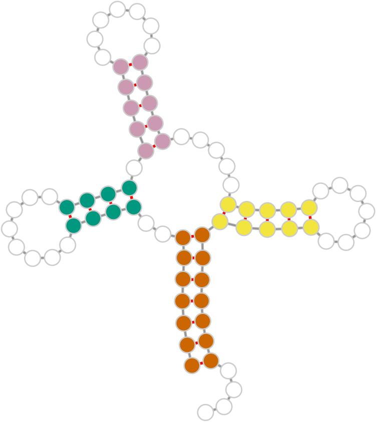
- RNA sequence (essentially a path P)
- RNA sequence, and list of pairs (essentially a matching using edges not in P, each edge is of the form CG or AU)
- RNA sequence, list of pairs and positions in 3-space of each nucleotide

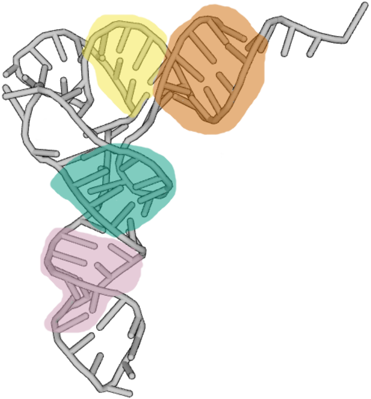


GCGGAUUUAGCUCAGUUGG
GAGAGCGCCAGACUGAAGA
UCUGGAGGUCCUGUGUUCG
AUCCACAGAAUUCGCACCA



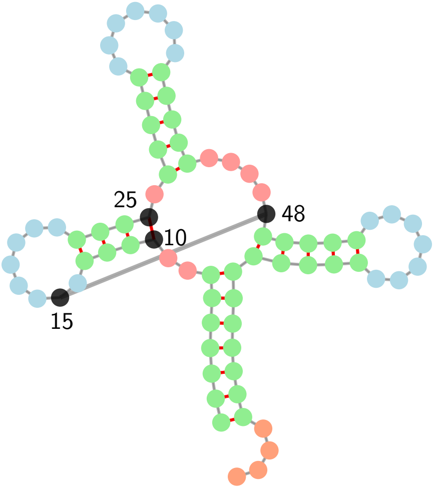
Tertiary to secondary



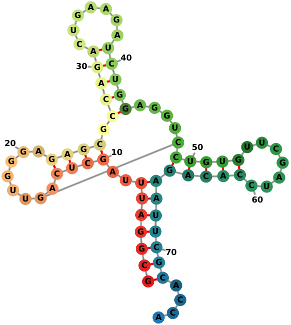


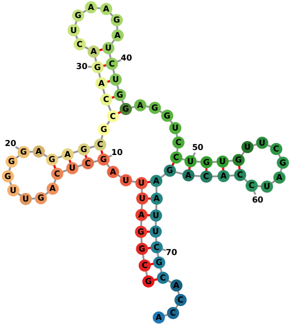
Pseudoknots

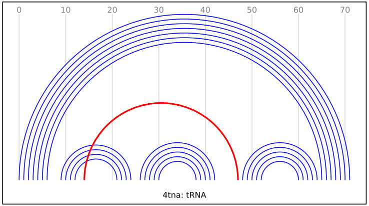
If a secondary structure contains two pairs $p = (i, j)$ and $p' = (i', j')$ such that $i < i' < j < j'$ then we say that the structure contains a pseudoknot.

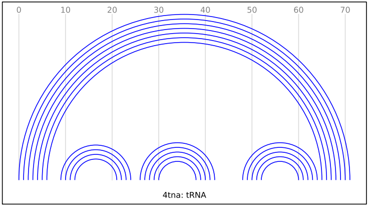


Assume structures contain no pseudoknots

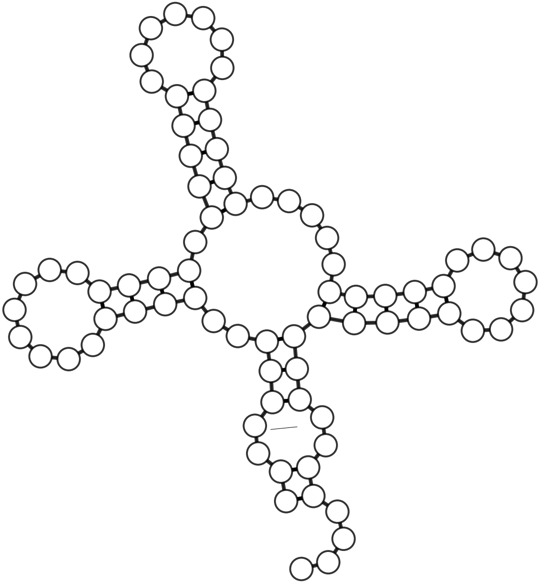




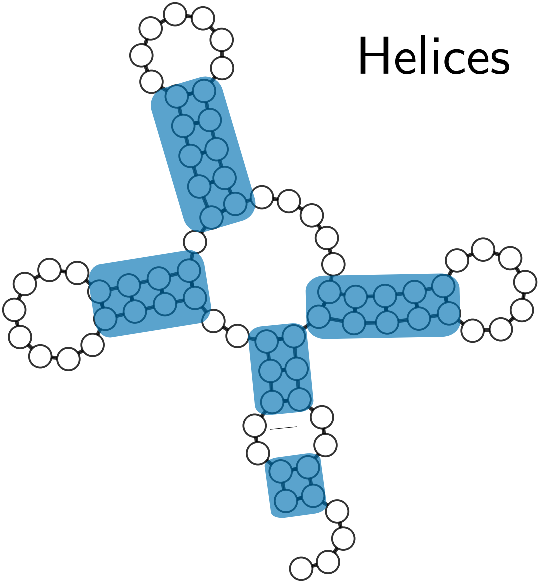




Some substructure terminology

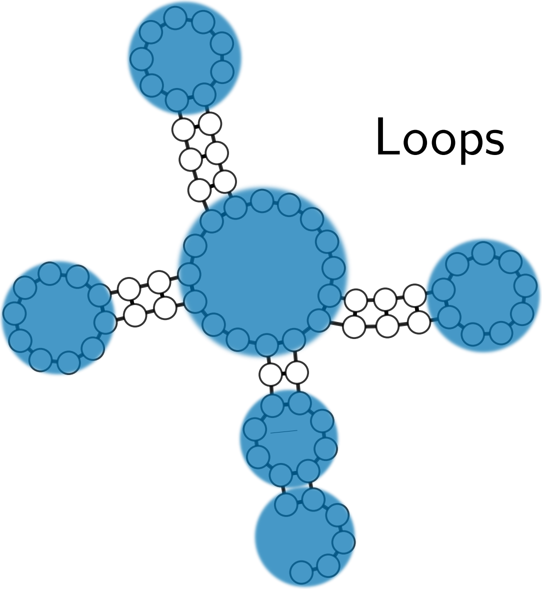


Some substructure terminology



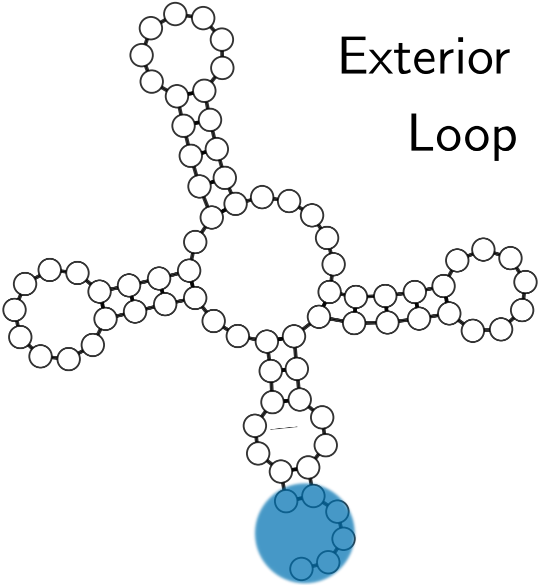
Helices

Some substructure terminology



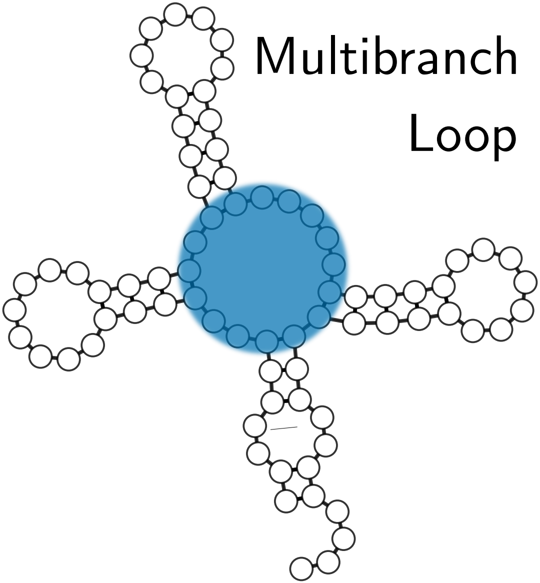
Loops

Some substructure terminology



Exterior Loop

Some substructure terminology



Multibranch
Loop

Nearest Neighbor Thermodynamic Model

Nearest Neighbor Thermodynamic Model

We can think of the Nearest Neighbor Thermodynamic Model (NNTM) as a way to compute energies of secondary structures via a table lookup. That is, given a secondary structure we can compute its free energy by adding the energy that each substructure contributes.

Given a particular RNA sequence S, the number of secondary structures it admits can be exponential on the length of S.

With the restriction that no pseudoknots appear, these structures can be recursively decomposed. Thus allowing for the use of dynamic programming to find the structure with minimum free energy in a reasonable time.

Energy of a multibranch loop

Under the NNTM model the energy assigned to a multibranch loop can be written as

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

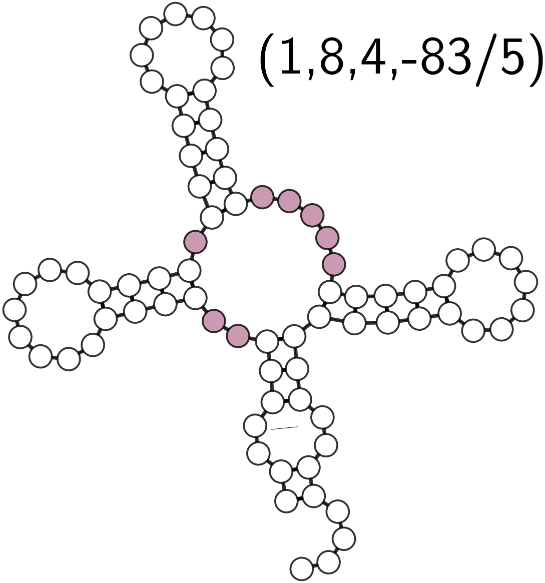
where a, b and c are parameters that have been estimated before.

Thus the energy of a secondary structure of a sequence is given as

$$\begin{aligned} & a \cdot (\#\text{multibranch loops}) + b \cdot (\#\text{unpaired nucleotides}) + \\ & c \cdot (\#\text{branching helices}) + w, \end{aligned}$$

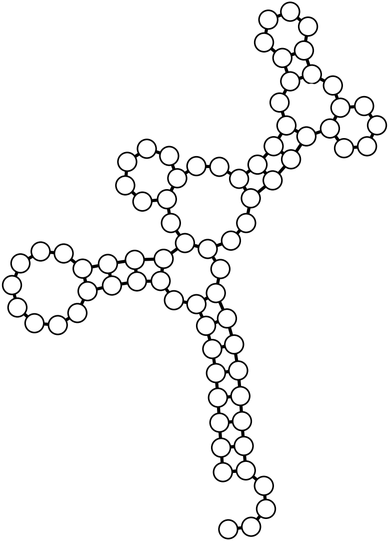
where w is the energy that arises from all other parameters in the NNTM model.

The signature (x, y, z, w) of a structure

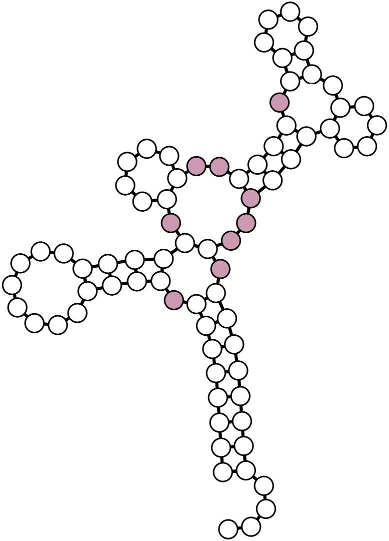


(1,8,4,-83/5)

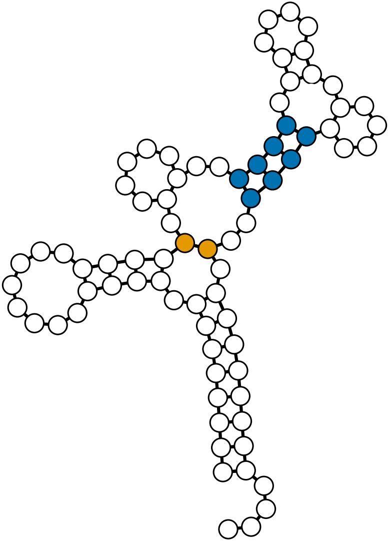
The signature (x, y, z, w) of a structure



The signature (x, y, z, w) of a structure

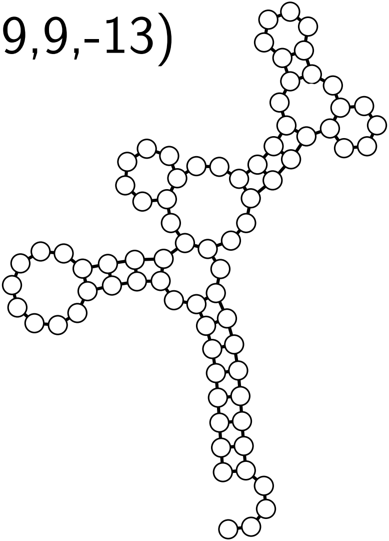


The signature (x, y, z, w) of a structure



The signature (x, y, z, w) of a structure

(3,9,9,-13)



Reformulation of NNTM

We will focus on trying to understand how the optimal secondary structure depends on the multibranch loop parameters.

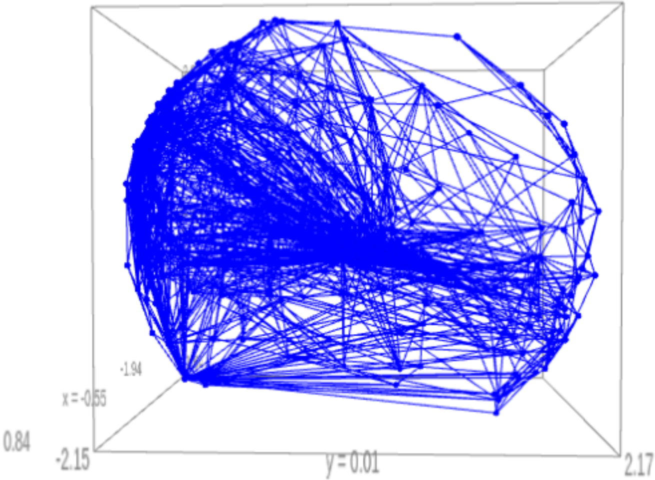
To this end, we reformulate the NNTM as follows. For a given secondary structure S we associate the following parametrized energy

$$\Delta G_S(a, b, c, d) = ax + by + cz + dw,$$

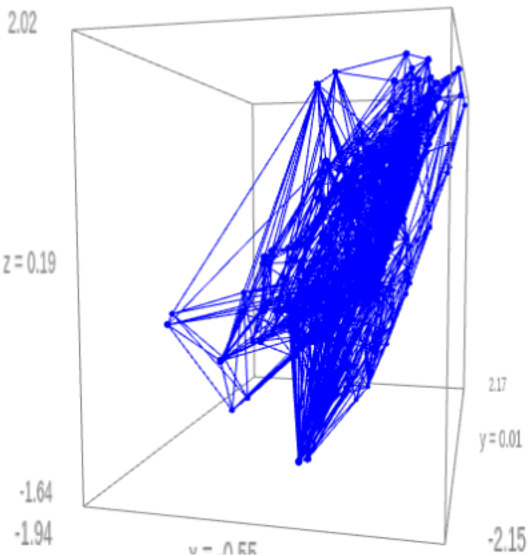
where (x, y, z, w) is the signature of S.

Geometry: Polytopes and normal fans

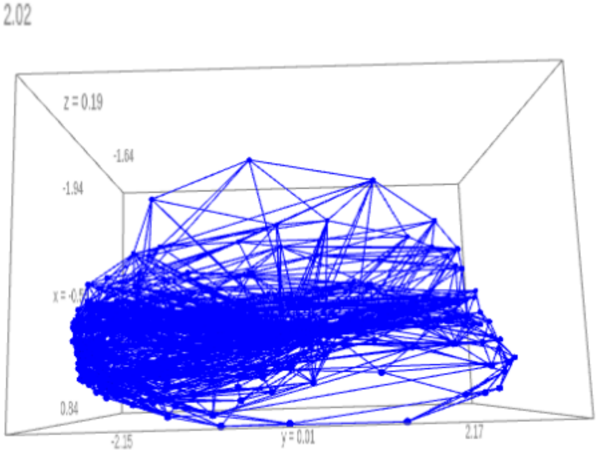
A branching polytope



A branching polytope



A branching polytope

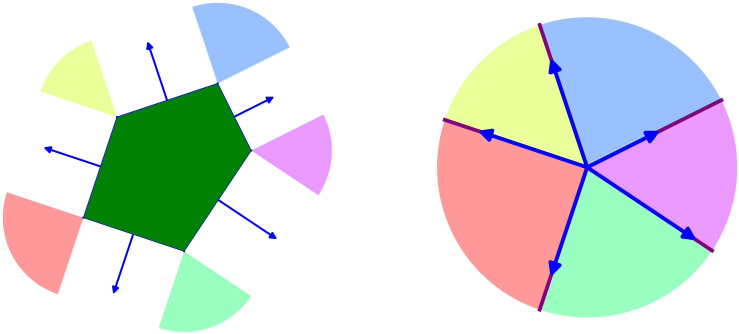


Normal fan

Given a polytope P and a non trivial face F of P we consider the set C of normal vectors $\mathbf{c}$ that are orthogonal to a supporting hyperplane of F and such that any point in F is a solution to the linear program

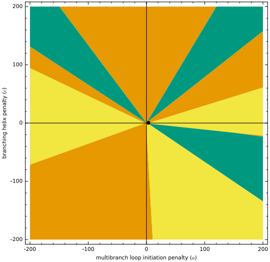
$$\max_{\mathbf{x} \in P} \mathbf{c}^{\mathsf{T}} \cdot \mathbf{x}.$$

The set C is a *cone* and the collection of all the cones is called the *normal fan* of P.

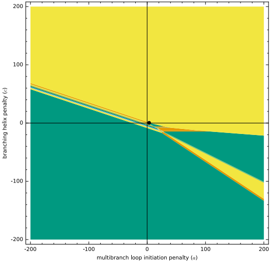


Which projection corresponds to a branching polytope?

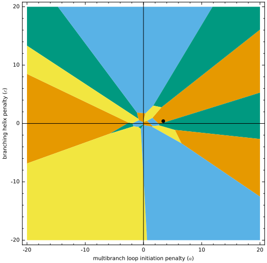
b = 0: 21 slices visible



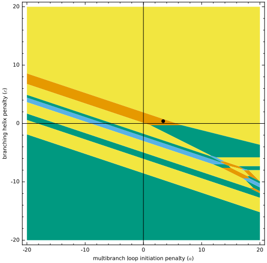
b = 0: 45 slices visible



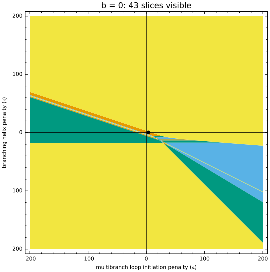
b = 0: 20 slices visible

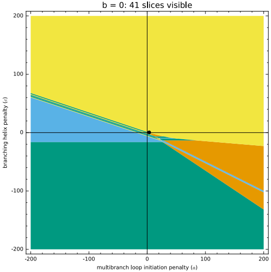


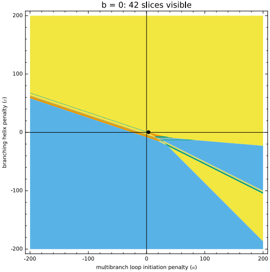
b = 0: 25 slices visible

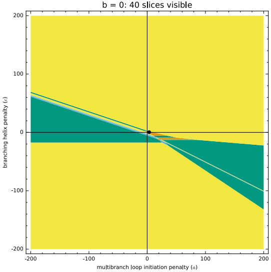


Strips, wedges and polygons.

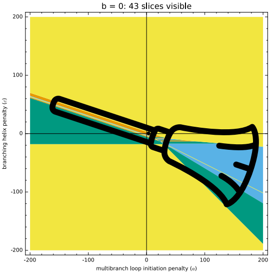




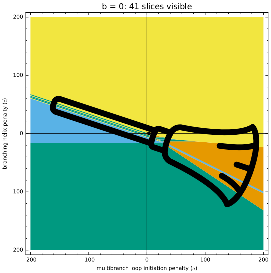




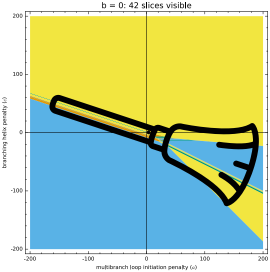
Strips, wedges and polygons.



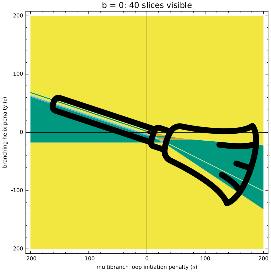
b = 0: 43 slices visible



b = 0: 41 slices visible



b = 0: 42 slices visible

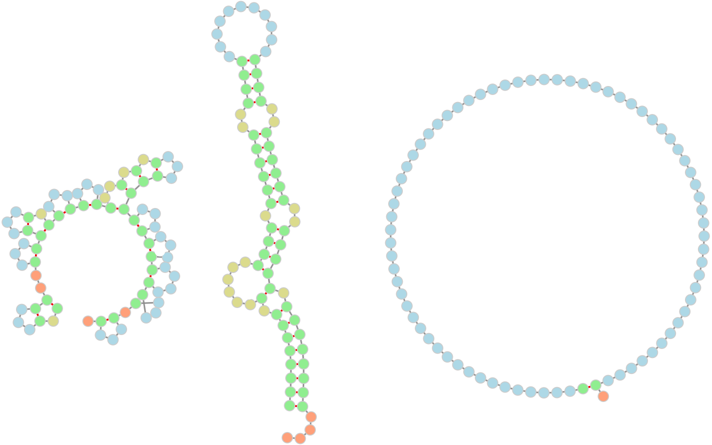


b = 0: 40 slices visible

Results

Structural results

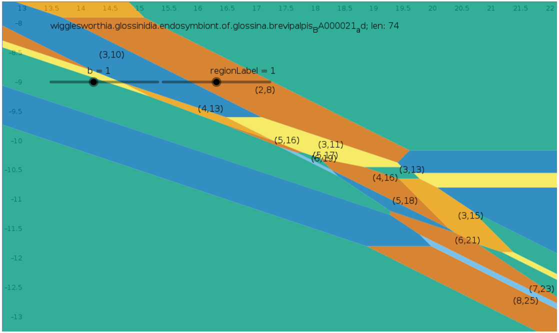
In any *ac*-slice of the normal fan, the region corresponding to $(0, 0, 0, w)$ is a wedge, in fact it contains a quadrant.



Structural results

The number of multibranch loops in an optimal structure increases as $a \to -\infty$. Similarly, the number branches in an optimal structure increases as $c \to -\infty$.

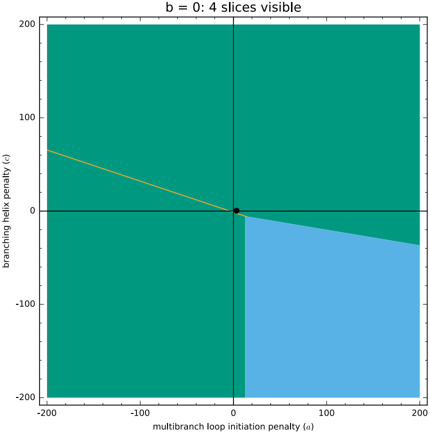
Intuitively, if we move horizontally and cross a boundary between regions, then the number of multibranch loops in the region to the left is greater than the number of multibranch loops in the region to the right. Similarly for the vertical direction.



Structural results

Let x_{max} be the largest value of x in a signature (x, y, z, w) appearing in a given ac-slice $\mathcal{S}$. Similarly define z_{max}.

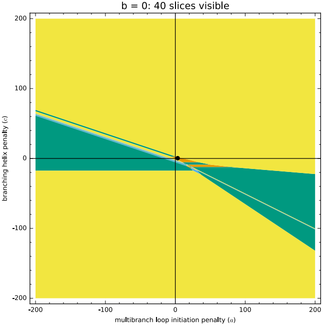
There exist no unbounded regions with positive slope if and only if there is a region with signature (x_{max}, y, z_{max}, w) in $\mathcal{S}$.



Structural results

Let $\mathcal{S}$ be an ac-slice and let $X = \{x : (x, y, z, w) \in \mathcal{S}\}$. For each $x \in X$ define $z_{\max}(x) = \max\{z : (x, y, z, w) \in \mathcal{S}\}$ and $z_{\min}(x) = \min\{z : (x, y, z, w) \in \mathcal{S}\}$.

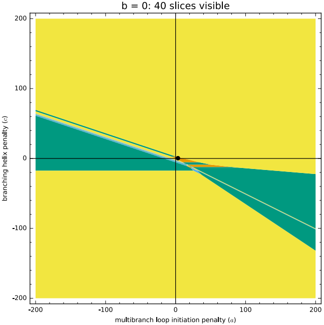
If R is an unbounded region associated to signature (x_0, y, z, w), then $z \in \{z_{\min}(x_0), z_{\max}(x_0)\}$.



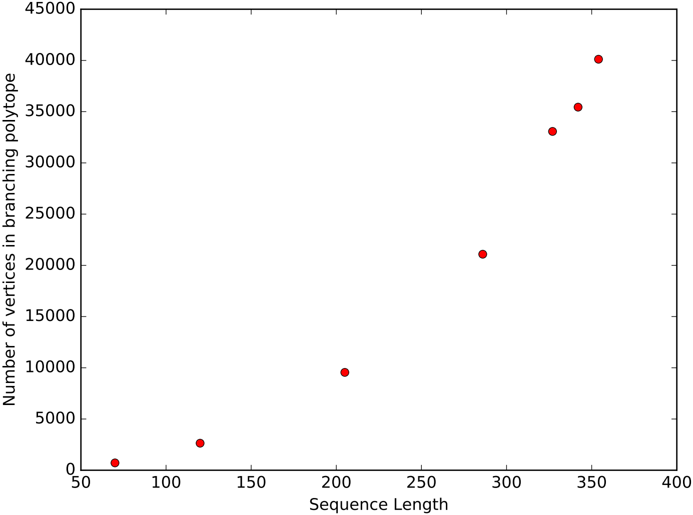
Structural results

The handle of the broomstick corresponds to structures having minimum branching, that is, structures were each multibranch loop is incident to exactly 3 helices.

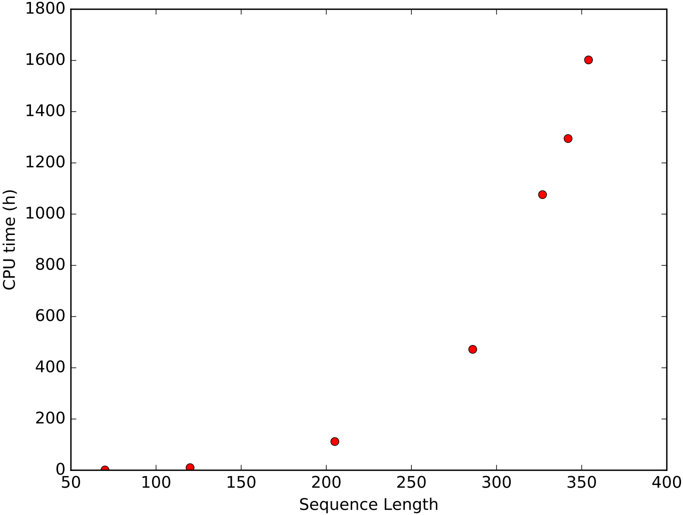
In other words, if there is x such that signature $(x, y, 3x, w)$ appears in an *ac*-slice, then this region is unbounded (it at least contains a ray in direction $(-3, 1)$).



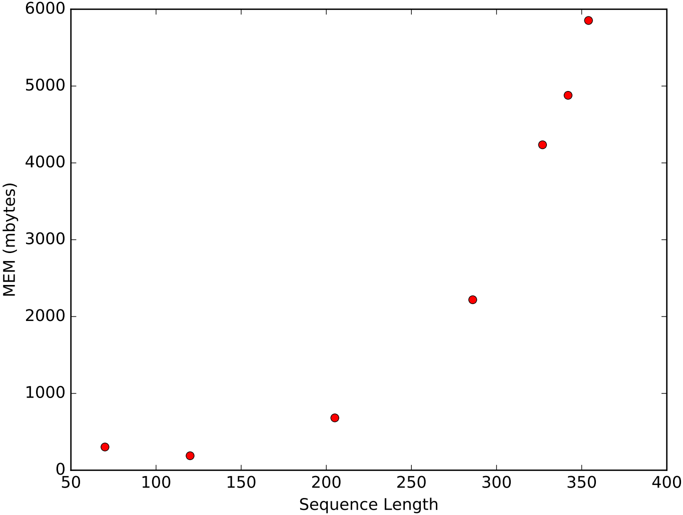
Size of polytopes



Computing times



Memory usage



Future steps

- It may be the case that several secondary structures are assigned the same signature. Does a *better* choice of the representing secondary structure improve on the accuracy of the prediction?
- How sensitive is the model to changing the parameters? Is there a set of parameters that would improve accuracy for a specific type of RNA?
- Can the structural information aid in optimizing the algorithm to compute the polytope for larger sequences? Is parallelization possible?

Branching Polytopes for RNA Sequences

Fidel Barrera-Cruz

Georgia Tech

Joint work with: Christine Heitsch (GATech), and
Svetlana Poznanović (Clemson)

2017 SIAM Annual Meeting
July 12, 2017

Thanks!