Topological classification and enumeration of RNA structures by genus

Andersen, Jørgen Ellegard; Penner, Robert C.; Reidys, Christian; Waterman, Michael S.

Published in:
Journal of Mathematical Biology

DOI:
10.1007/s00285-012-0594-x

Publication date:
2013

Document version:
Submitted manuscript

Citation for pulished version (APA):
Andersen, J. E., Penner, R. C., Reidys, C., & Waterman, M. S. (2013). Topological classification and enumeration of RNA structures by genus. *Journal of Mathematical Biology*, *67*(5), 1261-1278. https://doi.org/10.1007/s00285-012-0594-x

TOPOLOGICAL CLASSIFICATION AND ENUMERATION OF RNA STRUCTURES BY GENUS

J. E. ANDERSEN, R. C. PENNER, C. M. REIDYS, AND M. S. WATERMAN

Abstract. To an RNA pseudoknot structure is naturally associated a topological surface, which has its associated genus, and structures can thus be classified by genus. Based on earlier work of Harer-Zagier, we compute the generating function $\mathbf{D}_{g,\sigma}(z) = \sum_n \mathbf{d}_{g,\sigma}(n)z^n$ for the number $\mathbf{d}_{g,\sigma}(n)$ of those structures of fixed genus g and minimum stack size σ with n nucleotides so that no two consecutive nucleotides are basepaired and show that $\mathbf{D}_{g,\sigma}(z)$ is algebraic. In particular, we prove that $\mathbf{d}_{g,2}(n) \sim k_g\, n^{3(g-\frac{1}{2})}\gamma_2^n$, where $\gamma_2 \approx 1.9685$. Thus, for stack size at least two, the genus only enters through the sub-exponential factor, and the slow growth rate compared to the number of RNA molecules implies the existence of neutral networks of distinct molecules with the same structure of any genus. Certain RNA structures called shapes are shown to be in natural one-to-one correspondence with the cells in the Penner-Strebel decomposition of Riemann's moduli space of a surface of genus g with one boundary component, thus providing a link between RNA enumerative problems and the geometry of Riemann's moduli space.

Introduction

An RNA molecule is described by its primary structure, a linear string composed of the nucleotides **A**, **G**, **U** and **C**, referred to as the backbone. The number of nucleotides is called the length of the molecule. Nucleotides may pair according to the symmetric Watson-Crick rules: **A-U**, **G-C** and **U-G**. The predominance of such parings form the RNA secondary structure, where by definition, if nucleotides U and V are paired and X and Y are paired, then they cannot occur in the order $X - U - Y - V$ in the primary structure. The combinatorics and prediction of RNA secondary from primary structure was pioneered three decades ago by Michael Waterman [36, 37, 38, 39, 25].

Date: October, 2010.

JEA and RCP are supported by QGM (Centre for Quantum Geometry of Moduli Spaces, funded by the Danish National Research Foundation).

In fact, RNA is structurally less constrained than its chemical cousin DNA and folds into a variety of tertiary structures as shown by experimental findings as well as by comparative sequence analysis [41]. These structures are called pseudoknot structures, and their topology has been studied in [34, 35, 6, 20]. Folded RNA facilitates various biochemical tasks, for example, acting as a messenger linking DNA with proteins and catalyzing diverse reactions just as proteins themselves. Though most of the pairings in a folded RNA can typically be described by the secondary structure alone, pseudoknots occur rather often in practice and are known to be functionally important, for instance, in tRNAs, RNAseP [18], telomerase RNA [33] and ribosomal RNAs [16].

An RNA structure can be represented by drawing its backbone as a horizontal line containing vertices corresponding to nucleotides and each Watson-Crick base pair as a semi-circle or *chord* in the upper halfplane. Such a representation is called a *partial (linear) chord diagram*, i.e., a collection of chords attached to a backbone possibly containing isolated vertices.

Two distinct chords with respective endpoints $i_1 < j_1$ and $i_2 < j_2$ are "consecutively parallel" if $i_1 = i_2 - 1 \leq j_2 = j_1 - 1$, and consecutive parallelism generates the equivalence relation of "parallelism" whose equivalence classes are called *stacks*. A stack of size σ is such an equivalence class containing exactly σ consecutively parallel chords.

A partial chord diagram is called a *(linear) chord diagram*[1] if every vertex has an incident chord, so the number of vertices for a linear chord diagram is necessarily even. A chord connecting vertices which are consecutive along the backbone is called a 1-*chord*, and a chord connecting the first and last vertices is called a *rainbow*. A linear chord diagram in which every stack has cardinality one is called a *seed*, and a seed without 1-chords that contains the rainbow is called a *shape*.

Consider a graph G, for example, a (partial) linear chord diagram. Given an oriented edge e of G, let $v(e)$ denote the vertex to which e points. A *fatgraph* [21, 22, 23] is a graph together with a cyclic ordering on $\{e : v(e) = v\}$, for each vertex v of G. This additional structure gives rise to certain collection of cyclically ordered sequences of oriented edges called the *boundary cycles*, where an oriented edge e is followed by the next edge in the cyclic ordering at $v(e)$, but with the opposite

[1]These combinatorial structures occur in a number of instances in pure mathematics including finite type invariants of knots and links [4, 17], the representation theory of Lie algebras [7], the geometry of moduli spaces of flat connections on surfaces [2, 3], mapping class groups [1] and the Four-Color Theorem [5], and in applied mathematics including codifying the pairings among nucleotides in RNA molecules [27], or more generally the contacts of any binary macromolecule [24, 25, 40], and in the analysis of data structures [8, 9].

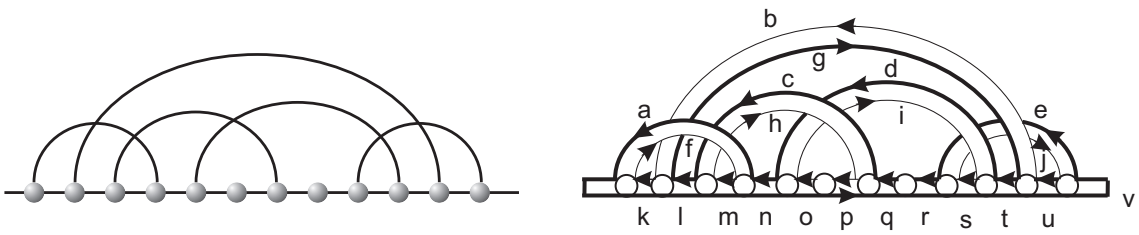

FIGURE 1. Computing the number of boundary components of partial chord diagram. The diagram contains $5 + 11$ edges and 12 vertices. We follow the cycles described in the text and observe that there are exactly two boundary cycles (bold and thin). The genus of the diagram is given by $1 - \frac{1}{2}(12 - 16 + 2) = 2$.

orientation, so that it points away from $v(e)$. In depicting a fatgraph, we shall always identify the cyclic ordering at a vertex with the counterclockwise orientation of the plane, according to which we shall represent the boundary cycle of G as a path alongside it with G on the left, cf. Figure 1.

Let r denote the number of distinct boundary cycles of the connected fatgraph G with v vertices and e edges. The Euler characteristic of G is $v - e = 2 - 2g - r$, where

$$g = 1 - \frac{1}{2}(e - v - r)$$

is called the *genus* of the fatgraph G. As illustrated in Figure 1, by fattening up the vertices into disks and the edges into bands connecting these disks, there results a topological surface $F(\mathbb{G})$ with r boundary components of genus g in the standard mathematical parlance; see [24] for details. In particular for a (partial) chord diagram, the backbone may be collapsed to a single vertex without affecting the Euler characteristic, whence the relationship

$$2 - 2g - r = 1 - m, \tag{0.1}$$

between the genus g, the number r of boundary cycles, and the number m of chords.

Let $\mathscr{C}_g(n), \mathscr{S}_g(n)$ and $\mathscr{T}_g(n)$ denote the respective collections of all linear chord diagrams, seeds and shapes of genus g on $2n$ vertices, i.e., with n chords. Furthermore, let $\mathbf{c}_g(n)$, $\mathbf{s}_g(n)$, $\mathbf{t}_g(n)$ denote the cardinalities of these sets, respectively, with generating functions $\mathbf{C}_g(z) = \sum_{n\geq 0} \mathbf{c}_g(n) z^n$, $\mathbf{S}_g(z) = \sum_{n\geq 0} \mathbf{s}_g(n) z^n$ and $\mathbf{T}_g(z) = \sum_{n\geq 0} \mathbf{t}_g(n) z^n$.

The objects of primary biological interest are *RNA σ-structures*, i.e., partial chord diagrams with minimum stack size σ that do not contain any 1-chords. The parameter σ derives from the fact that stacks of small cardinality are typically energetically unfavorable, and 1-chords are prohibited due to the tensile rigidity of the RNA sugar-phosphate backbone. Let $\mathbf{d}_{g,\sigma}(n)$ be the number of all RNA σ-structures of genus g with generating function $\mathbf{D}_{g,\sigma}(z) = \sum_{n\geq 0} \mathbf{d}_{g,\sigma}(n) z^n$.

We shall calculate $\mathbf{D}_{g,\sigma}(z)$ in Theorem 2 as

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

where $u_\sigma(z) = \frac{(z^2)^{\sigma-1}}{z^{2\sigma}-z^2+1}$. This expression for $\mathbf{D}_{g,\sigma}(z)$ is actually quite explicit owing to the fact that a three-term recursion is given for the coefficients [2] $\mathbf{c}_g(n)$ in [13] as recalled in Theorem 1. In Corollary 1, we compute $\mathbf{C}_g(z)$ in terms of a certain polynomial $P_g(z)$, which can likewise be recursively calculated, cf. section 2. In particular for $g = 0$, the $\mathbf{c}_0(n) = \binom{2n}{n}\frac{1}{n+1} = \frac{(2n)!}{(n+1)!n!}$ are given by the Catalan numbers, i.e., the numbers of triangulations of a polygon with $n + 2$ sides, with generating function $\mathbf{C}_0(z) = \frac{1-\sqrt{1-4z}}{2z}$.

In Theorem 2, we furthermore prove that $\mathbf{D}_{g,\sigma}(z)$ is algebraic over $\mathbb{C}(z)$, and for arbitrary but fixed g and $\gamma_2 \approx 1.9685$, we have

$$\mathbf{d}_{g,2}(n) \sim k_g\, n^{3(g-\frac{1}{2})}\gamma_2^n, \tag{0.2}$$

for some constant k_g. The exponential growth rate of 1.9685 shows that the number of RNA σ-structures grows much more slowly than the number of RNA sequences over the natural alphabet. This implies the existence of neutral networks [15, 28, 29], i.e., vast extended sets of RNA sequences all folding into a single RNA σ-structure. These neutral networks are of key importance in the context of neutral evolution of RNA sequences.

Our second main result is Corollary 4 which shows that Riemann's moduli space of a surface of genus g with one boundary component is naturally homeomorphic to the geometric realization of

[2] The numbers $\mathbf{c}_g(n)$ had been computed in another generating function over two decades ago by Harer-Zagier [13] in the equivalent guise of the number of side pairings of a polygon with $2n$ sides that produce a surface of genus g, namely,

$$1 + 2\sum_{n\geq 0}\sum_{2g\leq n} \frac{\mathbf{c}_g(n)}{(2n-1)!!}\, x^{n+1-2g}\, z^{n+1} = \left(\frac{1+z}{1-z}\right)^x,$$

a striking and beautiful formula.

all RNA shapes of genus g. This follows from a deep theorem of Penner-Strebel about a certain mapping class group invariant cell decomposition of Teichmüller space [21, 22, 32].

1. Further generating functions

Let $\mathscr{C}_g(n,m) \supseteq \mathscr{S}_g(n,m)$ denote the collections of all linear chord diagrams and seeds of genus $g \geq 0$ on $2n \geq 0$ vertices containing $m \geq 0$ 1-chords with respective generating functions

$$\mathbf{C}_g(x,y) = \sum_{m,n\geq 0} \mathbf{c}_g(n,m)x^n y^m,$$

$$\mathbf{S}_g(x,y) = \sum_{m,n\geq 0} \mathbf{s}_g(n,m)x^n y^m,$$

where $\mathbf{c}_g(n,m) = \mathbf{s}_g(n,m) = 0$ if $2g > n$ or if $m > n$.

Let $\mathscr{P}(n)$ denote the collection of all partial linear chord diagrams on n vertices. There is a natural projection ϑ from partial chord diagrams to seeds defined by collapsing each non-empty stack onto a single chord and removing any unpaired vertices

$$\vartheta\colon \sqcup_{n\geq 1} \mathscr{P}(n) \to \sqcup_{n\geq 1}\mathscr{S}(n),$$

which is surjective and preserves genus.

Furthermore, ϑ restricts to a surjection

$$\vartheta : \sqcup_{n\geq 0}\mathscr{C}_g(n,m) \to \sqcup_{n\geq 0}\mathscr{S}_g(n,m)$$

which collapses each stack to a chord and therefore preserves both the genus g and the number m of 1-chords. For any shape $\gamma \in \sqcup_{n\geq 0}\mathscr{S}(n,m)$, let

$$\mathscr{C}_\gamma(n,m) = \mathscr{C}(n,m) \cap \vartheta^{-1}(\gamma)$$

denote the intersection with the fiber $\vartheta^{-1}(\gamma)$ with its generating function $\mathbf{C}_\gamma(x,y)$.

Likewise, the projection ϑ restricts to a surjection

$$\vartheta : \sqcup_{n\geq 0}\mathscr{D}_{g,\sigma}(n) \to \sqcup_{n\geq 0}\mathscr{S}_g(n),$$

which preserves the genus. For any shape $\gamma \in \sqcup_{n\geq 0}\mathscr{S}(n)$, let

$$\mathscr{D}_{\gamma,\sigma}(n) = \mathscr{D}_\sigma(n) \cap \vartheta^{-1}(\gamma)$$

denote the intersection with the fiber $\vartheta^{-1}(\gamma)$ with its generating function $\mathbf{D}_{\gamma,\sigma}(z)$.

As a general notational point for any power series $R(z) = \sum a_i z^i$, we shall write $[z^i]R(z) = a_i$ for the extraction of the coefficient a_i of z^i.

2. The generating function $\mathbf{C}_g(z)$

A seminal result due to Harer and Zagier [13], cf. also [11, 12], computes a recursion and generating function for the number $\mathbf{c}_g(n)$ of linear chord diagrams of genus g with n chords as follows:

Theorem 1. [13] *The $\mathbf{c}_g(n)$ satisfy the recursion*

$$(n+1)\,\mathbf{c}_g(n) = 2(2n-1)\,\mathbf{c}_g(n-1) + (2n-1)(n-1)(2n-3)\,\mathbf{c}_{g-1}(n-2), \tag{2.1}$$

where $\mathbf{c}_g(n) = 0$ for $2g > n$.

The recursion eq. (2.1) translates into the ODE

$$z(1-4z)\frac{d}{dz}\mathbf{C}_g(z) + (1-2z)\mathbf{C}_g(z) = \Phi_{g-1}(z), \tag{2.2}$$

where

$$\Phi_{g-1}(z) = z^2\left(4z^3\frac{d^3}{dz^3}\mathbf{C}_{g-1}(z) + 24z^2\frac{d^2}{dz^2}\mathbf{C}_{g-1}(z) + 27z\frac{d}{dz}\mathbf{C}_{g-1}(z) + 3\mathbf{C}_{g-1}(z)\right)$$

with initial condition $\mathbf{C}_g(0) = 0$. The general solution is given by

$$\mathbf{C}_{g+1}(z) = \left(\int_0^z \frac{\Phi_g(y)}{(1-4y)^{3/2}}dy + C\right)\frac{\sqrt{1-4z}}{z}, \tag{2.3}$$

where

$$\begin{aligned}\Phi_g(z) &= 4z^5\frac{d^3}{dz^3}\mathbf{C}_g(z) + 24z^4\frac{d^2}{dz^2}\mathbf{C}_g(z) + 27z^3\frac{d}{dz}\mathbf{C}_g(z) + 3z^2\mathbf{C}_g(z)\\ &= \frac{Q_g(z)}{(1-4z)^{3g+5/2}}\end{aligned}$$

with $Q_g(z)$ a polynomial of degree at most $(3g+2)$, $Q_g(1/4) \neq 0$ and $[z^h]Q_g(z) = 0$ if $0 \leq h \leq 2g+1$. Analysis of the partial fraction expansion of $Q_g(z)$ then provides the following expression, which is implicit in [13].

Corollary 1. *For any $g \geq 1$ the generating function $\mathbf{C}_g(z) = \sum_{n\geq 0} \mathbf{c}_g(n) z^n$ is given by*

$$\mathbf{C}_g(z) = P_g(z)\frac{\sqrt{1-4\,z}}{(1-4z)^{3g}}, \tag{2.4}$$

where $P_g(z)$ is a polynomial with integral coefficients of degree at most $(3g-1)$, $P_g(1/4) \neq 0$, $[z^{2g}]P_g(z) \neq 0$ and $[z^h]P_g(z) = 0$ for $0 \leq h \leq 2g-1$.

The recursion eq. (2.1) permits the calculation of the polynomials $P_g(z)$, the first several of which are given as follows:

$$\begin{aligned}
P_1(z) &= z^2, \\
P_2(z) &= 21z^4\,(z+1) \\
P_3(z) &= 11z^6\,\left(158\,z^2+558\,z+135\right), \\
P_4(z) &= 143z^8\left(2339\,z^3+18378\,z^2+13689\,z+1575\right), \\
P_5(z) &= 88179z^{10}\,\left(1354\,z^4+18908\,z^3+28764\,z^2+9660\,z+675\right).
\end{aligned}$$

Conjecture 1. The polynomial $P_g(z)$ has all of its coefficients positive integers in the range $2g$ to $3g-1$ and is the generating polynomial for some as-yet unknown class of shapes.

A straightforward analysis [10] of the singularity of $\mathbf{C}_g(z)$ then gives:

Corollary 2. *For any $g \geq 1$ the generating function $\mathbf{C}_g(z)$ is algebraic over $\mathbb{C}(z)$ and has its unique singularity at $z = 1/4$ independent of genus. Furthermore, the coefficients of $\mathbf{C}_g(z)$ have the asymptotics*

$$[z^n]\mathbf{C}_g(z) \sim \frac{P_g(\frac{1}{4})}{\Gamma(3g-1/2)} n^{3g-\frac{3}{2}} 4^n. \tag{2.5}$$

3. RNA σ-structures of genus g

We extend the enumerative results of the previous section to RNA σ-structures by first specializing to seeds, which are then "inflated" by expanding chords into stacks and adding possible unpaired vertices.

Lemma 1. *If $g \geq 1$, then*

$$\mathbf{S}_g(z,u) = \frac{1+z}{1+2z-zu}\mathbf{C}_g\left(\frac{z(1+z)}{(1+2z-zu)^2}\right). \tag{3.1}$$

Proof. We first prove

$$\mathbf{C}_g(x,y) = \frac{1}{x+1-yx}\,\mathbf{C}_g\left(\frac{x}{(x+1-yx)^2}\right), \tag{3.2}$$

and to this end, choose $\xi \in \mathscr{C}_g(s+1,m+1)$ and label one of its 1-chords. Since we can label any of the $(m+1)$ 1-chords of ξ, $(m+1)$ $\mathbf{c}_g(s+1,m+1)$ different such labeled linear chord diagrams arise. On the other hand, to produce ξ with this labeling, we can add one labeled 1-chord to an element of $\mathscr{C}_g(s,m+1)$ by inserting a parallel copy of an existing 1-chord or by inserting a new labeled 1-chord in an element of $\mathscr{C}_g(s,m)$, where we may only insert the 1-chord between two vertices not already forming a 1-chord. It follows that we have the recursion

$$(m+1)\;\mathbf{c}_g(n+1,m+1) = (m+1)\;\mathbf{c}_g(n,m+1) + (2n+1-m)\;\mathbf{c}_g(n,m)$$

or equivalently the PDE

$$\frac{\partial \mathbf{C}_g(x,y)}{\partial y} = x\frac{\partial \mathbf{C}_g(x,y)}{\partial y} + 2x^2\frac{\partial \mathbf{C}_g(x,y)}{\partial x} + x\mathbf{C}_g(x,y) - xy\frac{\partial \mathbf{C}_g(x,y)}{\partial y}, \tag{3.3}$$

which is thus satisfied by $\mathbf{C}_g(x,y)$.

On the other hand,

$$\mathbf{C}^*_g(x,y) = \frac{1}{x+1-yx}\mathbf{C}_g\left(\frac{x}{(x+1-yx)^2}\right)$$

is also a solution of eq. (3.3), which specializes to $\mathbf{C}_g(x) = \mathbf{C}^*_g(x,1)$, and moreover, we have $\mathbf{c}^*_g(n,m) = [x^n y^m]\mathbf{C}^*_g(x,y) = 0$, for $m > n$. Indeed, the first assertion is easily verified directly, the specialization is obvious, and the fact that y only appears in the power series $\mathbf{C}^*_g(x,y)$ in the form of products xy implies that $\mathbf{c}^*_g(n,m) = 0$, for $m > n$. Thus, the coefficients $\mathbf{c}^*_g(n,m)$ satisfy the same recursion and initial conditions as $\mathbf{c}_g(n,m)$, and hence by induction on n, we conclude $\mathbf{c}^*_g(n,m) = \mathbf{c}_g(n,m)$, for $n,m \geq 0$. This proves that $\mathbf{C}_g(n,m)$ indeed satisfies eq. (3.2) as was claimed.

To complete the proof of eq. (3.1), we use that the projection ϑ is surjective and affects neither the genus nor the number of 1-chords, namely,

$$\mathbf{C}_g(x,y) = \sum_{m\geq 0}\sum_{\substack{\gamma \text{ having genus } g\\ \text{and } m \text{ 1-chords}}} \mathbf{C}_\gamma(x,y).$$

Furthermore, if a seed γ has s chords, of which t are 1-chords, then we have

$$\mathbf{C}_\gamma(x,y) \quad = \quad \left(\frac{x}{1-x}\right)^s y^t,$$

which shows that $\mathbf{C}_\gamma(x,y)$ depends only on the total number of chords and number of 1-chords in γ. Consequently,

$$\mathbf{C}_g(x,y) = \sum_{m\geq 0} \sum_{\substack{\gamma \text{ having genus } g \\ \text{and } m \text{ 1-chords}}} \mathbf{C}_\gamma(x,y) = \sum_{s\geq 0}\sum_{m=0}^{s} \mathbf{s}_g(s,m)\left(\frac{x}{1-x}\right)^s y^m = \mathbf{S}_g\left(\frac{x}{1-x}, y\right). \tag{3.4}$$

Setting $z = \frac{x}{1-x}$, i.e., $x = \frac{z}{1+z}$, and $u = y$, we arrive at

$$\mathbf{S}_g(z,u) = \frac{1+z}{1+2z-zu}\mathbf{C}_g\left(\frac{z(1+z)}{(1+2z-zu)^2}\right),$$

as required. □

Lemma 2. *For any seed γ with $s \geq 1$ chords and $m \geq 0$ 1-chords, we have*

$$\mathbf{D}_{\gamma,\sigma}(z) = (1-z)^{-1}\left(\frac{z^{2\sigma}}{(1-z^2)(1-z)^2 - (2z-z^2)z^{2\sigma}}\right)^s z^m.$$

In particular, $\mathbf{D}_{\gamma,\sigma}(z)$ depends only upon the number of chords and 1-chords in γ.

Proof. We shall construct $\sqcup_{n\geq 0}\mathcal{D}_{\gamma,\sigma}(n)$ with simple combinatorial building blocks. As a point of notation and as usual, if $\mathcal{X} = \sqcup_{n\geq 0}\mathcal{X}(n)$ is a collection of sets of partial matchings on $n \geq 0$ vertices, then we consider the corresponding generating function $\mathbf{X}(z) = \sum_{n\geq 0}\mathbf{x}(n)z^n$. In particular, we have the set $\mathcal{Z}$ consisting of a single vertex with generating function $\mathbf{Z}(z) = z$ and the set $\mathcal{R}$ consisting of a single arc and no additional vertices with generating function $\mathbf{R}(z) = z^2$.

Let $=$ denote set-theoretic bijection, $+$ disjoint union, $\times$ Cartesian product with iteration written as exponentiation, $\mathcal{I}$ the empty set, and $\textsc{Seq}(\mathcal{X}) = \mathcal{I} + \mathcal{X} + \mathcal{X}^2 + \cdots$, for any collection $\mathcal{X}$.

Define the set $\mathcal{L} = \textsc{Seq}(\mathcal{Z})$ consisting of any number $n \geq 0$ of isolated vertices and no chords, with its generating function $\mathbf{L}(z) = 1/(1-z)$, and the set $\mathcal{K}^\sigma$ comprised of a single stack with at least $\sigma \geq 1$ arcs and no additional vertices, with its generating function $\mathbf{K}^\sigma(z) = z^{2\sigma}/(1-z^2)$.

The collection $\mathcal{N}^\sigma = \mathcal{K}^\sigma \times \left(\mathcal{Z}\times\mathcal{L} + \mathcal{Z}\times\mathcal{L} + (\mathcal{Z}\times\mathcal{L})^2\right)$ of all single stacks together with a non-empty interval of unpaired vertices on at least one side thus has generating function

$$\mathbf{N}^\sigma(z) = \frac{z^{2\sigma}}{1-z^2}\left(2\frac{z}{1-z} + \left(\frac{z}{1-z}\right)^2\right).$$

Furthermore, the collection $\mathcal{M}^\sigma = \mathcal{K}^\sigma \times \textsc{Seq}(\mathcal{N}^\sigma)$ of all pairs consisting of a stack $\mathcal{K}^\sigma$ and a (possibly empty) sequence of neighboring stacks likewise has generating function

$$\mathbf{M}^\sigma(z) = \frac{\mathbf{K}^\sigma(z)}{1-\mathbf{N}^\sigma(z)} = \frac{\frac{z^{2\sigma}}{1-z^2}}{1-\frac{z^{2\sigma}}{1-z^2}\left(2\frac{z}{1-z} + \left(\frac{z}{1-z}\right)^2\right)},$$

where only intervals of isolated vertices as are necessary to separate the neighboring stacks have been inserted in $\mathcal{M}^\sigma$.

To complete the construction and count, we must still insert possible unpaired vertices at the remaining $2s+1$ possible locations, where there must be a non-trivial such insertion between the endpoints of each 1-chord. These insertions correspond to $\mathcal{L}^{2s+1-m}\times(\mathcal{Z}\times\mathcal{L})^m$, and we therefore conclude that $\sqcup_{n\geq 0}\mathcal{P}_\gamma(n) = (\mathcal{M}^\sigma)^s \times \mathcal{L}^{2s+1-m}\times(\mathcal{Z}\times\mathcal{L})^m$ has the asserted generating function

$$\begin{aligned}\mathbf{D}_{\gamma,\sigma}(z) &= \left(\frac{\frac{z^{2\sigma}}{1-z^2}}{1-\frac{z^{2\sigma}}{1-z^2}\left(2\frac{z}{1-z}+\left(\frac{z}{1-z}\right)^2\right)}\right)^s \left(\frac{1}{1-z}\right)^{2s+1-m}\left(\frac{z}{1-z}\right)^m \\ &= (1-z)^{-1}\left(\frac{z^{2\sigma}}{(1-z^2)(1-z)^2-(2z-z^2)z^{2\sigma}}\right)^s z^m.\end{aligned}$$

□

Our main result about enumerating RNA σ-structures follows.

Theorem 2. *Suppose $g,\sigma\geq 1$ and let $u_\sigma(z) = \frac{(z^2)^{\sigma-1}}{z^{2\sigma}-z^2+1}$. Then the generating function $\mathbf{D}_{g,\sigma}(z)$ is algebraic over $\mathbb{C}(x)$ and given by*

$$\mathbf{D}_{g,\sigma}(z) = \frac{1}{u_\sigma(z)z^2-z+1}\,\mathbf{C}_g\left(\frac{u_\sigma(z)z^2}{(u_\sigma(z)z^2-z+1)^2}\right). \tag{3.5}$$

In particular, for arbitrary but fixed g and $\gamma_2\approx 1.9685$, we have

$$[z^n]\mathbf{D}_{g,2}(z) \sim k_g\, n^{3(g-\frac{1}{2})}\gamma_2^n, \tag{3.6}$$

for some constant k_g depending only on g.

Proof. Since each element $\mathscr{D}_{g,\sigma}(n)$ projects to a unique seed γ with genus g and some number $m \geq 0$ of 1-chords, we have

$$\mathbf{D}_{g,\sigma}(z) = \sum_{m \geq 0} \sum_{\substack{\gamma \text{ having genus } g \\ \text{and } m \text{ 1-chords}}} \mathbf{D}_{\gamma,\sigma}(z). \tag{3.7}$$

According to Lemma 2, $\mathbf{D}_{\gamma,\sigma}(z)$ only depends on the number of chords and 1-chords of γ, and we can therefore express

$$\begin{aligned} \mathbf{D}_{g,\sigma}(z) &= \frac{1}{z-1}\, \mathbf{S}_g\left(\frac{z^{2g}}{(1-z^2)(1-z)^2 - (2z-z^2)z^{2\sigma}}, z \right) \\ &= \frac{1}{(1-z)+u_\sigma(z)z^2}\, \mathbf{C}_g\left(\frac{z^2\, u_\sigma(z)}{\big((1-z)+u_\sigma(z)z^2\big)^2} \right) \end{aligned}$$

using Lemma 1 in order to confirm eq. (3.5), where the second equality follows from direct computation. Let

$$\theta_\sigma(z) = \frac{z^2\, u_\sigma(z)}{\big((1-z)+u_\sigma(z)z^2\big)^2}$$

denote the argument of $\mathbf{C}_g$ in this expression.

Since any algebraic function is in particular D-finite as well as Δ-analytic [31], we conclude from Theorem 1 that

$$\mathbf{C}_g(z) = x_g\,(1-4z)^{-(3g-1/2)}(1+o(1)) \quad \text{for } z \to 1/4, \tag{3.8}$$

for some constant x_g. Since $\mathbf{C}_g(z)$ is algebraic over $K = \mathbb{C}(z)$, there exist polynomials $R_i(z)$, for $i = 1, \ldots, \ell$, such that $\sum_{i=1}^{\ell} R_i(z)\, \mathbf{C}_g(z)^i = 0$, whence $\sum_{i=1}^{\ell} R_i(\theta_\sigma(z))\, \mathbf{C}_g(\theta_\sigma(z))^i = 0$ as well. Setting $L = \mathbb{C}(\theta_\sigma(z))$, we thus have

$$[L(\mathbf{C}_g(\theta_\sigma(z))) \,:\, K] = [L(\mathbf{C}_g(\theta_\sigma(z))) \,:\, L] \cdot [L \,:\, K] < \infty,$$

i.e., $\mathbf{D}_{g,\sigma}(z)$ is algebraic over K. Pringsheim's Theorem [10] guarantees that for any $\sigma \geq 1$, $\mathbf{D}_{g,\sigma}(z)$ has a dominant real singularity $\gamma_\sigma > 0$.

In particular, for $\sigma = 2$, we verify directly that γ_2 is the unique solution of minimum modulus of $\theta_2(z) = 1/4$, which is strictly smaller than any other singularities of $\theta_2(z)$ and satisfies $\theta'(\gamma_2) \neq 0$. It follows that $\mathbf{D}_{g,2}(z)$ is governed by the supercritical paradigm [10], and hence $\mathbf{D}_{g,2}(z)$ has the singular expansion

$$\mathbf{D}_{g,2}(z) = k'_g\,(\gamma_2 - z)^{-(3g-1/2)}(1+o(1)) \quad \text{for } z \to \gamma_2, \tag{3.9}$$

for some constant k'_g.

For arbitrary but fixed g, we thus find the asymptotics

$$[z^n]\mathbf{D}_{g,2}(z) \sim k_g\, n^{3(g-1/2)}\, \gamma_2^n, \tag{3.10}$$

where $\gamma_2 \approx 1.9685$ as was claimed. □

4. RNA molecules and Riemann's moduli space

Lemma 1 implies that $\mathbf{S}_g(z,0)$ is the generating function for seeds of genus g with no 1-arcs. Since a shape is by definition simply such a seed together with a rainbow, the generating function $\mathbf{T}_g(z)$ for shapes of genus g satisfies $(1+z)\mathbf{T}_g(z) = z\mathbf{S}_g(z,0)$.

Proposition 1. *The generating function for shapes of genus g is the polynomial*

$$\mathbf{T}_g(z) = z(1+2z)^{6g-2}\, P_g\left(\frac{z(1+z)}{(1+2z)^2}\right). \tag{4.1}$$

In particular, a shape of genus g has at least $2g+1$ and at most $6g-1$ chords, so $\mathbf{T}_g(z)$ is a polynomial of degree $6g-1$ which is divisible by z^{2g+1}.

Proof. In view of Lemma 1 since $1 - 4\,\frac{z(1+z)}{(1+2z)^2} = \frac{1}{(1+2z)^2}$, we obtain

$$\mathbf{T}_g(z) \;=\; \frac{z}{1+2z}\,\mathbf{C}_g\left(\frac{z(1+z)}{(1+2z)^2}\right) \;=\; z(1+2z)^{6g-2}\, P_g\left(\frac{z(1+z)}{(1+2z)^2}\right).$$

□

Remark 1. The coefficient

$$[z^{2g+1}]\mathbf{T}_g(z) = \mathbf{c}_g(2g) = \frac{(4g)!}{4^g(2g+1)!}$$

is computed directly from the recursion eq. (2.1). Since $\lim_{z\to\infty} P_g\left(\frac{z(1+z)}{(1+2z)^2}\right) = P_g(1/4)$, the leading coefficient is given by $[z^{6g-1}]\mathbf{T}_g(z) = 2^{6g-2}\, P_g(1/4)$, where

$$P_g(1/4) = \left(\frac{9}{4}\right)^g \frac{\Gamma\left(g-1/6\right)\Gamma\left(g+1/2\right)\Gamma\left(g+1/6\right)}{6\pi^{3/2}\Gamma\left(g+1\right)}$$

can likewise be computed as the unique solution of another recursion

$$\begin{aligned} P_{g+1}(1/4) &= 4^{-4}(12g+6)(12g+2)(12g-2)P_g(1/4)/(3g+3) \\ &= \frac{9(g+1/2)(g+1/6)(g-1/6)}{4(g+1)}P_g(1/4), \end{aligned}$$

which follows with some work from eq. (2.1), with initial condition $P_1(1/4) = 1/16$.

For example, $P_1(z) = z^2$ gives

$$\mathbf{T}_1(z) \;=\; z^3(1+z^2) \;=\; z^3 + 2z^4 + z^5,$$

and $P_2(z) = 21z^4(1+z)$ gives

$$\begin{aligned}\mathbf{T}_g(z) &= 21z^5(1+z)^4\big[(1+2z)^2 + z(1+z)\big] \\ &= 21z^5 + 189z^6 + 651z^7 + 1134z^8 + 1071z^9 + 525z^{10} + 105z^{11}.\end{aligned}$$

Proposition 1 has a noteworthy implication for the folding of RNA structures of fixed genus, as follows.

Corollary 3. *RNA structures of fixed genus g can be computed in polynomial time.*

Remark 2. Eq. (2.1) as well as Corollary 2 provide evidence that the increase in time complexity passing from genus g to genus $g+1$ is $O(n^3)$. Clearly, since genus zero structures are RNA secondary structures which exhibit a time complexity of $O(n^3)$, we expect a $O(n^6)$ time complexity for folding genus one structures. Indeed in [26] a $O(n^6)$ time complexity folding of a certain class of "nested" genus 1-structures is presented.

Proposition 2. *For any $g \geq 1$, there is a bijection between RNA shapes of genus g and fatgraphs of genus g with a single boundary component each of whose vertices is of valence at least three except for a single vertex of valence one.*

Proof. Given a shape G_1, we may collapse its backbone in the natural way to produce a fatgraph G_2 with a single vertex as illustrated on the left in Figure 2. G_1 and G_2 have the same Euler characteristic, number of boundary components and hence genus. Notice that G_2 has a boundary cycle of length one arising from the rainbow of G_1, and this is its unique boundary cycle of length one since the shape G_1 can have no 1-chords. Furthermore, since a shape has no parallel chords, G_2 can have no boundary cycles of length two. It follows that other than its boundary cycle of length one coming from the rainbow, every other boundary cycle of G_2 must have length at least three. Notice that we may uniquely reconstruct the shape G_1 from the fatgraph G_2 by expanding its vertex to a backbone so that its unique boundary cycle of length one becomes a rainbow.

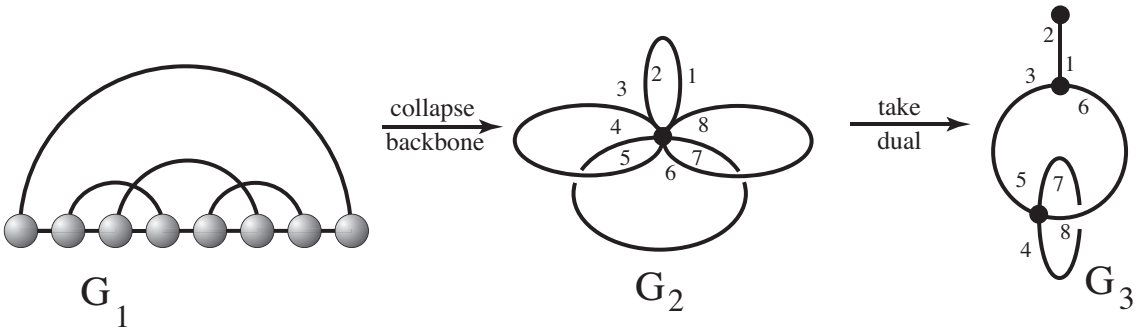

FIGURE 2. Collapse the backbone of the shape G_1 on the left to a vertex in order to produce the fatgraph G_2 in the middle with its labeled set of half-edges. Representing the permutation $i_1 \mapsto i_2 \mapsto \cdots \mapsto i_k \mapsto i_1$ as a cycle $(i_1, i_2, \ldots, i_k)$, G_2 is described by permutations $\sigma_2 = (1,2,3,4,5,6,7,8)$ and $\tau_2 = (1,2)(3,5)(4,7)(6,8)$. The dual fatgraph G_3 on the right is described by permutations $\sigma_3 = \sigma_2 \circ \tau_2 = (1,3,6)(2)(4,8,7,5)$ and $\tau_3 = \tau_2$. Notice that G_1 and G_2 have the same Euler characteristic -3, have 2 boundary components and have genus 1 . On the other hand, though G_2 and G_3 have the same genus, G_3 has only one boundary component (corresponding to the single vertex of G_2) and two vertices (corresponding to the two boundary components of G_2).

In general [22, 24], a fatgraph G with m edges may be described by a pair σ, τ of permutations on $2m$ objects identified with the half-edges of G, where σ is the composition of one disjoint k-cycle for each k-valent vertex of G corresponding to the cyclic orderings, and τ is the composition of m disjoint transpositions permuting the two half-edges contained in each edge. See Figure 2 for two examples. Furthermore in this representation, the boundary cycles of G correspond precisely to the cycles of the composition $\sigma \circ \tau$ as is also illustrated in Figure 2.

Suppose that G_2 is described in this manner by the pair σ_2, τ_2 of permutations, and let G_3 be the fatgraph corresponding to the pair $\sigma_3 = \sigma_2 \circ \tau_2$, $\tau_3 = \tau_2$. The boundary cycles of G_3 correspond to the vertices of G_2 and conversely. Letting v_i, e_i, r_i, g_i, respectively, denote the number of vertices, edges, boundary cycles and the genus of G_i, for $i = 2, 3$, we thus have $v_2 = r_3$, $v_3 = r_2$, and moreover $e_2 = e_3$ by construction, so we conclude $g_2 = g_3$. (In fact, G_2 and G_3 are related by duality in a closed surface of genus g.) In light of the constraints on G_2 already articulated since

it arises from the shape G_1, the fatgraph G_3 has all its vertices of valence at least three except for a unique vertex of valence one.

This provides a mapping from shapes to fatgraphs as asserted in the proposition. The inverse mapping is given by the same involution $\sigma \mapsto \sigma \circ \tau$, $\tau \mapsto \tau$ followed by expansion of the vertex to a backbone so that the cycle of length one becomes the rainbow. □

The collection of fatgraphs described in the previous proposition are precisely those arising in the Penner-Strebel cell decomposition of Riemann's moduli space [21, 32] for a surface of genus g with one boundary component. Furthermore, contraction of edges of fatgraphs corresponds to deletion of chords from shapes (amalgamating adjacent backbone edges incident on resulting isolated vertices so as to remain a shape), from which follows the striking consequence:

Corollary 4. *Riemann's moduli space of a surface of genus g with one boundary component is naturally homeomorphic to the geometric realization of set of all RNA shapes of genus g partially ordered by deletion of chords.*

One aspect of this insight is that the primary structure of an RNA molecule is compatible with only a certain collection of shapes that respect the Watson-Crick rules, and this in turn determines via the correspondence with fatgraphs a subspace of Riemann's moduli space that would have been otherwise inconceivably unmotivated. This stratification of moduli space by primary structure deserves further study and illustrates a sense in which the moduli space gives a suitable broad canvas for studying classes of RNA molecules in one space.

5. Discussion

Various filtrations of pseudoknot RNA structures have been suggested.

Haslinger and Stadler's bisecondary structures [14] are diagrams that can be written as pairs of secondary structures, one in the upper and one in the lower halfplane. Despite their simple definition, bisecondary structures turned out to be very difficult to analyze, and no generating function for them is known.

For the more general class of k-noncrossing RNA structures, i.e., diagrams in which there are no k mutually crossing arcs, explicit generating functions and simple asymptotic formulas for their coefficients have been obtained [27]. However, though their generating functions are D-finite and their numbers satisfy recursions with polynomial coefficients, for any odd k, logarithmic terms appear in the singular expansion. In particular for $k = 3$, they "almost" coincide with bisecondary structures in the sense that the corresponding exponential growth rates are very close. However, in contrast to bisecondary structures, 3-noncrossing structures are not necessarily planar. One prominent feature of k-noncrossing structures is that their exponential growth rate is an unbounded function of k, and the complexity of the crossings is manifest both in the exponential growth rate and in the subexponential factors.

The genus filtration discussed here[3] is of distinctively different nature. Additivity under prolongation of backbone assures that the exponential growth rate remains constant and identical to that of RNA secondary structures. Thus, higher genus effects only materialize in the subexponential factor, and Theorem 2 shows that this factor increases by $O(n^3)$ for each increase in genus. Furthermore, the generating function of σ-canonical structures of genus g is not only D-finite but also algebraic and therefore much simpler than that of k-noncrossing structures.

RNA structures of any genus $g \geq 1$ are completely determined by a *finite* set of shapes, obtained via collapsing all stacks into single arcs and removing all unpaired nucleotides. These operations evidently preserve genus, and any genus g structure can be obtained by inflating shapes into stems and inserting segments of isolated vertices. Thus, shapes are the key to folding topological structures. In [27], minimum free energy γ_1-structures, obtained by nesting and concatenating genus one shapes are folded, and their partition function is furthermore computed. The advantage of these topological structures over a common penalty for each crossing of gap-matrices [30] is that the "topology based" grammar naturally distinguishes different types of pseudoknots and admits different energy parameters for them. This additional freedom of parametrization leads to a substantial increase of sensitivity [27].

An intriguing observation derives from the relation between the shape polynomial $\mathbf{T}_g(z)$ and the polynomial $P_g(z)$. As explicated in Conjecture 1, Proposition 1 suggests that $P_g(z)$ is the generating polynomial for some subset (or quotient space) of shapes from which all shapes can

[3]There is physics literature initiated in [20] on RNA enumeration based on matrix models which relies on the genus of linear chord diagrams and provides a comparison of expected with observed genera [6, 34, 35].

be derived by some as-yet unknown process of inflation. If so, then this constitutes a significant enumerative compression to the g non-zero coefficients of $P_g(z)$ which hopefully can be utilized for the fast folding of pseudoknot structures.

References

1. J. E. Andersen, A. J. Bene, J.-B. Meilhan, R. C. Penner, "Finite type invariants and fatgraphs", *Adv. Math.* **225** (2010), 2117-2161.
2. J. E. Andersen, J. Mattes, N. Reshetikhin, "The Poisson Structure on the Moduli Space of Flat Connections and Chord Diagrams", *Topology* **35** (1996), 10691083.
3. —, "Quantization of the Algebra of Chord Diagrams," *Math. Proc. Camb. Phil. Soc.* **124** (1998), 451-467.
4. D. Bar-Natan, "On the Vassiliev knot invariants", *Topology* **34** (1995), 423-475.
5. —, "Lie algebras and the four colour problem", *Combinatorica* **17** (1997), 43-52.
6. M. Bon, G. Vernizzi, H. Orland, A. Zee, "Topological classification of RNA structures", *J. Mol. Biol.* **379** (2008), 900-911.
7. R. Campoamor-Stursberg and V. O. Manturov, "Invariant tensor formulas via chord diagrams", *J. Math. Sci.* **108** (2004), 3018-3029.
8. P. Flajolet, "Combinatorial aspects of continued fractions", *Discrete Math.* **32** (1980), 125-161.
9. P. Flajolet, J. Francon, J. Vuillemin, "Sequence of operations analysis for dynamic data structures", *J. Algorithms* **1** (1980), 111-141.
10. P. Flajolet and R. Sedgewick, Analytic Combinatorics, Cambridge University Press, Cambridge (2009).
11. P. Goulden and A. Nica, "A direct bijection for the Harer-Zagier formula", *J. Combinatorial Theory (A)* 111, 2005, 224-238.
12. A. Goupil and G. Schaeffer, " Factoring n-cycles and counting maps of given genus", *European J. Combin.*, **19**(7):819-834, 1998.
13. J. Harer and D. Zagier, "The Euler characteristic of the moduli space of curves", *Invent. Math.* **85** (1986), 457-485.
14. C. Haslinger and P.F. Stadler, "RNA structures with pseudo-knots", *Bull. Math. Biol.*, 61:437–467, 1999.
15. M. Kimura, The Neutral Theory of Molecular Evolution, Cambridge University Press, Cambridge (1983).
16. D.A.M. Konings and R.R. Gutell, "A comparison of thermodynamic foldings with comparatively derived structures of 16s and 16s-like rRNAs", *RNA*, 1:559–574, 1995.
17. M. Kontsevich, "Vassiliev's Knot Invariants." *Adv. Soviet Math.* **16** (1993), 137-150.
18. A. Loria and T. Pan, "Domain structure of the ribozyme from eubacterial ribonuclease", *RNA*, 2:551–563, 1996.
19. R. J. Milgram and R. C. Penner, "Riemann's moduli space and the symmetric groups", *Mapping class groups and moduli spaces of Riemann surfaces*, eds. C.-F. Bdigheimer and R. M. Hain, AMS Contemporary Math **150** (1993), 247-290.
20. H. Orland and A. Zee, "RNA folding and large N matrix theory", *Nucl. Phys. B* **620** (2002), 456-476.
21. R. C. Penner, "The Teichmuller space of a punctured surface", *Comm. Math. Phys.* (1987).
22. R. C. Penner, "Perturbative series and the moduli space of Riemann surfaces", *J. Diff. Geom.* **27** (1988), 35-53.
23. —, "Weil-Petersson volumes", *J. Diff. Geom.* **35** (1992), 559-608.

24. R. C. Penner, M. Knudsen, C. Wiuf, J. Andersen, "Fatgraph model of proteins", *Comm. Pure Appl. Math.* **63** (2010), 1249-1297.
25. R. C. Penner and M. S. Waterman, "Spaces of RNA secondary structures", *Adv. Math.* **101** (1993), 31-49.
26. C.M. Reidys, F.W.D. Huang, J.E. Andersen, R.C. Penner, P.F. Stadler and M.E. Nebel, "Topology and prediction of RNA pseudoknots", Bioinformatics (2011) doi: 10.1093/bioinformatics/btr090
27. C. M. Reidys, Combinatorial Computational Biology of RNA, Springer-Verlag, New York (2011).
28. C. M. Reidys, P. F. Stadler, P. K. Schuster, "Generic properties of combinatory maps and neutral networks of RNA secondary structrures", *Bull. Math. Biol.* **59** (1997), 339-397.
29. C. M. Reidys and P. F. Stadler, "Combinatorial landscapes", *SIAM Rev.* **44** (2002), 3-54.
30. Rivas, E. and Eddy, S. R. (1999), "A dynamic programming algorithm for RNA structure prediction including pseudoknots", *J. Mol. Biol.*, **285**, 2053–2068.
31. R. P. Stanley, "Enumerative Combinatorics", Cambridge Studies in Advanced Mathematics **49**, Cambridge University Press, Cambridge (1997).
32. K. Strebel. "Quadratic Differentials". Springer-Verlag, 1984.
33. D.W. Staple and S.E. Butcher, "Pseudoknots: RNA structures with diverse functions", *PLoS Biol. 3*, 6:956–959, 2005
34. G. Vernizzi, H. Orland, A. Zee, "Enumeration of RNA structures by matrix models", *Phys. Rev. Lett.* **94** (2005), 168103.
35. G. Vernizzi, P. Ribecca, H. Orland, A. Zee, "Topology of pseudoknotted homopolymers", *Phys. Rev. E* **73** (2006), 031902.
36. M. Waterman, Combinatorics of RNA hairpins and cloverleafs, Stud. Appl. Math. 60 (1979) 91–96.
37. M. Waterman, Secondary structure of single-stranded nucleic acids, Adv. Math. (Suppl. Studies) 1 (1978) 167–212.
38. J. Howell, T. Smith, M. Waterman, Computation of generating functions for biological molecules, SIAM J. Appl. Math. 39 (1980) 119–133.
39. M. Waterman, W. Schmitt, Linear trees and RNA secondary structure, Discr. Appl. Math. 51 (1994) 317–323.
40. M. S. Waterman, An Introduction Computational Biology, Chapman and Hall, New York (1995).
41. E. Westhof and L. Jaeger, "RNA pseudoknots", *Curr. Opin. Chem. Biol.*, 2:327–333, 1992.

Center for the Quantum Geometry of Moduli Spaces, Aarhus University, DK-8000 Aarhus C, Denmark

E-mail address: andersen@imf.au.dk

Center for the Quantum Geometry of Moduli Spaces, Aarhus University, DK-8000 Aarhus C, Denmark, and Departments of Math and Physics, Caltech, Pasadena, CA 91125 USA

E-mail address: rpenner@imf.au.dk

Institut for Mathematics and Computer Science, University of Southern Denmark, DK-5230 Odense, Denmark

E-mail address: duck@santafe.edu

Departments of Biological Sciences, Mathematics, Computer Science, University of Southern California, Los Angeles, CA 90089, USA

E-mail address: msw@usc.edu