J. Math. Biol. (2013) 67:1261–1278
DOI 10.1007/s00285-012-0594-x

Mathematical Biology

Topological classification and enumeration of RNA structures by genus

J. E. Andersen · R. C. Penner · C. M. Reidys · M. S. Waterman

Received: 21 May 2011 / Revised: 27 June 2012 / Published online: 2 October 2012

Abstract To an RNA pseudoknot structure is naturally associated a topological surface, which has its associated genus, and structures can thus be classified by the 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

J.E. Andersen and R.C. Penner are supported by QGM (Centre for Quantum Geometry of Moduli Spaces, funded by the Danish National Research Foundation).

J. E. Andersen · R. C. Penner
Center for the Quantum Geometry of Moduli Spaces, Aarhus University,
8000 Aarhus C, Denmark
e-mail: andersen@imf.au.dk

R. C. Penner
Departments of Math and Physics, Caltech, Pasadena, CA 91125, USA
e-mail: rpenner@imf.au.dk

C. M. Reidys (✉)
Institute for Mathematics and Computer Science, University of Southern Denmark,
5230 Odense, Denmark
e-mail: duck@santafe.edu

M. S. Waterman
Departments of Biological Sciences, Mathematics, Computer Science,
University of Southern California, Los Angeles, CA 90089, USA
e-mail: msw@usc.edu

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.

1 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 pairings 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 (Waterman and Schmitt 1994; Penner and Waterman 1993; Waterman 1979, 1978; Howell et al. 1980).

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 (Westhof and Jaeger 1992). These structures are called pseudoknot structures, and their topology has been studied in Vernizzi et al. (2005, 2006), Bon et al. (2008), Orland and Zee (2002). 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 (Loria and Pan 1996), telomerase RNA (Staple and Butcher 2005) and ribosomal RNAs (Konings and Gutell 1995).

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.

[1] These combinatorial structures occur in a number of instances in pure mathematics including finite type invariants of knots and links (Bar-Natan 1995; Kontsevich 1993), the representation theory of Lie algebras (Campoamor-Stursberg and Manturov 2004), the geometry of moduli spaces of flat connections on surfaces (Andersen et al. 1996, 1998), mapping class groups (Andersen et al. 2010) and the Four-Color Theorem (Bar-Natan 1997), and in applied mathematics including codifying the pairings among nucleotides in RNA molecules (Reidys 2011; Bon et al. 2008; Orland and Zee 2002), or more generally the contacts of any

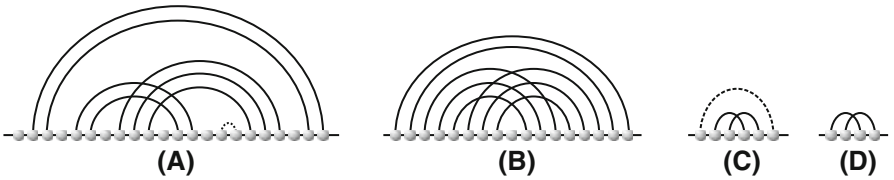

Fig. 1 The different diagram types: partial chord diagram with eight chords (**a**), chord diagram (**b**), seed (**c**) and shape (**d**). **a** contains a 1-arc (*dashed*) and (**c**) contains a rainbow (*dashed*). Note that (**b**) is σ-structure where $\sigma \geq 2$

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*. The class of "shapes" is quite similar to classes of diagrams discussed in Orland and Zee (2002), Vernizzi et al. (2005), Bon et al. (2008) called "irreducible diagrams". Please see Fig. 1 for examples of these notions.

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* (Penner 1987, 1988, 1992) 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 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. Fig. 2.

Let r denote the number of distinct boundary cycles of the connected fatgraph G with v vertices and e edges. The Euler characteristic (see Euler 1752) 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 Fig. 2, 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 Penner et al. 2010) 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

$$2 - 2g - r = 1 - m, \tag{1.1}$$

Footnote 1 continued
binary macromolecule (Penner et al. 2010; Penner and Waterman 1993; Waterman 1995), and in the analysis of data structures (Flajolet 1980; Flajolet et al. 1980).

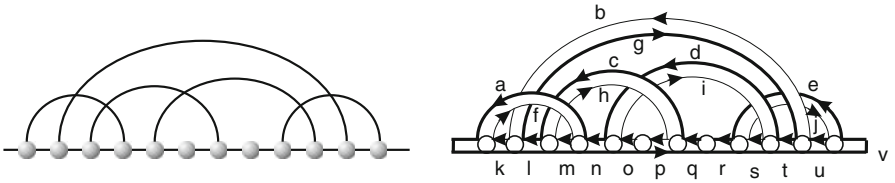

Fig. 2 Computing the number of boundary components of partial chord diagram. The diagram contains $5+11$ edges and 12 vertices. We follow the cycles described in the text and observe that there are exactly two boundary cycles (*bold and thin*). The genus of the diagram is given by $1-\frac{1}{2}(12-16+2)=2$. Similar figures occurred in this context in Penner (2004) and Vernizzi et al. (2005)

between the genus g, the number r of boundary cycles, and the number m of chords. Similar formulae can also be found in this context in Bon et al. (2008).

We remark that there are no shapes of genus zero, since a non-empty genus zero linear chord diagram must have a 1-chord.

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 $\mathbf{T}_g(z)=\sum_{n\geq 0}\mathbf{t}_g(n)z^n$, setting a standard use of fonts,which we will adopt through out the paper.

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

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

which is surjective and preserves genus.

Furthermore, ϑ restricts to a surjection

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

that 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 static subset of the fiber $\vartheta^{-1}(\gamma)$ with its generating function $\mathbf{C}^{\gamma}(x, y)$.

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.

Our choice of excluding only 1-chord diagrams is a technical one. None of the results change significantly when increasing the minimum arc-length to three. The derivation of the generating functions however becomes very tedious. See for example a similar analysis performed for the different generating function of k-noncrossing structures with minimum arc-length 4 in Reidys et al. (2010). An analysis analogous to that of Reidys et al. (2010) for topological RNA structures shows that only the exponential growth rate changes marginally, and the subexponential factor is unaffected.

Let $\mathscr{D}_{\sigma}(n)$ be the set of RNA σ-structures on n vertices and $\mathscr{D}_{g,\sigma}(n)$ the subset consisting of such structures of genus g.

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

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

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

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

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

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 Harer and Zagier (1986) 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. Sect. 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

[2] The numbers $\mathbf{c}_g(n)$ had been computed in another generating function over two decades ago by Harer and Zagier (1986) 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.

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

$$\mathbf{d}_{g,2}(n) \sim k_g \, n^{3(g-\frac{1}{2})} \gamma_2^n, \tag{1.2}$$

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 (Kimura 1983; Reidys et al. 1997; Reidys and Stadler 2002). Neutral networks appear in the context of analysing RNA evolution, and can be related to the quasi-species model.

The key observation is that there are many more sequences than structures, implying the existence of exponentially large sets of sequences all of which fold into a fixed RNA structure. Since neutral networks can be modeled as random graphs (Reidys et al. 1997) and random subsets of generalized n-cubes form typically large connected subgraphs (Reidys 2009; Jin and Reidys 2011), a sequence-to-structure mapping is likely to induce "interesting" networks by its preimages if it is a many-to-one mapping. The fact that the exponential growth rate of topological RNA structures is, independent of genus, much smaller than four implies the existence of preimages of fixed structures of exponential size. In fact, closer inspection shows that there exist exponentially many of such large preimages. Neutral networks are a generic phenomenon for maps from sequences to minimum free energy structures (Grüner et al. 1996a,b). The vast extended networks predicted by random graph theory also exist for pseudoknotted RNA structures explaining the success of inverse folding algorithms (Gao et al. 2010).

Evidently neutral networks play a key role for evolutionary optimization: a sequence population can neutrally evolve on such a network since in the presence of neutral neighbors mutants have a non-zero probability of being one of the neutral neighbors. Neutral networks allow us to lift important concepts such as that of the molecular quasispecies to the level of structural phenotypes (Reidys et al. 2001).

It is interesting to compare the formula (1.2) to formula (5) in Garg and Deo (2009), where the two right hand sides become identical, if we substitute n for L and γ_2 for $3-\alpha$ in (5) of Garg and Deo (2009). While the asymptotics in Garg and Deo (2009) is purely based on numerical studies, we give a mathematically rigorous proof of our asymptotics. It is however not clear how the right hand sides are related, although they do involve the same kinds of diagrams, namely partial chord diagrams. In Garg and Deo (2009), $1-\alpha$ is given the interpretation as a weight factor on each unpaired base, however the considered range is $0 \leq \alpha \leq 1$. Since our γ_2 is just below 2, this would produce an α just above 1 via the substitution discussed above. This would suggest the conjecture that the matrix model studied in Garg and Deo (2009) for $\alpha = 3-\gamma_2$ could be related to our generating function $d_{g,2}(n)$.

We want to stress that this paper does not purport to contribute to the mathematical theory of one-face maps. It is intended to connect that theory to problems in mathematical and computational biology, in particular the folding of RNA pseudoknot structures.

The application to canonical RNA pseudoknot structures and their exponential growth rates are new and of biological importance.

In addition, we wish to emphasize that Riemann's moduli space of a surface of genus g with one boundary component is naturally homeomorphic to the geometric realization of 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 (Penner 1987, 1988; Strebel 1984).

Various filtrations of pseudoknot RNA structures have been suggested. Haslinger and Stadler's bisecondary structures (Haslinger and Stadler 1999) 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 chords, explicit generating functions and simple asymptotic formulas for their coefficients have been obtained (Reidys 2011). 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 was initiated in Orland and Zee (2002) (see also Penner 2004). RNA enumeration methods based on matrix models which rely on the genus filtration of linear chord diagrams appeared in Bon et al. (2008), Vernizzi et al. (2005, 2006) and provide a comparison of expected with observed genera.

The enumeration problem for all partial chord diagrams was studied in Vernizzi et al. (2005) via matrix model techniques and the asymptotics given in formula (19) in that paper. A closed formula expression for the number of partial chord diagrams was given in dell'Erba and Zemba (2009) in terms of Stirling numbers of the first kind.

Additivity of genus 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 chords 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 Reidys (2011), minimum free energy γ_1-structures, obtained by nesting and concatenating genus one shapes are folded, and their partition function is furthermore computed. See also Pillsbury et al. (2005), where a genus-related algorithm is presented. The advantage of these topological structures over a common penalty for each crossing of gap-matrices

(Rivas and Eddy 1999) 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 (Reidys 2011).

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

A seminal result due to Harer and Zagier (1986), cf. also Goulden and Nica (2005), Goupil and Schaeffer (1998), 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 (Harer and Zagier 1986) *The* $\mathbf{c}_g(n)$ *satisfy the recursion*

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

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

The recursion Eq. (2.1) translates into the ODE

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

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

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

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[3] $[z^h]Q_g(z) = 0$ 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 Harer and Zagier (1986), but explicitly stated in Lando and Zvonkin (2004).

[3] 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.

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*

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

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 set of classes of shapes.*

We remark that $[z^{2g}]P_g(z)$ indeed is the number of shapes of genus g with $2g$ chords. For the g's for which we have calculated P_g, we observe that the coefficients are positive integers. In Corollary 1 and the remark just above it, it is suggested there that $P_g(z)$ is the generating polynomial for some set of classes of shapes from which all shapes can 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.

A straightforward analysis (Flajolet and Sedgewick 2009) of the singularity of $\mathbf{C}_g(z)$ then gives the following well known corollary, which was first obtained in Bender et al. (1988) (where the exact value of $P_g(\frac{1}{4})$ is also given):

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

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

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*

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

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

Proof We shall construct $\sqcup_{n\geq 0}\mathcal{D}_{\sigma}^{\gamma}(n)$ with simple combinatorial building blocks. 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 chord 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 $\sigma \geq 1$ chords 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

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

□

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*

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

For arbitrary but fixed g, we have

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

for some constant $k_{g,\sigma} > 0$ depending only on g and σ. In case of $\sigma = 2$, i.e., canonical RNA structures we have $\gamma_2 \approx 1.9685$.

Remark 1 The σ-dependence of our asymptotics lies exclusively in shifting the dominant singularity γ_σ, and we have verified that for $1 \leq \sigma \leq 10$, γ_σ is unique. In particular, Eq. (3.6) shows that the subexponential factor is independent of σ. We emphasize the case $\sigma = 2$ since the corresponding structures are canonical, i.e., they exhibit no isolated arcs. This minimum stack-size implies energetically favorable structures that are here of particular relevance.

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

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

According to Lemma 2, $\mathbf{D}_{\sigma}^{\gamma}(z)$ 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 (Stanley 1997), 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 (Flajolet and Sedgewick 2009) 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 (Flajolet and Sedgewick 2009), 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-chords. 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*

$$\mathbf{T}_g(z) = z(1+2z)^{6g-2}\, P_g\left(\frac{z(1+z)}{(1+2z)^2}\right) = \sum_{j=2g}^{3g-1} p_g^{(j)} z^{j+1}(1+z)^j (1+2z)^{2(3g-1-j)},$$

where

$$P_g(z) = \sum_{j=2g}^{3g-1} p_g^{(j)} z^j.$$

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 2 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

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

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

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

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

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

Remark 3 Equation (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 Reidys et al. (2011), a $O(n^6)$ time complexity folding of a certain class of "nested" genus 1-structures is presented. Vernizzi et al. (2005) reports recursion relations of time complexity $O(n^6)$ to generate RNA structures *of genus one* in the context of an RNA folding algorithm that is substantially different from the algorithm implied by our results. Vernizzi et al. (2005) does not consider loop-based energy models and is restricted to genus one RNA structures. Our results imply a loop-based $O(n^6)$ folding of structures that are "locally" genus restricted but have in general unbounded topological genus (Li and Reidys 2012). This algorithm employs the concept of γ-structures detailed in Reidys et al. (2011).

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 Fig. 3. 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

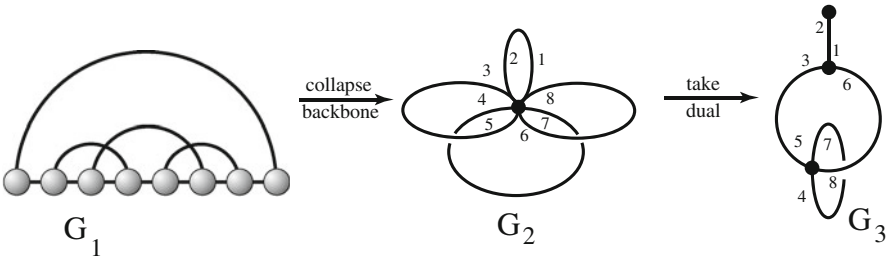

Fig. 3 Collapse the backbone of the shape G_1 on the *left* to a vertex in order to produce the fatgraph G_2 in the *middle* with its labeled set of half-edges. Representing the permutation $i_1 \mapsto i_2 \mapsto \cdots \mapsto i_k \mapsto i_1$ as a cycle $(i_1, i_2, \ldots, i_k)$, G_2 is described 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)

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.

In general (Penner 1988; Penner et al. 2010), 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 Fig. 3 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 Fig. 3.

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 (Penner 1987; Strebel 1984) 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 it follows that 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.

References

Andersen JE, Bene AJ, Meilhan J-B, Penner RC (2010) Finite type invariants and fatgraphs. Adv Math 225:2117–2161

Andersen JE, Mattes J, Reshetikhin N (1996) The poisson structure on the moduli space of flat connections and chord diagrams. Topology 35:1069–1083

Andersen JE, Mattes J, Reshetikhin N (1998) Quantization of the algebra of chord diagrams. Math Proc Camb Phil Soc 124:451–467

Bar-Natan D (1995) On the Vassiliev knot invariants. Topology 34:423–475

Bar-Natan D (1997) Lie algebras and the four colour problem. Combinatorica 17:43–52

Bender EA, Rodney Canfield E (1988) The asymptotic number of tree-rooted maps on a surface. J Comb Theory Ser A 48(2):156–164

Bon M, Vernizzi G, Orland H, Zee A (2008) Topological classification of RNA structures. J Mol Biol 379:900–911

Campoamor-Stursberg R, Manturov VO (2004) Invariant tensor formulas via chord diagrams. J Math Sci 108:3018–3029

dell'Erba MG, Zemba GR (2009) Thermodynamics of a model for RNA folding. Phys Rev E 79:011913

Euler L (1752) Elementa doctrinae solidorum. Novi Comm Acad Sci Imp Petropol 4:109–140

Flajolet P (1980) Combinatorial aspects of continued fractions. Discret Math 32:125–161

Flajolet P, Francon J, Vuillemin J (1980) Sequence of operations analysis for dynamic data structures. J Algorithms 1:111–141

Flajolet P, Sedgewick R (2009) Analytical combinatorics. Cambridge University Press, Cambridge

Gao JZM, Li LYM, Reidys CM (2010) Inverse folding of RNA pseudoknot structures. Algorithms Mol Biol 5:R27

Garg I, Deo N (2009) RNA matrix models with external interactions and their asymptotic behavior. Phys Rev E 79:061903

Goulden P, Nica A (2005) A direct bijection for the Harer–Zagier formula. J Comb Theory (A) 111:224–238

Goupil A, Schaeffer G (1998) Factoring n-cycles and counting maps of given genus. Eur J Comb 19(7): 819–834

Grüner WG, Strothmann D, Reidys CM, Weber J, Hofacker IL, Stadler PF, Schuster P (1996) Analysis of RNA sequence structure maps by exhaustive enumeration II. Neutral Netw Chem Mon 127:375–389

Grüner WG, Strothmann D, Reidys CM, Weber J, Hofacker IL, Stadler PF, Schuster P (1996) Analysis of RNA sequence structure maps by exhaustive enumeration I. Neutral Netw Chem Mon 127:355–374

Harer J, Zagier D (1986) The Euler characteristic of the moduli space of curves. Invent Math 85:457–485

Haslinger C, Stadler PF (1999) RNA structures with pseudo-knots. Bull Math Biol 61:437–467

Jin EY, Reidys CM (2011) Random induced subgraphs of Cayley graphs induced by transpositions. Discret Math 21(311):2496–2511

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Konings DAM, Gutell RR (1995) A comparison of thermodynamic foldings with comparatively derived structures of 16s and 16s-like r RNAs. RNA 1:559–574

Kontsevich M (1993) Vassiliev's knot invariants. Adv Sov Math 16:137–150

Lando SK, Zvonkin AK (2004) Graphs on surfaces and their applications: with an appendix by Don B. Zagier. Encyclopaedia of Mathematical Sciences, 141. Low-Dimensional Topology, II. Springer-Verlag, Berlin

Li TJX, Reidys CM (2012) The genus filtration of γ-structures. Math Biosci (submitted)

Loria A, Pan T (1996) Domain structure of the ribozyme from eubacterial ribonuclease. RNA 2:551–563

Milgram RJ, Penner RC (1993) Riemann's moduli space and the symmetric groups. In: Bödigheimer C-F, Hain RM (eds) Mapping class groups and moduli spaces of Riemann surfaces. AMS contemporary math, vol 150. pp 247–290

Orland H, Zee A (2002) RNA folding and large N matrix theory. Nucl Phys B 620:456–476

Penner RC (1987) The Teichmuller space of a punctured surface. Commun Math Phys

Penner RC (1988) Perturbative series and the moduli space of Riemann surfaces. J Diff Geom 27:35–53

Penner RC (1992) Weil–Petersson volumes. J Diff Geom 35:559–608

Penner RC (2004) Cell decomposition and compactification of Riemann's moduli space in decorated Teichmüller theory. In: Tongring N, Penner RC (eds) Woods hole mathematics-perspectives in math and physics. World Scientific, Singapore, pp 263–301 (arXiv)

Penner RC, Knudsen M, Wiuf C, Andersen J (2010) Fatgraph model of proteins. Comm Pure Appl Math 63:1249–1297

Penner RC, Waterman MS (1993) Spaces of RNA secondary structures. Adv Math 101:31–49

Pillsbury M, Orland H, Zee A (2005) Steepest descent calculation of RNA pseudoknots. Phys Rev E 72:011911

Pillsbury M, Taylor JA, Orland H, Zee A (2005) An algorithm for RNA pseudoknots. arXiv: cond-mat/0310505v2

Reidys CM, Huang FWD, Andersen JE, Penner RC, Stadler PF, Nebel ME (2011) Topology and prediction of RNA pseudoknots. Bioinformatics. doi:10.1093/bioinformatics/btr090

Reidys CM (2011) Combinatorial computational biology of RNA. Springer, New York

Reidys CM, Wang RR, Zhao AYY (2010) Modular, k-noncrossing diagrams. Electr J Comb 1(17):R76

Reidys CM, Stadler PF, Schuster PK (1997) Generic properties of combinatory maps and neutral networks of RNA secondary structrures. Bull Math Biol 59:339–397

Reidys CM, Stadler PF (2002) Combinatorial landscapes. SIAM Rev 44:3–54

Rivas E, Eddy SR (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. J Mol Biol 285:2053–2068

Reidys CM, Forst CV, Schuster P (2001) Replication and mutation on neutral networks. Bull Math Biol 63:57–94

Reidys CM (2009) Large Components of Random induced subgraphs of n-cubes. Discret Math 309:3113–3124

Stanley RP (1997) Enumerative combinatorics. Cambridge studies in advanced mathematics, vol 49. Cambridge University Press, Cambridge

Strebel K (1984) Quadratic differentials. Springer, Berlin

Staple DW, Butcher SE (2005) Pseudoknots: RNA structures with diverse functions. PLoS Biol 3(6): 956–959

Vernizzi G, Orland H, Zee A (2005) Enumeration of RNA structures by matrix models. Phys Rev Lett 94:168103

Vernizzi G, Ribecca P, Orland H, Zee A (2006) Topology of pseudoknotted homopolymers. Phys Rev E 73:031902

Waterman M (1979) Combinatorics of RNA hairpins and cloverleafs. Stud Appl Math 60:91–96

Waterman M (1978) Secondary structure of single-stranded nucleic acids. Adv Math (Suppl Stud) 1:167–212

Howell J, Smith T, Waterman M (1980) Computation of generating functions for biological molecules. SIAM J Appl Math 39:119–133

Waterman M, Schmitt W (1994) Linear trees and RNA secondary structure. Discret Appl Math 51:317–323

Waterman MS (1995) An introduction computational biology. Chapman and Hall, New York

Westhof E, Jaeger L (1992) RNA pseudoknots. Curr Opin Chem Biol 2:327–333