MODULI SPACES AND MACROMOLECULES

R. C. PENNER

Abstract. Techniques from moduli spaces are applied to biological macromolecules. The first main result provides new *a priori* constraints on protein geometry discovered empirically and confirmed computationally. The second main result identifies up to homotopy the natural moduli space of several interacting RNA molecules with the Riemann moduli space of a surface with several boundary components in each fixed genus. Applications to RNA folding prediction are discussed. The mathematical and biological frameworks are surveyed and presented from first principles.

Contents

Introduction 2
1. Moduli Spaces 6
1.1. Conformal quadrilaterals 7
1.2. Triangles in neutral geometries 8
1.3. Elliptic curves 10
1.4. Riemann moduli space 13
1.5. Fatgraphs 16
1.6. Flat G-connections 22
2. Protein 24
2.1. Chemistry and geometry 24
2.2. Protein folding 26
2.3. Fatgraph model 28
2.4. $SO(3)$-graph connections 30
2.5. Further remarks on protein 35
3. Sugar 37
4. RNA 39
4.1. Chemistry and topology 40
4.2. RNA Folding 43
4.3. Chord diagrams, seeds and shapes 47
4.4. Further remarks on RNA 51
Appendix. Matrix models 56
References 59

AMS Subject Classification: 92-02, 92C40, 92C05, 30F60, 32G15, 53C05.
Keywords: macromolecule, protein, RNA, surface, fatgraph, graph connection.

Bill Goldman, Piotr Sułkowski and especially Fenix Huang, Nadya Morozova and Scott Wolpert provided helpful comments and critical readings. Further thanks are due to Ebbe Andersen for assistance with the figures and for many useful comments.

Introduction

This paper surveys recent progress in applying techniques from moduli spaces in mathematics and physics to study the geometry and topology of three families of macromolecules of interest in biology, namely, RNAs, proteins and polysaccharides. Our intended audience includes mathematicians whose expertise does not necessarily extend to either moduli spaces or macromolecules and who may be interested in this nascent application of geometry to problems in biology. In such application, it is imperative to stay humbly focused on the relevant biology and avoid the temptation to invent mathematics for its own sake, which is a perfectly legitimate but entirely different undertaking.

These techniques in essence capture the combinatorics of interacting families of one-dimensional objects and have already proven their utility in theoretical physics including string theory, where the one-dimensional strings occur at the Planck scale. Here we apply these same combinatorial tools to macromolecules that occur at the scale of hundreds of Ångström, larger by some 25 orders of magnitude. A childishly simple but nevertheless profound remark is that combinatorics is insensitive to scale, so it is only natural that these techniques from high energy physics for studying interactions of one-dimensional strings should apply to macromolecules as well. In fact, it is already a slippery slope of dimension as the foundations of string theory were originally developed before the emergence of quantum chromodynamics in order to describe hadrons such as neutrons, some 5 orders of magnitude smaller than macromolecules, and then dropped overnight [123] to gravitational strings partly justified by precisely the same invariance under scaling, albeit wildly decreasing there and increasing here.

A *moduli space* in its philosophical or linguistic sense is the *variety of all possibilities* whose mathematical explication evidently requires further definition. Two important examples are the moduli spaces of Riemann surfaces and the closely related moduli spaces of flat connections on principal G-bundles over surfaces for some Lie group G. Among other more elementary examples to set the stage, these two families of moduli spaces are surveyed in the next section. The important point here is that each of these two families admits an elementary combinatorial formulation which can then be applied to macromolecules.

Specifically with $G = SO(3)$, the group of rigid rotations of 3-space $\mathbb{R}^3$, the combinatorial model of flat G-connections allows us to probe the geometry of proteins, namely as explained in detail later, the geometry of hydrogen bonds among peptide units in a protein. Our main result on proteins is the entirely empirical finding that these rotations

cluster into only about thirty percent of the volume of $SO(3)$, and moreover within this region there is a further aggregation into 30 sub-regions or clusters. This gives a new classification for the geometry of hydrogen bonding that unifies and extends those already known. Consequences of these new *a priori* constraints in protein science are discussed later. There is furthermore a numerical simulation using so-called Density Functional Theory for the quantum system of two peptide units which partly reproduces this empirical finding as will also be discussed.

It is not the geometry but rather the topology we describe for RNA. In fact, there is a natural decomposition of the Riemann moduli space for a surface F whose cells are in one-to-one correspondence with homotopy classes of appropriate graphs embedded in F. There is moreover a natural combinatorial model based on chord diagrams for the moduli space of r interacting RNA molecules which have genus g in a suitable sense. The striking theorem is that the Riemann moduli space of a surface F of genus g with r boundary components is combinatorially isomorphic with this RNA moduli space up to homotopy. The proof of this remarkable isomorphism is perhaps a let down since it is a purely combinatorial identification, however, the depth of structure of the Riemann moduli space, which plays central roles throughout mathematics and physics, carries over to the RNA setting since much of this structure can in fact be described combinatorially though with unclear significance for RNA. Moreover, there are tools in quantum field theory called matrix models which have been successfully employed in mathematics and physics for explicit computations involving this combinatorial version of Riemann moduli space. Matrix models can also therefore be profitably applied to combinatorial aspects of RNA.

By now there is a small group of mathematicians, physicists, bioinformaticians and biologists employing these methods to analyze RNA and protein. We can only hope that this survey paper will attract still others since there is very much more that can be done. For example for polysaccharide, also called sugar or carbohydrate, there is a clear applicability of both the geometric and topological techniques as we shall very briefly explain, but a serious biological or biophysical application of these methods has yet to be undertaken.

Gromov asks the question in [52] "Is there mathematics in biology?" and then goes on to give affirmative examples as we believe also are the studies here. There are furthermore entire fields of combinatorics and computer science [9, 132] dedicated to problems related to so-called sequence or multiple sequence alignment, which entails the effective comparison of two or more arbitrary words in a fixed alphabet of letters, in practice 20 amino acid letters for protein and 4 nucleic acid

letters for RNA as will be explained. These sequence alignment questions lie at the heart of what can be called *computational biology* or in their application *bioinformatics* as opposed to *mathematical biology* as Gromov presumably intends in his question. A fundamental obstruction to mathematics in biology is the ansatz that no statement in biology is always true except for this one, which may be a simple consequence of diverse attempts to overcome shared terrestrial challenges through natural selection. The biological compared to the mathematical ethos thus allows only for *theorems with exceptions* which is of course anathema to the Tao of Mathematics.

We close this introduction with a "day in the life of a cell" just to very briefly explain the roles and interactions of macromolecules and other aspects. An excellent introductory reference to cellular molecular biology is [6]. DNA is the brains of the operation instructing the duplication of snippets of one or the other of its helical strands to the chemically similar RNA. So-called mRNA arising from this transcription of the DNA after splicing and editing contains the genetic coding for cellular protein expression whose instruments include proteins as well as the tRNA and rRNA active in the ribosome. The resulting array of translated proteins constitutes the workhorses and machinery of cellular life with each activity in the whole enterprise principally powered by the dephosphorylation of ATP to ADP. In addition to extensive mechanical duties throughout the organism, the proteins are like a flock of handlers to the celebrity RNA being shuffled importantly about the cell. A zoo of other RNAs including so-called miRNA, siRNA, snRNA, scRNA and snoRNA also participates in diverse regulatory activities, and various other RNAs presumably take part in further cellular enterprises as well. In fact, the active or mature protein or RNA is more than its simple polymer of amino or nucleic acids, for protein is altered by glycosylation, phosphorylation and methylation and RNA is spliced and edited, methylated and pseudouridylated among other modifications to their biologically active forms. It is actually the complex of several types of mature macromolecules together that begins to faithfully describe the complicated and true biology.

The dynamics of the cell is of course also stupendously complex. In addition to the electromagnetic and other forces, let us emphasize that the cell is usually an essentially aqueous environment. Hydrogen bonding as discussed later between macromolecules and with the ambient water molecules thus plays a crucial role in dynamics. The *hydrophobicity* of a compound measures its tendency to be energetically favorably hidden from water, and the differing hydrophobicities

of amino acids provide a major force of protein dynamics for instance comprising both entropic and electromagnetic aspects.

There is actually a fourth compound sometimes considered a macromolecule which we mention here for completeness, namely, lipids or fats. Certain lipid molecules have one end hydrophobic and the other hydrophilic and so combine in water into lipid bilayers of two stacked one hydrophobic end upon the other. These form bilipid sheets exhibiting only hydrophilic exterior in contact with water that form cell boundaries such as the cell wall itself or the intracellular vesicles delivering protein and other substances to their appointed locations. These lipid bilayers contain impurities such as cholesterol rendering them somewhat fluid. The exterior of the cell wall furthermore contains a distribution of a number of glycoconjugates of proteins and lipids which have been implicated in cell signaling among other functions. It may be interesting to model these impurities or other species within lipid surfaces as so-called configuration spaces, namely, as collections of distinct points within the surface, the simplest of all moduli spaces.

We begin by introducing moduli spaces via examples leading up to our two key combinatorial formulations: the moduli space of Riemann surfaces and the moduli space of flat connections on a surface. We next survey chemical and geometric aspects of proteins in order to probe their experimentally determined crystallographic structures using graph connections and present our principal discoveries on the geometry of protein backbone hydrogen bonding. Both combinatorial formulations are next discussed in the context of sugar though no serious applications in biology are described; this digression among other remarks aptly illustrates that our methods surely have traction for further utility as well as making familiar those sugars that occur in the RNA backbone. We turn finally to RNA with a survey of chemical and topological aspects in order to apply tools from the Riemann moduli space leading up to its identification with an appropriate moduli space of RNA structures. Each chapter on protein and RNA moreover ends with various remarks and speculations, and the paper itself concludes with an appendix briefly describing matrix models.

It is a pleasure to acknowledge and thank my friends and excellent collaborators on some of the material presented here, specifically Ebbe Andersen, Sigeo Ihara, Michael Knudsen, Alexey Finkelstein, Jens Jensen, Jakob Nielsen, Poul Nissen, Joanna Sułkowska, Takashi Tsuboi, Carsten Wiuf and especially Jørgen Ellegaard Andersen on the protein projects and the RNA initiatives including also Nikita Alexeev, Leonya Chekhov, Bertrand Eynard, Fenix Huang, Christian Reidys, Piotr Sułkowski, Peter Zograf and especially Mike Waterman.

1. Moduli Spaces

The typical context of a moduli space involves a system of differential equations often derived from geometrical or physical considerations as well as a group action on its space of solutions reflecting the inherent symmetries. It is the quotient of the solution space by the symmetry group that is the moduli space, and the dicey issue in many examples is in what sense to take this quotient so as to produce a tractable object. In various contexts, both the solution space and the group are infinite-dimensional though the quotient moduli space is finite-dimensional as a rule. Furthermore, the group action typically has finite isotropy on a suitably small subset of the solution space, so the quotient moduli space is not truly a manifold but rather a mild generalization to what is called an *orbifold* in geometry or a *stack* in algebraic geometry as discussed later. Moduli spaces are often non-compact.

Our goal in the next several sections is simply to give illustrative examples of moduli spaces including the two families of key importance here, namely, the Riemann moduli space of an orientable topological surface F of interest in both mathematics and theoretical physics, and the moduli space of flat G-connections on such a surface F, where G is some fixed Lie group, of general interest in gauge theories and in Chern-Simons theory in particular. We shall apply the former to study the topology of RNA, the latter to the geometry of proteins for $G = SO(3)$ and more speculatively both tools to the structure of polysaccharides.

Before this series of examples and owing to the dependency of our two key families on an underlying surface F, let us first digress to discuss surfaces in general. There are two versions of surfaces we must consider in order to handle all of our biological manifestations, namely, punctured surfaces (with isolated distinguished points and no boundary) and bordered surfaces (with boundary and no punctures where each boundary component contains a unique distinguished point).

A closed and connected orientable topological surface F_g is uniquely determined by its genus $g \geq 0$. Suppose that P is a finite non-empty set of distinct points in F_g and define the punctured surface $F_g^s = F_g - P$, where we shall require that the Euler characteristic $2 - 2g - s < 0$ is negative. Sometimes it is useful to regard points in P as punctures removed from F_g to form F_g^s and at other times as distinguished points in F_g. For bordered surfaces, suppose that Q is a finite non-empty set of open disks in F_g, the closure of any two of which have disjoint neighborhoods in F_g, and consider $F_{g,r} = F_g - \cup Q$ where we require that $g + r \geq 2$. We furthermore demand that each boundary component of $F_{g,r}$ comes equipped with a single distinguished base point. Examples

of punctured and bordered surfaces are given in Figures 4 and 6. To be sure, there is a more elaborate discussion simultaneously allowing both punctures and boundary components with one or more distinguished points, but we shall not highlight these here.

In order to proceed with convenience, let us choose for each non-negative g, r, s a specific oriented smooth surface still written F_g^s or $F_{g,r}$. The point here is that we can and shall speak of Riemannian metrics on these smooth manifolds, specifically finite-area complete metrics whose boundary is geodesic in the latter case. Special aspects [113, 2] of working in real dimension two imply that this choice of a smooth surface representing its topological type is immaterial to later considerations and that two diffeomorphisms of surfaces are homotopic if and only they are isotopic [15, 39]. We shall thus blur this distinction between differentiable and topological surfaces and between homotopy and isotopy in the sequel.

1.1. **Conformal quadrilaterals.** Consider a Euclidean rectangle

$$R = R_{h,w} = \{(x, y) \in \mathbb{R}^2 : 0 \leq x \leq w \text{ and } 0 \leq y \leq h\}$$

with its Riemannian metric ρ_E inherited from the Euclidean metric $ds^2 = dx^2 + dy^2$ on $\mathbb{R}^2$. Define the *modulus* of R to be

$$\frac{h^2}{A} = \frac{h^2}{hw} = \frac{h}{w},$$

where $A = hw$ denotes the area. Two metrics on R are said to be *conformal* if they have the same angles but not necessarily distances, and a metric on R conformal to ρ_E is thus described by scaling $f\rho_E$ for some $f : R \to \mathbb{R}_+$. It is not hard to see that two Euclidean rectangles are conformal if and only if their moduli coincide. The moduli space of conformal classes of Euclidean rectangles is thus identified with the space $\mathbb{R}_+$ of all moduli, i.e., the modulus is a complete invariant of conformal classes of Euclidean rectangles. Notice the non-compactness as the modulus goes to zero or infinity.

More interesting is the infinite-dimensional space of all Riemannian metrics on some closed topological disk Q with four distinguished boundary points. Given a metric ρ_0 on Q, again a conformal metric ρ is given by scaling ρ_0 by some $f : Q \to \mathbb{R}_+$. Call two opposite sides of Q the top and bottom of the quadrilateral, let h_ρ denote the ρ-length of the shortest arc connecting them, let A_ρ denote the ρ-area of Q and define the modulus of ρ to be $\mu_\rho = \frac{h_\rho^2}{A_\rho}$. It is easy to derive a conformal invariant of ρ_0 from the modulus, namely, the *conformal modulus* $\mu = \sup_\rho \mu_\rho$, the supremum taken over all conformal metrics $\rho = f\rho_0$.

It is an amazing yet elementary fact [1] that this supremum defining the conformal modulus is finite, it is realized as a maximum by an extremal metric and there is a homeomorphism $Q \to R$ to a Euclidean rectangle of the same modulus mapping boundary distinguished points to vertices and top/bottom to top/bottom so that the push forward[1] of the extremal metric is Euclidean. Thus, the conformal modulus is a complete invariant of conformal classes of metrics on abstract topological quadrilaterals.

It is characteristic of moduli spaces that there is some über space– the space of Euclidean rectangles or the space of conformal classes of metrics on abstract quadrilaterals in this example–supporting a group action–here the $\mathbb{R}_+$-action by homothety on rectangles or push forward on conformal classes of metrics on quadrilaterals. In general, a moduli space can often have multiple descriptions by several possible über spaces as in this example.

1.2. **Triangles in neutral geometries.** Our next example is the moduli space E_3 of equivalence classes of triangles in the Euclidean plane $\mathbb{R}^2$, where two triangles are equivalent if they are congruent. Here the über space is the set of all plane triangles supporting the action of the group of orientation preserving isometries generated by rigid rotations and translations in the plane. Euclid's side-side-side congruence theorem states that three side lengths determine a triangle up to congruence. To prove this directly, translations act transitively on $\mathbb{R}^2$ whence we may assume that two congruent triangles share a vertex, and rotations act transitively on lines through this vertex whence we may assume that the two congruent triangles share an edge; the opposite vertex can then be uniquely determined from the angles adjacent to this edge which of course coincide for the two isometric triangles.

In order to apply this theorem, let us consider a still larger über space $\tilde{E}_3$ of all congruence classes of triangles in the plane where the triangles come equipped with a labeling of their sides into first, second and third in a linear ordering compatible with the counter clockwise cyclic order coming from the orientation of the plane. The topology on $\tilde{E}_3$ is induced from the metric topology on the $\mathbb{R}^6$ coordinates of the three vertices in $\mathbb{R}^2$. According to Euclid's theorem, this space $\tilde{E}_3$ is parametrized by the collection of all ordered triples (a, b, c) of lengths

[1]Recall that if $f : M \to N$ is a diffeomorphism from the Riemannian manifold (M, g) to any smooth manifold N, then we can *push forward* the metric on M to produce the Riemannian metric $f_*(g)_p(X, Y) = g_{f^{-1}(p)}(df^{-1}(X), df^{-1}(Y))$ on N, where $p \in N$ and X, Y are tangent vectors to N at p. Thus, f is an isometry between the Riemannian manifolds (M, g) and $(N, f_*(g))$.

of the labeled sides, where these lengths of course must satisfy all three strict triangle inequalities $a < b + c$, $b < a + c$ and $c < a + b$. These triangle inequalities cut out an open cone homeomorphic to $\tilde{E}_3$ lying inside the positive orthant $\mathbb{R}^3_+$ in $\mathbb{R}^3$, and the quotient by the cyclic permutation of entries thus describes

$$E_3 \approx \{(a, b, c) \in \mathbb{R}^3_+ : a < b+c, b < a+c, c < a+b\}/\text{cyclic permutation}$$

itself up to homeomorphism, where E_3 inherits the quotient topology under the cyclic group action. Notice that in $\tilde{E}_3$ it makes sense to discuss the length of the first edge while the corresponding statement in E_3 itself is without meaning. Also observe that equilateral triangles play a special role in that they are the fixed points of the cyclic permutation on $\tilde{E}_3$; a neighborhood in E_3 of a corresponding point is naturally described as the finite quotient of an open set in $\mathbb{R}^3$.

Closely related is the analogous moduli space H_3 of equivalence classes of triangles in the hyperbolic plane $\mathcal{U} = \{z = x+iy \in \mathbb{C} : y > 0\}$ equipped with its Poincaré metric $ds^2 = \frac{dx^2+dy^2}{y^2}$. Recall [17, 48] that this complete metric has constant Gauss curvature -1 on $\mathcal{U}$, geodesics or straight lines for this metric are either semicircles perpendicular to the boundary $\mathbb{R} = \overline{\mathcal{U}} - \mathcal{U} \subset \mathbb{C}$ or vertical half-lines with endpoints in $\mathbb{R}$ and this upper half plane $\mathcal{U}$ with its Poincaré metric gives a model for hyperbolic geometry.

Furthermore, the Lie group of orientation preserving isometries of $\mathcal{U}$ is called the *Möbius group*

$$PSL_2(\mathbb{R}) = \left\{ \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) : a, b, c, d \in \mathbb{R} \text{ and } ad-bc = 1 \right\} \Big/ \left\{ \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \sim \left(\begin{smallmatrix} -a & -b \\ -c & -d \end{smallmatrix}\right) \right\}$$

acting on $\mathcal{U}$ by fractional linear transformations

$$\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) : z \mapsto \frac{az+b}{cz+d}.$$

A more uniform treatment of the ideal points is to introduce the circle $\mathbb{S}^1 = \mathbb{R} \cup \{\infty\}$ at infinity, so that each geodesic is determined by a pair of points in $\mathbb{S}^1$, and in fact the action of $PSL_2(\mathbb{R})$ extends in the natural way to this circle[2] in effect setting $0 = \frac{0}{1}$ and $\infty = \frac{1}{0}$.

There is an important trichotomy on $A \in PSL_2(\mathbb{R})$:

[2]The Cayley transform $\mathcal{U} \to \mathbb{D}$ from upper half plane to the open unit disk $\mathbb{D} \subset \mathbb{C}$ given by $z \mapsto \frac{z-i}{z+i}$ maps $\mathcal{U}$ to the Poincaré disk $\mathbb{D}$ with the unit circle in $\mathbb{C}$ as ideal boundary and push forward Riemannian metric $ds^2 = 4\frac{dx^2+dy^2}{(1-|z|^2)^2}$. $\mathcal{U}$ is a good model of the hyperbolic plane for computing while $\mathbb{D}$ is a good one for visualizing.

A is *hyperbolic* if $|trace\ A| > 2$; then A has two fixed points in $\mathbb{S}^1$ and acts as translation along the geodesic spanned by them;

A is *parabolic* if $|trace\ A| = 2$; then A has a single fixed point in $\mathbb{S}^1$ and acts as translation along circles tangent to $\mathbb{S}^1$ at that point;

A is *elliptic* if $|trace\ A| < 2$; then A has no fixed points in $\mathbb{S}^1$ and acts as rotation about a point of $\mathcal{U}$.

The proof is an easy exercise using just the equation for fixed points and the solution of the quadratic equation. Using rotations and translations precisely as in the Euclidean case, congruence of triangles sharing side lengths is seen to hold also in the hyperbolic case. We therefore find homeomorphic moduli spaces $H_3 \approx E_3$ of congruence classes of triangles parametrized exactly as before.

The existence of these extra parabolic transformations in the hyperbolic case will be exploited in the sequel, and we briefly elaborate here. A Euclidean circle in $\overline{\mathcal{U}}$ which is tangent to $\mathbb{R}$ at a point $p \in \mathbb{R} \subset \mathbb{S}^1$ is called a *horocycle centered at* p together with the exceptional case of horocycles centered at $\infty \in \mathbb{S}^1$ which are lines in $\mathcal{U} \subset \mathbb{C}$ with constant imaginary part. Examples of horocycles are given in Figure 2. A parabolic transformation fixes p and translates along each horocycle centered at p, the prototypical case being $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right) : z \mapsto z + 1$ for $p = \infty$.

Another aspect of hyperbolic geometry that is novel compared to the Euclidean case is that two distinct geodesics can be asymptotic to a common point of $\mathbb{S}^1$ leading in particular to so-called ideal triangles, namely, triples of geodesics disjoint in $\mathcal{U}$ and pairwise asymptotic to distinct points of $\mathbb{S}^1$. Examples of ideal triangles are given in Figure 2, each of which has hyperbolic area π from the Gauss-Bonnet Theorem.

1.3. **Elliptic curves.** Consider a discrete subgroup Λ in $\mathbb{C}$ of rank 2. The quotient of $\mathbb{C}$ by Λ is a flat torus or *elliptic curve* which comes equipped with a distinguished point corresponding to $0 \in \mathbb{C}$. Up to conformal equivalence, we may assume that Λ is generated by the unit $1 \in \mathbb{C}$ and $\tau \in \mathcal{U} \subset \mathbb{C}$.

The full Möbius group $PSL_2(\mathbb{R})$ of orientation preserving isometries contains the discrete subgroup $PSL_2(\mathbb{Z}) \subset PSL_2(\mathbb{R})$ with integral entries called the *modular group* which acts naturally on $\tau \in \mathcal{U}$ by fractional linear transformation and hence on lattices. It is not difficult to show that a maximal open region in $\mathcal{U}$ whose interior meets each orbit of $PSL_2(\mathbb{R})$ exactly once, a so-called "fundamental domain", is given by $\{z = x + iy : |x| \leq \frac{1}{2} \text{ and } x^2 + y^2 \geq 1\}$ as on the left in Figure 1.

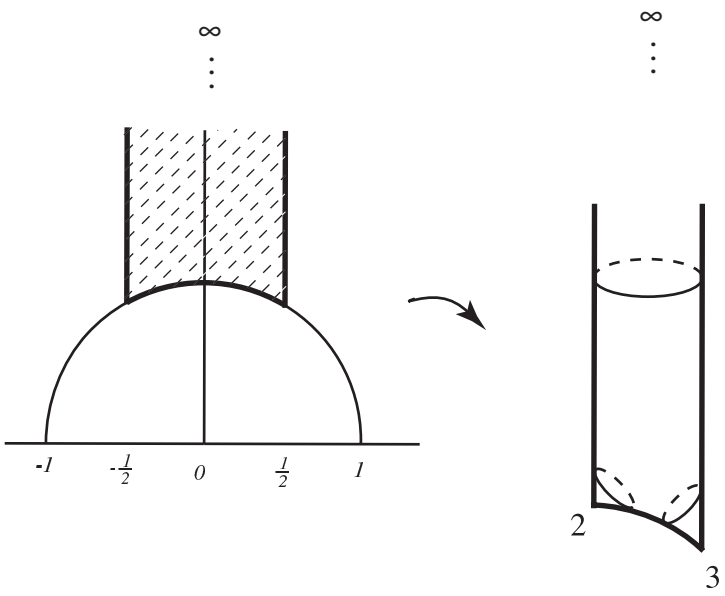


FIGURE 1. The modular curve.

The *moduli space of elliptic curves* is the quotient $\mathcal{M} = \mathcal{U}/PSL_2(\mathbb{Z})$ illustrated in Figure 1 on the right and is also called the *modular curve*. As the terminology would suggest, this is perhaps the most basic non-trivial moduli space of all appearing across a gamut of fields including number theory, dynamics and theoretical physics, topology and geometry and whose study goes back to Gauss. The points labeled 2 and 3 on the right in the figure naturally arise from the points i and $\frac{\sqrt{3}\pm 1}{2}$ in $\mathcal{U}$ with isotropy subgroups in $PSL_2(\mathbb{Z})$ of these respective orders not unlike the equilateral triangles in §1.2.

To get a better understanding of the modular group, we next produce a family $\mathcal{H}$ of horocycles in $\mathcal{U}$ which is invariant under its action. This inductive construction begins with the collection of horocycles h_n of Euclidean diameter one centered at $n \in \mathbb{Z} \subset \mathbb{R}$, for each $n \in \mathbb{Z}$, so h_n is tangent to $h_{n\pm 1}$ and is disjoint from the other horocycles. We also add the horizontal line at unit height as the horocycle h_∞ centered ∞, which is tangent to each h_n.

For the inductive step of the construction, two hororycles h_n, h_{n+1} centered at consecutive points determine a triangular region bounded by the interval $[n, n+1] \subset \mathbb{R}$ together with the horocyclic segments connecting the centers of the horocycles to the point of tangency of h_n and h_{n+1}. There is a unique horocycle contained in such a region simultaneously tangent to h_n and h_{n+1} as well as $\mathbb{R}$, and we let $h_{n+\frac{1}{2}}$

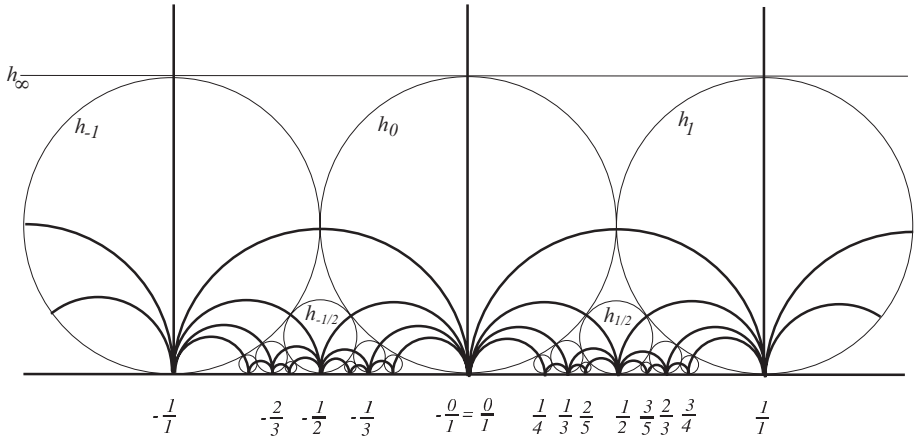


FIGURE 2. Horocycles $\mathcal{H}$ and Farey tesselation $\mathcal{F}$.

denote this horocycle. At the first go, these are evidently tangent to the real axis at the half-integer points $n + \frac{1}{2}$ and of diameter $\frac{1}{4}$, but in general the new center *is not* Euclidean equidistant to the two nearby old centers. Continue recursively in this manner to add horocycles tangent to the real axis and tangent to pairs of consecutive tangent horocycles to produce a family of horocycles $\mathcal{H}$ in $\mathcal{U}$. See Figure 2.

Lemma 1.1 (Farey-Cauchy Lemma)**.** *There is a unique horocycle in $\mathcal{H}$ centered at each extended rational point $\bar{\mathbb{Q}} = \mathbb{Q} \cup \{\infty\} \subset \mathbb{S}^1$. Furthermore, the horocycles in $\mathcal{H}$ centered at distinct points $\frac{p}{q}, \frac{r}{s} \in \bar{\mathbb{Q}}$ are tangent to one another if and only if $ps - qr = \pm 1$, and in this case, the horocycle in $\mathcal{H}$ tangent to these two horocycles is centered at $\frac{p+r}{q+s} \in \bar{\mathbb{Q}}$.*

It is not difficult to prove this inductively starting with the second sentence. The colorful history here is that this result was not proved but rather discovered by the mineralogist J. Farey thus solving the long-standing problem of giving a one-to-one enumeration of the rational numbers. After Farey published his empirical findings, Cauchy quickly supplied the inductive proofs.

The *Farey tesselation* is the collection $\mathcal{F}$ of hyperbolic geodesics in $\mathcal{U}$ that connect centers of tangent horocyles in $\mathcal{H}$, cf. Figure 2. It is invariant under the action of the modular group in that each element of $PSL_2(\mathbb{Z})$ extends to a mapping $\bar{\mathbb{Q}} \to \bar{\mathbb{Q}}$ which in turn induces a map on the collection of geodesics in $\mathcal{U}$ via the diagonal action on endpoints of geodesics, and this action on geodesics preserves the Farey tessellation. In fact, the action of $PSL_2(\mathbb{Z})$ on the set of *oriented* geodesics in $\mathcal{F}$ is simply transitive, so the Farey tessellation allows one

to quite effectively visualize the modular group $PSL_2(\mathbb{Z})$. For example, an especially beautiful combinatorial fact is that the continued fraction expansion of $\frac{p}{q} \in \mathbb{Q}$ can be read off from the sequence of right and left turns in $\mathcal{F}$ of a direct path in $\overline{\mathcal{U}}$ connecting $i \in \mathcal{U}$ to $\frac{p}{q} \in \mathbb{S}^1$.

1.4. **Riemann moduli space.** Our next example of a family of moduli spaces is most profound for its prevalence in low-dimensional classical and quantum topology, certain aspects of algebraic geometry and in string theoretic physics for example, namely, the *Riemann moduli space* $M(F)$ of all suitable classes of conformal structures on a fixed topological type of surface F. We shall begin with the case of possibly punctured surfaces $F = F_g^s$ with $s \geq 0$ and afterwards treat bordered surfaces. For a conformal structure on a manifold M of dimension $2n$, the transition functions of a covering of M by charts lie in the so-called structure group $O(2n) \times \mathbb{R}_+$, where $O(2n)$ denotes the group of orthogonal matrices of rank $2n$, while a complex structure on M has structure group the general linear group $GL_n(\mathbb{C})$. In our special case $M = F$ of complex dimension $n = 1$, conformal and complex structures thus coincide since their structure groups $O(2) \times \mathbb{R}_+ \approx \mathbb{C} \approx GL_1(\mathbb{C})$ agree. A *Riemann surface* is a surface with complex or conformal structure.

Let $Diff_+(F)$ be the infinite-dimensional group of orientation preserving diffeomorphisms of F and $Con(F)$ and $Com(F)$ denote the respective infinite-dimensional spaces of conformal and complex structures on F. $Diff_+(F)$ acts by push forward on both $Con(F)$ and $Com(X)$, and the quotient

$$M(F) = Con(F)/Diff_+(F) \approx Com(F)/Diff_+(F)$$

is the *Riemann moduli space* of F. Both über spaces $Con(F)$ and $Com(F)$ as well as the group $Diff_+(F)$ are infinite-dimensional.

One can continue with much more detail to make precise these infinite-dimensional spaces and groups and finite-dimensional Riemann moduli spaces as we shall discuss later, however, we shall take a different tack based on the Uniformization Theorem [20], a celebrated and game-changing result due to Koebe, Klein and Poincaré, which asserts that every simply connected Riemann surface is conformally equivalent to one of the three domains: the upper half plane $\mathcal{U}$, the complex plane $\mathbb{C}$, or the Riemann sphere, and in any case admits a Riemannian metric of constant respective Gauss curvature $-1, 0, 1$. It is the first case that is pertinent here to the study of a possibly punctured surface $F = F_g^s$, which must have negative Euler characteristic by the Gauss-Bonnet Theorem, and for which we must impose the further technical condition if $s > 0$ that the Riemannian metric is complete with finite area.

The Uniformization Theorem thus implies that the universal cover of F is none other than $\mathcal{U}$ with the fundamental group $\pi_1 = \pi_1(F)$ as the group of deck transformations acting upon it by orientation preserving isometry. That is, there is a representation

$$\rho : \pi_1 \to PSL_2(\mathbb{R})$$

of the fundamental group π_1 of the surface F as hyperbolic isometries, and $F = \mathcal{U}/\rho(\pi_1)$ as a Riemannian manifold with the metric inherited from the hyperbolic metric on $\mathcal{U}$. Of course, ρ must be injective, and we further require that the image group $\Gamma = \rho(\pi_1) < PSL_2(\mathbb{R})$ must be discrete, that is, the identity $I \in \Gamma$ is isolated from $\Gamma - \{I\}$ in the topology of $PSL_2(\mathbb{R})$ in order that the quotient is actually a surface.

To guarantee that the induced metric on $F = \mathcal{U}/\Gamma$ is indeed complete and finite-area when $s > 0$, there is a further parabolicity condition: if $g \in \pi_1$ is a simple loop surrounding a puncture p of F_g^s, then $\rho(g) \in PSL_2(\mathbb{R})$ must be parabolic. The paradigm for this is the puncture labeled ∞ in Figure 1. Likewise, one can easily imagine interesting examples of subgroups of the discrete group $PSL_2(\mathbb{Z})$ with finite index by choosing a connected sub polygon of the Farey tessellation as fundamental domain and identifying edges in pairs respecting the orientation and taking care to satisfy the parabolicity condition. A point at infinity in the fundamental domain for Γ in $\mathcal{U}$ corresponds to a fixed point of a parabolic transformation in $\rho(\Gamma)$ representing a simple curve about the corresponding puncture.

The *Teichmüller space* of the possibly punctured surface $F = F_g^s$ of negative Euler characteristic is the quotient

$$T(F) = \ \mathrm{Hom}'\big(\pi_1, PSL_2(\mathbb{R})\big)/PSL_2(\mathbb{R}),$$

where Hom' denotes the space of all injective homomorphisms

$$\rho : \pi_1 \to \Gamma < PSL_2(\mathbb{R})$$

whose image is a discrete group Γ satisfying the parabolicity condition, where $PSL_2(\mathbb{R})$ acts in the natural way by conjugacy on representations. The tuple of values taken by a representation on a generating set for π_1 provides a topological embedding into a product of copies of $PSL_2(\mathbb{R})$, so the topology on the space of representations and hence its quotient Teichmüller space is naturally induced from that of the Lie group $PSL_2(\mathbb{R})$. In fact, the Teichmüller space $T(F_g^s)$ is homeomorphic to an open ball and in fact analytically equivalent to a complex domain of real dimension $6g - g + 2s$ admitting numerous natural and significant metrics including the Weil-Petersson (Kähler Hermitian) and Teichmüller (Finsler) metrics for example, cf. [70, 96].

$Diff_+(F)$ has its canonical subgroup $Diff_0(F)$ of diffeomorphisms which are homotopic to the identity. The quotient

$$MC(F) = MC(F_g^s) = Diff_+(F_g^s)/Diff_0(F_g^s)$$

is called the *mapping class group* and is a finitely presented discrete group of truly paramount importance that is highly studied [44]. In other words, $MC(F)$ is simply the group of homotopy classes of orientation preserving homeomorphisms of F.

Finally, here is our rigorous entirely finite-dimensional definition: the *Riemann moduli space* of the possibly punctured surface F is the quotient

$$M(F) = T(F)/MC(F)$$

of Teichmüller space by the mapping class group. The distinction is that at a point in Teichmüller space, the über space from this point of view, one may speak of the hyperbolic length of a particular curve, whereas in moduli space one cannot as there is only the $MC(F)$-orbit of the curve. Again, $MC(F)$ acts with finite isotropy on $T(F)$, so the moduli space is an orbifold [44], that is, a space much like a manifold which is locally modeled by open sets in some $\mathbb{R}^n$ but supporting finite group actions in the current context. Low-dimensional exemplars of non-manifold points in an orbifold are provided by the points labeled 2 and 3 in Figure 1. This is all again reminiscent of the simple example of congruence classes of triangles discussed in §1.2.

It is no accident that the modular curve seems to be a paradigm[3] for the Riemann moduli space since an elliptic curve is a flat torus with its origin as distinguished point, that is, a surface of type F_1^1. The conformal structure of the punctured torus is uniformized by some subgroup $\Gamma < PSL_2(\mathbb{R})$ calculable from the modulus $\tau \in \mathcal{U}$ of the lattice in terms of the Weierstrauss $\wp$ function, and the hyperbolic metric on our surface descends from $\mathcal{U}$.

There are two important compactifications which we mention parenthetically sticking to surfaces $F = F_g^s$ without boundary only for convenience:

[3]There is actually a small correction only in this case F_1^1 in that the mapping class group $MC(F_1^1)$ contains a central element, the so-called elliptic involution ε, which is of order two and thereby uniquely determined. One takes the quotient in this special case to arrive at our familiar upper half plane $\mathcal{U} = T(F_1^1)/\varepsilon$ supporting the action of the modular group $PSL_2(\mathbb{Z}) = MC(F_1^1)/\varepsilon$.

Algebraically, the moduli space $M(F)$ of Riemann surfaces homeomorphic to F has its fundamental [35] *Deligne-Mumford compactification* $\bar{M}(F)$ which is a projective algebraic variety containing $M(F)$ as a dense open set. $\bar{M}(F)$ evidently plays a basic role.

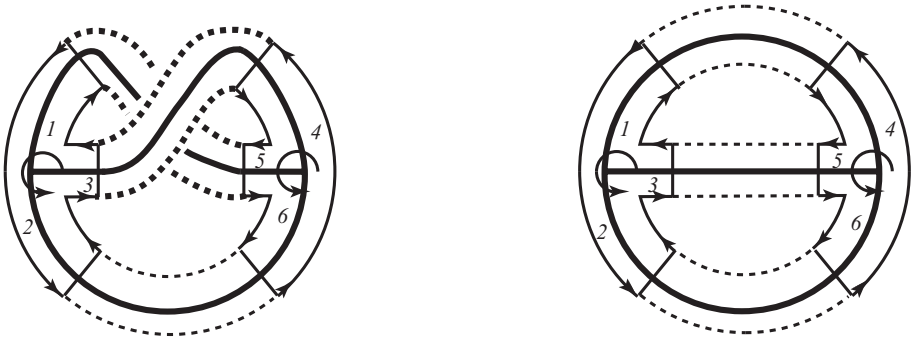
William Thurston [126, 45] described a compactification of the Teichmüller space by so-called projective measured laminations in the surface, the space of which is a piecewise-linear sphere of dimension $6g - 7 + 2s$ compactifying the open ball that is $T(F)$ to a closed ball on which $MC(F)$ acts continuously albeit ergodically on the boundary sphere. This naturally extended and revitalized work of Jakob Nielsen [93] from the 1940s.

Other geometrical aspects of note include that the Ricci flow [55] on the surface carries complete finite-area metrics to those of constant curvature, and the Riemann moduli spaces themselves thus provide the space of solutions to Einstein's field equations in this simplified 2D case of two spatial dimension. Furthermore [127], the Ricci flow describes the renormalization-group evolution of 2D sigma models.

Finally in the case of bordered surfaces $F = F_{g,r}$, we demand that metrics on F have geodesic boundary. In effect, gluing two copies of F along their boundaries produces a closed surface to which we can apply the previous discussion. The Uniformization Theorem provides a discrete subgroup Γ of $PSL_2(\mathbb{R})$ and a subset $\Omega \subset \mathcal{U}$ with geodesic boundary so that $F = \Omega/\Gamma$. The Teichmüller space $T(F)$ is then defined to be conjugacy classes of injective homomorphisms of π_1 onto discrete groups Γ consisting entirely of hyperbolic transformations. For the definition of the mapping class group $MC(F)$ in the bordered case, we demand that homeomorphisms must fix the boundary distinguished points setwise and homotopies of homeomorphisms must fix them pointwise. The Riemann moduli space $M(F) = T(F)/MC(F)$ is then defined just as before.

1.5. **Fatgraphs.** In general for punctured surfaces, there is an analogue of the Farey tessellation that gives a suitable cellular decomposition of a slightly elaborated version of Teichmüller space which is invariant under the action of the mapping class group $MC(F)$. In order to describe this, we must introduce a mild generalization of graphs, called "fatgraphs" (sometimes "ribbon graphs" or "maps") as follows.

A *fatgraph* is a one-dimensional CW complex τ together with a collection of cyclic orderings on the half-edges incident on each vertex, where a half-edge is one of the two complementary components to an interior point of an edge. The number of half-edges incident on a



Fatgraph and skinny suface for the once punctured torus F_1^1

Fatgraph and skinny surface for the thrice punctured sphere F_0^3

FIGURE 3. Two fattenings of a single graph and their skinny surfaces. Each vertex of valence k contributes a non-convex polygon of $2k$ sides illustrated with solid lines, and each edge contributes one quadrilateral respecting the orientations at its endpoints illustrated with dashed lines.

fixed vertex is called its *valence*. Fatgraphs in turn determine skinny surfaces $F(\tau)$ with boundary as depicted in Figure 3. Furthermore, the boundary components of $F(\tau)$ determine closed edge paths on τ themselves called the *boundary cycles* of the fatgraph. Notice that one can compute the genus g from the formula for Euler characteristic $2-2g-s=m-n$, where τ has n edges, m vertices and s boundary cycles. In order to retrieve the punctured surface $F=F_g^s$ from $F(\tau)$, we may adjoin one once-punctured disk to $F(\tau)$ along its boundary to each boundary component of $F(\tau)$. We then have the inclusions $\tau \subset F(\tau) \subset F$ all homotopy equivalences, and we say that the fatgraph τ is a *spine* of F in this case. Examples of fatgraph spines $\tau \subset F$ are given in Figure 4.

There is a representation of a fatgraph $\tau=\tau_{\sigma,\iota}$ as a data type that is especially amenable to computer representation, namely, a fatgraph τ is uniquely determined by a pair $\sigma, \iota \in \Sigma_{2n}$ of permutations on $2n$ letters, where n is the number of edges of τ, σ is an arbitrary permutation and ι is an involution; furthermore, fatgraph isomorphism classes are precisely the double cosets in Σ_{2n}. Namely, $\sigma \in \Sigma_{2n}$ is a disjoint union of k cycles $(i_1, i_2, \ldots, i_k)$, for certain k, and each such cycle determines a vertex of τ with its incident half-edges labeled $i_1, i_2, \ldots, i_k$ in counterclockwise order; as an involution, $\iota \in \Sigma_{2n}$ is a product of some number of disjoint transpositions (j_1, j_2), and we adjoin one edge connecting the half-edges with these labels j_1 and j_2 for each transposition in

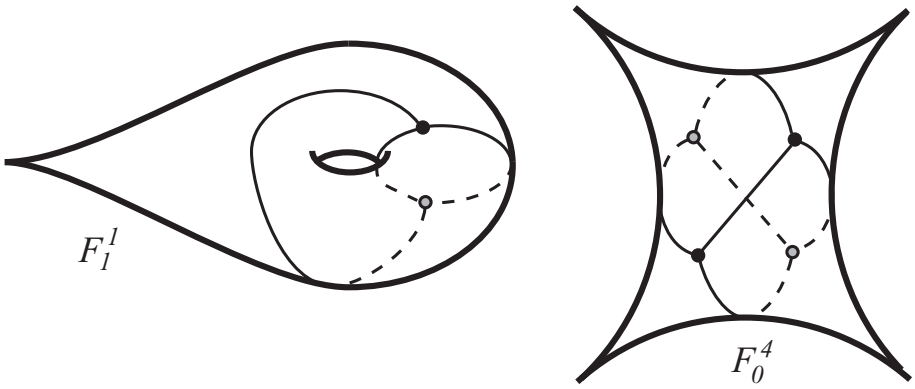


FIGURE 4. Examples of fatgraph spines for punctured surfaces.

τ. For example with the labeling of half-edges in Figure 3, in either case we have $\sigma = (1,2,3)(4,5,6)$ a product of two 3-cycles with $\iota_1 = (1,5)(2,6)(3,4)$ for F_1^1 and $\iota_2 = (1,4)(2,6)(3,5)$ for F_0^3. Higher valence vertices are handled similarly. Notice that fixed points of ι correspond to univalent vertices, 2-cycles of σ to bivalent vertices and fixed points of σ to isolated points of τ. Another especially nice aspect of this notation is that the boundary cycles of $\tau = \tau_{\sigma,\iota}$ are none other than the cycles of the composition $\rho = \sigma \circ \iota$. Indeed in an appropriate sense we shall later exploit, Poincaré duality on the closed surface F_g is simply described by $(\sigma, \iota) \leftrightarrow (\rho, \iota)$.

Though we must treat univalent vertices in the sequel, let us first consider fatgraphs each of whose vertices has valence at least three. There is a face relation $\tau' < \tau_1$ on such fatgraphs generated by contracting an edge of the trivalent fatgraph τ_1 with distinct endpoints to produce τ' as illustrated in Figure 5. Expanding the unique four-valent vertex of τ' in the unique distinct way produces another fatgraph $\tau_2 > \tau'$, and we say that τ_1 and τ_2 differ by a *flip*.

By a *metric* on a fatgraph τ, we mean the assignment of some non-negative real number $\mu(e)$ to each edge e of τ so that there are no essential cycles in τ each of whose constituent edges has vanishing μ value, a restriction we shall call the *no-vanishing cycle condition*. A metric μ on τ thus has a (possibly empty) forest $\Phi \subset \tau$ on which it vanishes. Each component of Φ can be collapsed to a distinct vertex in order to produce another fatgraph τ_Φ to which τ contracts, and the metric μ on τ induces the identical strictly positive metric on τ_Φ. It is useful to pass to projective classes of positive metrics on τ, which are naturally parametrized by points of the open $(e-1)$-dimensional

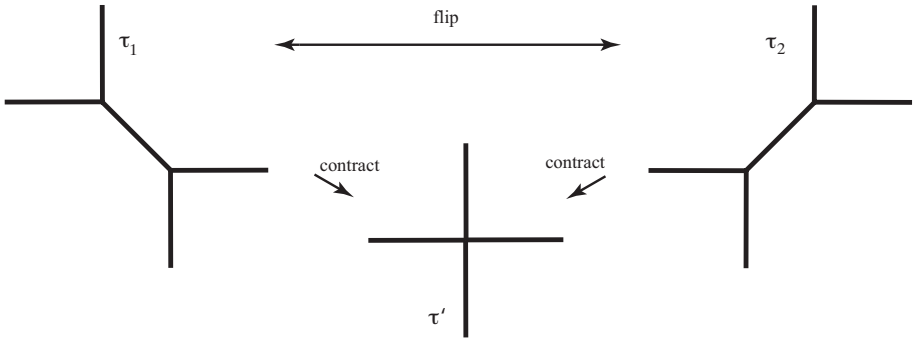


FIGURE 5. Flips and contractions.

simplex Δ^{e-1}, where e is the number of edges of τ. We may furthermore identify the open face of Δ^{e-1} corresponding to the vanishing of barycentric coordinates on the edges in the forest Φ with the open simplex for τ_Φ. In this manner, the space of all projective metric fatgraph spines in F naturally inherits the structure of a union of open simplices together with certain of their faces, namely those faces corresponding to forests, where the face relation is generated by the contraction of edges with distinct endpoints.

We must also introduce a mild generalization of $T(F)$: the *decorated Teichmüller space* is simply $\tilde{T}(F_g^s) = T(F_g^s) \times \mathbb{R}_+^s$. The depth of this definition lies in the interpretation of the coordinates $\mathbb{R}_+^s$ as hyperbolic lengths in $F = \mathcal{U}/\Gamma$ of horocycles as follows. Consider a horocycle h in $\mathcal{U}$ centered at the fixed point of a parabolic transformation in Γ. It is tantamount to completeness of the hyperbolic metric in F near the corresponding puncture that h projects to a closed curve in F, which is also called a horocycle but now in F as opposed to $\mathcal{U}$. If h is short enough, then a horocycle in F is a simple closed curve separating the corresponding puncture from the rest of F. The fiber coordinates in decorated Teichmüller space are taken to be the hyperbolic lengths of a collection of horocycles in F, one such not necessarily embedded horocycle about each puncture. By permuting punctures of F and hence their coordinates, the usual $MC(F)$ action on $T(F)$ extends to $\tilde{T}(F)$. It is again useful to projectivize $\tilde{T}(F_g^s)$ to produce $T(F) \times \Delta^{s-1}$, so in particular the projectivized decorated Teichmüller space is canonically identified with the Teichmüller space $T(F)$ itself if $s = 1$.

Theorem 1.2. [125, 58, 99] *Suppose that* $2g-2+s>0$. *Then there is a* $MC(F_g^s)$*-invariant cell decomposition of projectivized decorated Teichmüller space* $T(F_g^s)\times\Delta^{s-1}$ *which is isomorphic to the combinatorial space of all homotopy classes of projective metric fatgraph spines* $\tau\subset F$, *each of whose vertices has valence at least three, under the face relation generated by contraction.*

The hyperbolic description of this decomposition [99] is pointwise different but combinatorially identical with the conformal one [125, 58]. It is important to emphasize that the same combinatorial type of fatgraph occurs infinitely often as the index of a cell in $T(F)\times\Delta^{s-1}$ from its many different homotopy classes of embeddings as a spine. In effect, there is a tiling of $T(F_g^s)\times\Delta^{s-1}$, where the tiles are in one-to-one correspondence with homotopy classes of fatgraph spines of F_g^s, and this tiling is natural in the sense that it is invariant under the obvious action of $MC(F_g^s)$. That is, the *finite collection* of abstract fatgraphs of the correct combinatorial type for F_g^s describe a tiling of a fundamental domain for the action of $MC(F_g^s)$; in fact, it is the underlying cells indexed by combinatorial types modulo their symmetry groups that comprise the decorated moduli space $\tilde{M}(F)=(T(F)\times\Delta^{s-1})/MC(F)$.

Theorem 1.2 exactly generalizes the classical Farey tessellation from elliptic curves to arbitrary multiply punctured surfaces with negative Euler characteristic, namely, for F_1^1, each and every ideal triangle in the Farey tessellation $\mathcal{F}$ has its corresponding fatgraph combinatorially identical with the one illustrated in Figure 3 while Figure 4 identifies a particular homotopy class of spine in the surface and hence determines a particular ideal triangle complementary to $\mathcal{F}$. It is important to emphasize that the contractions and flips of Figure 5 are performed on spines within the surface. According to the no-vanishing cycle condition, we can contract any single edge of a trivalent fatgraph spine in F_1^1 to produce a fatgraph spine with a single vertex of valence four and pass thereby to the three codimension one faces of a complementary ideal triangle to $\mathcal{F}$; however, we cannot contract two such edges since any two edges comprise a cycle, and this is reflected by the fact that the vertices of Farey tessellation lie at infinity, i.e., do not lie in $T(F)$.

We must still go beyond this result a bit to treat surfaces $F_{g,r}$ with $r\geq 1$ boundary components and no punctures. To this end combinatorially, a univalent vertex of a fatgraph uniquely determines its incident edge called a *tail*, and each boundary component of a fatgraph for a bordered surface must have exactly one incident tail as illustrated in Figure 6 whose univalent vertex must be different from the boundary distinguished point. A metric does not assign any value to a tail, and

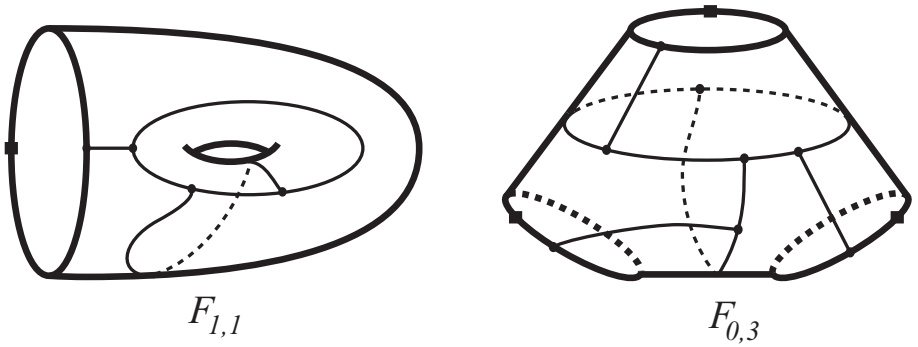


FIGURE 6. Examples of fatgraphs with tails as spines for bordered surfaces. Boundary distinguished points are marked with ■ icons.

only non-tail edges can be contracted still subject to the no-vanishing cycle condition.

To this end geometrically, we have already chosen in each boundary component a distinguished point and consider complete finite-area metrics with geodesic boundary of constant Gauss curvature -1 on $F_{g,r}$ to define the decorated Teichmüller space $\tilde{T}(F_{g,r})$ with the action of its mapping class group $MC(F_{g,r})$ of homeomorphisms setwise fixing the collection of boundary distinguished points modulo homotopies fixing them pointwise.

Theorem 1.3. [102] *Suppose that $g + r - 1 > 0$. Then there is a $MC(F_{g,r})$-invariant cell decomposition of decorated Teichmüller space $\tilde{T}(F_{g,r})$ which is homotopy equivalent to the combinatorial space of all isotopy classes of projective metric fatgraph spines $\tau \subset F$ with tails whose univalent vertices lie in the boundary, with exactly one in each boundary component, and otherwise vertices have valence at least three under the face relation generated by contraction of non-tail edges.*

It is fair to say that Teichmüller theory was traditionally a topic in complex analysis with a beautiful and extensive theory of $Com(F) \approx Con(F)$ based on Banach manifolds in the work of Ahlfors and Bers [21, 22, 70] and their school which makes precise our first definition of the Riemann moduli space. The definition can also be formulated in terms of algebraic geometry [53, 91, 60, 35] though this is again rather involved. While the complex analytic viewpoint prevailed in the Ahlfors-Bers school, hyperbolic geometry was employed though not

centrally until Thurston and his school in the 1970s and 1980s meanwhile brought various geometric techniques into play providing a whole new perspective.

The combinatorial treatment described here has proven useful computationally in multiple contexts in part owing to the applicability of techniques for enumerating fatgraphs from quantum field theory called *matrix models* as discussed in the Appendix. Let us also just mention that Theorems 1.2 and 1.3 furthermore provide a new description of the mapping class group $MC(F)$ as the stabilizer of any object in an associated mapping class groupoid generated by flips as well as explicit representations of both group and groupoid as rational mappings on appropriate coordinates [99, 104]. These give prototypical examples of so-called cluster varieties [47].

1.6. **Flat G-connections.** Good general references for the material in this section and beyond are [78, 89].

Fix a Lie group G and fix a possibly punctured or bordered suface F. A *principal G-bundle* over F is a fiber bundle $p : P \to F$ with a right action of G on P so that G acts freely and transitively on each fiber. Principal bundles are always locally trivial. An *(Ehresmann) connection* on P is a "horizontal" subbundle $H \subset T_*P$ of the tangent bundle with the kernel $Ker(dp)$ of dp "vertical" which is invariant under the G-action such that $H \oplus Ker(dp) = T_*P$. For example, the trivial G-bundle over F is given by the projection $p : F \times G \to F$ onto the first factor with the action of G by right multiplication, and the trivial connection is given by pull back $p^*(T_*F)$.

Two principal G-bundles are isomorphic if there is a G-equivariant bundle isomorphism covering the identity map of F, and two bundles with connections are isomorphic if moreover H pushes forward to H'. A connection is said to be *flat* if the bundle H is comprised of tangent spaces to a foliation of P.

The *group of gauge transformations* $\mathcal{G}(P)$ of P is the group of principal G-bundle isomorphisms of P. $\mathcal{G}(P)$ is naturally isomorphic to the group of equivariant smooth maps $P \to G$ with respect to the conjugation action, and in particular, if the principal G-bundle P is trivial, then the group of equivariant maps $P \to G$ is simply identified with the group of all maps $F \to G$.

The *holonomy* of H along a closed curve $\gamma : [a, b] \to F$ is the element $g \in G$ so that $gx = y$, where x is any chosen point in the fiber of $P \to F$ over $\gamma(a) = \gamma(b)$, and the unique lift of γ to P starting from x whose tangent vectors lie in H has its endpoint with fiber coordinate y. More generally, lifting vectors horizontally along paths in this manner

tangent to H is called *parallel transport*. If the connection is flat, then the holonomy is well-defined on homotopy classes of curves and indeed upon choosing a base point in F for the fundamental group $\pi_1 = \pi_1(F)$ gives rise to a homomorphism $\pi_1 \to G$. This choice of base point corresponds to an inner automorphism of P, so a flat connection on the principal G-bundle $P \to F$ gives rise to a well defined element of $\mathrm{Hom}(\pi_1, G)/G$.

Conversely, such a homomorphism $\rho : \pi_1 \to G$ for a chosen base point in F in turn gives rise to the flat G-bundle $(\tilde{F} \times G)/\pi_1$, where $\gamma \in \pi_1$ acts diagonally on the universal cover $\tilde{F}$ as deck transformations and on G as left multiplication by $\rho(\gamma)$. This bundle comes equipped with a flat connection inherited from the trivial connection on $\tilde{F} \times G$, and its holonomy is given by ρ.

Consider the moduli space $M(F, G)$ of all flat connections on principal G-bundles $P \to F$ modulo the group of all gauge transformations $\mathcal{G}(P)$. Flat connections for a fixed F form an affine space modeled on the Banach space of 1-forms on P with values in the Lie algebra of G that satisfy the Maurer-Cartan equation. Techniques of global analysis thus again provide rigorous definition for $M(F, G)$, or at least for the über-space of all flat connections on all principal G-bundles. Morevoer, the elementary constructions above based on holonomy provide mappings between $M(F, G)$ and $\mathrm{Hom}(\pi_1, G)/G$, and we have:

Theorem 1.4. *For any Lie group G and any surface F, we have*

$$M(F, G) = \mathrm{Hom}(\pi_1(F), G)/G,$$

that is, the moduli space of flat G-connections on principal G-bundles over F is naturally identified with the representation variety.

There is the small quirk of terminology to point out that it is not the quotient Riemann moduli space $M(F) = T(F)/MC(F)$ but rather Teichmüller space $T(F)$ itself which corresponds to the moduli space $M(F, PSL_2(\mathbb{R}))$ of flat connections. Indeed, Teichmüller space is actually a component of $M(F, PSL_2(\mathbb{R}))$ according to [51], and likewise at least for real reductive groups G, there is an analogous so-called Hitchin component [62] of $M(F, G)$. In any case, the mapping class group $MC(F)$ acts on the moduli space $M(F, G)$ as a representation variety in the current context, and the dynamics, which is only partly understood, depends upon the specific Lie group G.

Our final interpretation of $M(F, G)$ follows immediately from its description as a representation variety, and we suppose that F has at least one puncture or boundary component and choose any fatgraph spine $\tau \subset F$. The next idea is very simple: instead of representing the

fundamental group π_1 of F, let us instead represent the fundamental path groupoid of τ in G.

Namely, a *G-graph connection* on the graph underlying τ is the assignment $g(e) \in G$ to each oriented edge e of τ so that $g(\bar{e}) = g(e)^{-1}$ if $\bar{e}$ is the reverse orientation to e. Two such assignments $g(e), h(e) \in G$ are regarded as equivalent if there is $k_v \in G$, for each vertex v of τ, so that $g(e) = k_v h(e) k_w^{-1}$ for each oriented edge e of τ with initial point v and terminal point w.

Corollary 1.5. *For any Lie group G, any surface F, and any graph τ homotopy equivalent to F, we have*

$$M(F, G) = \{G - \text{graph connections on } \tau\},$$

that is, the moduli space is naturally identified with the collection of all G-graph connections on τ.

The value taken by a graph connection on an oriented edge simply describes the parallel transport along it, and the holonomy of the graph connection g on τ along the closed oriented edge path $e_1 - e_2 - \cdots - e_n$ in τ is the ordered product $g(e_1)g(e_2)\cdots g(e_n) \in G$.

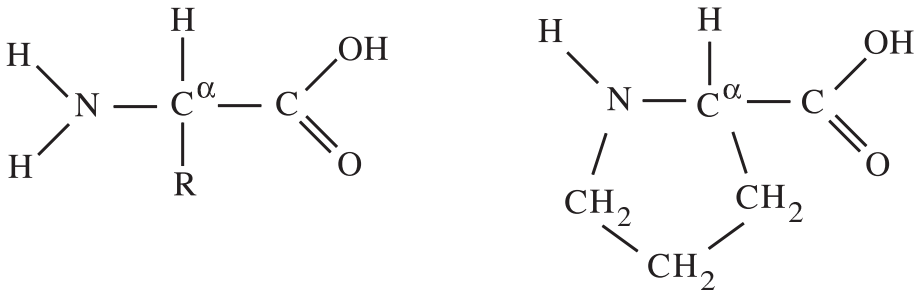
In order to get straight to the combinatorics, we have ignored the foundational complexity in effect taking $\mathrm{Hom}(\pi_1, G)/G$ as our *operative definition* of $M(F, G)$. The key difficult and critical issue is to control the topology of moduli space $M(F, G)$ as the quotient by the group $\mathcal{G}(P)$ of gauge transformations. This has an expansive history based on global analysis in Banach manifolds and on important work of Mumford, Narasimhan-Seshadri, Atiyah-Bott, Hitchin, Simpson, Donaldson and Marsden-Weinstein, Kempf, Ness, Kirwin among others. We refer the interested reader to the forthcoming Bulletin of the American Mathematical Society article by Goldman et al.

2. Protein

The geometric and chemical nature of proteins required here is surveyed in [105] as we first recall, and the fatgraph model of proteins is discussed. This is not our key result on protein, rather, using $SO(3)$-graph connections, we have discovered [106] new geometric constraints on proteins as we shall explain effectively passing from topology in [105] to geometry in [106]. The marvelous monograph [46] provides the broader view of protein structure and dynamics.

2.1. **Chemistry and geometry.**

A protein is a special type of polypeptide, and a polypeptide is a linear polymer of amino acids. An amino



a) Typical amino acid b) Proline

FIGURE 7. Amino acids. On the left, residue R is one of a number of possibilities and C^α is the first or "αth" carbon; the exceptional imino acid Proline is depicted on the right. We shall for simplicity here also refer to Proline as amino acid.

acid is one of 20 special molecules[4] of the general structure illustrated in Figure 7.

In any case, the OH on the right hand side in Figure 7 of an amino acid can condense off a water molecule with an H on the left hand side of another amino acid thus combining two or many amino acids into a polymer linked by new inter-amino bonds C$=\!=$N called "peptide bonds" as in Figure 8. These are drawn as partial bonds between C_{i-1} with O_{i-1} including N_i because it is a consequence of quantum chemistry–and an amazing geometric fact about this hybridized bond– that the six atoms C_{i-1} and N_i plus O_{i-1}, C^α_{i-1}, C^α_i and H_i comprising the so-called "peptide unit" actually lie in a common plane, namely, the centers of mass of the Bohr models for these atoms are coplanar. The angles are also nearly 120°, so the geometry of peptide units in Figure 8 is roughly accurate[5]. Furthermore, the angles at each C^α_i are tetrahedral. The "backbone" is the sequence of molecules N-C^α-C-N-C^α-$\cdots$-C-N-C^α-C, and the only moduli for the geometry of the

[4]We shall take this number of 20 since these are the "classical" gene-encoded amino acids including Proline which occur widely across all species although two others have been more recently discovered: There is a 21st (Selenocysteine) which also occurs widely across species but rarely compared to the other 20 as well as a 22nd (Pyrrolysine) arising only in methanogenic archaea and certain exotic bacteria.

[5]The peptide units are all drawn with the C^αs on opposite sides of the line of the peptide bond. This is the so-called *trans* conformation of the protein, and it is usually energetically favorable to keep these carbon atoms thus separated. The other possible conformation is called *cis*, and it occurs relatively commonly only for Proline-preceding peptide bonds, where the energetics is mitigated by the steric implications of the massive carbon groups in the Proline itself.

backbone are thus the so-called conformational angles[6] φ, ψ, one pair of these dihedral angles for each C^α again as illustrated.

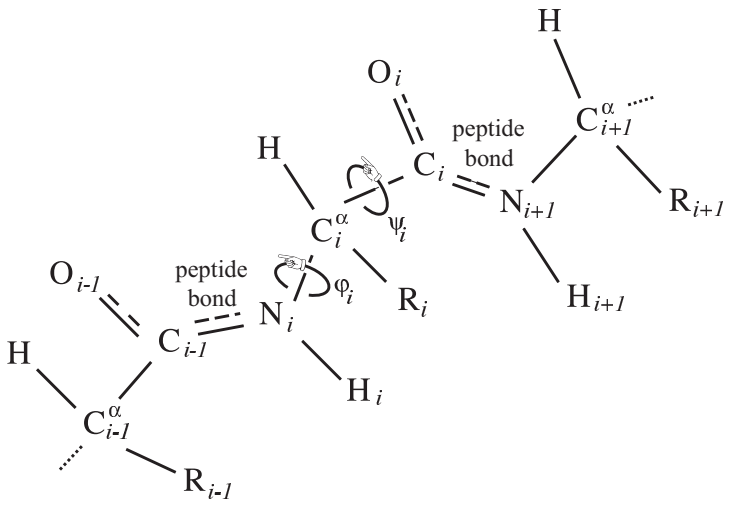


FIGURE 8. Polypeptide, backbone, peptide unit and conformational angles from [105].

The "primary structure" of a polypeptide is given by the sequence of amino acid residues that comprise it in the linear order from its N- to C-terminus. This word in the 20-letter alphabet of residues uniquely determines the chemical structure. A reasonable operative length to keep in mind for a protein is a few hundred amino acid residues though lengths of biologically active proteins can in fact range from just a few into the tens of thousands.

2.2. **Protein folding.** Certain rare polypeptides have the property that they crystalize–not in any mathematical sense–but in the sense that there is a least energy state whose energy level is well separated at normal biological conditions from its competitors; this state is moreover often densely situated in space in that it cannot be penetrated by ambient water molecules. This is perhaps a good attempt to define "protein" but only for moderately sized so-called globular or water soluble proteins which are sometimes implicitly considered. Larger water soluble proteins are commonly comprised of a number of moderately sized and often repeating constituents. The other two main classes of proteins beyond the water soluble ones are fibrous (like collagen

[6]The histogram of these two backbone conformational angles in a torus $S^1 \times S^1$ over some representative subset varies from one amino acid to another and is the so-called Ramachandran plot of important utility throughout protein science.

whose primary structures are typically repeating) and transmembrane (which cross cellular lipid bilayers and therefore have their own special properties for example sometimes having several comparable low energy conformations in order to mechanically pump ions efficiently across cell membranes). At any rate, certain polypeptides are selected for expression in an organism, and these are its natural proteins. Their folding and function must be sufficiently reliable that the organism can depend upon them through thick and thin and thus must indeed therefore have these well separated least energy states.

Proteins fold into characteristic shapes, and most moderately sized globular proteins do so spontaneously without further instruction from the cell. The prediction of the folded structure from the primary structure is the famous "protein folding problem" rightfully ballyhooed as the fundamental problem in molecular biology since a comprehensive solution to it would in principle allow drug invention and testing to be performed largely on the computer at a minuscule cost and difficulty compared to current laboratory methods just for example. Accurate predictions of protein folding and protein interactions from first physical principles are difficult because many forces are at play: electromagnetic and ionic forces coming from charges on the different residues, van der Waals forces and hydrophobicity of residues and protein regions among others, all this taking place in the essentially aqueous environment that is the organism–let alone consideration of the full Schrödinger equations to describe the quantum system. The computer modeling of classical physical systems approximating folding or other evolution is called "Molecular Dynamics, and there is a vast literature, cf. [114, 18, 81, 65, 40, 130, 122]. There are also numerical methods of "Density Functional Theory" approximate solutions to the many-body quantum systems, cf. [63, 79, 97, 28, 66]. Both approaches are computationally extremely intensive.

The repository [19] of all experimental data on folded proteins started slowly in the 1970s and is called the Protein Data Bank (PDB). Each PDB file provides the complete set of spatial coordinates of each constituent atom in the corresponding protein. As an aside, we mention that another procedure of predicting folded protein structure called "Homology Modeling" comes from comparing primary structure with the known structures in PDB, cf. [34, 86, 16, 90, 129, 76]. The PDB is readily available online and grows daily. However, individual files among the 100,000 or so entries vary in quality and idiosyncrasy. The novice is encouraged to probe this very accessible data base, but any systematic processing of this library is best undertaken with guidance to overcome various quirks.

Some biologists would argue that these PDB files do not truly represent the structure in the cell since first the proteins are isolated and respectively either crystalized or dissolved in specific buffers incorporating heavy isotopes and only then analyzed using either X-ray crystallography or NMR techniques. Further criticisms come from the equal standing as PDB files of diverse experimental conditions rendered by different laboratories as well as the equally diverse methods of processing the raw data–electron cloud densities or pulse sequences of ensembles–to finally produce the PDB file. Two aspects of this data processing are "validation" and "refinement", namely, comparing spatial coordinates determined from the raw data to idealized conformations and then improving this fit according to some scoring procedure. It is clear that *a priori* geometric constraints such as planarity of the peptide unit for instance play a critical role in this basic processing from raw data to PDB file.

One of the key stabilizing forces for folded proteins as well as one of the key driving forces of hydrophobicity is provided by hydrogen bonding or H-bonding as follows. An electronegative atom is one that tends to attract electrons, and examples of such atoms include C,N,O in this order of increasing attraction. When one electronegative atom approaches another one which is chemically bonded to a hydrogen atom, the two atoms can share the electron cloud of the H atom and thus attract one another through an H-bond; the former is called the "acceptor" and the latter the "donor" of the H-bond. These H-bonds are roughly ten times weaker than the covalent bonds between adjacent atoms in the backbone and residues, however, H-bonds are still relatively energetically expensive, for example exceeding van der Waals forces, and hence are deeply involved in protein folding. Electronegative atoms in the residues can also participate in H-bonds.

Two special motifs of H-bonding are especially prevalent in practice, namely, alpha helices and beta strands as illustrated in Figure 9, where the H-bonds are oriented from donor to acceptor. Notice that beta strands can naturally combine via further H-bonds into beta sheets of multiple strands. These patterns of bonding sufficiently saturate the protein for these energetically important H-bonds that fully 60-70 percent of the backbone participates in them. The pattern of alpha helices and beta strands is called the "secondary structure" of the protein, and its "tertiary structure" is the complete set of spatial coordinates of each constituent atom as recorded in a PDB file.

2.3. **Fatgraph model.** There is a topological model of proteins as fatgraphs introduced in [105] which associates a fatgraph $\tau = \tau(\mathcal{P})$ to a

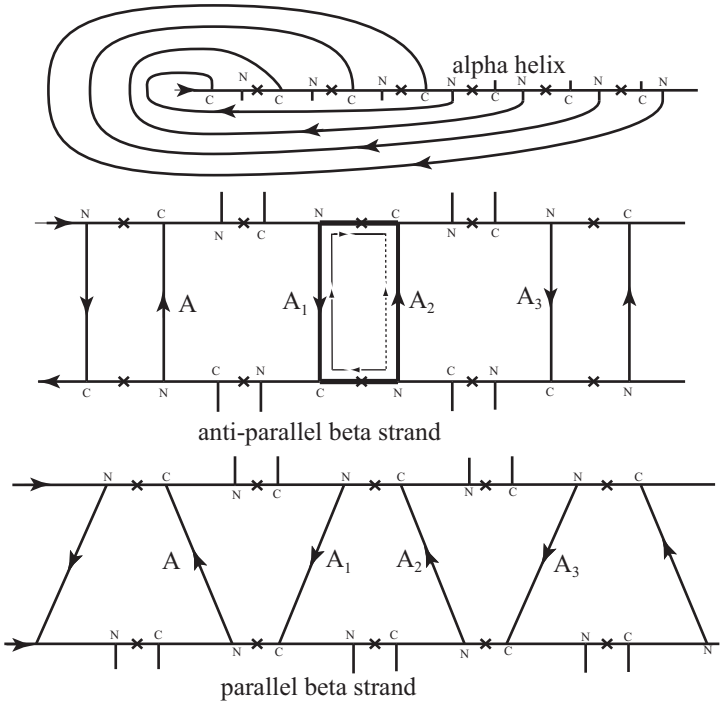


FIGURE 9. Alpha helices and beta strands. Each $\times$ represents a C^α. Specific H-bonds will be discussed later.

folded protein $\mathcal{P}$ that is so natural that we just applied aspects of it without explanation or apology in the depiction of alpha helices and beta strands in Figure 9 as is typically done in protein science: to each peptide unit is assigned a fatgraph building block for the backbone consisting of a small horizontal segment with one smaller vertical segment at each end, one pointing up and the other down. The endpoints of the building blocks represent C^αs, and the first and second vertical segments along the backbone respectively represent C and N. These can be combined at their endpoints either preserving or reversing the pattern of up and down as was convenient in Figure 9 along the backbone, and we then add one edge to τ for each H-bond in the natural way oriented from donor to acceptor.

In fact, τ must now be enhanced to allow two types of edges, one type as before and the new type contributing a once-twisted band instead of an untwisted one as usual to the skinny surface $F(\tau)$. This permits the backbone depictions as in Figure 9 and evidently leads to possibly non-orientable skinny surfaces $F(\tau)$ associated to the pattern $\mathcal{P}$ of H-bonding. It furthermore turns out to be necessary to turn to this category of not necessarily orientable surfaces using fatgraphs with

twisted and untwisted edges in order that the topological type of $F(\tau)$ changes only a bit under experimental errors or uncertainties in the data for $\mathcal{P}$. The details are given in [105] which also describes the natural twisting on the edges of τ arising from the H-bonds, e.g., those with arrows on them in Figure 9, that we have yet to discuss here.

To explain this key construction from [105], let $\overrightarrow{PQ}$ denote the displacement vector from P to Q and recall from above that the unit vector parallel to the displacement vector $\overrightarrow{C_{i-1}N_i}$ from C_{i-1} to N_i lies in the plane of the peptide unit, itself oriented naturally counter-clockwise from $\overrightarrow{C_{i-1}N_i}$ to $\overrightarrow{C^\alpha_{i-1}C_{i-1}}$. This gives a vector in an oriented plane in $\mathbb{R}^3$ or equivalently a so-called *positive 3-frame* (u, v, w) of mutually perpendicular unit vectors where the third $w = u \times v$ is the cross product of the first two. Thus, the pair of peptide units participating in an H-bond[7] gives a pair of positive 3-frames (u_ℓ, v_ℓ, w_ℓ), where $\ell = a$ for acceptor and $\ell = d$ for donor, and hence also determines the unique rotation of $\mathbb{R}^3$ carrying $(u_d, v_d, w_d) \mapsto (u_a, v_a, w_a)$.

2.4. $SO(3)$**-graph connections.** We have seen that an H-bond between peptide units gives rise to an element of the Lie group

$$SO(3) = \{3 \times 3 \text{ matrices } A : A^t A = I \text{ and } \det A = 1\},$$

where superscript ^{t} denotes the transpose, det the determinant and I the 3-by-3 identity matrix. Under the assumption of idealized geometry, it is an exercise to compute the rotation in $SO(3)$ for two peptide units consecutive along the backbone in terms of the backbone conformational angles:

Lemma 2.1. [105] *Consider two consecutive peptide units with the backbone conformational angles φ, ψ between them under the idealized geometry that angles in the peptide units are exactly 120° and the angles at the C^αs are exactly tetrahedral. Then the element $A \in SO(3)$ mapping the 3-frame of one peptide unit*[8] *to the next one along the*

[7]The question naturally arises of how to recognize an H-bond in practice, i.e., from the PDB file of the protein. There is a standardized method [75] of recognizing and classifying H-bonds from a PDB file called Dictionary of Secondary Structure for Proteins (DSSP) which employs a crude approximation of the energy of an H-bond and takes a threshold of -0.5 kcal/mole below which the H-bond is regarded to exist. Refinements of DSSP as well as other methods for recognizing H-bonds are also used.

[8]When the peptide unit before the acceptor C^α is in the *cis* conformation, the matrix A must be pre-multiplied by a diagonal matrix with entries $(1, -1, -1)$.

backbone is given by

$$A = B(\varphi)C(\varphi+\psi)\begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} & 0 \\ \frac{\sqrt{3}}{2} & \frac{1}{2} & 0 \\ 0 & 0 & -1 \end{pmatrix},$$

where

$$B(\varphi) = \begin{pmatrix} \frac{2}{3} - \frac{C_1^2}{3} + \frac{S_1^2}{6} & -2\big[\frac{\sqrt{2}C_1}{3} + \frac{S_1^2}{4\sqrt{3}}\big] & 2\big[\frac{C_1S_1}{2\sqrt{3}} - \frac{S_1}{3\sqrt{2}}\big] \\ 2\big[\frac{\sqrt{2}C_1}{3} - \frac{S_1^2}{4\sqrt{3}}\big] & \frac{2}{3} - \frac{C_1^2}{3} - \frac{S_1^2}{6} & -2\big[\frac{C_1S_1}{6} + \frac{S_1}{\sqrt{6}}\big] \\ 2\big[\frac{C_1S_1}{2\sqrt{3}} + \frac{S_1}{3\sqrt{2}}\big] & 2\big[\frac{S_1}{\sqrt{6}} - \frac{C_1S_1}{6}\big] & \frac{2}{3} + \frac{C_1^2}{3} - \frac{S_1^2}{3} \end{pmatrix}$$

for $C_1 = \cos\varphi$ *and* $S_1 = \sin\varphi$ *and*

$$C(\varphi+\psi) = \begin{pmatrix} 1 - \frac{3}{2}S_2^2 & \frac{\sqrt{3}}{2}S_2^2 & \sqrt{3}C_2S_2 \\ \frac{\sqrt{3}}{2}S_2^2 & 1 - \frac{1}{2}S_2^2 & -C_2S_2 \\ -\sqrt{3}C_2S_2 & C_2S_2 & 1 - 2S_2^2 \end{pmatrix}$$

for $C_2 = \cos\frac{\varphi+\psi}{2}$ *and* $S_2 = \sin\frac{\varphi+\psi}{2}$.

Recall (cf. [106]) that $SO(3)$ supports the metric

$$d(A,B) = \big|\arccos\frac{\mathrm{trace}(AB^t) - 1}{2}\big|, \text{ for } A, B \in SO(3),$$

which is invariant under both right and left multiplication. In fact, we define an edge in our model of τ in [105] to be untwisted[9] if $d(I,A) < d(I,B)$, where $A \in SO(3)$ maps $(u_d, v_d, w_d) \mapsto (u_a, v_a, w_a)$ and B maps $(u_d, v_d, w_d) \mapsto (u_a, -v_a, -w_a)$; the fatgraph of [105] is thus quite natural and reproduces the diagrams usually drawn with physical but until now without precise mathematical interpretation in protein science. The paper [105] goes on to use the topological type of the resulting possibly non-orientable surface $F(\tau)$ to predict certain aspects of protein classification, however, the paper [107] shows that the primary structure does a better job of this than the topological type of $F(\tau)$. We shall not further discuss the topological considerations of [105] here yet shall rely critically in the sequel on the geometric construction of the $SO(3)$ graph connection of a folded protein that was just described.

A fortunate consequence is that we shall not be forced to further consider non-orientable surfaces and can take any convenient up or down depiction of the backbone to which we add one ordinary edge as before for each H-bond in order to determine a *graph* τ (not a fatgraph though we may draw it as such) together with an $SO(3)$ graph connection upon it. Specifically, a positive 3-frame $\mathcal{F} = (u, v, w)$ determines

[9]There is a little more to it depending upon the *cis* or *trans* conformation, but we shall leave it at that here and refer the interested reader to the original text.

a matrix $A \in SO(3)$ whose respective columns are given by u, v, w in the standard vector basis; this matrix A maps the standard coordinate 3-frame $(\vec{i}, \vec{j}, \vec{k})$ to $\mathcal{F}$. For any two peptide units with corresponding positive frames $\mathcal{F}_i, \mathcal{F}_j$ and matrices A_i, A_j, the rotation $A_j A_i^{-1}$ thus maps $\mathcal{F}_i$ to $\mathcal{F}_j$. However, this is not a useful invariant since it changes if we rotate the entire protein in space. There is a standard trick to correct this deficiency by transforming both $\mathcal{F}_i$ and $\mathcal{F}_j$ by A_i^{-1} so that $\mathcal{F}_i$ becomes the standard 3-frame, $\mathcal{F}_j$ becomes $A_i^{-1}(\mathcal{F}_j)$, and the rotation $A_{i,j} = A_j^{-1} A_i$ mapping the former to the latter is our true invariant value of the $SO(3)$ graph connection associated to the H-bond between two peptide units. Notice that for three peptide units i, j, k, we have the identity $A_{i,k} = A_{i,j} A_{j,k}$.

Since edges corresponding to H-bonds can be canonically oriented from donor to acceptor, we do not really need the full fatgraph building blocks described before.

Construction 2.1. Given a folded protein $\mathbb{P}$, take a "backbone" interval K with integer endpoints in $\mathbb{R}$ whose consecutive integral points are in one-to-one correspondence with the consecutive peptide units in $\mathbb{P}$. If there is an H-bond from the ith (donor) to the jth (acceptor) peptide units, then add to K a semicircular arc in the upper half plane oriented from $i \in K$ to $j \in K$. This determines a graph $\tau(\mathbb{P})$ with its non backbone edges oriented. Assign to each oriented edge the matrix $A_{i,j}$ discussed before associated to peptide units i and j participating in H-bonding and further assign to each backbone interval oriented from i to $i+1$ the matrix $A_{i,i+1}$ rotating consecutive 3-frames, cf. Lemma 2.1. This determines an $SO(3)$ graph connection $A_{\mathbb{P}}(e) \in SO(3)$, for e an oriented edge of $\tau(\mathbb{P})$.

Recall from Gauss that a rotation $A \in SO(3)$ of $\mathbb{R}^3$ is equivalently described in "angle-axis" form as rotation by some angle $\theta \geq 0$ about some unit vector $\vec{\omega}$ with the understandings that $\theta, \vec{\omega}$ is equivalent to $-\theta, -\vec{\omega}$ and $\theta = 0$ corresponds to $A = I$. By associating the product $\theta\vec{\omega} \in \mathbb{R}^3$, we may thus draw figures in $SO(3)$ depicted as the ball of radius π centered at the origin in $\mathbb{R}^3$ where antipodal points in the boundary of this ball are to be identified.

Main Discovery 2.1. [106] *Apply Construction 2.1 to a representative subset of H-bonds from the PDB to extract roughly one million H-bonds using high-quality data. Then the histogram in $SO(3)$ of the associated H-bond rotations actually occurs in only about 30 percent (with most of the data in fact occurring in only about 20 percent) of the volume of $SO(3)$ as depicted in Figure 10. The data in this region of H-bond*

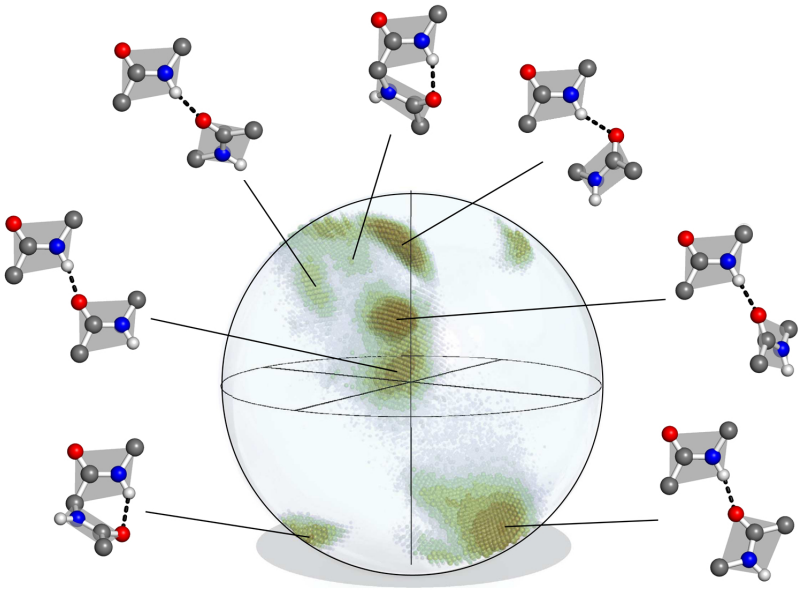
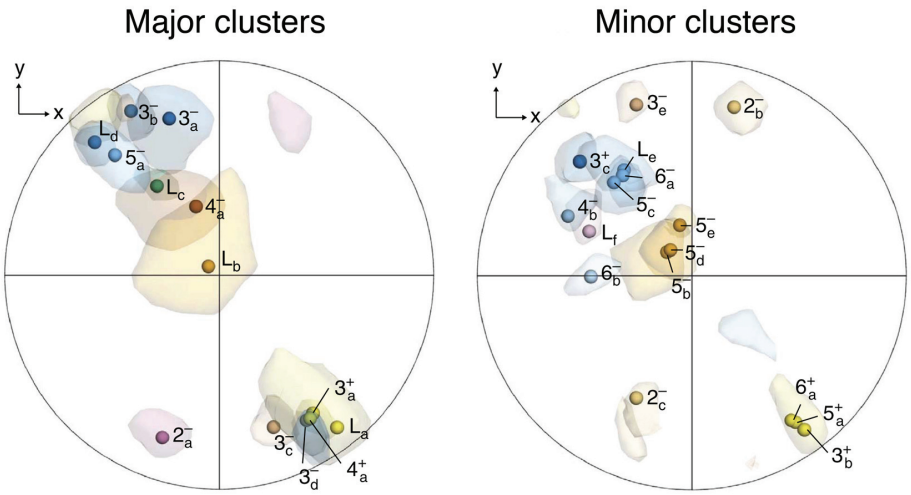


FIGURE 10. Histogram in $SO(3)$ of a representative set of roughly one million H-bonds together with several sample rotations of peptide units adapted from [106].

rotations in $SO(3)$ that do arise further cluster into 30 regions depicted in Figure 11 providing a new classification for H-bonds.

Using $SO(3)$ graph connections to probe the geometry of H-bonds in the PDB, we have thus discovered new *a priori* geometric constraints on folded proteins, namely, roughly 70 percent by volume of the possible H-bond conformational space is avoided. It is important to emphasize that nearly the entire space $SO(3)$ arises in principle for some configuration of peptide units with an appropriate translation though some several percent by volume of $SO(3)$ is ruled out presumably for steric reasons. Nature is thus more conservative than Geometry. For example, just as the restrictions on backbone conformational angles in Ramachandran plots are widely used to refine and validate PDB files as determined from raw experimental data in practice, so too these new constraints on H-bond geometry should be imposed as well. The utility of this will take some time yet to confirm, but the phenomenon of clustering and the attendant classification it entails are so robust that they are plainly visible to the naked eye already in the raw data in Figure 10. Moreover, the same basic clusters occur for other data



Cluster	Population	Volume	Rotational space Mode	Ave. translation	Ave. dist.	Max dist.
2_a^-	16,327	0.50%	3.03 (− 0.28, − 0.78, 0.56)	(− 0.73, − 2.27, 1.50)	0.268	1.253
2_b^-	1,249	0.20%	2.86 (0.22, 0.86, 0.46)	(− 0.45, − 2.52, − 1.13)	0.285	1.059
2_c^-	110	0.03%	2.46 (− 0.33, − 0.73, − 0.60)	(− 1.11, − 2.01, − 1.70)	0.393	1.283
3_a^-	164,832	1.38%	2.50 (− 0.29, 0.92, − 0.28)	(2.05, − 3.66, − 0.27)	0.341	1.729
3_b^-	16,761	0.79%	2.89 (− 0.45, 0.83, − 0.33)	(2.56, − 3.46, 1.46)	0.338	1.429
3_c^-	7,706	0.48%	2.55 (0.31, − 0.87, − 0.38)	(2.03, − 3.76, 0.01)	0.292	1.135
3_d^-	5,490	0.57%	2.86 (0.45, − 0.74, − 0.50)	(2.48, − 3.28, − 1.57)	0.416	1.508
3_e^-	321	0.08%	2.78 (− 0.29, 0.90, 0.32)	(2.09, − 3.13, − 1.97)	0.240	0.537
4_a^-	504,642	2.42%	1.08 (− 0.32, 0.93, − 0.17)	(2.88, − 3.77, 0.19)	0.214	2.045
4_b^-	1,969	0.68%	2.16 (− 0.84, 0.41, 0.36)	(3.55, − 3.01, − 0.13)	1.222	3.139
5_a^-	16,500	1.21%	2.41 (− 0.63, 0.73, 0.26)	(3.16, − 3.35, − 1.25)	0.394	2.797
5_b^-	3,661	0.67%	0.57 (− 0.64, 0.60, − 0.48)	(3.11, − 3.34, 1.23)	0.499	1.967
5_c^-	3,406	0.29%	2.05 (− 0.56, 0.67, 0.50)	(3.77, − 2.44, − 1.58)	0.239	1.26
5_d^-	1,907	0.30%	0.51 (− 0.62, 0.74, 0.26)	(3.3, − 3.41, − 0.40)	0.296	1.218
5_e^-	295	0.08%	0.77 (− 0.23, 0.96, − 0.18)	(3.01, − 3.57, − 0.52)	0.247	1.195
6_a^-	1,964	0.68%	1.95 (− 0.52, 0.75, 0.40)	(3.11, − 3.07, − 0.89)	0.880	3.141
6_b^-	1,308	0.31%	1.49 (− 0.99, − 0.01, 0.12)	(3.33, − 2.72, 1.56)	0.420	1.987
2_a^+	266	0.12%	2.71 (0.57, − 0.75, − 0.34)	(2.59, − 3.80, − 0.67)	0.668	2.001
3_a^+	6,965	1.10%	2.55 (0.54, − 0.79, − 0.29)	(2.29, − 4.02, − 0.40)	0.563	3.139
3_b^+	2,088	0.46%	2.80 (0.59, − 0.80, 0.08)	(2.63, − 3.81, 0.85)	0.452	1.546
3_c^+	707	0.29%	2.41 (− 0.68, 0.69, − 0.24)	(2.60, − 4.02, 1.00)	0.493	1.485
4_a^+	6,848	1.34%	2.54 (0.52, − 0.83, − 0.21)	(2.38, − 3.92, 0.21)	0.605	3.136
5_a^+	4,525	1.10%	2.67 (0.57, − 0.80, − 0.18)	(2.43, − 3.90, 0.36)	0.668	3.120
6_a^+	1,325	0.52%	2.67 (0.54, − 0.81, − 0.23)	(2.48, − 3.72, 0.69)	0.881	2.380
L_a	242,357	7.62%	2.82 (0.61, − 0.79, 0.00)	(2.44, − 3.90, 0.48)	0.591	3.141
L_b	127,879	5.97%	0.21 (− 0.78, 0.63, 0.06)	(3.00, − 3.51, 0.13)	0.591	2.512
L_c	13,727	1.81%	1.63 (− 0.56, 0.80, − 0.20)	(2.77, − 3.60, − 0.57)	0.513	2.322
L_d	8,747	0.93%	2.79 (− 0.65, 0.70, − 0.30)	(2.35, − 4.04, 1.26)	0.383	1.436
L_e	1,221	0.30%	1.96 (− 0.51, 0.79, 0.35)	(2.84, − 3.25, − 1.52)	0.352	1.235
L_f	808	0.24%	1.84 (− 0.82, 0.35, − 0.45)	(2.79, − 3.21, 2.06)	0.370	1.268

FIGURE 11. Depiction of the 30 clusters together with their population, percentage volume of $SO(3)$, point of highest density or so-called mode given in angle-axis coordinates, average translation vector, average and maximum distance from mode again adapted from [106]. The notation N_x^ϵ for a cluster indicates that $\Delta = \epsilon N$ as explained in the text, and L denotes long-range $|\Delta| > 6$, with alphabetical subscripts $x = a, b, \ldots$ distinguishing the various clusters. Major/minor clusters are simply those of largest/smallest population.

sets extracted from the PDB using various cutoffs of primary sequence identity and PDB file quality; see the original paper for details.

It turns out that certain density features of the original histogram can be explained by separating the data into 11 subsets depending upon the signed distance $\Delta = j - i - 1$ along the backbone from the donor i to the acceptor j taking values $2 \leq |\Delta| \leq 6$ and "long-range" with $|\Delta| > 6$. These subsets are separately grouped into clusters according to methods discussed in [106] leading to 30 regions in $SO(3)$. H-bonds that are short-range along the backbone, i.e., those with $|\Delta| \leq 6$, have of course been well-studied in terms of the backbone conformational angles φ, ψ. Our results confirm and refine the existing classifications of so-called helices, turns and hairpins, and moreover provide new classes. Furthermore, this is the first systematic study and classification of long-range H-bond geometries.

Two planes in space are of course related not only by a rotation but also a translation, and it is natural to wonder if there might be a further refinement of the rotation clustering by translations. For length $\Delta = -3$, there was found an evident sub-clustering of the main region into two sub-clusters which depend upon the translation, and this distinction is incorporated into the 30 clusters of Discovery 2.1. It is also interesting that the translations provide no further aggregations of density among rotations other than this case of $\Delta = -3$.

Purely computational verification of the empirical discovery above is based on Density Functional Theory, which we recall gives an approximate solution to the Schrödinger equation, where we compute minimum energies for two peptide units engaged in an H-bond. The resulting solution again lies in a region roughly thirty percent of the volume of $SO(3)$ nicely roughly agreeing with that already discovered albeit without the fine detail of the 30 clusters. We refer the interested reader to the original paper.

2.5. **Further remarks on protein.** We close this discussion applying $SO(3)$ graph connections to protein with several mostly mathematical remarks. Let us first just mention that the $SO(3)$ graph connection method has been elaborated in [94] to provide a useful graphical technique for studying and comparing protein structures which is then applied to a complex of the protein streptavidin and the vitamin biotin highlighting the importance of so-called water bridges, another manifestation of H-bonding involving the ambient water molecules.

It follows from Lemma 2.1 that H-bonds which are short-range along the backbone can be computed directly in terms of the conformational angles, but the situation is more interesting for long-range H-bonds.

Referring back to Figure 8, the ideal backbone conformations given in degrees are known to be $(\phi, \psi) = (-135, 150)$ for antiparallel beta and $(-120, 135)$ for parallel beta. The $SO(3)$ graph connection of H-bond rotations clearly has no holonomy in the sense that the product of holonomies along any closed edge path in the fatgraph must give the identity since the graph connection is itself derived from the 3-frames. Thus, the two paths illustrated on the anti-parallel beta strand have the same holonomy, and this gives a recursion for A_2 in terms of A_1 based on the anti-parallel backbone matrices computed for the ideal geometry from Lemma 2.1. Similar remarks apply in the parallel case.

For example, recognition that the long-range clusters called L_b is associated with parallel beta strands comes from the fact that its point of highest density is very near a fixed point of the ideal dynamical system in the metric geometry of $SO(3)$. This is interesting for several reasons including demonstrating that the intrinsic Lie geometry of $SO(3)$ appears to be an effective tool for biophysics, i.e., this is the right metric for this question about proteins. Such dynamical systems on a Cartesian product of several copies of $SO(3)$ may warrant further study in treating other motifs comprised of several H-bonds.

As was already mentioned, our $SO(3)$ graph connections have trivial holonomy, yet the holonomy of general $SO(3)$ graph connections can certainly be non-trivial. One such example arises from elements of $SO(3)$ occurring as averages over ensembles of graph connections, which could quite reasonably arise experimentally.

To a fatgraph τ arising from Construction 2.1 with K chords, we assign an open $(K-1)$-simplex Δ_τ with the natural face relation as before. Explicitly, let us consider some positive real number assigned to each chord which we shall call its "free energy", and let us furthermore take the quotient by an overall homothetic scaling by $\mathbb{R}_+$. Setting these projective free energies to zero on an edge corresponds to erasing that edge and passing to the corresponding face of the simplex naturally enough. It is much like the structure in §1.5, where contraction of an edge there corresponds to erasure of a chord here. Of course, the value of the graph connection is lost when its projective free energy vanishes, i.e., when it is erased.

Consider a fatgraph τ arising from Construction 2.1. Any graph connection on τ can always be represented by one that is identically $I \in SO(3)$ on the backbone edges as is easily confirmed, and we shall assume this realization henceforth. The union

$$\mathcal{P}_N = \cup_\tau \, \Delta_\tau \times M(\tau, SO(3))$$

of all $SO(3)$ graph connections $M(\tau, SO(3))$ on all possible graphs τ of oriented chords supporting projective free energies on a backbone interval containing N integral points thus provides a *moduli space for proteins* on N peptide units.

In fact, H-bonds can bifurcate, though this is not terribly common especially for donors, and this could be incorporated into the definition of $\mathcal{P}_N$, i.e., usually no more than one semicircle pointing towards or two pointing away from any integer point. This moduli space $\mathcal{P}_N$ is so far only of speculative utility, but nevertheless this new concept of *metric (fat)graph connection* seems to be an interesting one in gauge theory. There are also other physical restrictions on proteins, for example, the so-called steric constraints that constituent atoms cannot overlap which depend upon the primary structure of the protein.

Let us finally emphasize that this idea of applying $SO(3)$-graph connections to physical chemistry surely goes far beyond the application here to protein H-bonds between peptide units.

3. Sugar

Polysaccharide is a typically not linear but rather treelike polymer of monosaccharides (one constituent sugar) and oligosaccharides (a few, usually 3-6 with some exceptional cases of 7-9 sugars). Monosaccharides have the chemical formula $(CH_20)^n$ for some $n = 3, \ldots, 7$ and come in two chemically distinct families: ketoses containing C=O and aldoses containing CH=O. Monosaccharides come in both open linear and closed cyclic forms though it is the cyclic form that prevails in solution as well as in our attention here. The linear form of a monosaccharide is concisely described by its Fischer projection (1891) and the cyclic form by its Haworth diagram (1928) as illustrated in Figure 12 for glucose, ribose and fructose. It is surely an affirming point that the Haworth diagram itself is already a fatgraph meant to describe a molecule containing a carbon-oxygen ring lying in a cylinder in space. The C and O atoms lying in the cycle of a cyclic form are called the backbone atoms of the monosaccharide. Chemists number the C in the Haworth diagram as illustrated in the figure in each case. The backbone C next to the O which *does not* carry just the CH_2OH group is called the anomeric carbon for both aldoses and ketoses as also illustrated in the figure, and it plays a special role as we shall see.

For each of these two chemical families, there are two basic dichotomies of conformation: a monosaccharide is in the D-form if the bottom OH in the Fischer projection, just above the terminal CH_2OH, lies to the right, and it is in the L-form if it is the mirror image of

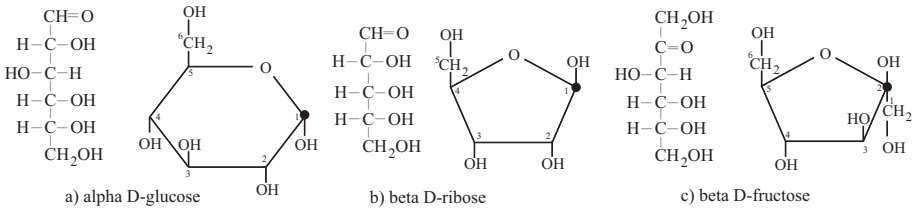


FIGURE 12. Typical aldose and ketose monosaccharides in Fischer and Haworth projections. The bold vertex denotes the anomeric carbon in either case, and the standard ordering on carbons is indicated.

the D-form with this OH therefore lying to the left. The second distinction between so-called alpha and beta forms captures whether the OH carried by the anomeric carbon lies on the same or different side, in the Haworth diagram up or down, as the nearby CH_2OH group. Both alpha and beta conformations typically occur while the D-form is drastically more common in nature than the L-form, which occurs only negligibly, and we have therefore illustrated in Figure 12 both conformations alpha and beta but only D-forms.

As to the polymer, mono- and oligosaccharides assemble into oligo- and polysaccharides through the formation of covalent so-called glycosidic bonds between the OH on the anomeric carbon and another OH on the backbone ring of another mono- or oligosaccharide while condensing off a water molecule. There is an entirely obvious fatgraph to build by simply connecting Haworth fatgraph diagrams of mono- and oligosaccharides with one edge for each glycosidic bond in the obvious way as we did for H-bonds in protein.

This associates a natural fatgraph $\tau(S)$ to a polysaccharide S. There is moreover a graph $\bar{\tau}(S)$ arising from $\tau(S)$ by collapsing each cycle in the Haworth diagram of each constituent monosaccharide to a distinct point. Because of the special bonding which always involves an anomeric carbon, it is not difficult to show that $\bar{\tau}(S)$ must be a tree. For example, a cellular storage device for glucose called glycogen consists of linear strands of glucose with glycosidic bonds, i.e., edges of the fatgraph, including the anomer OH at position 1 and the OH of the 4th carbon, where long such linear strands of 1,4 linkages branch every 10-14 glucose units via 1,6 linkages to form a tree.

The topological type of this tree $\bar{\tau}(S)$ is one invariant of the polysaccharide S. The fatgraph $\tau(S)$ without the collapse furthermore has in particular its boundary cycles, and for example, the number ℓ_k of

boundary cycles of edge-length k for each k–the so-called *length spectrum* of $\tau(S)$ or S–provide further polysaccharide fatgraph invariants.

It is important to stress the vital fact that the geometry of the Haworth diagram is approximately correct for a monosaccharide in space but not quite. Indeed, the backbone atoms satisfy non-generic planarity conditions: for 5 backbone atoms (like the vertices of an envelope) and for 6 backbone atoms (like the vertices of a lawn chair), there is a plane containing O and 3 backbone C. In any case, it is again encouraging that there is thus a canonical positive 3-frame associated with each monosaccharide which could be useful in studying the geometry of polysaccharide interactions.

In contrast, matrix models may not be of much use in realistic biology since interesting biologically active roles of sugars, cellular signaling for example, depend on exquisitely refined structure that has been evolutionarily selected, and certainly not on bulk properties of ensembles of polysaccharides as captured by matrix models. However, here is a problem in physical chemistry to which matrix models may apply:

Take a test tube mostly filled with water and add a selection of oligosaccharides according to some specified recipe. Let it be shaken and then left alone until a steady state emerges of the ensemble of mostly cyclic polysaccharides S in the test tube. We speculate that there is a version of matrix models suited to polysaccharides for studying these stable ensembles and predicting the attributes of their constituent fatgraphs. Computing the average asymptotic topological attributes of the ensemble in terms of the original recipe would be a goal beyond the abilities of current techniques in computational chemistry towards this end, which would require [131] Monte Carlo methods for Molecular Dynamics. It is not difficult to start building attributes of polysaccharides into a matrix model, but a natural and new matrix model type to describe polysaccharides has yet to be invented.

4. RNA

RNA and protein share certain properties in that each is a linear polymer of acid residues, 20 amino acids in the former case and 4 nucleic acids in the latter in the standard models. These residues interact by H-bonds in the former case and by Watson-Crick [134] or other rules in the latter–though a distinction for RNA is that only certain residue pairs can interact. On the other hand, a key difference is that it is the *geometry* of protein interactions that has been successfully probed as we have already discussed while it is the *topology* of RNA interactions that is both deeply connected with the Riemann moduli

spaces and admits effective description and computation with matrix models. A good reference for RNA chemistry and structure is [92] with [115, 132, 109, 74] providing excellent and comprehensive treatments of the RNA bioinformatics and beyond. Most of the material in this section comes from [11, 117, 104, 14].

4.1. **Chemistry and topology.** An RNA molecule is a linear polymer of nucleic acids, and a nucleic acid is one of the four[10] compounds: Adenine, Guanine, Cytosine or Uracil.

Just as for protein, there is a backbone this time given by an alternating sequence of ribose sugars[11] and phosphate groups, where each sugar is covalently bonded to a nucleic acid residue, which is also called a "base". It is the 3rd carbon in the sugar that connects to the subsequent phosphate and the 5th carbon that connects to the previous phosphate along the backbone so as before for protein, the RNA backbone comes with a canonical orientation determined by the chemistry called "from the 5' to 3' end". The "primary structure" of the RNA molecule, which completely determines it chemically, is the word in this 4-letter alphabet $\{A, G, C, U\}$ oriented from the 5' end to the 3' end of the molecule. The classical[12] Watson-Crick bonds or "base pairs" are $C - G$ arising from three H-bonds and $A - U$ arising from two. The two nucleic acid residues participating in a Watson-Crick base pair are essentially planar, and these planes stack one on top of another in a tightly folded RNA molecule. These and other aspects as well are illustrated in Figure 13a, where the backbone including the ribose sugars familiar from the previous section are depicted in blue and the nucleic acid residues in green. A chemically more detailed illustration of the residues is given in Figure 17.

In contrast to our discussion of protein, however, it is not the geometry of RNA but rather its topology that is studied here. This reflects both the relative paucity of data for folded RNA compared to protein as well as the admittedly controversial view that in some circumstances RNA has continuous moduli which proteins typically do

[10]This is just the simplest version. Indeed, a mature RNA undergoes chemical modification of these four, and furthermore, there is a fifth player called Inosine, which is part of Crick's so-called wobble model.

[11]In fact, in order to participate in the RNA backbone as well as in other important cellular mechanisms, a ribose sugar as illustrated in Figure 12b must first be phosphorylated via an ADP-ATP cycling reaction catalyzed by a so-called ribokinase enzymatic protein.

[12]Crick's wobble model allows also the non-standard pairing $G - U$ as well as $I - X$, for $X = A, C, U$. Another geometrically more refined version of base pairing will also be discussed later.

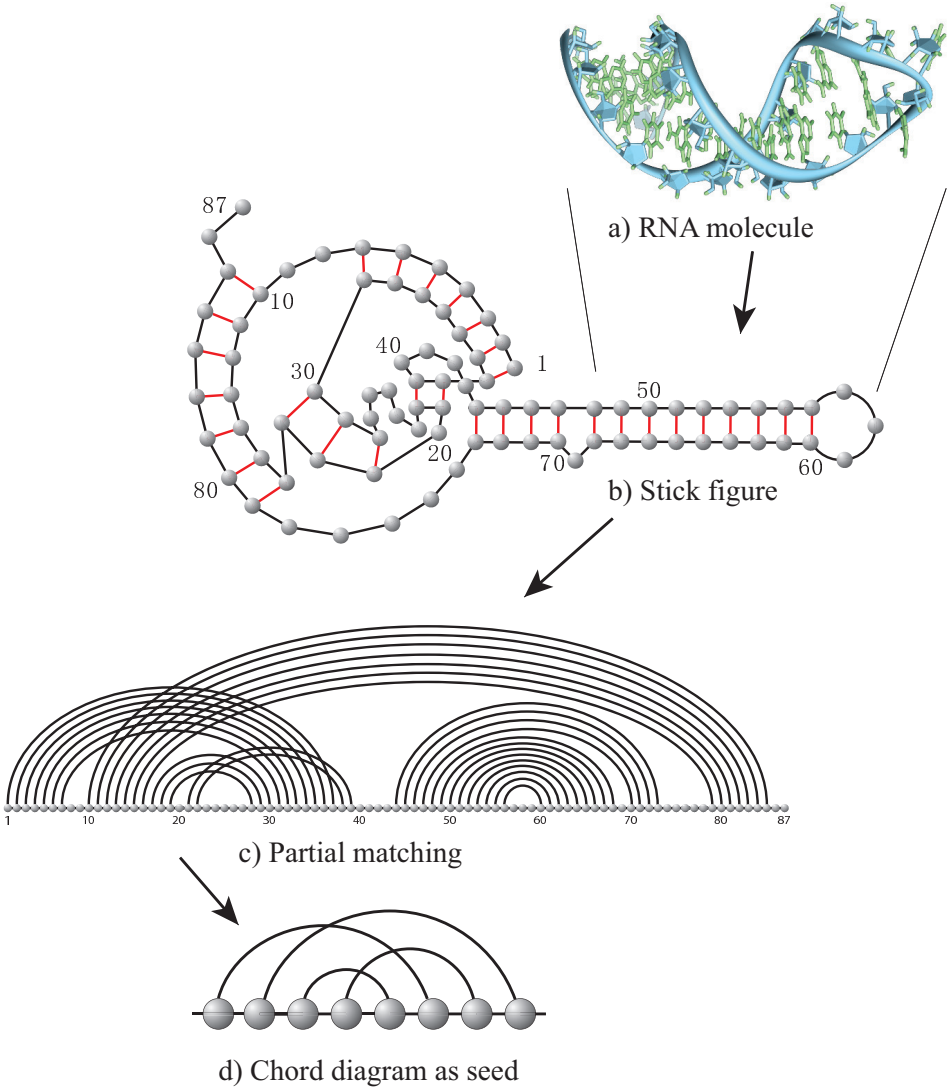


FIGURE 13. Seed of a pseudoknotted RNA molecule. Part a) © Vossman from Wikipedia taken from w:PDB ids 1sv model 1 imaged using w:UCSF Chimera. Part b) models the Watson-Crick bonds as red lines in a stick figure, part c) as a partial matching on a linear backbone, and part d) presents the associated seed.

not. The geometry of RNA is regarded as invisible to our model, and it is the pattern of H-bonding as a topological surface that is recorded as a fatgraph. Furthermore, evident hybrids of the methods could be developed to handle both the topology and geometry.

We shall find that there is a fundamental identification up to homotopy of the Riemann moduli space $M(F_{g,r})$ of a bordered surface with the moduli space of $r \geq 1$ many interacting RNA molecules with genus g as we shall define it thus directly relating two otherwise quite disparate topics. This identification is understood only on the combinatorial level presented here. Nevertheless, much of the deep structure of Riemann moduli spaces can be described purely combinatorially [104] and so can indeed be carried over straightaway to the biophysics but with unclear meaning. It is an exciting but perhaps naively optimistic hope for a deeper geometric or dynamical confluence between Riemann surfaces and RNA. Nevertheless, matrix model techniques from quantum field theory are available to enumerate and compute varied aspects of fatgraph complexes in general and these RNA moduli spaces in particular as we shall also explain.

Here is the basic model of an RNA complex[13] which originated independently in [108, 101] and [95, 25]: Consider a collection of $b \geq 1$ pairwise disjoint oriented intervals lying in the real line $\mathbb{R} \subset \mathbb{C}$, each component of which is called a *backbone*: a *chord diagram* C on these backbones is comprised of a collection of $n \geq 0$ semi-circles called *chords* lying in the upper half plane whose endpoints lie at distinct interior points in the backbones *so that the resulting diagram is connected*. This description of C with its chords in the upper half plane automatically determines a corresponding fatgraph structure on C, and as such, the skinny surface associated to C has its genus $g(C) \geq 0$ and number $r(C) \geq 1$ of boundary components. Chords in C can represent interactions between their endpoints such as Watson-Crick bonds in an RNA molecule as illustrated in Figure 13, or they may represent still other binary interactions as will be discussed later.

This chord diagram C is already a combinatorial simplification since by definition a chord diagram cannot contain isolated vertices (i.e., it is a "complete matching" rather than a "partial matching" in the parlance of classical combinatorics) in contrast to the evident biophysics as illustrated in Figure 13b and c. A chord $c \subset C$ with endpoints $x < y$ is called a *non-crosser* if x and y lie in a common backbone and there is no chord in C with exactly one endpoint between x and y. A *stack* in a chord diagram C is a collection of $m \geq 1$ chords $c_1, \ldots, c_m$ where c_j has endpoints x_j, y_j so that $x_1 < \ldots < x_m$ and $y_m < \ldots y_1$ each comprise consecutive vertices lying in a common backbone of C. Geochemically,

[13]Complexes comprised purely of RNA are not common in nature although so-called antisense RNA [26, 50, 98] provides an important example. Nevertheless, more biologically realistic models of several RNAs in complex with proteins or sugars of course then restrict to the purely RNA sub-complex described here.

a stack represents a helix of base pairs, and for example, the Watson-Crick bonded nucleic acids indeed "stack" upon one another in space as already mentioned.

We simplify C further still by removing all non-crossers and collapsing each stack to a single chord in order to define the essential combinatorial abstraction called a "seed" as depicted in Figure 13d. The utility of seeds arises from basic combinatorics: the set of all partial matchings on a collection of backbones can be efficiently enumerated from seeds by a process of "inflation" that inserts isolated points as well as planar sub-diagrams consisting entirely of non-crossers and expands arcs to stacks. One can furthermore specify two parameters: the minimum size of a stack and the minimum distance along the backbone between endpoints of a chord. The interested reader is referred to [14].

4.2. **RNA Folding.** The enumeration of partially matched RNA structures of fixed genus g was obtained in [128] using matrix models, and a closed-form expression for the number of such partial chord diagrams was given in [38] in terms of Stirling numbers of the first kind. Chord diagrams without isolated vertices of fixed genus g are likewise computed in [11], cf. the Appendix. Moreover, second-order, non-linear, algebraic partial differential equations on generating functions for chord diagrams as well as partial chord diagrams based on lengths of boundary cycles are derived in [7]. These lead to efficient enumerative methods in both the orientable and non-orientable settings. The scheme for partial matchings is furthermore formulated as a matrix model whose planar limit is computed using techniques of free probability.

In [14], the minimum stack size σ of RNA structures is taken into account, the generating function of those structures is computed and it is found that the genus only enters through a sub-exponential factor for $\sigma \geq 2$. This slow growth rate compared to the number of RNA molecules implies the existence of so-called neutral networks of distinct RNA molecules with the same structure of any genus.

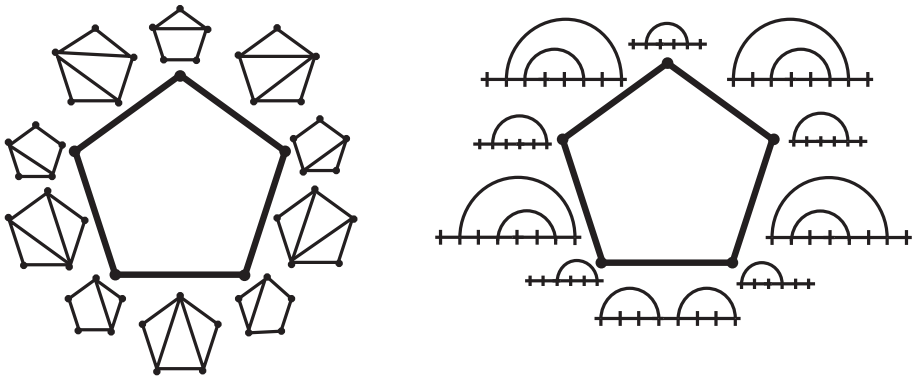
Enumeration of seeds is already a combinatorially interesting problem especially owing to the utility of matrix models to actually compute in practice as well as the intimate relationship with the Riemann moduli spaces of bordered surfaces that we shall describe. However, beyond the purely combinatorial aspect of enumeration lies the folding problem for RNA of interest biologically: fix the primary structure of a single RNA molecule or a family of them and predict the partial matching that corresponds to its pattern of Watson-Crick or other bonding. In the associated bioninformatics problem, the folding is determined by minimizing some specified free energy function.

The vast majority of folded single RNA molecules appears to have genus zero, hence a planar chord diagram C on one backbone with all arcs non-crossers. In this special case of planar chord diagrams on a single backbone, there is the ansatz of a particular free energy function due to Waterman [133] in the 1970s which is roughly modeled on the idea that as for protein the H-bonds should be as saturated as possible since they are quite energetically favorable especially in this instance when they occur in larger stacks–now subject to the Watson-Crick rules however–and that the backbone itself has a tensile strength and cannot bend much. This leads to dynamic programming methods to minimize this free energy, namely, compute the minimum free energy (MFE) solutions as well as their corresponding Boltzman partition functions. These methods are $O(N^3)$ in time and $O(N^2)$ in space, where N is the number of bases in the RNA. In fact, the method provides a collection of solutions with free energy near the minimum, say in the five-percent band of energy around the MFE, and the "correct solution" always lies in this five-percent band. In this limited sense, the *planar RNA folding problem for a single backbone* is sometimes regarded as solved.

However, there are a number of examples of non-planarity in the Watson-Crick bonding, and they are typically of biological significance. A chord c in the chord diagram C is said to form a *pseudoknot*[14] if the chord diagram $C - \{c\}$ has a different genus than that of C, and C is said to be *pseudoknotted* if it contains a pseudoknot. This is a precise definition of the term pseudoknot as used in the literature here based on the fatgraph topology. It is a misnomer since there is no natural pseudoknot-free sub diagram of a general chord diagram, but the planar sub-bonding of a folded RNA molecule is nevertheless often called its "secondary structure".

Example 4.1. (Moduli space of all RNA secondary structures [108]) Consider a polygon F, namely, an unpunctured surface of genus zero with one boundary component containing n distinguished points, where we shall assume that $n \geq 3$. There is a standard CW complex $\mathcal{A} = \mathcal{A}(F)$ called the *arc complex* of F where there is one $(k-1)$-simplex in $\mathcal{A}$ for each k-tuple of distinct diagonals in F which are pairwise disjoint except perhaps at their endpoints, and the face relation is generated by erasure of diagonals as before. One of the many interpretations of the Catalan numbers [124] is given by the number of top-dimensional cells in $\mathcal{A}$, i.e., the number of triangulations of F.

[14]The term comes from the fact that the backbone of an RNA molecule is after all usually not a closed curve but rather an interval and hence cannot be knotted, and the term pseudoknot arose in practice to describe non-planarity.



a) arc families b) seed-like partial matchings

FIGURE 14. Seed-like planar partial matchings on one backbone with $n-1$ unbonded sites are in bijective correspondence with arc families in a polygon with n sides.

The arc complex of a pentagon is illustrated in Figure 14a and is itself also a pentagon. That the arc complex $\mathcal{A}(F)$ is a sphere S^{n-4} for a polygon F with $n \geq 4$ sides is the classical case of a more general result [103]. The classical case apparently [121] was known to Hassler Whitney, and [108] gives a novel elementary proof of it. In fact, the analogous arc complex for a general surface F contains a dense open subset that is equivariantly identified [99, 102] with the Teichmüller space $T(F)$, so the quotient of the arc complex by the mapping class group $MC(F)$ provides a combinatorial compactification of the moduli space $M(F)$ of Riemann surfaces, cf. [103].

In order to exploit these combinatorics to describe the binary kind of unmatched bonding that occurs for example with classical Watson-Crick RNA, we say that a partial matching B on the collection of integral vertices lying in a single backbone interval is *seed-like* if no two bonds in B are parallel and no chord in B has its endpoints either consecutive or consecutive-but-one. As illustrated in Figure 14b and shown in [108] in somewhat greater generality up to homotopy, the geometric realization of the partially ordered set of seed-like planar partial matchings with $f \geq 4$ free or unmatched vertices is homeomorphic to a sphere S^{f-3} of dimension $f-3$.

A fundamental aspect of the full folding problem for RNA including pseudoknots, even on only one backbone, is that once the Pandora's Box of higher genus is opened, the problem of computing MFE solutions becomes NP complete [5, 84] provided that the energy function is based upon the size of stacks. Thus, we go from a reasonable polynomial time

solution directly to NP hard in one go. One must therefore limit the topology in some way in order to permit viable algorithms to probe RNA pseudoknot folding, and this can be done in a number of ways.

The classification of chord diagrams by genus is applied to design a folding algorithm of RNA structures in [117]. The key idea in this work is to consider irreducible shapes of genus γ as building blocks under concatenation and nesting, where the number of irreducible shapes is finite. The notion of shape irreducibility is derived from the concept formulated in [77] plus work [73, 118] on pseudoknot shapes, or equivalently as we shall see, corresponds simply to having top-dimensional dual in the Riemann moduli space of the associated bordered surface. Irreducibility is furthermore equivalent to the notion of primitivity introduced by [25] as inspired by the work in [37]. For $\gamma = 1$, there are four basic irreducible shapes as first presented in [111, 25] or equivalently already in the 1990s as the top-dimensional cells in $M(F_{1,1})$.

The γ-structures [117] whose irreducible shapes have genus at most $\gamma = 1$ are built by concatenating and nesting these four basic irreducible shapes. Associating a specific energy to each irreducible shape allows one to compute a specific MFE akin to that of [133] and again minimize using dynamic programing algorithms for the associated multiple context-free grammar providing also the Boltzman partition function and base pairing probabilities; this is implemented in the software package `gfold` which runs in $O(N^6)$ time and $O(N^4)$ space in the number N of bases. This algorithm `gfold` is different in kind from the work in [110] which only generates RNA structures of genus one without any loop-based energy model, cf. also [49]. Further combinatorics of γ-structures are studied in [56, 82].

A chord diagram with two backbones is sometimes called an "interaction structure", and these are important in principle for studying antisense RNA [26, 50, 98] for example. The analogue [13] of `gfold` for γ-interaction structures, this time with $\gamma = 0$ having seven irreducible seeds, again runs in $O(N^6)$ time and $O(N^4)$ space. The generating function of γ-interaction structures is shown to be algebraic in [112], which implies that their number and its asymptotics can be computed by singularity analysis.

In the literature, there are other such heuristics [61, 119] as well as other approaches [115, 72] to predicting RNA folding with pseudoknotting. For the general class of so-called 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 [115].

4.3. **Chord diagrams, seeds and shapes.** Recall that the so-called Catalan numbers

$$c_n = \frac{1}{n+1}\binom{2n}{n} = \frac{(2n)!}{(n+1)!\,n!} = \prod_{k=2}^{n} \frac{n+k}{k}, \qquad \text{for } n \geq 0,$$

count

$$c_n = \#\{\text{triangulations of a polygon with } n+2 \text{ sides}\}.$$

They furthermore satisfy the recursion $c_{n+1} = \sum_{i=0}^{n} c_i c_{n-i}$ with $c_0 = 1$. In effect, we shall study here the possibly higher genus and possibly many backbone generalization of this classical sequence. Let $c_{g,b}(n)$ denote the number of fatgraph isomorphism classes of chord diagrams of genus g with n chords on b labeled backbones with generating function

$$C_{g,b}(z) = \sum_{n \geq 0} c_{g,b}(n)\ z^n, \text{ for } g \geq 0.$$

Two chords c and c' in a chord diagram C with respective endpoints x, y and x', y' are *parallel* if x, x', as well as y, y', lie in a common backbone with no chord endpoints in between, where $x < x'$ and $y' < y$. Parallelism generates an equivalence relation whose equivalence classes are called *stacks* as before.

Suppose that the endpoints x, y of a chord c in C lie in a common backbone β. If there are no chord endpoints between x, y, then c is called a *1-arc* on β. At the other extreme, if all chord endpoints in β lie between x and y, then c is called a *rainbow* on β.

A *seed* is a chord diagram where every stack has cardinality one and each 1-arc is a rainbow. A seed is a *shape* provided every backbone has a rainbow. For a planar chord diagram, an innermost chord with both endpoints on a single backbone is necessarily a 1-arc, so the only seeds of genus zero on one backbone are the empty diagram with no chords and the diagram with a single chord given by the rainbow, and only the latter is also a shape.

Proposition 4.2. *Other than the case $g = 0, b = 1$ just discussed, a seed of genus g on b backbones with n chords must satisfy the constraints $2g + b - 1 \leq n \leq 6g - 6 + 5b$, and these inequalities are sharp. In particular, there are only finitely many seeds or shapes for fixed g, b.*

Proof. The lower bound on n follows from $2g + b - 1 \leq n + 1 - r$ according to the Euler characteristic plus the obvious constraint $r \geq 1$. Conversely, a seed which saturates this lower bound has $r = 1$.

If the skinny surface associated to a seed has more than one boundary component, then there must be a chord with different boundary components on its two sides by connectivity of chord diagrams; removing

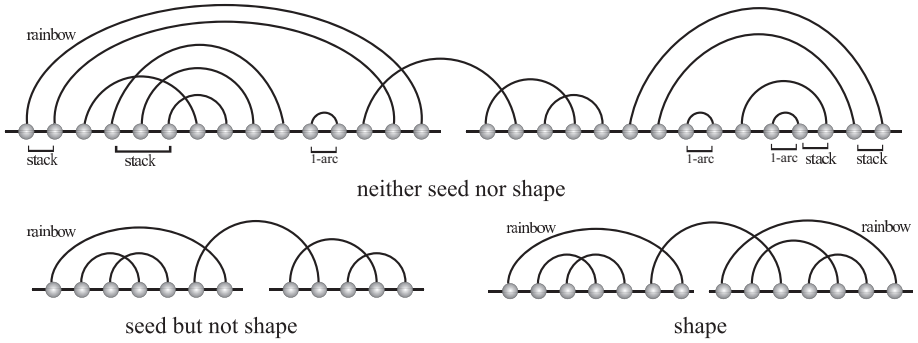


FIGURE 15. Stacks, 1-arcs, rainbows, shapes, seeds from [11].

this chord decreases r by exactly one preserving g again from consideration of the Euler characteristic. Define the "length" of a boundary cycle to be the number of chords it traverses counted with multiplicity. If there are ν_ℓ boundary cycles of length ℓ, for each ℓ, then $2n = \sum_\ell \ell\nu_\ell$ since each side of each chord is traversed exactly once in the boundary. It follows that $2n = 2(b+r+2g-2) \geq \nu_1+2\nu_2+3(r-\nu_1-\nu_2) \geq 3r-2b$ since $2\nu_1+\nu_2 \leq 2b$ (except in the excluded case for which $2\nu_1+\nu_2 \leq 4b$). Thus, we have $4(b+g-1) \geq r$, and there can thus be at most $4g+4b-5$ such removals of chords to produce a seed with $r = 1$ providing the upper bound on n. □

Let $c_{g,b}(n)$, $s_{g,b}(n)$ and $t_{g,b}(n)$, respectively, denote the number of isomorphism classes of connected chord diagrams, seeds and shapes of genus $g \geq 0$ with $n \geq 0$ chords on $b \geq 1$ backbones. In each alphabetic case of $X = C, S, T$, consider generating functions

$$X_{g,b}(z) = \sum_{n\geq 0} x_{g,b}(n)\ z^n.$$

Whereas $C_{g,b}(z)$ is a formal power series, $S_{g,b}(z)$ and $T_{g,b}(z)$ are polynomials by Proposition 4.2 for each fixed g, b. Indeed, $C_0(z) = C_{0,1}(z)$ is the generating function for the Catalan numbers, the recursion for which gives $C_0(z) = 1 + z[C_0(z)]^2$, and hence $C_0(z) = \frac{1-\sqrt{1-4z}}{2z}$ by solving the quadratic.

Theorem 4.3. *The generating functions for seeds and chord diagrams are related by*

$$C_{g,b}(z) = [C_0(z)]^b\ S_{g,b}\left(\frac{C_0(z)-1}{2-C_0(z)}\right),$$

$$S_{g,b}(z) = \left[\frac{z+1}{1+2z}\right]^b C_{g,b}\left(\frac{z(1+z)}{(1+2z)^2}\right),$$

and the generating functions for seeds and shapes are related by

$$(1+z)^b\ T_{g,b}(z) = z^b\ S_{g,b}(z).$$

Proof. Writing simply C_0 for $C_0(z)$ and using $C_0 - 1 = zC_0^2$, we have

$$\frac{C_0-1}{2-C_0} = \frac{C_0-1}{1-(C_0-1)} = zC_0^2 \sum_{i\geq 0}(zC_0^2)^i = \sum_{j\geq 1}(zC_0^2)^j.$$

The jth term in the sum corresponds to inflating a single arc in a seed to a stack of cardinality $j \geq 1$ as well as inserting a genus zero diagram immediately preceding along the backbone each of the resulting $2j$ chord endpoints. Still another factor C_0 arises from the insertion of a genus zero diagram following the last endpoint of the seed on each backbone accounting for the further factor C_0^b. The seed is connected if and only if so too is the resulting chord diagram proving the first formula.

For the second formula, direct calculation shows that $z = \frac{u(1+u)}{(1+2u)^2}$ inverts the expression $u = \frac{C_0(z)-1}{2-C_0(z)} = \frac{1-\sqrt{1-4z}}{2\sqrt{1-4z}}$. The first formula therefore reads

$$S_{g,b}(u) = \left[C_0\left(\frac{u(1+u)}{(1+2u)^2}\right)\right]^{-b} C_{g,b}\left(\frac{u(1+u)}{(1+2u)^2}\right).$$

Direct computation substituting $z = \frac{u(1+u)}{(1+2u)^2}$ into $C_0(z) = \frac{1-\sqrt{1-4z}}{2z}$ gives $C_0\big(\frac{u(1+u)}{(1+2u)^2}\big) = \frac{1+2u}{1+u}$, and the expression for $S_{g,b}$ then follows.

The third formula is truly elementary since a shape is by definition simply a seed where there is a rainbow on each backbone. □

Suppose that C is a shape of genus g on b backbones. Removal of certain chords can separate C, and removal of any chord of C other than a pseudoknot preserves the genus by definition. Thus, we may remove any non-separating, non-pseudoknot and non-rainbow chord in order to produce another shape from C of the same genus g.

Let us build a combinatorial space $\mathcal{R}_{g,b}$, where there is one $(n-1)$-dimensional simplex for each such shape C with n chords. Certain faces of C correspond to shapes of the same genus as just discussed, and these faces are to be identified with the simplices associated to these sub shapes. The other faces of the simplex for C are simply absent to produce the non-compact space $\mathcal{R}_{g,b}$.

As before in more biophysical words, assign to each chord in a shape a real-valued free energy of the bond and projectivize by the homothetic action of $\mathbb{R}_+$ on all these weights simultaneously where erasing an edge corresponds to the vanishing of its projectivized free energy. The *moduli space $\mathcal{R}_{g,b}$ of RNA shapes* can alternatively be described

simply as the space of all projectively weighted shapes of genus g on b backbones, where vanishing projective free energy of an arc corresponds to its removal, with its natural combinatorial structure.

Theorem 4.4. *Provided $g + b - 1 > 0$, the moduli space $\mathcal{R}_{g,b}$ of RNA shapes of genus g on b backbones is combinatorially isomorphic to the Riemann moduli space $M(F_{g,b})$ for a surface $F_{g,b}$ of genus g with b boundary components up to homotopy.*

In light of this result, the explicit calculation of the generating function for shapes discussed here thus simultaneously gives as a consequence the numbers of cells of fixed dimension in the Riemann moduli spaces of bordered surfaces.

Proof. Given a shape C on $b \geq 1$ backbones, we may collapse each backbone to a distinct point in the natural way to produce a fatgraph τ with b vertices. C and τ have the same Euler characteristic, number of boundary components and hence genus.

Notice that τ has b boundary cycles of length one arising from the rainbows of C, and these are the unique boundary cycles of length one since a shape has no other 1-arcs. Furthermore, since a shape has no two parallel chords, τ can have no boundary cycles of length two. It follows that other than its boundary cycles of length one coming from the rainbows, every other boundary cycle of τ must have length at least three. We may uniquely reconstruct the shape C from the fatgraph τ by expanding each vertex to a backbone so that its unique boundary cycle of length one becomes a rainbow.

As already discussed, the fatgraph τ with m edges may be described by a pair σ, ι of permutations on $2m$ letters identified with the half-edges of τ, where σ is the composition of one disjoint k-cycle for each k-valent vertex of τ corresponding to the cyclic orderings, and ι is the composition of m disjoint transpositions permuting the two half-edges contained in an edge. Furthermore in this representation, the boundary cycles of τ correspond to the cycles of the composition $\rho = \sigma \circ \iota$.

Let τ' be the fatgraph corresponding to the composition $\rho = \sigma \circ \iota$ and the same ι. The boundary cycles of τ' correspond to the vertices of τ and conversely. Let v, e, r, g and v', e', r', g' denote the respective numbers of vertices, edges, boundary cycles and the genus of τ and τ'. Thus, $v = r'$, $r = v'$, and moreover $e = e'$ by construction, so we conclude $g = g'$. In fact, τ and τ' are related by Poincaré duality in the closed surface of genus g. In light of the constraints on τ already articulated since it arises from the shape C, the fatgraph τ' has all

its vertices of valence at least three except for a unique tail on each boundary component.

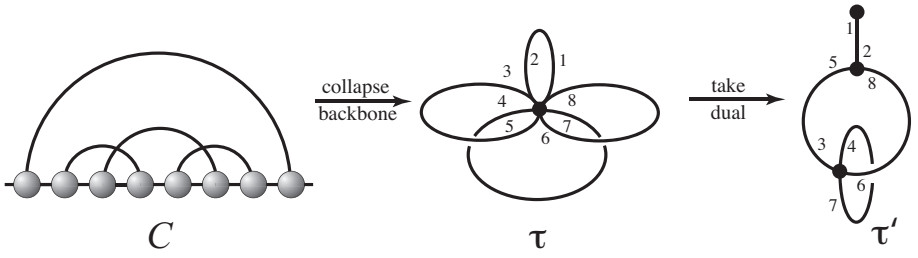


FIGURE 16. Dual fatgraph τ' of a shape C from [11].

An example on one backbone is illustrated in Figure 16. Beginning with the shape C, collapse its backbone to a vertex to produce the fatgraph τ which is described by the cycle $\sigma = (1, 2, 3, 4, 5, 6, 7, 8)$ for its single vertex plus involution $\iota = (1, 2)(3, 5)(4, 7)(6, 8)$ with one transposition for each edge of τ. It follows that $\rho = \sigma \circ \iota = (2, 5, 8)(3, 7, 6, 4)$ fixing 1. The depicted fatgraph τ' whose vertices are described by ρ and whose edges are again given by ι is dual to τ.

This provides a mapping from shapes C to the required fatgraphs τ' as asserted. The inverse mapping is given by the same involution $\sigma \mapsto \sigma \circ \iota$, $\iota \mapsto \iota$ followed by expansion of the vertex to a backbone so that each cycle of length one becomes a rainbow. Note that the face relation of removal of chords in the RNA moduli space $\mathcal{R}_{g,b}$ is dual to contraction of edges in the Riemann moduli space $M(F_{g,b})$. □

4.4. **Further remarks on RNA.** There is another well-known [83, 8] approach to RNA combinatorics and geometry as follows. Early on, in addition to observing the predicted geometry of Watson-Crick base pairs, a different geometry for H-bonds between base pairs was observed by Hoogsteen [64]. Subsequent analysis furthermore showed that H-bonds also form with the backbone ribose sugars in a structured RNA molecule. The model due to Leontis-Westhof [83] which we next describe provides a classification of this more elaborate RNA bonding. There is a triangular region t around each nucleic acid base that supports the H-bonding as illustrated in Figure 17A, where the edge of t adjacent to the nucleic acid base supports the usual Watson-Crick base pairs themselves, which we recall are comprised of two or three H-bonds. In the so-called *anti* conformation, the edge of t towards the 5' end supports the Hoogsteen base pairs and towards the 3' end supports the sugar bonds with these roles reversed if t is in the

less common *syn* conformation. A crucial point is that this description is roughly faithful biophysically thus providing a starting point for describing the geometry.

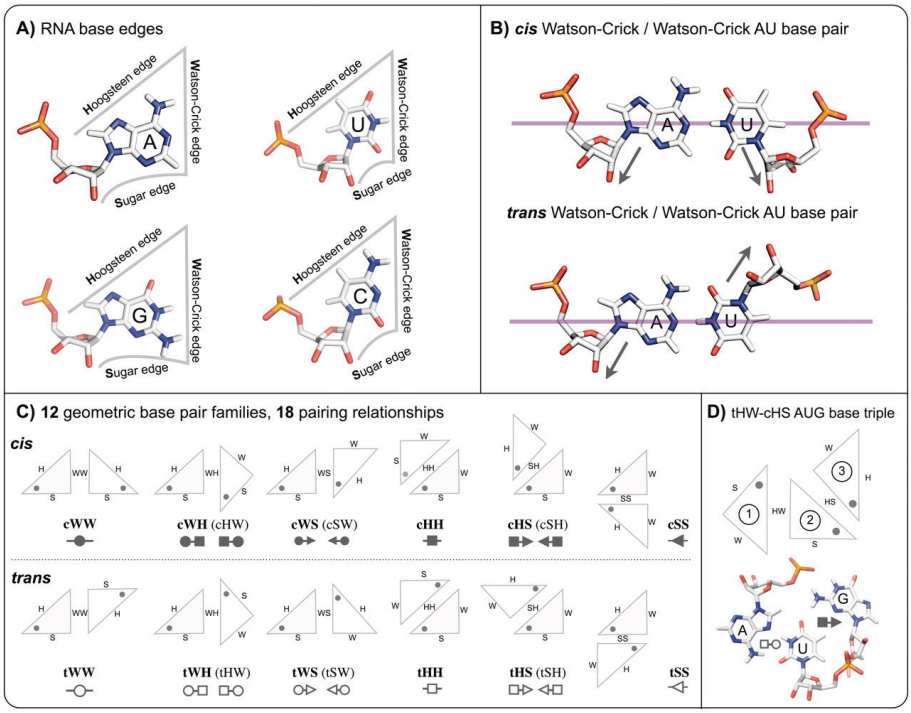


FIGURE 17. Original figure from [8] with permission summarizing the Leontis-Westhof base pairing classification. (A) Each RNA nucleotide displays three edges for base pairing interactions represented by triangles as shown. (B) For each pair of edges, nucleotides can pair in two distinct ways designated cis and trans and related by 180° rotation of one nucleotide about the magenta axis. (C) Schematic representations of the 12 basic base pair families associating circles with W edges, squares with H edges and triangles with S edges. Filled in and open symbols respectively represent cis and trans base pairs. The 12 base pair families result in 18 ordered base pairing relations. (D) A representative regular base triple denoted AUG tHW/cHS.

The analogue of chord diagrams in this context replaces each 3-valent vertex of a chord diagram by a 5-valent vertex, where the three non-backbone half-edges at any vertex occur in either the clockwise order *SWH* or *HWS* starting from the 5' end depending on the respective *syn* or *anti* conformation and where *H* stands for Hoogsteen, *W* for Watson-Crick and *S* for sugar bonds. We add chords to this model of the backbone for the various types of bonds illustrated in Figure 17B in the natural way and call this the *trichord RNA diagram*. This

fatgraph model of RNA naturally supports an $SO(3)$ graph connection giving the plane of a suitably defined ideal Leontis-Westhof triangle which could be compared to RNA physical chemistry analyses in the literature. It may be interesting to likewise study the planes determined by the ribose sugars along the backbone. There is furthermore a matrix model enumerating trichord diagrams which might be computed.

This trichord model of a single RNA molecule has been employed in a fascinating recent study by Ebbe Andersen and Piotr Sułkowski [10] as follows. Consider the subinterval of the backbone that contains the first M bases starting from the 5' end. Add the chords both of whose endpoints occur in this interval to produce a sequence of trichord diagrams T_M. Taken together, the bonds are sufficiently pervasive that the genus of T_M in general grows roughly linearly in M however with a finite set of jumps at certain residues. An exciting aspect of this study is that these jump points of genus seem to be preserved across species for ribosomal RNA; as might be expected, these conserved jump points occur at the so-called ribosomal domain boundaries, but they also occur elsewhere along the RNA with as yet undetermined significance.

We turn away now from the Leontis-Westhof RNA base pair classification and nomenclature and specialize to only one backbone. In this case with $g \geq 1$, we set $c_g(n) = c_{g,1}(n)$ and have the remarkable formula of Harer-Zagier [59]:

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

or equivalently, the $c_g(n)$ satisfy the recursion

$$(n+1)\, c_g(n) = 2(2n-1)\, c_g(n-1) + (2n-1)(n-1)(2n-3)\, c_{g-1}(n-2).$$

From this recursion, it is not difficult to show that the generating function for $g \geq 1$ is of the form $C_g(z) = P_g(z)/(1-4z)^{3g-\frac{1}{2}}$, where $P_g(z)$ is an integral polynomial divisible by z^{2g} but no higher power and of degree at most $(3g-1)$ with $P_g(\frac{1}{4}) \neq 0$. The first few $P_g(z)$ are

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

Needless to say, polynomials like these with non-negative integral coefficients might reasonably be expected to be generating polynomials for some as yet unknown fatgraph structure.

Moreover, in complete generality for an arbitrary number of backbones as long as it is not the Catalan case $g = 0$, $b = 1$, then techniques of so-called topological recursion (cf. the Appendix) can be used [43] to prove that indeed

$$C_{g,b}(z) = \frac{P_{g,b}(z)}{(1-4z)^{3g-3+\frac{5}{2}b}} ,$$

where $P_{g,b}$ is a polynomial with $P_{g,b}(\frac{1}{4}) \neq 0$. This gives no clue, however, as to the non-negative integrality of the coefficients of $P_{g,b}(z)$, a fact which has tantalizingly resisted proof for some time but which holds pervasively in computer experiments and which we conjecture should follow from a suitable fatgraph interpretation of the coefficients.

There is in fact a huge literature on this Harer-Zagier recursion, and there has recently been its first constructive proof in [30] that nevertheless fails to derive it from first combinatorial principles. We shall describe the main ideas from [30] and their consequences for RNA here and refer the interested reader to [30] and the references therein.

A fatgraph τ with only one boundary component is sometimes called a "unicellular map". Such are therefore by definition dual to a fatgraph of the same genus with a single vertex, i.e., a chord diagram on a single backbone. The paper [30] introduces a novel bijection between unicellular maps of genera g and $g - k$ by a specific slicing/gluing process on vertices as follows.

The slicing process splits a vertex into $2k + 1$ separate vertices while edges are preserved, thereby reducing the genus by k, for some k, and the gluing process is the inverse map of gluing $2k + 1$ distinct vertices into one. Both processes preserve unicellularity. Indeed, suppose that the unicellular map τ is described as a fatgraph by the two permutations σ for its vertices and ι for its edges and we $\rho = \sigma \circ \iota$ corresponding to the single boundary cycle. The key point is that given a half-edge r, the permutations ρ and σ each induce a natural strict linear ordering on certain sets of half-edges, namely,

$$r <_\rho \rho(r) <_\rho \cdots <_\rho \rho^{2n-1}(r) \text{ and } r <_\sigma \sigma(r) <_\sigma \cdots <_\sigma \sigma^k(r).$$

A half-edge $r >_\gamma \sigma(r)$ where $\sigma(r)$ is not the $<_\sigma$-minimum half-edge at the vertex is called a *trisection*. The trisection records σ- and ρ-order violations and is considered an indicator of genus since there are in fact exactly $2g$ trisections in a unicellular map of genus g. Iteratively slicing a unicellular map of genus g finally produces a planar tree.

This bijection [30] has both enumerative and algorithmic application since pseudoknotted RNA structures can be processed as if they were pseudoknot-free plus additional gluing information. This provides in particular linear time uniform sampling and generating algorithms for RNA structures of fixed genus [67, 68] and facilitates a novel stochastic context-free grammar [116] for RNA structures including pseudoknots.

In an entirely different direction based on trisection, signed permutations and the reversal action on them have been widely studied in evolutionary biology, cf. [57]. A fatgraph formalism for them has been introduced in [69] including an extension of the trisection technique.

Despite our caveat in the first paragraph of the Introduction to humbly respect the biology, we are aware of the danger that the current speculations for RNA threaten to represent bioinformatics or biophysics for its own sake. The enumeration of RNA seeds or shapes certainly likewise threatens to provide mathematics for its own sake with the suggestion of biological relevance coming from the prediction of RNA folding based on an appropriate free energy. This may indeed be the case for certain of the mRNA, tRNA or rRNAs involved in protein expression, however, regulatory RNAs such as snoRNAs must surely fold according to other as yet undetermined rules or mechanisms which might still be illuminated by topological considerations.

Appendix. Matrix models

Matrix model techniques from quantum field theory [23, 87, 36, 4] have provided effective computational tools in practice in numerous instances of geometry and theoretical physics over the last 30 years including [100, 80, 88, 71, 42]. Gian-Carlo Rota prophetically christened his journal Advances in Applied Mathematics issue one of volume one with these methods hoping [120] to transport these powerful combinatorial tools from high energy physics to mainstream mathematics.

We shall describe here an application of these methods to an effective computation of the $c_{g,b}(n)$ in §4.3 from [11], which is surveyed in [12]. These quantities also arise from the multi-resolvents of the simpler Gaussian density, but our methods encode at once $c_{g,b}(n)$ for all $b \geq 1$. Suppose that C is a disjoint union of chord diagrams and let

$b(C)$ denote its number of backbones,
$n(C)$ denote its number of chords,
$r(C)$ denote its number of boundary cycles,
$Aut(C)$ denote its automorphism group possibly permuting oriented backbones, and
$g(C)$ denote the sum of the genera of its components.

For any tuple $v_1, v_2, \ldots, v_K$ of non-negative integers, define

$$P_{v_1,\ldots,v_K}(s,t,N) = \frac{1}{\prod_k v_k!} \int_{\mathcal{H}^N} e^{-\mathrm{tr}\ H^2/2} \prod_k \Big(s\ \mathrm{tr}\ (tH)^k \Big)^{v_k}\, dM$$

for the integral over the $N \times N$ Hermitian matrices $\mathcal{H}^N$ with respect to the normalized Haar measure

$$dM(H) = \frac{1}{2^{N/2}\pi^{N^2/2}} \left(\bigwedge_{i=1}^{N} dH_{ii} \right) \wedge \left(\bigwedge_{i<j} dReH_{ij} \wedge dImH_{ij} \right),$$

where tr denotes the trace.

Lemma A.1 *For any tuple $v_1, \ldots, v_K$, parameters s, t and natural number N, we have*

$$P_{v_1,\ldots,v_K}(s,t,N) = \sum \frac{N^{r(C)}\ s^{b(C)}\ t^{2n(C)}}{\#Aut(C)},$$

where the sum is over all isomorphism classes of disjoint unions of chord diagrams with v_k backbones having k incident half-chords, for $k = 1, \ldots, K$, and $\#$ denotes cardinality.

Proof. The special case $s = t = 1$ is completely analogous to that of Theorem 2.1 of [100] which gives an elementary and self-contained proof. The replacement of $\frac{1}{k}$ tr H^k there by tr H^k here corresponds to replacing fatgraphs there by chord diagrams here insofar as distinguishing one sector at each vertex of the fatgraph to represent the location of the backbone in the chord diagram kills any cyclic permutation about the vertices. Since $b(C) = \sum_k v_k$ and $2n(C) = \sum_k kv_k$, the general case then follows. □

An easy argument as on p. 49 of [100] then gives:

Theorem A.2 *For the* partition function

$$Z(s,t,N) = \int_{\mathcal{H}^N} e^{-N\text{tr } V(s,t,H)}\, dM$$

with potential

$$V(s,t,H) = \frac{1}{2}H^2 - \frac{stH}{1-tH} = \frac{1}{2}H^2 - s\sum_{k\geq 1}(tH)^k,$$

we have

$$log\ Z(s,t,N) = \sum \frac{N^{2-2g(C)}\ s^{b(C)}\ t^{2n(C)}}{\#Aut(C)},$$

where the sum is over all (connected) chord diagrams C.

Meanwhile in theoretical physics, there has been spectacular progress over the last decade in actually computing matrix models in general using a new technique called topological recursion [32, 33, 42]. In order to apply this in the current situation, our partition function $Z = \int_{\mathcal{H}^N} e^{-N\text{tr } V(H)}\, dM$ can be written

$$Z = \exp \sum_{g=0}^{\infty} N^{2-2g} F_g,$$

in terms of the so-called free energies F_g in genus g which are given by

$$F_g(s,t) = \frac{B_{2g}}{2g(2g-2)} + \sum_{b\geq 1} \frac{s^b}{b!}\ C_{g,b}(t^2),\ \text{for } g \geq 0,$$

where the constant terms, with appropriate modifications for $g = 0, 1$, reproduce the Gaussian free energies [54], B_{2g} denoting Bernoulli numbers. The free energies had previously been computed in the two cases $g = 0$ [27, 85] and $g = 1$ [31].

The topological recursion is applied in [11] to compute F_g explicitly for $g \leq 3$ and in principle in complete generality for any $g \geq 0$. For example in genus two, we find after much computation that

$$\begin{aligned}
F_2 = &-\frac{t^4(1-\sigma)^2}{240\delta^4(1-\delta-4\sigma+3\sigma^2)^5(1+\delta-4\sigma+3\sigma^2)^5} \\
&\Big[160\delta^4(1-3\sigma)^4(1-\sigma)^6 - 80\delta^2(1-3\sigma)^6(1-\sigma)^8 \\
&+ 16(1-3\sigma)^8(1-\sigma)^{10} + \delta^{10}(-16+219\sigma-462\sigma^2+252\sigma^3) \\
&+ 10\delta^6(1-3\sigma)^2(1-\sigma)^4(-16-126\sigma-423\sigma^2+2286\sigma^3-2862\sigma^4+1134\sigma^5) \\
&+ 5\delta^8(1-\sigma)^2(16+189\sigma-2970\sigma^2+9549\sigma^3-11286\sigma^4+4536\sigma^5)\Big],
\end{aligned}$$

where $\sigma = t(a+b)/2$, $\delta = t(a-b)/2$ and a, b satisfy

$$\begin{aligned}
0 &= a + b + \tfrac{st(at+bt-2)}{\left((at-1)(bt-1)\right)^{3/2}}, \\
16 &= (a-b)^2 + \tfrac{4s\left((2-\frac{(a+b)t}{2})(at+bt-2)+2abt^2-3t(a+b)+4\right)}{\left((at-1)(bt-1)\right)^{3/2}}.
\end{aligned}$$

This can be solved exactly in one of s, t and perturbatively in the other to extract specific $C_{g,b}(z)$ as in [11], for example

$$\begin{aligned}
C_{2,1}(z) &= \frac{21z^4}{(1-4z)^{\frac{11}{2}}}(z+1), \\
C_{2,2}(z) &= \frac{z^5}{(1-4z)^8}(1485+6096z+1696z^2), \\
C_{2,3}(z) &= \frac{6z^6}{(1-4z)^{\frac{21}{2}}}(15015+137934z+197646z^2+27592z^3), \\
C_{2,4}(z) &= \frac{144z^7}{(1-4z)^{13}}(38675+620648z+2087808z^2+1569328z^3+134208z^4).
\end{aligned}$$

References

[1] L. V. Ahlfors, Conformal Invariants: Topics in Geometric Function Theory, American Mathematical Society, Chelsea Publishing, 1973.

[2] L. V. Ahlfors and L. Sario, Riemann Surfaces, Princeton University Press, 1960.

[3] I. J. R. Aitchison and A. J. G. Hey, Gauge Theories in Particle Physics, 2nd edition, Institute of Physics Publishing, Bristol and Philadelphia, 1989.

[4] The Oxford Handbook of Random Matrix Theory, eds. G. Akemann, J. Baik, P. Di Francesco, Oxford University Press, 2011.

[5] T. Akutsu, Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots, *Discr. Appl. Math.* **104** (2000), 45-62.

[6] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J. D. Watson, Molecular Biology of the Cell, third edition, Garland Publishing, 1994.

[7] N. V. Alexeev, J. E. Andersen, R. C. Penner, P. G. Zograf, Enumeration of chord diagrams on many backbones and their non-orientable analogs, preprint (2014).

[8] Comprehensive survey and geometric classification of base triples in RNA structures, A. S. A. Almarken, A. I. Petrov, J. Stombaugh, C. L. Zirbel, N. B. Leontis, *Nucleic Acids Research* **40** (2012), 1407-1423.

[9] S. F. Altschul, M. S. Boguski, W. Gish, J. C. Wootton, Issues in searching molecular sequence databases, *Nature Genetics* **6** (1994), 119-129.

[10] Ebbe S. Andersen and Piotr Sułkowski, private communication (2015).

[11] J. E. Andersen, L. O. Chekhov, R. C. Penner, C. M. Reidys, P. Sułkowski, Topological recursion for chord diagrams, RNA complexes, and cells in moduli spaces, *Nuclear Physics, Section B* **866** (2012), 414-443.

[12] J. E. Andersen, L. O. Chekhov, R. C. Penner, C. M. Reidys, P. Sułkowski, Enumeration of RNA complexes via random matrix theory, *Biochemical Society Transactions* **41** (2013), 652-655.

[13] J. E. Andersen, J. E. Huang, R. C. Penner, C. M. Reidys, Topology of RNA-RNA interaction structures, *Journal of Computational Biology* **19** (2012), 928-943.

[14] J. E. Andersen, R. C. Penner, C. M. Reidys, M. S. Waterman, Topological classification and enumeration of RNA structures by genus, *Journal of Mathematical Biology* (2012), 1-18.

[15] R. Baer, Isotopie von Kurven auf orientierbaren, geschlossenen Flächen und ihr Zusammenhang mit der topologischen Deformation der Flächen, *Journal für die Reine und Angewandte Mathematik* **159** (1928), 101-116.

[16] D. Baker, and A. Sali, Protein structure prediction and structural genomics, *Science* **294** (2001), 93-96.

[17] A. Beardon, The Geometry of Discrete Groups, Graduate Texts in Mathematics **91**, Springer, 1983.

[18] O. M. Becker, A. D. Mackerell Jr., B. Roux, M. Watanabe, Computational Biochemistry and Biophysics, Marcel Dekker, 2001.

[19] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. E. Meyer Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi, The Protein Data Bank: A Computer-based Archival File For Macromolecular Structures, *Journal of Molecular Biology* **112** (1977), 535.

[20] L. Bers, Uniformization, moduli, and Kleinian groups, *The Bulletin of the London Mathematical Society* **4** (1972), 257-300.
[21] L. Bers, Quasiconformal mappings, with applications to differential equations, function theory and topology. *Bulletin of the American Mathematical Society* **83** (1977), 1083-1100.
[22] L. Bers, Finite dimensional Teichmüller spaces and generalizations, *Bulletin of the American Mathematical Society* **5** (1981), 131-172.
[23] D. Bessis, C. Itzykson, J. B. Zuber, Quantum field theory techniques in graphical enumeration, *Advances in Applied Mathematics* **1** (1980), 109-157.
[24] J. Birman and B. Wajnryb, Errata: Presentations of the mapping class groups, *Israel Journal Math.* **88** (1994), 425-427.
[25] M. Bon, G. Vernizzi, H. Orland, A. Zee, Topological classification of RNA structures, *Journal of Molecular Biology* **379** (2008), 900-911.
[26] S. Brantl, Antisense-RNA regulation and RNA interference, *Biochimica et Biophysica Acta–Gene Structure and Expression* **1575** (2002), 15-25.
[27] E. Brezin, C. Itzykson, G. Parisi, J. Zuber, Planar diagrams, *Communications in Mathematical Physics* **59** (1978), 35-51.
[28] K. Burke, J. Werschnik, E. K. U. Gross, Time-dependent density functional theory: Past, present, and future, *The Journal of Chemical Physics* **123** (2005), 062206.
[29] A. J. Casson, D. P. Sullivan, M. A. Armstrong, C. P. Rourke, G. E. Cooke, The Hauptvermutung Book: A Collection of Papers on the Topology of Manifolds, edited by A. A. Ranicki, *K-Monographs in Mathematics*, 1996.
[30] G. Chapuy, A new combinatorial identity for unicellular maps via a direct bijective approach, *Advances in Applied Mathematics* **47** (2011), 874-893.
[31] L. Chekhov, Genus one correlation to multi-cut matrix model solutions, *Theoretical Mathematical Physics* **141** (2004), 1640-1653.
[32] L. Chekhov, B. Eynard, Hermitean matrix model free energy: Feynman graph technique for all genera, *Journal of High Energy Physics* **0603** (2006).
[33] L. Chekhov, B. Eynard, N. Orantin, Free energy topological expansion for the 2-matrix model, *Journal of High Energy Physics* **0612** (2006).
[34] C. Chothia and A. M. Lesk, The relation between the divergence of sequence and structure in proteins, *EMBO Journal* **5** (1986), 823-8266.
[35] P. Deligne and D. Mumford, Irreducibility of the space of curves of a given genus. *Inst. Hautes Études Scientifique Publications Mathématique* **36** (1979), 75-110.
[36] P. Di Francesco, P. Ginsparg, J. Zinn-Justin, 2D gravity and random matrices, *Physics Reports* **254** (1995), 1-133.
[37] F. J. Dyson, The s matrix in quantum electrodynamics *Physical Review* **75** (1949), 1736-1755.
[38] M. G. dell'Erba, G. R. Zemba, Thermodynamics of a model for RNA folding, *Physical Review E* **79** (2009).
[39] D. B. A. Epstein, Curves on 2-manifolds and isotopies, *Acta Mathematica* **116** (1966), 83-107.
[40] D. J. Evans and G. P. Morriss, Statistical Mechanics of Nonequilibrium Liquids, Second Edition, Cambridge University Press, 2008.

[41] B. Eynard, Invariants of spectral curves and intersection theory of moduli spaces of complex curves, *Communications in Number Theory and Physics* **08** (2014), 541-588.
[42] B. Eynard and N. Orantin, Invariants of algebraic curves and topological expansion, *Communications in Number Theory and Physics* **1** (2007), 347-452.
[43] Bertrand Eynard, private communcation (2015).
[44] B. Farb and D. Margalit, A Primer on Mapping Class Groups, Princeton University Press, 2011.
[45] A. Fathi, F. Laudenbach, V. Poenaru, Travaux de Thurston sur les surfaces. *Asterisque* **66-67**, Societe Mathématique de France, Paris, 1979.
[46] Alexei Finkelstein and Oleg Ptitsyn, Protein Physics: a Course of Lectures, Academic Press, 2002.
[47] V. Fock and A. Goncharov, Moduli spaces of local systems and higher Teichmüller theory, *Publications Mathématiques de l'IHÉS* **103** (2006), 1-211.
[48] L. R. Ford, Automorpic Functions, American Mathematical Society, Chelsea Publishing 2004.
[49] I. Garg and N. Deo, RNA matrix model with external interactions and their asymptotic behaviors *Physical Review E* **79**, (2009).
[50] J. Georg and W. R. Hess, *cis*-Antisense RNA, another level of gene regulation in bacteria, *Microbiology and Molecular Biology Reviews* **75** (2011), 286-300.
[51] W. M. Goldman, Topological components of spaces of representations, *Inventiones Mathematicae* **93** (1988), 557-607.
[52] M. Gromov, Crystals, proteins, stability and isoperimetry, *Bulletin of the American Mathematical Society* **48** (2011), 229-257.
[53] A. Grothendieck, Techniques de construction en géométrie analytique. I. Description axiomatique de l'espace de Teichmüller et de ses variantes, *Séminaire Henri Cartan* **13**, Exposés 7/8 (Paris: Secrétariat Mathématique).
[54] S. Gukov, P. Sułkowski, A-polynomial, B-model, and Quantization, *Journal of High Energy Physics* **1202** (2012).
[55] R. Hamilton The Ricci flow on surfaces, *Contemporary Math.* **71** (1988), 237-261.
[56] H. S. W. Han, T. J. X. Li, C. M. Reidys, Combinatorics of γ-structures, *Journal of Computational Biology* (2013).
[57] S. Hannenhalli and P. A. Pevzner, Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals, *Journal of the ACM* **46** (1999), 1-27.
[58] J. L. Harer, Stability of the Homology of the Mapping Class Groups of Orientable Surfaces, *Annals of Mathematics* **121** (1985), 215-249.
[59] J. Harer and D. Zagier, The Euler characteristic of the moduli space of curves, *Inventiones Mathematicae* **85** (1986), 457-485.
[60] J. Harris and I. Morrison, Moduli of Curves. Springer Verlag, Berlin, 1998.
[61] C. Haslinger and P. F. Stadler, RNA structures with pseudo-knots, *Bulletin of Mathematical Biology* **61** (1999), 437-467.

[62] N. J. Hitchin, Lie groups and Teichmüller space, *Topology* **31** (1992), 449-473.
[63] P. Hohenberg and W. Kohn, Inhomogeneous electron gas, *Physical Review B* **136** (1964) 864-871. .
[64] K. Hoogsteen, The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine, *Acta Crystllographica* **16** (1963), 907-916.
[65] W. G. Hoover, Computational Statistical Mechanics, Elsevier, 1991.
[66] W. Hu, L. Lin, C. Yang, DGDFT: A massively parallel method for large scale density functional theory calculations, *Journal of Chemical Physics* **143** (2015), 124110.
[67] F. Huang, C. M. Reidys, M. E. Nebel, Generation of RNA pseudoknot structures with topological genus filtration, *Mathematical Biosciences* **2** (2013).
[68] F. Huang and C. M. Reidys, Shape of topological RNA structures, preprint (2014).
[69] F. Huang and C. M. Reidys, A topological framework for signed permutations, preprint (2015).
[70] J. H. Hubbard, Teichmüller Theory And Applications To Geometry, Topology, And Dynamics Matrix Press, 2006.
[71] K. Igusa, Combinatorial Miller-Morita-Mumford classes and Witten cycles, *Algebraic and Geometric Topolgy* **4** (2004), 473-520.
[72] H. Isambert and E. D. Siggia, Modeling RNA folding paths with pseudoknots: Application to hepatitis virus ribozyme, *Proceedings of the National Academy of Sciences USA* **97** (2000), 6515-6520.
[73] E. Y. Jin and C. M. Reidys, Combinatorial design of pseudoknot RNA, *Advances in Applied Mathematics* **42** (2009), 35-151.
[74] N. C. Jones and P. A. Pevzner, An Introduction to Bioinformatics Algorithms, MIT Press, 2004.
[75] W. Kabsch and C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers* **22** (1983), 2577-2637.
[76] S. Kaczanowski and P. Zielenkiewicz, Why similar protein sequences encode similar three-dimensional structures?, *Theoretical Chemistry Accounts* **125** (2010), 643-650.
[77] D. Kleitman, Proportions of irreducible diagrams, *Studies in Applied Mathematics* **49** (1970, 297-299.
[78] Kobayashi and Nomizu, Foundations of Differential Geometry, vol 1, Wiley Classics.
[79] W. Kohn and L. J. Sham, Self-Consistent Equations Including Exchange and Correlation Effects, *Physical Review* **140** (1965), 1133-1138.
[80] M. Kontsevich, Intersection theory on the moduli space of curves and the matrix Airy function, *Communications in Mathematical Physics* **147** (1992), 1-23.
[81] A. Leach, Molecular Modeling: Principles and Applications 2nd Edition, Prentice Hall, 2001.
[82] T. J. X. Li and C. M. Reidys, The genus filtration of γ-structures, *Mathematical Biosciences* **241** (2013), 24-33.

[83] N. B. Leontis and E. Westhof, Geometric nomenclature and classification of RNA base pairs, *RNA* **7** (2001), 499-512.
[84] R. B. Lyngsø and C. N. Pedersen, RNA pseudoknot prediction in energy-based models, *Journal of Computational Biology* **7** (2000), 409-427.
[85] M. Mariño, Les Houches lectures on matrix models and topological strings, In: Applications of Random Matrices in Physics (Les Houches Lecture Notes) NATO Sci. Ser. **221** (2005) Springer, New York, 319-378.
[86] M. A. Marti-Renom, A. C. Stuart, A. Fiser, R. Sanchez, F. Melo, A. Sali, Comparative protein structure modeling of genes and genomes, *Annual Reviews of Biophysics and Biomolecular Structure* **29** (2000), 291-325.
[87] M. L. Mehta, Random Matrices, Academic Press, 2004.
[88] G. Mondello, Combinatorial classes on $\overline{\mathcal{M}}_{g,n}$ are tautological, *International Mathematical Research Notes* **44** (2004), 2329-2390.
[89] J. W. Morgan, An Introduction to Gauge Theories in Gauge Theory and the Topology of Four-Manifolds, eds. R. Friedman and J. W. Morgan, American Math Society, 1998.
[90] D. M. Mount, Bioinformatics: Sequence and Genome Analysis, Second Edition, Cold Spring Harbor Laboratory Press, 2005.
[91] D. Mumford, J. Fogarty, F. Kirwan, Geometric Invariant Theory, Ergebnisse der Mathematik und ihrer Grenzgebiete **34**, Springer-Verlag, Berlin, 1994.
[92] S. Neidle, Principles of Nucleic Acid Structure, Elsevier, 2008.
[93] J. Nielsen, Surface transformation classes of algebraically finite type, Matematisk-fysiske meddelelser, Munksgaard, 1944.
[94] Y. Ohta, H. Kodama, A. Sugiyama,M. Matsuoka, H. Doi, T. Tsuboi, J. E. Andersen, R. C. Penner, S. Ihara, $SO(3)$ Rotation in the backbone reveals the shape of protein function, preprint (2012).
[95] H. Orland and A. Zee, RNA folding and large N matrix theory, *Nuclear Physics B*, **620** (2002), 456-476.
[96] Athanase Papadopoulos editor, Handbook of Teichmüller theory, volumes I-IV *IRMA Lectures in Mathematics and Theoretical Physics* **11** (2007), **13** (2009), **17** (2012), **19** (2014), European Mathematical Society.
[97] R. G. Parr and W. Yang, Density-Functional Theory of Atoms and Molecules, Oxford University Press, 1989.
[98] V. Pelechano and L. M. Steinmetz, Gene regulation by antisense transcription, *Nature Reviews Genetics* **14** (2013), 880-893.
[99] R. C. Penner, The decorated Teichmüller space of punctured surfaces, *Communications in Mathematical Physics* **113** (1987), 299-339.
[100] R. C. Penner, Perturbative series and the moduli space of Riemann surfaces, *Journal of Differential Geometry* **27** (1988), 35-53.
[101] R. C. Penner, Cell decomposition and compactification of Riemann's moduli space in decorated Teichmüller theory, in Woods Hole Mathematics-perspectives in math and physics, editors Nils Tongring and R. C. Penner, World Scientific Publishing Company (2004).
[102] R. C. Penner, Decorated Teichmüller space of bordered surfaces, *Communications in Analysis and Geometry* **12** (2004), 793-820.
[103] R. C. Penner, The structure and singularities of quotient arc complexes, *Journal of Topology* **1** (2008), 527-550.

[104] R. C. Penner, Decorated Teichmüller Theory, QGM Lecture Note Series **1**, European Mathematical Society, 2011.
[105] R. C. Penner M. Knudsen, C. Wiuf, J. E. Andersen, Fatgraph model of proteins, *Communication in Pure and Applied Math* **63** (2010), 1249-1297.
[106] R. C. Penner, E. S. Andersen, J. L. Jensen, A. K. Kantcheva, M. Bublitz, P. Nissen, A. M. H. Rasmussen, K. L. Svane, B. Hammer, R. Rezazadegan, N. C. Nielsen, J. T. Nielsen, J. E. Andersen, Hydrogen bond rotations as a uniform structural tool for analyzing protein architecture, *Nature Communications* **5** (2014).
[107] R. C. Penner Michael Knudsen, Carsten Wiuf, Jørgen E. Andersen, An Algebro-Topological Description of Protein Domain Structure, *PLOS one* (2011).
[108] R. C. Penner and M. S. Waterman, Spaces of RNA secondary structures, *Advances in Mathematics* **101** (1993), 31-49.
[109] P. A. Pevzner, Computational Molecular Biology: An algorithmic approach, MIT Press, 2000.
[110] M. Pillsbury, H. Orland, A. Zee, Steepest descent calculation of RNA pseudoknots, *Physical Reviews E* **72** (2005).
[111] M. Pillsbury, J. A. Taylor, H. Orland, A. Zee, An Algorithm for RNA Pseudoknots, (2005), arXiv:cond-mat/0310505.
[112] J. Qin and C. M. Reidys, On topological RNA interaction structures, *Journal of Computational Biology* **20** (2013), 495-513.
[113] T. Radó, Uber den Begriffric ander Riemannschen Fläche, *Acta Scientiarum Mathematicarum Szegediensis* **2** (1925), 101-121.
[114] D. C. Rapaport, The Art of Molecular Dynamics Simulation, Cambridge University Press,1996.
[115] C. M. Reidys, Combinatorial Computational Biology of RNA, Springer-Verlag, New York, 2011.
[116] C. M. Reidys and F. Huang. A stochastic context-free grammar for topological RNA pseudo-knot structures, preprint (2015).
[117] C. M. Reidys, F. W. D. Huang, J. E. Andersen, R. C. Penner, P. F. Stadler, M. E. Nebel, Topology and prediction of RNA pseudoknots, *Bioinformatics* **27** (2011), 1076-1085.
[118] C. M. Reidys and R. Wang, Shapes of RNA pseudoknot structures, *Journal of Computational Biology* **17** (2010), 1575-1590.
[119] E. Rivas and S. R. Eddy, A dynamic programming algorithm for rna structure prediction including pseudoknots, *Journal of Molecular Biology* **285** (1999), 2053-2068.
[120] Gian-Carlo Rota, private communication (1989).
[121] Gian-Carlo Rota, private communication (1992).
[122] T. Schlick, Molecular Modeling and Simulation. Springer Verlag, 2002.
[123] J. Schwarz, Resuscitating Superstring Theory, *New Scientist* (16Nov1987).
[124] R. P. Stanley, Catalan Numbers, Cambridge University Press, 2015.
[125] K. Strebel, Quadratic Differrentials, Ergebnisse der Mathematik und ihrer Grenzgebiete, Springer Verlag, 1984.
[126] W. P. Thurston, On the geometry and dynamics of diffeomorphisms of surfaces, *Bulletin of the American Mathematical Society* **19** 1988, 417-431.

[127] A. A. Tseytlin, Sigma model renormalization group flow, "central charge" action, and Perelman's entropy, *Physical Review D* **75** (2007), 064024(6).
[128] G. Vernizzi, H. Orland, A. Zee, Enumeration of RNA structures by matrix models, *Physical Review Letters* **94**, 168103.
[129] B. Wallner and A. Elofsson, All are not equal: A benchmark of different homology modeling programs, *Protein Science* **14** (2005), 1315-1327.
[130] A. Warshel, Multiscale Modeling of Biological Functions: From Enzymes to Molecular Machines, *Angewandte Chemie* **12** (2014), 10020-10031.
[131] Arieh Warshel, private communication (2015).
[132] M. S. Waterman, An Introduction Computational Biology, Chapman and Hall, New York, 1995.
[133] M. S. Waterman, Secondary structure of single-stranded nucleic acids, *Advances in Mathematics Supplementary Studies* **1** (1978), 167-212.
[134] J. D. Watson and F. H. C. Crick, Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid, *Nature* **171** (1953), 737-738.

Institut des Hautes Études Scientifiques
Le Bois-Marie 35, route de Chartres
91440 Bures-sur-Yvette, France
rpenner@ihes.fr

Departments of Mathematics and Physics Theory
California Institute of Technology
Pasadena, CA 91125 USA
rpenner@caltech.edu