Reprinted from ADVANCES IN MATHEMATICS
All Rights Reserved by Academic Press, New York and London

Vol. 101, No. 1, September 1993
Printed in Belgium

Spaces of RNA Secondary Structures

R. C. PENNER* AND MICHAEL S. WATERMAN†

*Department of Mathematics, University of Southern California,
University Park, Los Angeles, California 90089-1113*

We prove two topological theorems in physical chemistry. Namely, we introduce a hybrid of transverse and tangential measures on train tracks to prove sphericity of various simplicial complexes which arise from certain idealized models of physical chemistry. These complexes are at once identified with Thurston's space of projective geodesic laminations on an ideal polygon and with the analogue of a compactification (described elsewhere) of the moduli space of a punctured Riemann surface. The physical structures we study are various sub-collections of the set of all possible planar chemical bonds among the sites of a linear macromolecule. Each such collection we consider has a natural partial ordering, and the geometric realizations of appropriate posets are shown to be topological spheres. Such a topological statement encodes a wealth of combinatorial data, as we briefly discuss. In fact, our primary motivation here is to study secondary structures on RNA. This imposes the further restriction that there can be at most one base-pair supported at a given site of underlying linear macromolecule, and imposing this restriction leads to the class of "binary macromolecules." Our main results here assert the sphericity of certain topological spaces of both arbitrary and binary macromolecules, and it is the latter which we hope may have applications to RNA. Our techniques are largely elaborations of elementary topological techniques from Techmüller theory and the theory of train tracks.

1. MACROMOLECULES AND RNA

By the *linear macromolecule M_m of length m*, we mean the interval $[1, m] \subset \mathbb{R}$ together with the specification of the $m \geqslant 1$ integral points $[1, m] \cap \mathbb{Z}$. Each point of $[1, m] \cap \mathbb{Z}$ is called a *site* of M_m, and in our model below for RNA or DNA, the sites correspond to nucleotides. An unordered pair $\beta = \{s, s'\} \subset \mathbb{Z}$ of (distinct) sites of M_m so that $m-1 > |s-s'| > 1$ is called a *bond* on M_m, and the bond β is said to be *supported at the sites s, s'*. The usual terminology in biochemistry refers to β as a "base-pair," which is realized by one or more hydrogen bonds; here we abuse this terminology and refer to β itself as a bond.

* Partially supported by the National Science Foundation.

† Partially supported by the National Institutes of Health and the National Science Foundation.

0001-8708/93 $9.00

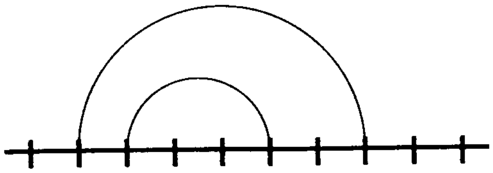

FIGURE 1

By a *secondary structure* on M_m, we mean a non-empty collection S of bonds with the property that if $\{s_1, s_1'\}$, $\{s_2, s_2'\} \in S$ where $s_1 < s_1'$ and $s_2 < s_2'$, then $s_1 \leqslant s_2 \leqslant s_1'$ if and only if $s_1 \leqslant s_2' \leqslant s_1'$. This restriction on families of bonds will be explained momentarily.

We require several different formulations of the combinatorics of secondary structures in the sequel and begin by defining the "bond picture" of a secondary structure as follows. If $S = \{\beta_i\}_{i=1}^n$ is a secondary structure on M_m, then consider drawing a collection of n semi-circles in $\mathbb{R}^2 \supset \mathbb{R} \supset M_m$, where to the bond $\beta_i = \{s_i, s_i'\}$, we draw the upper semi-circle of the circle with endpoints s_i, s_i' and center in $\mathbb{R} \subset \mathbb{R}^2$. To illustrate this definition, we draw in Fig. 1 the bond picture associated to the secondary structure $S = \{\{2, 8\}, \{3, 6\}\}$ on the linear macromolecule M_{10}. The restriction above on the bonds in a secondary structure guarantees that any two semi-circles in the associated bond picture intersect in a single site if at all.

Define an *arbitrary macromolecule* to consist of a secondary structure $S = \{\beta_i\}_{i=1}^n$ on M_m for some $m \geqslant 1$ and $n \geqslant 1$. One imagines collapsing to a point each semi-circle in the bond picture associated to a secondary structure S on M_m in the natural way so as to obtain a planar diagram which is intended to model the "actual" chemical bonds determining a macromolecular structure. See Fig. 2 for this "collapsed bond picture" corresponding to the secondary structure considered in Fig. 1.

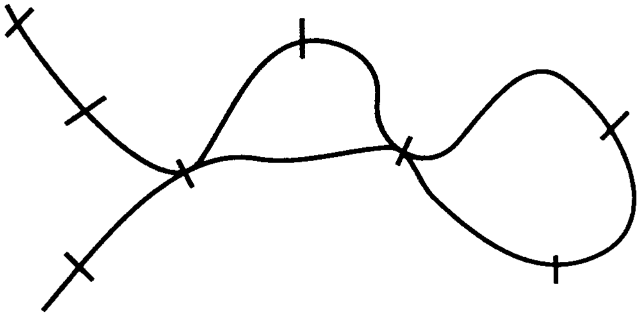

FIGURE 2

A primary motivation for our investigations here is to study the secondary structures (in the sense of [StW, Wa]) on RNA, and in this situation, one should further restrict the allowed bonding for physical-chemical reasons as follows. We say that an arbitrary macromolecule S is *binary* provided that for each site s of M_m there is at most one bond in S supported at s. By definition, the secondary structures considered on RNA satisfy this condition, and more complicated bonding corresponds to "tertiary structure."

Another "actual" restriction on a secondary structure corresponding to RNA arises from the "rigidity" of the sugar-phosphate backbone of the underlying linear macromolecule; namely, the sites comprising the support of a bond cannot be too close together. Indeed, experimental evidence [Le] suggests that if $\beta = \{s, s'\}$ is a bond, then $|s - s'|$ is at least three or four (that is, a "hairpin" cannot be too "tight"), whereas our condition on secondary structures requires only that the sites supporting a single bond cannot be consecutive. We do not handle any further such combinatorial restrictions here, and must consider binary macromolecules as a simplified model for secondary structures on RNA. Of course, by ignoring the restrictions on the secondary structure imposed by the allowable bonding between particular nucleotides (the "primary structure"), our model is already simplistic. Furthermore, as noted above, our very notion of a bond is simplistic from the point of view of RNA insofar as a base-pair actually corresponds to several hydrogen bonds between nucleotides.

On the other hand, we also consider here arbitrary macromolecules with no bound on the number of bonds supported at a given site, and we again do not impose any combinatorial restriction arising from "rigidity" of the underlying backbone. Our main results come, therefore, in two essentially different combinatorial guises: arbitrary macromolecules and binary macromolecules, and it is the latter which we hope may have applications to RNA.

2. Spaces of Macromolecules

Let us fix the number m of sites and define

$$\mathscr{S}_m = \{\text{arbitrary secondary structures on } M_m\}.$$

For an example, we enumerate the set $\mathscr{S}_5$ in Fig. 3. (The configuration of the figure will be explained below). The finite set $\mathscr{S}_m$ admits a natural partial ordering $\preccurlyeq$ induced by inclusion of sets, so $S \preccurlyeq S'$ if and only if the bond picture associated to S arises from that associated to S' by erasing some collection of semi-circles.

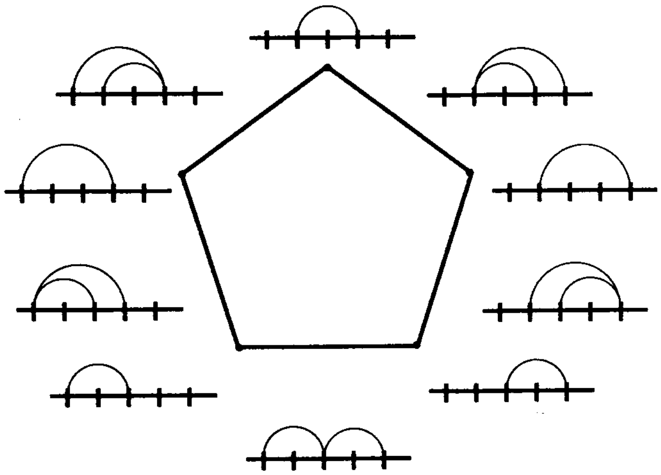

FIGURE 3

Next, we briefly discuss a standard construction in topology which will be required in the sequel. This construction is called the "geometric realization of a poset" and associates a simplicial decomposition of a topological space to a given poset. We give a definition tailored to our needs below and consider some examples here, referring the reader to the text [Mu], for instance, for more details and background material. To describe this construction, suppose that X is a finite set, and let $\mathscr{P}$ be some poset whose elements are subsets of X, where the partial ordering in $\mathscr{P}$ is induced by inclusion of subsets of X, and make the assumption that if $A \in \mathscr{P}$ and $\emptyset \neq B \subset A$, then $B \in \mathscr{P}$ as well. We build a triangulated topological space $\tilde{\mathscr{P}}$ by first defining a collection $\tilde{\mathscr{P}}^0 \subset \tilde{\mathscr{P}}$ of points (called "0-simplices"), then a collection of line segments $\tilde{\mathscr{P}}^1$ (called "1-simplices") in $\tilde{\mathscr{P}}$ with vertices among $\tilde{\mathscr{P}}^0$, then a collection $\tilde{\mathscr{P}}^2$ of triangles (called "2-simplices") in $\tilde{\mathscr{P}}$ whose boundary edges are among the line segments $\tilde{\mathscr{P}}^1$, and so on, finally taking the space $\tilde{\mathscr{P}}$ itself to be the union of all the simplices considered.

Indeed, there is one 0-simplex in $\tilde{\mathscr{P}}$ for each singleton $A \in \mathscr{P}$. Suppose inductively that all of the p-simplices in Y have been defined for each $p \leqslant N-1$ and proceed to define the N simplices in $\tilde{\mathscr{P}}$ as follows. If $A \in \mathscr{P}$ consists of $N+1$ elements of X, then there is a corresponding N-simplex in $\tilde{\mathscr{P}}$; writing $A = \{x_1, ..., x_{N+1}\}$, we identify each $A - \{x_i\}$ for $i = 1, ..., N+1$ with its corresponding $N-1$ simplex in $\tilde{\mathscr{P}}^{N-1}$ to describe the inclusion into $\tilde{\mathscr{P}}$ of the N-simplex associated to A.

Let $\tilde{\mathscr{S}}_m$ be the triangulated topological space arising as the geometric realization of the poset $\mathscr{S}_m$. As an example of geometric realization, we illustrate the space $\tilde{\mathscr{S}}_5$ in Fig. 3; the pentagon in the figure is the geometric realization $\tilde{\mathscr{S}}_5$ of $\mathscr{S}_5$, and next to each simplex (i.e., point or line segment) of the pentagon, we illustrate the corresponding secondary structure on M_5. This explains the configuration of the figure, as was promised above.

To better understand $\tilde{\mathscr{S}}_m$, fix some $S \in \mathscr{S}_m$, and arbitrarily associate a positive real number $x_\beta \in \mathbb{R}$, called the *bond strength of* β, to each bond $\beta \in S$; an assignment of a bond strength to each bond in S is called simply a *bond strength on* S itself. There is a natural equivalence relation on the collection of all possible bond strengths on S, where two bond strengths $(x_\beta : \beta \in S)$ and $(y_\beta : \beta \in S)$ are equivalent if there is some $t \in \mathbb{R}_+$ so that $x_\beta = t y_\beta$ for each $\beta \in S$. An equivalence class of bond strenghts on S is called a *projective bond strength on* S, so a projective bond strength on a secondary structure S is an assignment of ratios of bond strengths to the set of bonds comprising S. Thus, if S consists of $N+1 \geqslant 1$ bonds, then the collection of all projective bond strengths on S is naturally identified with an N-simplex.

Identifying the simplex of all projective bond strengths on S with the simplex in $\tilde{\mathscr{S}}_m$ corresponding to $S \in \mathscr{S}_m$ in the natural way, we find that $\tilde{\mathscr{S}}_m$ is just the collection of all pairs $(S, [\mathrm{x}])$, where $S \in \mathscr{S}_m$ and $[\mathrm{x}]$ is a projective bond strength on the bonds comprising S.

Suppose that $\{s_1, s'_1\}$, $\{s_2, s'_2\}$ are bonds in $S \in \mathscr{S}_\infty$ with $s_1 < s'_1$ and $s_2 < s'_2$. We say $\{s_1, s'_1\}$, $\{s_2, s'_2\}$ are *parallel* provided $|s_1 - s_2| = 1 = |s'_1 - s'_2|$ and consider the equivalence relation generated by parallelism on the set of bonds comprising S. (The collection of bonds comprising a parallelism class is called a "helix" in the biological literature.) Let $\{\Pi_l\}_{l=1}^L$ denote the equivalence classes, where Π_l consists of p_l bonds, for each $l = 1, ..., L$, and define

$$p(S) = \sum_{l=1}^{L} (p_l - 1).$$

Let $q(S)$ denote the number of bonds $\{s, s'\} \in S$ so that $|s - s'| = 2$, set $r(S) = p(S) + q(S)$, and let $f(S)$ denote the number of ("free") sites which do not occur in the support of any bond in S.

Turning finally to binary macromolecules, suppose $S \in \mathscr{S}_\infty = \bigcup_{m \geqslant 1} \mathscr{S}_m$ is a binary macromolecule, and notice that if $S \in \mathscr{S}_m$, then $m \leqslant 3f(S) + 2p(S)$. In particular, the poset

$$\mathscr{B}_f^r = \{\text{binary } S \in \mathscr{S}_\infty : f(S) = f \text{ and } r(S) \leqslant r\},$$

is finite for each $r \geqslant 0$. There is a corresponding infinite poset

$$\mathscr{B}_f = \bigcup_{r \geqslant 0} \mathscr{B}_f^r,$$

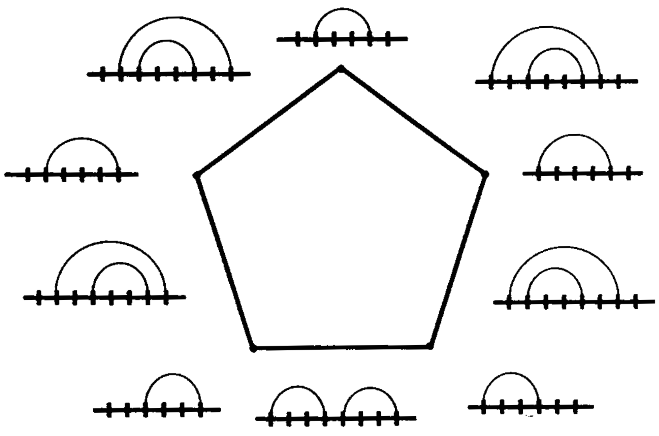

FIGURE 4

for each $f \geqslant 1$, and the chain

$$\mathscr{B}_f^0 \subset \mathscr{B}_f^1 \subset \cdots \subset \mathscr{B}_f$$

of posets gives rise to a corresponding chain

$$\tilde{\mathscr{B}}_f^0 \subset \tilde{\mathscr{B}}_f^1 \subset \cdots \subset \tilde{\mathscr{B}}_f$$

of simplicial inclusions of geometric realizations. As an example, we illustrate $\mathscr{B}_0^4$ in Fig. 4, where the configuration is explained as before.

3. Rooted Fattrees

Given an abstract rooted tree τ and an embedding $\tau \subset \mathbb{R}^2$ of τ in the plane, there is a further structure induced on τ. Namely, for each vertex v of τ there is an induced (counter-clockwise) cyclic ordering on the collection of all edges of τ incident on v. We refer to such specifications of cyclic orderings on the edges incident on a common vertex v, for each vertex v of τ, as a *fattening* of τ. Abstractly, we define a *rooted fattree* τ to be a rooted tree together with a fattening. We regard two rooted fattrees as identical if there is an isomorphism of underlying rooted trees (so the isomorphism must map the root to the root) which respects the cyclic orderings. Thus, an embedding of a rooted tree in the plane uniquely determines a rooted fattree. Conversely, it is elementary to see that a fattening of an abstract rooted tree uniquely determines an isotopy class of embeddings of the underlying rooted tree into $\mathbb{R}^2$. Define the *valence* of a vertex in a rooted fattree τ to be the number of distinct edges incident on it, and let $v_k(\tau)$

denote the number of k-valent vertices of τ for each $k \geqslant 1$. In particular, we let $u(\tau) = v_1(\tau) - 1$ denote the number of univalent vertices excluding the root and let $b(\tau) = v_2(\tau)$ denote the number of bivalent vertices. We also let $e(\tau)$ denote the total number of edges of τ and $i(\tau)$ denote the number of internal edges of τ.

LEMMA 1. *If τ is a rooted fattree, then*

$$u(\tau) = +1 + \sum_{k \geqslant 2} (k-2)\, v_k(\tau)$$

$$e(\tau) = +1 + \sum_{k \geqslant 2} (k-1)\, v_k(\tau)$$

$$i(\tau) = -1 + \sum_{k \geqslant 2} v_k(\tau).$$

Proof. For the first identity, consider two copies of τ and identify each corresponding pair of univalent vertices to a single bivalent vertex to produce a "fatgraph" G (cf. [P1]) with bivalent vertices as in Fig. 5; G is called the "double" of τ along its univalent vertices. The Euler characteristic $\chi(G)$ of the underlying graph is evidently $1 - u(\tau)$, while the very definition of Euler characteristic of G gives $\chi(G) = -\Sigma_{k \geqslant 2} (k-2)\, v_k(G)$, as was asserted.

Insofar as the double G obviously has $\sum_{k \geqslant 1} k v_k(\tau) = 1 + u(\tau) + \sum_{k \geqslant 2} k v_k(\tau)$ edges and this is clearly twice $e(\tau)$, the remaining identities are elementary consequences of the first identity. Q.E.D.

Given some $u \geqslant 2$, consider the collection $\mathscr{T}_u$ of all isomorphisms classes of rooted fattrees τ with $u(\tau) = u$, where the root of τ is univalent and the other endpoint of the edge of τ containing the root is a vertex of valence at least three. Notice that $\mathscr{T}_u$ is countably infinite for each u; on the other hand, for each $b \in \mathbb{Z}$, there are only a finite number of elements of $\mathscr{T}_u$ with $b(\tau) \leqslant b$ by the first part of the previous lemma.

In fact, each $\mathscr{T}_u$ is a poset for each $u \geqslant 2$. To define the partial ordering $\preccurlyeq$ on $\mathscr{T}_u$, we suppose that $\tau \in \mathscr{T}_u$ and define another $\tau' \in \mathscr{T}_u$ as follows. Consider an internal edge e of τ whose endpoints correspond to vertices v^1, v^2 of respective valences v_1 and v_2. We may "contract" e to a single point, coalescing v^1 and v^2 to a single vertex of valence $v_1 + v_2 - 2$ as illustrated in Fig. 6; also illustrated in Fig. 6 is the natural induced fattening on the resulting rooted tree. We let τ' denote the resulting rooted fattree and say that τ' arises from τ by an *elementary move along* the internal edge e.

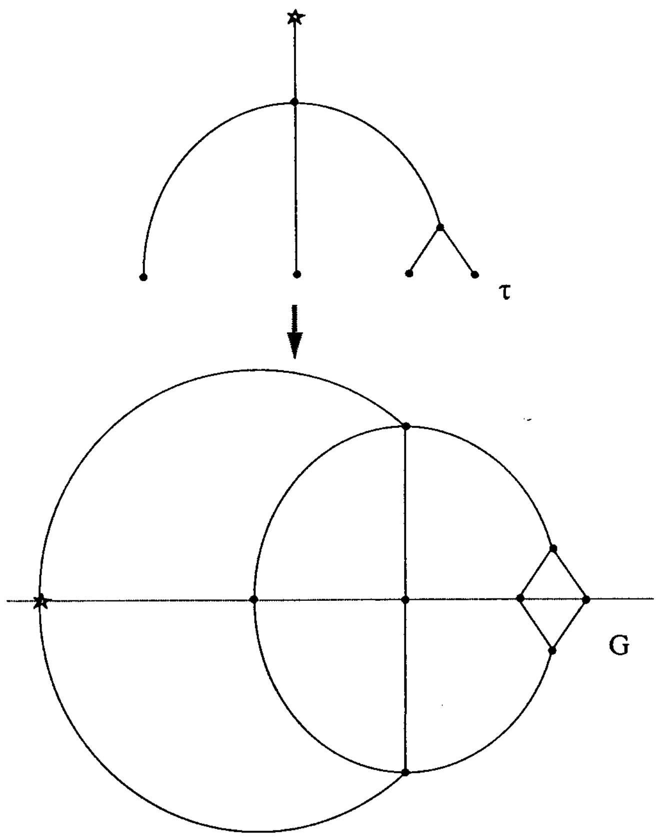

FIGURE 5

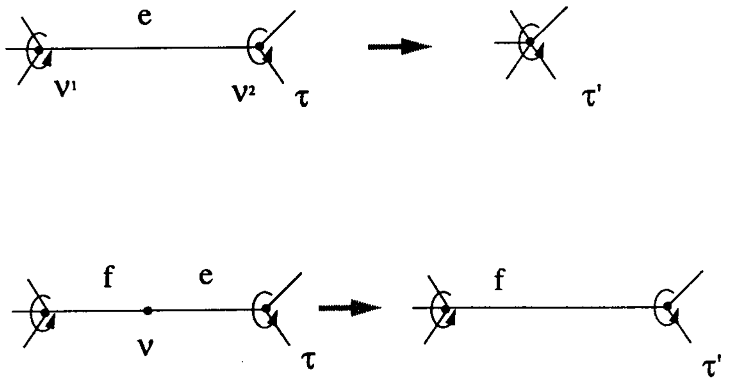

FIGURE 6

We stipulate that $\tau' \preccurlyeq \tau$ and let $\preccurlyeq$ denote the induced partial ordering on $\mathcal{T}_u$. Thus, $\tau_1 \preccurlyeq \tau_2$ if and only if one can pass from τ_2 to τ_1 by a finite sequence of elementary moves. In particular, given any $\tau_2 \in \mathcal{T}_u$, there is a well-defined $\tau_1 \in \mathcal{T}_u$ where $b(\tau_1) = 0$ obtained by applying elementary moves along all the internal edges of τ_2 with a bivalent endpoint. In fact, we also consider the finite sub-posets

$$\mathcal{T}_u^b = \{\tau \in \mathcal{T}_u : b(\tau) \leqslant b\}$$

consisting of rooted fattrees with $u+1$ univalent vertices (including the root) and at most b bivalent vertices.

The rest of this section is dedicated to the proof of

THEOREM 2. *There are canonical isomorphisms of posets*

$$\sigma : \mathcal{B}_f^r \to \mathcal{T}_f^r,$$
$$\tau : \mathcal{S}_m \to \mathcal{T}_{m-1}^0.$$

Proof. The mapping σ is described in [ScW, Proposition 1], which we next recall. The rooted fattree $\sigma = \sigma(S)$ associated to $S \in \mathcal{B}_f^r$ is constructed inductively from the bond picture of S. To describe this construction, choose some vertical interval I in the upper halfspace in $\mathbb{R}^2$, and take its upper endpoint to be the root. Adjoin to I disjointly embedded edges running from the lower endpoint of I to each outermost bond and each "accessible" free site, as in Fig. 7. One proceeds constructing σ in this way (as is also illustrated in Fig. 7) to produce the desired $\sigma = \sigma(S) \in \mathcal{T}_f^r$.

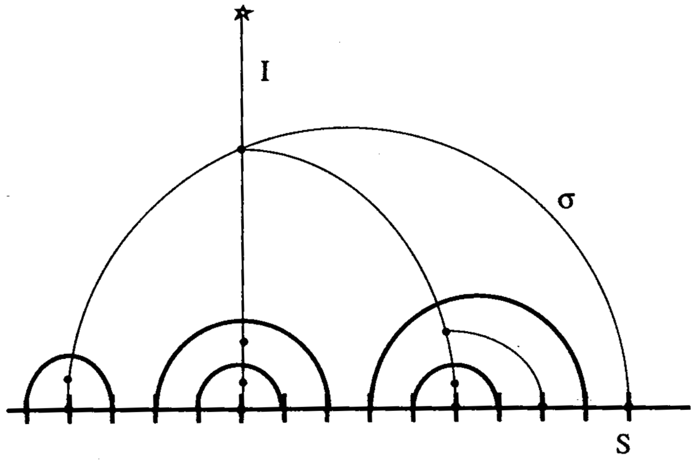

FIGURE 7

To see that σ is a bijection, suppose we are given a rooted fattree $\sigma \in \mathcal{T}^r_f$, and embed σ in the upper halfspace in $\mathbb{R}^2$ in such a way that the root lies at positive height in $\mathbb{R}^2$ and each of the other univalent vertices lies on $\mathbb{R} \subset \mathbb{R}^2$. To draw the bond picture of the corresponding secondary structure S, we draw one semi-circle for each interior edge of σ, and there is an essentially unique way to draw these semi-circles so as to be disjointly embedded since σ is a tree. Mapping the univalent vertices as well as the endpoints of the semi-circles to consecutive integer points in the natural way gives the bond picture for a corresponding secondary structure $S \in \mathcal{B}^r_f$.

The maps in the previous two paragraphs are evidently inverses and respect the partial orderings, so $\sigma: \mathcal{B}^r_f \to \mathcal{T}^r_f$ is indeed an isomorphism of posets.

Turning to the map τ, suppose that $S \in \mathcal{S}_m$, and consider the bond picture of S. To construct $\tau = \tau(S) \in \mathcal{T}^0_{m-1}$, begin as before with a vertical interval in $\mathbb{R}^2$ whose upper endpoint is taken as the root of τ. Adjoin to I disjointly embedded edges running from the lower endpoint of I to each outermost bond and each "accessible" interval in $\mathbb{R} \subset \mathbb{R}^2$ whose endpoints comprise the intersection of the interval with the collection of sites, as in Fig. 8. Continue constructing τ in this way (as illustrated in Fig. 8) to produce the desired $\tau = \tau(S) \in \mathcal{T}^0_{m-1}$.

To describe the inverse of τ, suppose $\tau \in \mathcal{T}^0_{m-1}$, and embed τ in $\mathbb{R}^2$ in the upper halfspace as before so that the root lies at positive height in $\mathbb{R}^2$ and each of the other univalent vertices lies in $\mathbb{R} \subset \mathbb{R}^2$. To draw the bond picture of the corresponding secondary structure S, put one site between each pair of consecutive univalent vertices of τ lying in $\mathbb{R}$ as well as one site before the first and one site after the last univalent vertices of τ lying in $\mathbb{R}$

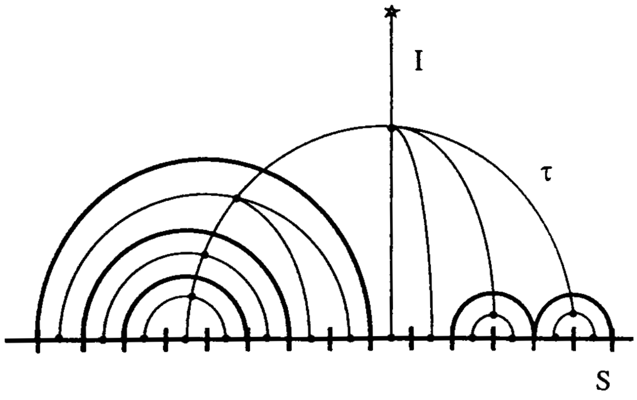

FIGURE 8

as in Fig. 8. There is again one bond in S for each internal edge of τ, and we may uniquely draw the bond picture as before.

As above, the maps in the previous two paragraphs are evidently inverses and respect the partial orderings, so $\tau: \mathscr{S}_m \to \mathscr{T}^0_{m-1}$ is indeed an isomorphism of posets. Q.E.D.

We sketch in Section 6 below a treatment of arbitrary macromolecules which is analogous to our treatment of binary macromolecules in Theorem 2.

4. The Arc Complex

We show in this section that $\mathscr{T}^0_u$ is isomorphic to a certain poset $\mathscr{A}_u$, called an "arc poset," whose geometric realization $\tilde{\mathscr{A}}_u$ is a topological sphere of dimension $u-3$. The sphericity of $\tilde{\mathscr{A}}_u$ is in a sense well-known (as we discuss later), but we include a simple proof of this fact here.

To define the poset $\mathscr{A}_u$, consider a planar polygon P_u of u sides, and choose a distinguished edge of P_u once and for all. Define an *arc family* in P_u to be (the homotopy class of) a (non-empty) collection of arcs disjointly embedded in P_u with endpoints among the vertices of P so that there is at most one arc connecting a given pair of vertices and each arc connects non-consecutive and distinct vertices. See Fig. 9 for the arc families in the pentagon P_5.

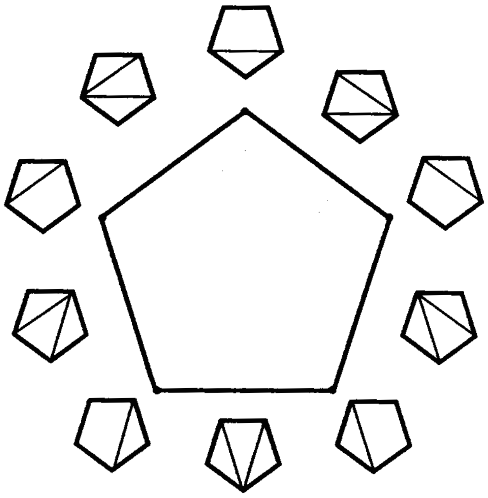

Figure 9

Identify the vertices of P_u once and for all with a set V. Given an arc family in $\Delta \subset P_u$, we may associate a collection of unordered pairs of elements of V in the natural way, where we associate to an arc $\alpha \in \Delta$ its unordered pair of endpoints, and to Δ itself we associate the set whose elements are the unordered pairs associated to the component arcs. Thus, $\mathscr{A}_u$ is a poset of the type considered in Section 2.

Proposition 3. *There is an isomorphism of posets*

$$F\colon \mathscr{A}_u \to \mathscr{T}^0_{u-1}.$$

Proof. The map F is essentially a recapitulation of the map τ in Theorem 2; see Fig. 10 for the rooted fattrees corresponding to the arc families in Fig. 9. Actually, this is the polygonal case of the duality between "fatgraphs" and "ideal cell decompositions" (cf. [P2]). Q.E.D.

Our main result for this section is

Theorem 4. *For $u \geqslant 4$, the geometric realization $\tilde{\mathscr{A}}_u$ of the poset $\mathscr{A}_u$ is a topological sphere of dimension $u-4$.*

Proof. Rather than study $\tilde{\mathscr{A}}_u$ directly, we consider the "deprojectivization" $\bar{\mathscr{A}}_u$, defined to consist of all pairs $(\Delta, \mathbf{x})$, where $\mathbf{x} = (x_\alpha : \alpha \in \Delta)$ is a

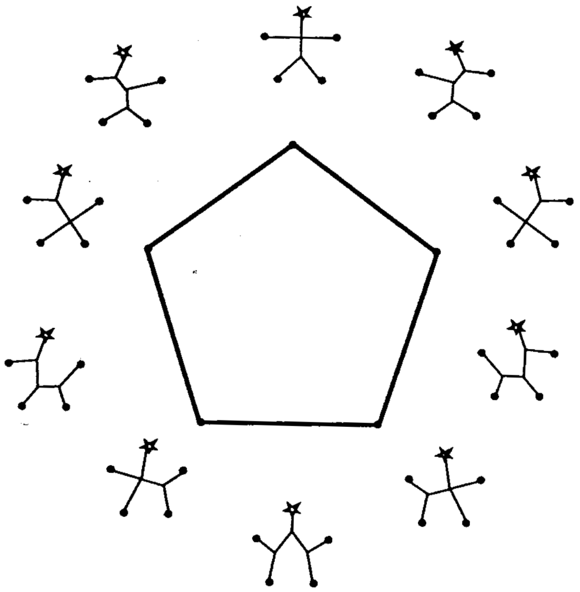

Figure 10

bond strength (*not* a projective bond strength) on Δ; cf. Section 2. We show that $\bar{\mathscr{A}}_u$ is naturally homeomorphic to $\mathbb{R}^{u-3} \approx \bar{\mathscr{A}}_u$, where $\approx$ denotes the equivalence relation of being homeomorphic, in such a way that $\mathbf{0} \in \mathbb{R}^{u-3}$ corresponds to the empty weighted arc family $\varnothing$ in P_u. Thinking of $\bar{\mathscr{A}}_u$ as the cone over $\tilde{\mathscr{A}}_u$ from $\varnothing$, it is then evident that

$$\tilde{\mathscr{A}}_u \approx (\bar{\mathscr{A}}_u - \{\varnothing\})/\mathbb{R} \approx (\mathbb{R}^{u-3} - \{\mathbf{0}\})/\mathbb{R} \approx S^{u-4},$$

where S^n denotes the n-dimensional sphere.

Thus, it suffices to prove that $\bar{\mathscr{A}}_u \approx \mathbb{R}^{u-3}$, and to this end, we choose a triangulation Δ_* of P_u once and for all; that is, Δ_* is an arc family so that each component of $P_u - \Delta_*$ is a triangle, and one finds that there are $u-3$ arcs in Δ_*. (In fact, there are $\binom{2u-4}{u-2} - \binom{2u-4}{u-3}$ different such triangulations we might choose here, using the standard notation for binomial coefficients; see [Kn, Problem 2.2.1, No. 4].)

Given a vector $\mathbf{X} = (X_\alpha : \alpha \in \Delta_*)$ where each $X_\alpha \in \mathbb{R}$, we define a corresponding $(\Delta, \mathbf{x}) \in \bar{\mathscr{A}}_u$ (so that $\mathbf{X} = \mathbf{0}$ if and only if $\Delta = \varnothing$), and there are cases depending on the signs of the component X_α''s as well as other conditions. Let us formally associate $X_\beta = 0$ to each frontier arc $\beta \subset P_u$ so that to each triangle T complementary to Δ_* in P_u is associated the triple $\{X_1, X_2, X_3\}$ of real numbers assigned to its frontier edges. There are the following possibilities to consider for $\{X_1, X_2, X_3\}$:

(a) $X_i \geqslant 0$, for $i = 1, 2, 3$, and $X_i \leqslant X_j + X_k$ whenever $\{i, j, k\} = \{1, 2, 3\}$; that is, the X_i satisfy all three weak triangle inequalities;

(b) $X_i \geqslant 0$, for $i = 1, 2, 3$, and some triangle inequality among the X_i fails, say $X_1 \geqslant X_2 + X_3$. Notice that if $X_1 \geqslant X_2 + X_3$ and $X_2 \geqslant X_1 + X_3$ (for instance), then $X_3 \leqslant 0$, while if all three triangle inequalities fail, then X_1, X_2, $X_3 \leqslant 0$;

(c) $X_1 \leqslant 0$ and X_2, $X_3 \geqslant 0$ (for instance);

(d) $X_1 \geqslant 0$ and X_2, $X_3 \leqslant 0$ (for instance);

(e) $X_i \leqslant 0$, for $i = 1, 2, 3$.

The construction of a weighted arc family inside the triangle T in each of these cases is indicated in the corresponding part of Fig. 11.

To combine these weighted arc families on the complementary triangles into a weighted arc family on P_u itself, consider a neighborhood of an arc $\alpha \in \Delta_*$. There are two cases depending on whether $X_\alpha \leqslant 0$ or $X_\alpha \geqslant 0$. In the former case, simply combine the two arcs parallel to α into a single parallel arc with weight $x_\alpha = |X_\alpha|$; see Fig. 12(a). One imagines taking a corridor about each arc of width $\frac{1}{2}|X_\alpha|$ and sewing these two corridors together in the natural way to get a corridor about α of width $|X_\alpha|$.

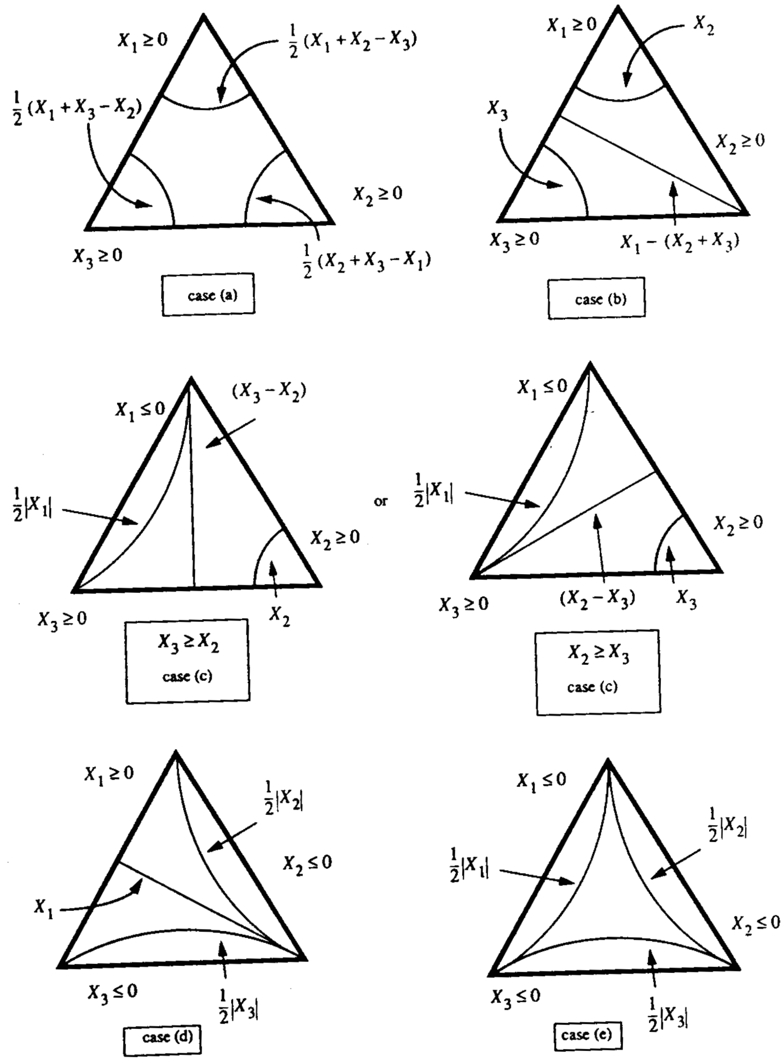

FIGURE 11

Suppose $X_\alpha \geqslant 0$, let T_1, T_2 denote the triangles in P_u on either side of α, and choose an orientation on α. We have constructed above in T_i a weighted arc family and let $t_i^1, ..., t_i^{n_i}$ (where $n_i \leqslant 3$ by construction) denote the arcs in T_i meeting α enumerated in their order of occurrance along α, for $i=1, 2$, and let the weight on t_i^j be denoted w_i^j. Imagine embedding in P_u a corridor of width w_i^j about each arc t_i^j. There is a unique way to

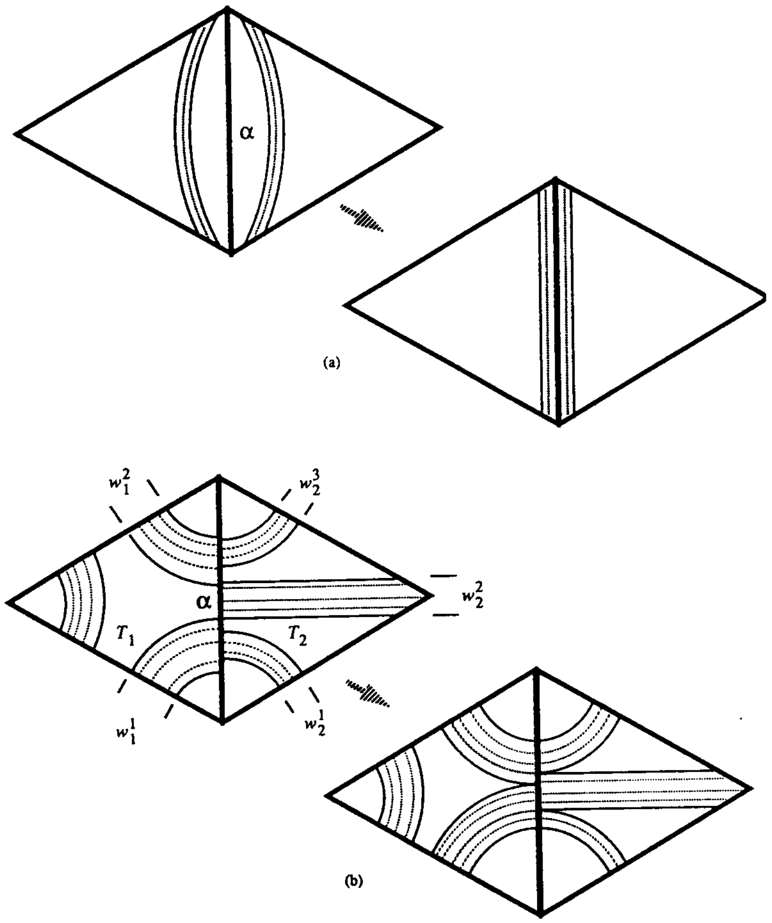

FIGURE 12

combine these corridors near α, as illustrated in Figure 12(b), since the collections of corridors on either side of α have the same total width

$$\sum_{j=1}^{n_1} w^j_1 = X_\alpha = \sum_{j=1}^{n_2} w^j_2,$$

as in Kirchhoff's laws of electricity. (See [PH, Construction 1.7.7] for more details.)

Combining the arc corridors in each complementary triangle, we finally construct a family of corridors in P_u, where each corridor consists of a family of parallel copies of an arc in β in P_u. By construction, the various

β may be disjointly embedded, and the resulting arc family in P_u is denoted Δ. The width of the corridor about β gives a corresponding bond strength x_β for each such β. Thus, $(\Delta, \mathbf{x}) \in \bar{\mathscr{A}}_u$ is canonically associated to $\mathbf{X}$, as was promised above.

The inverse map associates $\mathbf{X}$ to some $(\Delta, \mathbf{x}) \in \bar{\mathscr{A}}_u$ by the assignment

$$X_\alpha = \begin{cases} -x_\alpha, & \text{if } \alpha \in \Delta_*; \\ \sum\limits_{\alpha \in \Delta_*} \sum\limits_{p \in \alpha \cap \Delta} x_\alpha, & \text{if } \alpha \notin \Delta_*, \end{cases}$$

where the inner sum in case $\alpha \notin \Delta_*$ is only over transverse intersection points (of geodesic representatives).

The maps of the previous two paragraphs are evidently continuous and mutual inverses, completing the proof. Q.E.D.

Remark. These coordinates are a hybrid of Thurston's transverse and tangential coordinates on train tracks (cf. [PH]); this global coordinatization of the space $\mathscr{A}_u$ of measured laminations on a polygon is new. At the same time, we have described in [P3] a compactification of the moduli space of an arbitrary (i.e., not necessarily planar) punctured Riemann surface, and the proof of Theorem 4 first arose in that context as a very special case of our conjecture that this new compactification is an orbifold; in fact, here (and only in this setting) this compactification is identified with the space of projectively measured geodesic laminations, which is well known to be spherical.

5. Combinatorics of Secondary Structures

We finally bring together the previous results into explicit combinatorial assertions about secondary structures. Our main result for this paper, which follows, is an immediate corollary of Theorem 2, Proposition 3, and Theorem 4:

Theorem 5. *For any $m \geqslant 4$, $\tilde{\mathscr{S}}_m$ is a topological sphere of dimension $m-4$, and for any $f \geqslant 3$, $\tilde{\mathscr{B}}_f^0$ is a topological sphere of dimension $f-3$.*

We derive some explicit combinatorial consequences below. Prefatory to this, we refer the reader to [Mu] for a discussion of simplicial homology groups and chain contractions and have

Lemma 6. *For each $r \geqslant 0$, the natural inclusion*

$$\mathscr{B}_f^r \subset \mathscr{B}_f^{r+1}$$

is a chain homotopy equivalence, for any $f \geqslant 3$.

Proof. By Theorem 2, it suffices to show that $\mathscr{T}^b_u \subset \mathscr{T}^{b+1}_u$ is a chain homotopy equivalence, and one simply takes a chain homotopy $\mathscr{T}^b_u \to \mathscr{T}^{b+1}_u$ induced by inserting a bivalent vertex along a specified branch. Q.E.D.

Our next result (which follows immediately from the previous two results) is a natural extension of Theorem 5.

THEOREM 7. *For each $m \geqslant 4$, $\mathscr{S}_m$ is a triangulated topological $(m-4)$-dimensional sphere. For each $r \geqslant 0$ and $f \geqslant 3$, $\mathscr{B}^r_f$ is a homology $(f-3)$-dimensinal sphere, and in particular, $\mathscr{B}^0_f$ is a triangulated topological $(f-3)$-dimensional sphere.*

Of course, Theorem 7 is a topological synthesis of a wealth of explicit combinatorial information about secondary structures. To give an example of an explicit combinatorial result about secondary structures which follows from Theorem 7, we define the *total valence* of a fattree τ as $V(\tau) = \sum_{k \geqslant 2} v_k(\tau)$ and have

COROLLARY 8. *For any $m \geqslant 4$, we have*

$$\sum_{S \in \mathscr{S}_m} (-1)^{V(\tau(S))} = \begin{cases} 0, & \text{if } m \text{ is odd;} \\ 2, & \text{if } m \text{ is even,} \end{cases}$$

and for any $r \geqslant 0$ and $f \geqslant 3$, we have

$$\sum_{S \in \mathscr{B}^r_f} (-1)^{V(\sigma(S))} = \begin{cases} 2, & \text{if } f \text{ is odd;} \\ 0, & \text{if } f \text{ is even.} \end{cases}$$

Proof. We first consider the poset $\mathscr{S}_m$ and must compute the dimension δ of the simplex corresponding to $S \in \mathscr{S}_m$. According to the definitions, $\delta = i(\tau(S)) - 1$, so by the third part of Lemma 1, $\delta = V(\tau(S)) - 2$. In the same way, the dimension of the cell corresponding to $S \in \mathscr{B}^r_f$ is found to be $V(\sigma(S)) - 2$.

Thus, the left-hand side of the equation in Corollary 8 is simply the definition of the Euler characteristic whereas the Euler characteristic of the N-dimensional sphere vanishes for N odd and equals two for N even. Q.E.D.

6. ANOTHER TREATMENT OF ARBITRARY MACROMOLECULES

We sketch here a treatment of arbitrary macromolecules which is analogous to our treatment above of binary macromolecules and must reformulate the very definition of bonds as follows. We alter the definition of bond in Section 1 to allow bonding between consecutive sites (still requiring, however, that a bond on M_m cannot have its support at 1, $m \in M_m$). Furthermore, we weaken the definition of secondary structure in Section 1 to allow multiple bonds with the same support. Thus, the bond picture now allows semi-circles whose endpoints are consecutive integers as well as allowing multiple "semi-circles" connecting a given pair of sites, and we let $\mathscr{C}_m$ denote the collection of these more general secondary structures on M_m. Again, each $\mathscr{C}_m$ is an infinite poset in the natural way, and we let $\tilde{\mathscr{C}}_m$ denote its geometric realization.

We must also consider various subspaces of $\tilde{\mathscr{C}}_m$, as follows. Given $C \in \mathscr{C}_m$, let $v(C)$ denote the number of bonds $\{s, s'\} \in C$ so that $|s - s'| = 1$. We also say two bonds in $C \in \mathscr{C}_m$ are *parallel* if they have the same support, suppose that there are t_l elements of the parallelism equivalence classes, for $l = 1, ..., L$, and define

$$t(C) = \sum_{l=1}^{L} (t_l - 1)$$

as in Section 2.

Set $w(C) = v(C) + t(C)$, and define the finite sub-poset

$$\mathscr{C}^w_m = \{C \in \mathscr{C}_m : w(C) \leqslant w\} \subset \mathscr{C}_m,$$

for each $w \geqslant 0$, so, in particular, $\mathscr{C}^0_m$ is just the familiar poset $\mathscr{S}_m$. Again, there is a chain

$$\mathscr{C}^0_m \subset \mathscr{C}^1_m \subset \cdots \subset \cdots \subset \mathscr{C}_m$$

of inclusions if posets. Let $\tilde{\mathscr{C}}^w_m$ denote the geometric realization of the poset $\mathscr{C}^w_m$, so there is a corresponding chain

$$\tilde{\mathscr{C}}^0_m \subset \tilde{\mathscr{C}}^1_m \subset \cdots \subset \tilde{\mathscr{C}}_m$$

of simplicial inclusions of topological spaces.

The obvious analogue of the map τ in Theorem 2 establishes an isomorphism

$$\mathscr{C}^w_m \to \mathscr{T}^w_{m-1}$$

of posets, the natural inclusions

$$\mathscr{C}^w_m \subset \mathscr{C}^{w+1}_m$$

are chain homotopy equivalences as in Lemma 6 for $w \geqslant 0$ and $m \geqslant 4$, and we have

THEOREM 9. *For each $m \geqslant 4$ and $w \geqslant 0$, $\tilde{\mathscr{C}}^w_m$ is a homology $(m-4)$-dimensional sphere.*

REFERENCES

[Kn] D. E. KNUTH, "The Art of Computer Programming," Vol. I, Addison–Wesley, Reading, MA, 1968.

[Le] B. LEWIN, "Genes, IV," Oxford Univ. Press, London/New York, 1990.

[Mu] J. MUNKRES, "Elements of Algebraic Topology," Addison–Wesley, Reading, MA, 1984.

[P1] R. C. PENNER, Perturbative series and the moduli space of Riemann surfaces, *J. Differential Geom.* **27** (1988), 35–53.

[P2] R. C. PENNER, Weil–Petersson volumes, *J. Differential Geom.* **35** (1992), 559–608.

[P3] R. C. PENNER, The Poincaré dual of the Weil–Petersson Kähler two-form, *Comm. Anal. and Geom.* **1** (1993), 43–70.

[PH] R. C. PENNER WITH J. L. HARER, "Combinatorics of Train Tracks," Ann. of Math. Stud., Vol. 125, Princeton Univ. Press, Princeton, NJ, 1992.

[ScW] W. R. SCHMITT AND M. S. WATERMAN, Plane trees and RNA secondary structure, *Discrete Appl. Math.*, in press.

[StW] P. R. STEIN AND M. S. WATERMAN, On some new sequences generalizing the Catalan and Motzkin numbers, *Discrete Math.* **26** (1978), 261–272.

[Wa] M. S. WATERMAN, Secondary structure of single-stranded nucleic acids, *in* "Studies in Foundations and Combinatorics," Advances in Mathematics Supplementary Studies, Vol. 1, 1978.

Printed by Catherine Press, Ltd., Tempelhof 41, B-8000 Brugge, Belgium