Fatgraph Models of Proteins

R. C. PENNER
Departments of Mathematics and Physics/Astronomy
University of Southern California
Center for the Topology and Quantization of Moduli Spaces
Aarhus University

MICHAEL KNUDSEN
Bioinformatics Research Center
Aarhus University

CARSTEN WIUF
Bioinformatics Research Center and Centre for Membrane Pumps
in Cells and Disease—PUMPKIN
Aarhus University

AND

JØRGEN ELLEGAARD ANDERSEN
Center for the Topology and Quantization of Moduli Spaces
Aarhus University

Abstract

We introduce a new model of proteins that extends and enhances the traditional graphical representation by associating a combinatorial object called a fatgraph to any protein based upon its intrinsic geometry. Fatgraphs can easily be stored and manipulated as triples of permutations, and these methods are therefore amenable to fast computer implementation. Applications include the refinement of structural protein classifications and the prediction of geometric and other properties of proteins from their chemical structures.

Introduction

A *fatgraph* G is a graph in the usual sense of the term together with cyclic orderings on the half-edges about each vertex (see Section 2.2 for the precise definition). They arose in mathematics [26] as the combinatorial objects indexing orbicells in a certain decomposition of Riemann's moduli space [26, 29] and in physics [4, 30] as index sets for the large-N limit of certain matrix models. A basic geometric point is that a fatgraph G uniquely determines a corresponding surface $F(G)$ with a boundary that contains G as a deformation retract. Fatgraphs have already proved useful in geometry [14, 16, 22, 26], theoretical physics [8, 18], and modeling RNA secondary structures [27], for example.

Communications on Pure and Applied Mathematics, Vol. LXIII, 1249–1297 (2010)

A *protein* P is a linear polymer of amino acids (see Section 1 for more precision), and their study is a central theme in contemporary biophysics [1, 10]. Our main achievement in this paper is to introduce a model of proteins that naturally associates a fatgraph $G(P)$ to a protein P based upon the spatial locations of its constituent atoms. The idea is that the protein is roughly described geometrically as the concatenation of a sequence of planar polygons called *peptide units* meeting at tetrahedral angles at pairs of vertices and twisted by pairs of dihedral angles between the polygons. To each peptide unit, we associate a positively oriented orthonormal 3-frame and a fatgraph building block, and we concatenate these building blocks using these 3-frames in a manner naturally determined by the geometry of the Lie group SO(3). There are furthermore *hydrogen bonds* between atoms contained in the peptide units, and these are modeled by including further edges connecting the building blocks so as to determine a well-defined fatgraph $G(P)$ from P. Thus, the fatgraph $G(P)$ derived from the protein P captures the geometry of the protein "backbone" and the geometry and combinatorics of the hydrogen bonding along the backbone; elaborations of this basic model are also described that capture further aspects of protein structure.

The key point is that topological or geometric properties of the fatgraph $G(P)$ can be taken as properties or "descriptors" of the protein P itself. A fundamental aspect not usually relevant in applying fatgraphs is that this construction of $G(P)$ is based on actual experimental data about P in which there are uncertainties and sometimes errors as well. Furthermore, the notion that the protein P is comprised of atoms at fixed relative spatial locations, which is the basic input to our model, is itself a biological idealization of the reality that a given protein at equilibrium may have several closely related coexisting geometric incarnations. In order that the protein descriptors arising from fatgraphs be meaningful characteristics of proteins in light of these remarks, we shall be forced to go beyond the usual situation and consider fatgraphs G whose corresponding surfaces $F(G)$ are nonorientable. This is easily achieved combinatorially by including in the definition of a fatgraph a coloring of its edges by a set with two elements.

The desired result of *robust* protein descriptors, i.e., properties of $G(P)$ that do not change much under small changes in the relative spatial locations of the atoms constituting P, is a key attribute of our construction; for example, the number of boundary components and the Euler characteristic of $F(G(P))$ are such robust invariants, and we give a plethora of further numerical and nonnumerical examples. Another key point of our construction rests on the fact that biophysicists *already* often associate a graph to a protein P based upon its hydrogen and chemical bonding, and our model succeeds in reproducing this usual graphical depiction of a protein but now with its enhanced structure as a fatgraph $G(P)$; i.e., the graph underlying $G(P)$ is the one usually associated to P in biophysics. Furthermore, an important practical point is that fatgraphs can be conveniently stored and manipulated on the computer as triples of permutations.

Since this is a math paper whose central purpose is to introduce fatgraph models of proteins, we shall not dwell on biophysical applications; nevertheless, we feel compelled to include here several such applications as follows. Certain proteins decompose naturally into "domains" or "globules," roughly 115,000 of which have so far been determined experimentally and categorized into several thousand classes (cf. [12, 15, 23, 25]), and we concentrate here for definiteness on the CATH classification [25] of domains. Our most basic robust descriptors of a domain P are given by the topological types of the surface $F(G(P))$ computed with various thresholds of potential energy imposed on the hydrogen bonds (see Section 3.4 for details). We show here that the topological types of $F(G(P))$ for several such potential energy thresholds uniquely determine P among all known protein globules. Other such "injectivity results" for globules based on various robust protein descriptors are also presented. Further classification prediction results are analyzed; specifically, the prediction of domain from the topological type and other robust fatgraph invariants using a random forest method [7] is described in the two examples of *glycosyltransferase* and *pectate lyase C-like* with satisfactory accuracy, and a further study of the topology of $F(G(P))$ in the latter case is presented through the entire hierarchy of domains.

This paper is organized as follows: Section 1 introduces an abstract definition of *polypeptides*, which gives a precise mathematical formulation of the biophysics of a protein required for our model; a more detailed discussion of proteins from first principles is given in the beautiful book [10], which we heartily recommend. Section 2 introduces the notion of fatgraphs required here, whose corresponding surfaces may be nonorientable and contains basic results about them. In particular, a number of results, algorithms, and constructions are presented showing that our methods are amenable to fast computer implementation.

Section 3 is the heart of the paper and describes the fatgraph associated to a polypeptide structure in detail. Background on SO(3) graph connections is given in Section 3.1 and applied in Section 3.2, where we explain how the fatgraph building blocks associated with peptide units are concatenated. Section 3.3 discusses the addition of edges corresponding to hydrogen bonds, thus completing the basic construction of the fatgraph model of a polypeptide structure. Section 3.4 discusses this basic model and its natural generalizations and extensions for proteins and beyond. An alternative description of this model, which is more physically transparent but less mathematically tractable, is given in Appendix A, and the standard structural motifs of "alpha helices" and "beta strands" are discussed in this alternative model.

Robust invariants of fatgraphs are defined and studied in Section 4 providing countless meaningful new protein descriptors. Section 5 gives the empirical results mentioned above after first discussing certain practical aspects of implementing our methods. Finally, Section 6 contains closing remarks including several further biophysical applications of our methods that will appear in companions and sequels to this paper.

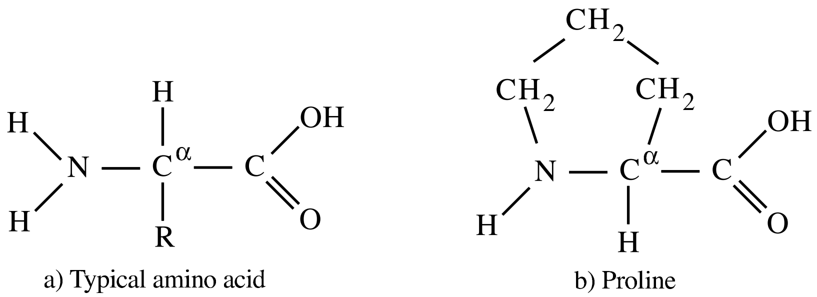

FIGURE 1.1. Chemical structure of amino acids.

1 Polypeptides

There are 20 *amino acids*,[1] 19 of which have the basic chemical structure illustrated in Figure 1.1a), where H, C, N, and O, respectively, denote hydrogen, carbon, nitrogen, and oxygen atoms, and the *residue* R is one of 19 specific possible submolecules; the one further amino acid called proline has the related chemical structure containing a ring CCCCN of atoms illustrated in Figure 1.1b). The residue ranges from a single hydrogen atom for the amino acid called glycine to a submolecule comprised of 19 atoms for the amino acid called tryptophan. All 20 amino acids are composed exclusively of H, C, N, and O atoms except for the amino acids called cysteine and methionine, each of which also contains a single sulfur atom.

In either case of Figure 1.1, the submolecule COOH depicted on the right-hand side is called the *carboxyl group*, and the NH_2 depicted on the left-hand side in Figure 1.1a) or the NHC on the left-hand side in Figure 1.1b) is called the *amine group*. The carbon atom bonded to the carboxyl and amine groups is called the *alpha carbon atom* of the amino acid, and it is typically denoted C^{α}. The alpha carbon atom is bonded to exactly one further atom in the residue, either a hydrogen atom in glycine or a carbon atom, called the *beta carbon atom*, in all other cases.

As illustrated in Figure 1.2, a sequence of L amino acids can combine to form a *polypeptide*, where the carbon atom from the carboxyl group of the i^{th} amino acid forms a *peptide bond* with the nitrogen atom from the amine group of the $(i+1)^{\text{st}}$ amino acid together with the resulting condensation of a water molecule comprised of an OH from the carboxyl group of the former and an H from the amine group of the latter for $i = 1, 2, \ldots, L-1$. The nature of this peptide bond and the accuracy of the implied geometry of Figure 1.2 will be discussed presently, and the further notation in the figure will be explained later.

The *primary structure* of a polypeptide is the ordered sequence $R_1, R_2, \ldots, R_L$ of residues or of amino acids occurring in this chain, i.e., a word in the 20-letter

[1] Strictly speaking, these 20 molecules are the "standard gene-encoded" amino acids, i.e., those amino acids determined from RNA via the genetic code; in fact, there are a few other nonstandard gene-encoded amino acids that are relatively rare in nature and which we shall ignore here.

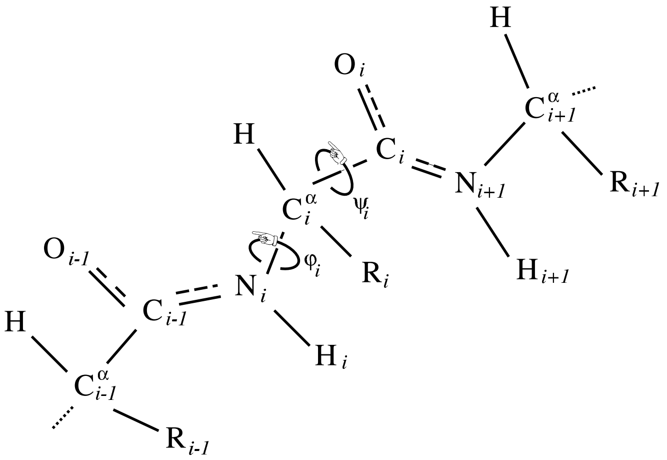

FIGURE 1.2. A polypeptide.

alphabet of amino acids of length L, which ranges in practice from $L = 3$ to $L \approx 30{,}000$. The carbon and nitrogen atoms that participate in the peptide bonds together with the alpha carbon atoms form the *backbone* of the polypeptide, which is described by

$$N_1 - C_1^{\alpha} - C_1 - N_2 - C_2^{\alpha} - C_2 - \cdots - N_i - C_i^{\alpha} - C_i - \cdots - N_L - C_L^{\alpha} - C_L,$$

indicating the standard enumeration of atoms along the backbone. The first amine nitrogen atom and the last carboxyl carbon atom, respectively, are called the N and C *termini* of the polypeptide.

The i^{th} *peptide unit* for $i = 1, 2, \ldots, L-1$ Is comprised of the consecutively bonded atoms $C_i^{\alpha} - C_i - N_{i+1} - C_{i+1}^{\alpha}$ in the backbone together with the oxygen atom O_i from the carboxyl group bonded to C_i and one further atom, namely, the remaining hydrogen atom H_{i+1} of the amine group except for proline, for which the further atom is the carbon preceding the nitrogen of the amine group in the proline ring.

This describes the basic chemical structure of a polypeptide, where the further physicochemical details about residues, for example, can be found in any standard text and will not concern us here.

There are several key geometrical facts about polypeptides as follows, where we refer to the center of mass of the Bohr model of a nucleus as the *center* of the atom and to the line segment connecting the centers of two chemically bonded atoms as the *bond axis*.

FACT 1.1 *For any polypeptide, there are the following geometric constraints*:

Fact A: *Each peptide unit is planar*; *i.e., the centers of the six constituent atoms of the peptide unit lie in a plane, and furthermore, the angles between the bond axes in a peptide unit are always fixed at 120°.*

Fact B: *At each alpha carbon atom* C_i^α*, the four bond axes* (*to hydrogen,* C_i*,* N_i*, and the residue, i.e., to the hydrogen atom of glycine or to the beta carbon atom in all other cases*) *are tetrahedral.*[2]

Fact C: *In the plane of each peptide unit, the centers of the two alpha carbons occur on opposite sides of the line determined by the bond axis of the peptide bond, except occasionally for the peptide unit preceding proline.*

We must remark immediately that these geometric facts are only effectively true; that is, the peptide unit is *almost* planar and the angles between bond axes in a peptide unit are *nearly* 120°, for example, in Fact A; thus, the depiction in Figure 1.2 of the peptide unit is nearly geometrically accurate. In nature, thermal and other fluctuations do slightly affect the geometric absolutes stated in Fact 1.1, but we shall nevertheless take these facts as geometric absolutes in constructing our model.

Fact A is fundamental to our constructions, and it arises from purely quantum effects: the planar character is provided by the "sp^2 hybridization" of electrons in the C_i and N_{i+1} atoms in the i^{th} peptide unit, and the peptide unit is rigid because of additional bonding with O_i of the two p-electrons from C_i, N_{i+1} not involved in the sp^2 hybridization. This complexity of shared electrons is why the peptide bond and the bond between C_i and O_i are often drawn as "partial double bonds" as in Figure 1.2. In contrast, Fact B is a standard consequence of the valence of carbon atoms in the Bohr model absent any quantum mechanical hybridization of electrons.

As a point of terminology, Fact C expresses that except for proline, the peptide unit occurs in what is called the "transconformation," and the complementary possibility (with the centers of the alpha carbon atoms in a peptide unit on the same side of the line determined by the axis of the peptide bond) is called the "cis-conformation." This geometric constraint follows from the simple fact that in the cis-conformation, the two "large" alpha carbon atoms in the peptide unit would be so close together as to be energetically unfavorable. In contrast, for cis-proline, the two conformations are comparable since in either case, two carbons (either the two alpha carbons or one alpha and the delta carbon in the proline ring) must be close together; nevertheless, cis-proline, as opposed to trans-proline, occurs only about 10 percent of the time in nature since the latter is still somewhat energetically favorable. Peptide units preceding residues other than proline also occur in the cis-conformation but only extremely rarely. This exemplifies a general trend: somewhat energetically unfavorable conformations do occur but more rarely than favorable ones, and extremely energetically unfavorable conformations occur extremely rarely if at all.

The mechanism underlying Fact C is that atoms cannot "bump into each other," or more precisely, their centers cannot be closer than their van der Waals radii

[2] Another geometric constraint on any gene-encoded protein is that when viewed along the bond axis from hydrogen to C_i^α, the bond axes occur in the cycle ordering corresponding to C_i, residue, N_i. This imposes various chiral constraints on proteins but plays no role in our basic fatgraph model.

allow, and this is called a *steric constraint*, which will be pertinent to subsequent discussions.

Facts A and B together indicate the basic geometric structure of a polypeptide: a sequence of planar peptide units meeting at tetrahedral angles at the alpha carbon atoms; these planes can rotate rather freely about the axes of these tetrahedral bond axes, and this accounts for the relative flexibility of polypeptides. For a polypeptide at equilibrium in some environment, the dihedral angle along the bond axis of $\mathrm{N}_i - \mathrm{C}_i^\alpha$ (and $\mathrm{C}_i^\alpha - \mathrm{C}_i$) between the bond axis of $\mathrm{C}_{i-1} = \mathrm{N}_i$ (and $\mathrm{N}_i - \mathrm{C}_i^\alpha$) and the bond axis of $\mathrm{C}_i^\alpha - \mathrm{C}_i$ (and $\mathrm{C}_i = \mathrm{N}_{i+1}$) is called the *conformational angle* φ_i (and ψ_i, respectively); see Figure 1.2. Illustrating the physically possible pairs $(\varphi_i, \psi_i) \in \mathbb{S}^1 \times \mathbb{S}^1$, steric constraints for each amino acid can be plotted in what is called a Ramachandran plot; cf. Figure 3.3; in particular, for any polypeptide at equilibrium in any environment, φ_i is bounded away from 0 because of steric constraints involving C_{i-1} and C_i.

This completes our discussion of the intrinsic physicochemical and geometric aspects of polypeptides underlying our model. The remaining such aspect of importance to us depends critically upon the ambient environment in which the polypeptide occurs.

An *electronegative* atom is one that tends to attract electrons, and examples of such atoms include C, N, and O in this order of increasing such tendency. When an electronegative atom approaches another electronegative atom that is chemically bonded to a hydrogen atom, the two electronegative atoms can share the electron envelope of the hydrogen atom and attract one another through a *hydrogen bond*. A hydrogen bond has a well-defined potential energy determined on the basis of electrostatics that can be computed from the spatial locations of its constituent atoms and the physical properties of its environment.[3]

For example, the O_i or $\mathrm{N}_{i+1} - \mathrm{H}_{i+1}$ in one peptide unit can form a hydrogen bond with the $\mathrm{N}_{j+1} - \mathrm{H}_{j+1}$ or O_j in another peptide unit, respectively, where $i \neq j$ owing to rigidity and fixed lengths of 1.3–1.6 Å of bond axes. For another example, many of the remarkable properties of water arise from the occurrence of hydrogen bonds among HOH and OH_2 molecules. The absolute potential energy of hydrogen bonds is rather large, so a polypeptide in a given environment seeks to saturate as many hydrogen bonds as possible subject to steric and other physicochemical and geometric constraints. For example, in an aqueous environment, the oxygen and nitrogen atoms in the peptide units of a polypeptide might

[3] For instance, in the standard method called DSSP [17] where r_{XY} denotes the distance between the centers of atoms $X, Y \in \{\mathrm{H}, \mathrm{N}, \mathrm{O}\}$ in Å and the location of H is determined from idealized geometry and bond lengths in practice, the assignment of potential energy to the hydrogen bond between O and NH in a water environment is given by $q_1 q_2\{r_{\mathrm{ON}}^{-1} + r_{\mathrm{CH}}^{-1} - r_{\mathrm{OH}}^{-1} - r_{\mathrm{CN}}^{-1}\} \times 332$ kcal/mole, where $q_1 = 0.42$ and $q_2 = 0.20$ based on the respective assignment of partial charges -0.42e and $+0.20$e to the carboxyl carbon and amine nitrogen with e representing the election charge. This is obviously only a rough but standard approximation of the actual electrostatics that is built into the DSSP definition.

form hydrogen bonds with one another or with the ambient water molecules of their environment, and there may also occur hydrogen bonding involving atoms comprising the residues or the alpha carbons.

Suppose that a polypeptide is at equilibrium, i.e., at rest, in some environment. Its *tertiary structure* in that environment is the specification of the spatial coordinates of the centers of all of its constituent atoms. Furthermore, fix some *energy cutoff* and regard a pair O_i and N_j of backbone atoms as being hydrogen bonded if the potential energy discussed above is less than this energy cutoff; a standard convention is to take the energy cutoff to be −0.5 kcal/mole.[4] The *secondary structure*[5] of the polypeptide at equilibrium in an environment is the specification of hydrogen bonding as determined by an energy cutoff among its constituent backbone atoms O_i and N_j for $i, j = 1, 2, \dots, L$.

Certain polypeptides occur as the "proteins" that regulate and effectively define life as we know it. The collective knowledge of protein primary structures is deposited in the manually curated SWISS-PROT data bank [2], which contains about 400,000 distinct entries, and the computer-curated UNI-PROT data bank [31], which contains about 6,000,000 entries. These data are readily accessible at `www.ebi.ac.uk/swissprot` and `www.uniprot.org`, respectively. The collective knowledge of protein tertiary structure is deposited in the Protein Data Bank (PBD) [3], which contains roughly 55,000 proteins at this moment, where the atomic locations of each of the constituent atoms of each of these proteins is recorded; each entry in the PDB, i.e., each protein, thus comprises a vast amount of data. Atomic locations in the PDB should be taken with an experimental uncertainty of 0.2 Å, and the conformational angles φ and ψ computed from them should be taken with an experimental uncertainty of 15°–20°; however, the unit displacement vectors of bond axes along the backbone, upon which our model is based, are substantially better determined [11]. It is worth emphasizing that the quality of data in the PDB varies wildly from one entry to another, so these nominal experimental thresholds give only a lower bound to the indeterminacy.

Upon postulating definitions of the various secondary structure elements in terms of properties of the atomic locations, protein secondary structure can be calculated from tertiary structure. A standard such method is called the Dictionary of Secondary Structures for Proteins (DSSP) [17], and proprietary software for these calculations and DSSP files for each PDB entry can be found at `http://swift.cmbi.ru.nl/gv/dssp`. Hydrogen bond strengths and various conformational angles are also output as part of the calculations of DSSP.

[4] Other methods [19, 20] of determining hydrogen bonds are also employed.

[5] This is a slight abuse of terminology as biologists might call this rather "supersecondary structure"; we shall explain this distinction further when it is appropriate.

2 Fatgraphs

2.1 Surfaces

According to the classification of surfaces [21], a compact and connected surface F is uniquely determined up to homeomorphism by the specification of whether it is orientable together with its genus $g = g(F)$ and number $r = r(F)$ of boundary components, or equivalently, by either g or r and its Euler characteristic

$$\chi = \chi(F) = \begin{cases} 2 - 2g - r & \text{if } F \text{ is orientable,} \\ 2 - g - r & \text{if } F \text{ is nonorientable.} \end{cases}$$

It is useful to define the *modified genus* of a connected surface F to be

$$g^* = g^*(F) = \begin{cases} g & \text{if } F \text{ is orientable,} \\ \frac{g}{2} & \text{if } F \text{ is nonorientable,} \end{cases}$$

so the formula $\chi = 2 - 2g^* - r$ holds in either case.

Recall [21] that the *orientation double cover* of a surface F is the oriented surface $\widetilde{F}$ together with the continuous map $p : \widetilde{F} \to F$ so that for every point $x \in F$ there is a disk neighborhood U of x in F, where $p^{-1}(U)$ consists of two components on each of which p restricts to a homeomorphism and where the further restrictions of p to the boundary circles of these two components give both possible orientations of the boundary circle of U. Such a covering $p : \widetilde{F} \to F$ always exists, and its properties uniquely determine $\widetilde{F}$ up to homeomorphism and p up to its natural equivalence. In particular, if F is connected and orientable, then $\widetilde{F}$ has two components with opposite orientations, each of which is identified with F by p. Furthermore, provided F is connected, F is nonorientable if and only if $\widetilde{F}$ is connected, and a closed curve in F lifts to a closed curve in $\widetilde{F}$ if and only if a neighborhood of it in F is homeomorphic to an annulus as opposed to a Möbius band.

2.2 Fatgraphs and Their Associated Surfaces

Consider a finite graph G in the usual sense of the term comprised of vertices $V = V(G)$ and edges $E = E(G)$ that do not contain their endpoints and where an edge is not necessarily uniquely determined by its endpoints; in other words, G is a finite one-dimensional CW complex. Our standard notation will be $v = v(G) = \#V$ and $e = e(G) = \#E$, where $\#X$ denotes the cardinality of a set X. To avoid cumbersome cases in what follows, we shall assume that no component of G consists of a single vertex or a single edge with distinct endpoints. Removing a single point from each edge produces a subspace of G, each component of which is called a *half-edge*. A half-edge that contains $u \in V$ in its closure is said to be *incident* on u, and the number of distinct half-edges incident on u is the *valence* of u.

A *fattening* on G is the specification of a cyclic ordering on the half-edges incident on u for each $u \in V$, and an X-*coloring* on G is a function $E \to X$ for any set X.

A *fatgraph* G is a graph endowed with a fattening together with a coloring by a set with two elements, where we shall refer to the two colors on edges as "twisted" and "untwisted." A fatgraph G uniquely determines a surface $F(G)$ with boundary as follows:

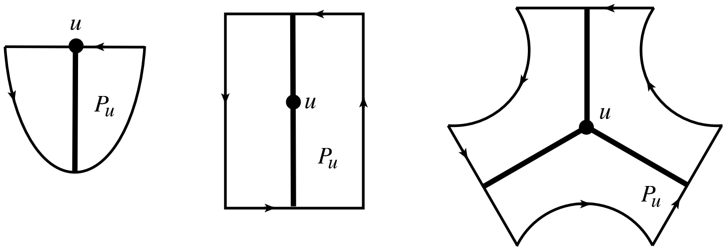

FIGURE 2.1. The polygon P_u associated with a vertex u.

CONSTRUCTION 2.1 For each vertex $u \in V$ in G of valence $k \geq 2$, we associate an oriented surface diffeomorphic to a polygon P_u of $2k$ sides containing in its interior a single vertex of valence k. Each edge incident on this vertex is also incident on a univalent vertex contained in every other side of P_u, and these are identified with the half-edges of G incident on u so that the induced counterclockwise cyclic ordering on the boundary of P_u agrees with the fattening of G about u. For a vertex u of valence $k = 1$, the corresponding surface P_u contains u in its boundary. See Figure 2.1. The surface $F(G)$ is the quotient of the disjoint union $\bigsqcup_{u \in V} P_u$, where the frontier edges, which are oriented with the polygons on their left, are identified by a homeomorphism if the corresponding half-edges lie in a common edge of G; this identification of oriented segments is orientation preserving if and only if the edge is twisted. The graphs in the polygons P_u for $u \in V$ combine to give a fatgraph embedded in $F(G)$ with its univalent vertices in the boundary, which is identified with G in the natural way so that we regard $G \subseteq F(G)$.

Our standard notation will be to set

$$\begin{aligned} r(G) &= r(F(G)) \quad &&\text{(number of boundary components of } F(G)), \\ g^*(G) &= g^*(F(G)) \quad &&\text{(modified genus of } F(G)). \end{aligned}$$

It is often convenient to regard a fatgraph more pictorially by considering the planar projection of a graph embedded in 3-space, where the cyclic ordering is given near each vertex by the counterclockwise ordering in the plane of projection and edges can be drawn with arbitrary under/over crossings; we also depict untwisted edges as ordinary edges and indicate twisted edges with an icon ×, or more generally, take this as defined modulo 2 so that an even number of icons × represents an

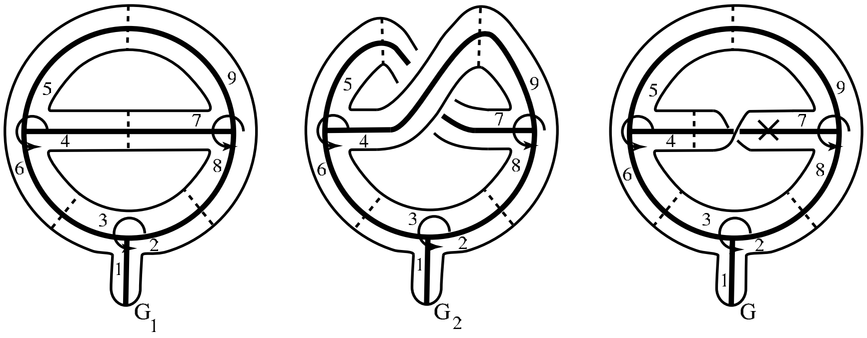

FIGURE 2.2. The surface associated to a fatgraph.

untwisted edge and an odd number represents a twisted edge. Several examples of fatgraphs and their corresponding surfaces are illustrated in Figure 2.2, where the bold lines indicate the planar projection of the fatgraph, the dotted lines indicate the gluing along edges of polygons, and the further notation in the figure will be explained later.

The graph G is evidently a strong deformation retract of $F(G)$, so the Euler characteristic is $\chi(F(G)) = \chi(G) = v(G) - e(G)$, and the boundary components of $F(G)$ are composed of the frontier edges of $\bigsqcup_{u\in V} P_u$ that do not correspond to half-edges of G.

PROPOSITION 2.2 *Suppose that G is a fatgraph and $X, Y \subseteq E(G)$ are disjoint collections of edges. Change the color, twisted or untwisted, of the edges in X and delete from G the edges in Y to produce another fatgraph G', whose cyclic orderings on half-edges are induced from those on G in the natural way. Then $|r(G) - r(G')| \leq \#X + \#Y$.*

PROOF: By the triangle inequality, it suffices to treat the case that $X \cup Y = \{f\}$, and we set $r = r(G)$. If $f \in E(G)$ is incident on a univalent vertex, then neither changing the color of nor deleting f alters r, so we may assume that this is not the case. Consider an arc a properly embedded in $F(G)$ meeting f in a single transverse intersection and otherwise disjoint from G. Rather than changing the color on f to produce G', let us instead cut $F(G)$ along a and then reglue along the two resulting copies of a, reversing orientation to produce a surface homeomorphic to $F(G')$. If the endpoints of a occur in a common boundary component of $F(G)$, then the change of color on f either leaves r invariant or increases it by 1, and if they occur in different boundary components, then the change of color on f necessarily decreases r by 1. For the remaining case, rather than removing the edge f to produce G', let us instead consider cutting $F(G)$ along a to produce a surface homeomorphic to $F(G')$. If the endpoints of a occur in the same boundary component of $F(G)$, then cutting on a either leaves r invariant or increases it by 1, and if they occur in different boundary components, then the cut on a decreases r by 1. □

We say that a fatgraph G is *untwisted* if all of its edges are untwisted, and this is evidently a sufficient but not a necessary condition for $F(G)$ to be orientable.

Remark 2.3. Suppose that G is an untwisted fatgraph. Let us emphasize that the genus of $F(G)$ is *not* the classical genus of the underlying graph, i.e., the least genus orientable surface in which the underlying graph can be embedded. Rather, the classical genus of the underlying graph is the least genus of an orientable surface $F(G)$ arising from all possible fattenings on the underlying graph.

We say that two fatgraphs G_1 and G_2 are *strongly equivalent* if there is an isomorphism of the graphs underlying G_1 and G_2 that respects the cyclic orderings and preserves the coloring and that they are *equivalent* if there is a homeomorphism from $F(G_1)$ to $F(G_2)$ that maps $G_1 \subseteq F(G_1)$ to $G_2 \subseteq F(G_2)$. It is clear that strong equivalence implies equivalence and that equivalence implies that the corresponding surfaces are homeomorphic; neither converse holds in general.

Given a vertex u of G, define the *vertex flip* of G at u by reversing the cyclic ordering on the half-edges incident on u and adding another icon × to each half-edge incident on u. In particular, a vertex flip on a univalent vertex simply adds an icon × to the edge incident upon it.

PROPOSITION 2.4 *Two untwisted fatgraphs are equivalent if and only if they are strongly equivalent. Two arbitrary fatgraphs G_1 and G_2 are equivalent if and only if there is a third fatgraph G that arises from G_1 by a finite sequence of vertex flips so that G and G_2 are strongly equivalent. In particular, if G arises from G_1 by a vertex flip, then G and G_1 are equivalent.*

PROOF: In case G_1 and G_2 are untwisted, a homeomorphism from $F(G_1)$ to $F(G_2)$ mapping G_1 to G_2 restricts to a strong equivalence of G_1 and G_2, and the converse follows by construction in any case, as already observed, thus proving the first assertion.

The third assertion follows since a flip on a vertex u of G_1 corresponds to simply reversing the orientation of the polygon P_u in the construction of $F(G)$, i.e., in our graphical depiction, removing the neighborhood of u from the plane of projection, turning it upside down in 3-space, and then replacing it in the plane of projection at the expense of twisting one further time each incident half-edge of G. This evidently extends to a homeomorphism of $F(G_1)$ to $F(G)$ that maps G_1 to G but does not preserve coloring.

Since strong equivalence implies equivalence by construction and equivalence of fatgraphs is clearly a transitive relation, if there is such a fatgraph G as in the statement of the proposition, then G_1 and G_2 are indeed equivalent. For the converse, we may and shall assume that G_1 and G_2 are connected.

Consider a fatgraph G with v vertices and e edges, and choose a maximal tree T of G. There are $1 - \chi(G) = 1 - v + e$ edges in $G - T$ since we may collapse T

to a point without changing $v - e$, which is therefore the Euler characteristic of the collapsed graph comprised of a single vertex and one edge for each edge of $G - T$.

We claim that there is a composition of flips of vertices in G that results in a fatgraph with any specified twisting on the edges in T. To see this, consider the collection of all functions from the set of edges of G to $\mathbb{Z}/2$, a set with cardinality 2^e. Vertex flips act on this set of functions in the natural way, and there are evidently 2^v possible compositions of vertex flips. The simultaneous flip of all vertices of G acts trivially on this set of functions and corresponds to reversing the cyclic orderings at all vertices, so only 2^{v-1} such compositions may act nontrivially. Insofar as $2^e/2^{v-1} = 2^{1-v+e}$ and there are $1 - v + e$ edges of $G - T$ by the previous paragraph, the claim follows.

Finally, suppose that G_1 and G_2 are equivalent and let $\phi : F(G_1) \to F(G_2)$ be a homeomorphism of surfaces that restricts to a homeomorphism of G_1 to G_2. Performing a vertex flip on G_1 and identifying edges before and after in the natural way produces a fatgraph in which T is still a maximal tree and which is again equivalent to G_2, according to previous remarks, by a homeomorphism still denoted ϕ, which maps T to the maximal tree $\phi(T) \subset G_2$. By the previous paragraph, we may apply a composition of vertex flips to G_1 to produce a fatgraph G so that an edge of the maximal tree $T \subset G$ is twisted if and only if its image under ϕ is twisted.

Adding an edge of $G-T$ to T produces a unique cycle in G, and a neighborhood of this cycle in $F(G)$ is either an annulus or a Möbius band with a similar remark for edges of $G_2-\phi(T)$. Since ϕ restricts to a homeomorphism of the corresponding annuli or Möbius bands in $F(G)$ and $F(G_2)$, an edge of $G-T$ is twisted if and only if its image under ϕ is twisted. It follows that G and G_2 are strongly equivalent as desired. □

2.3 Fatgraphs and Permutations

We shall adopt the standard notation for a permutation on a set S writing $(s_1, s_2, \dots, s_k)$ for the cyclic permutation $s_1 \mapsto s_2 \mapsto \cdots \mapsto s_k \mapsto s_1$ on distinct elements $s_1, s_2, \dots, s_k \in S$, called a *transposition* if $k = 2$, and shall compose permutations σ and τ on S from right to left, so that $\sigma \circ \tau(s) = \sigma(\tau(s))$. An *involution* is a permutation τ so that $\tau \circ \tau = 1_S$, where 1_S denotes the identity map on S. Two permutations are *disjoint* if they have disjoint supports, so disjoint permutations necessarily commute.

Fix a fatgraph G. A *stub* of G is a half-edge that is not incident on a univalent vertex of G. There are exactly two nonempty connected fatgraphs with no stubs, namely, the two we have proscribed consisting of a single vertex with no incident half-edges and a single edge with distinct endpoints.

A fatgraph G determines a triple $(\sigma(G), \tau_u(G), \tau_t(G))$ of permutations on its set $S = S(G)$ of stubs as follows:

CONSTRUCTION 2.5 For each vertex u of G of valence $k \geq 2$ with incident stubs $s_1, s_2, \dots, s_{k(u)}$ in a linear ordering compatible with the cyclic ordering given by the fattening on G, consider the cyclic permutation $(s_1, s_2, \ldots, s_{k(u)})$. By construction, the cyclic permutations corresponding to distinct vertices of G are disjoint. The composition

$$\sigma(G) = \prod_{\substack{\{\text{vertices } u \in V : \\ u \text{ has valence } \geq 2\}}} (s_1, s_2, \ldots, s_{k(u)})$$

is thus well-defined independently of the order in which the product is taken, and likewise for the compositions of transpositions

$$\tau_u(G) = \prod_{\substack{\{\text{pairs of distinct stubs } h, h' \text{ contained} \\ \text{in some untwisted edge of } G\}}} (h, h'),$$

$$\tau_t(G) = \prod_{\substack{\{\text{pairs of distinct stubs } h, h' \text{ contained} \\ \text{in some twisted edge of } G\}}} (h, h').$$

Notice that $\sigma(G)$ has no fixed points because we have taken the product over vertices of valence at least 2, and $\tau_u(G)$ and $\tau_t(G)$ are disjoint involutions whose fixed points are the stubs corresponding to the univalent vertices of G.

For example, enumerating the stubs of the fatgraphs G_1, G_2, and G_3 as illustrated in Figure 2.2, we have:

$$\begin{aligned}
&\sigma(G_1) = \sigma(G_2) = \sigma(G_3) = (1,2,3)(4,5,6)(7,8,9), \\
&\tau_u(G_1) = (2,8)(3,6)(4,7)(5,9), \quad \tau_t(G_1) = 1_S, \\
&\tau_u(G_2) = (2,8)(3,6)(4,9)(5,7), \quad \tau_t(G_2) = 1_S, \\
&\tau_u(G_3) = (2,8)(3,6)(5,9), \quad \tau_t(G_3) = (4,7).
\end{aligned}$$

Remark 2.6. There is another treatment of fatgraphs as triples of permutations on the set of all half-edges instead of stubs, where the univalent vertices are expressed as fixed points of the analogue of σ. Moreover, there is a transposition in the analogue of $\tau_u \circ \tau_t$ corresponding to each half-edge, but the formulation we have given here, which treats univalent vertices as "endpoints of half-edges rather than endpoints of edges," does not require these additional transpositions. Since our model will have a plethora of univalent vertices, we prefer the more "efficient" version described above, which is just a notational convention for permutations.

Define a *labeling* on a fatgraph G with N stubs to be a linear ordering on its stubs, i.e., a bijection from the set of stubs of G to the set $\{1, 2, \ldots, N\}$.

PROPOSITION 2.7 *Fix some natural number $N \geq 2$. The map $G \mapsto (\sigma(G), \tau_u(G), \tau_t(G))$ of Construction* 2.5 *induces a bijection between the set of strong equivalence classes of fatgraphs with N stubs and the set of all conjugacy classes of triples (σ, τ_u, τ_t) of permutations on N letters, where σ is fixed-point free and τ_u and τ_t are disjoint involutions.*

PROOF: The assignment $G \mapsto (\sigma(G), \tau_u(G), \tau_t(G))$ induces a mapping from the set of labeled fatgraphs with N stubs to the set of triples of permutations on $\{1, 2, \dots, N\}$ in the natural way. This induced mapping has an obvious two-sided inverse, where the labeled fatgraph is constructed directly from the triple of permutations; we are here using our convention that no component of G is a single vertex or a single edge with distinct univalent endpoints. A strong equivalence of labeled fatgraphs induces a bijection of $\{1, 2, \dots, N\}$ that conjugates their corresponding triples of permutations to one another and conversely, so the result follows. $\square$

CONSTRUCTION 2.8 Suppose that G is a fatgraph with triple (σ, τ_u, τ_t) of permutations on its set S of stubs determined by Construction 2.5. Construct a new set $\overline{S} = \{\overline{s} : s \in S\}$ and a new permutation $\overline{\sigma}$ on $\overline{S}$ where there is one k-cycle $(\overline{s}_k, \dots \overline{s}_2, \overline{s}_1)$ in $\overline{\sigma}$ for each k-cycle $(s_1, s_2, \dots, s_k)$ in σ. Construct from τ_u a new permutation $\overline{\tau}_u$ on $\overline{S}$, where there is one transposition $(\overline{s}_1, \overline{s}_2)$ in $\overline{\tau}_u$ for each transposition (s_1, s_2) in τ_u, and construct yet another new permutation $\overline{\tau}_t$ on $S \sqcup \overline{S}$ from τ_t, where there are two transpositions $(\overline{s}_1, s_2)$ and $(s_1, \overline{s}_2)$ in $\tilde{\tau}_t$ for each transposition (s_1, s_2) in τ_t. Finally, define permutations on $S \sqcup \overline{S}$ by

$$\sigma' = \sigma \circ \overline{\sigma},$$

$$\tau' = \tau_u \circ \overline{\tau}_u \circ \overline{\tau}_t,$$

where the order of composition on the right-hand side is immaterial because the permutations are disjoint in each case.

PROPOSITION 2.9 *Suppose that Construction* 2.5 *assigns the triple* (σ, τ_u, τ_t) *of permutations to the fatgraph* G *with set* S *of stubs, let* σ' *and* τ' *be determined from them according to Construction* 2.8, *and consider the untwisted fatgraph* G' *determined by Construction* 2.5 *from the triple* $(\sigma', \tau', 1_{S \sqcup \overline{S}})$. *Then* $F(G')$ *is the orientation double cover of* $F(G)$, *and the covering transformation is described by* $s \leftrightarrow \overline{s}$. *In particular, provided* $F(G)$ *is connected,* $F(G')$ *is connected if and only if* $F(G)$ *is nonorientable. Furthermore, there is a one-to-one correspondence between the boundary components of* $F(G')$ *and the orientations on the boundary components of* $F(G)$; *i.e.,* $F(G')$ *has twice as many boundary components as* $F(G)$.

PROOF: The surface $F(G')$ has the required properties of the orientation double cover by construction, so the first two claims follow from the general principles articulated in Section 2.1. Since each boundary component of $F(G)$ evidently has a neighborhood in $F(G)$ homeomorphic to an annulus, the final assertion follows as well. $\square$

PROPOSITION 2.10 *Adopt the hypotheses and notation of Proposition* 2.9 *and consider the composition* $\rho' = \sigma' \circ \tau'$.

(i) *The orientations on the boundary components of* $F(G)$ *are in one-to-one correspondence with the cycles of* ρ'. *More explicitly, suppose that*

$$s_1^1 s_1^2 s_2^1 s_2^2 \cdots s_n^1 s_n^2$$

is the ordered sequence of stubs traversed by an oriented edge-path in G representing a boundary component of $F(G)$ with some orientation, where s_j^1, s_j^2 are contained in a common edge of G and perhaps $s_j^1 = s_j^2$ if they are contained in an edge incident on a univalent vertex for $j = 1, 2, \dots, n$. Erasing the bars on elements from the corresponding cycle of ρ' produces the sequence $(s_1^2, s_2^2, \dots, s_n^2)$ of stubs of G serially traversed by the corresponding oriented boundary component of $F(G)$, called a reduced cycle *of ρ'.*

(ii) *There is the following algorithm to determine whether G is connected in terms of the associated triple (σ, τ_u, τ_t) of permutations. For any linear ordering on S, let X be the subset of S in the reduced cycle of ρ' containing the first stub.* (*) *If $X = S$, then G is connected, and the algorithm terminates. If $X \neq S$, then consider the existence of a least stub $s \in X - S$ so that $\tau(s) \in X$. If there is no such stub s, then G is not connected, and the algorithm terminates. If there is such a stub s, then update X by adding to it the subset of S in the reduced cycle of ρ' containing s. Go to* (*).

PROOF: Let us first consider the case that $\tau_t = 1_{S \sqcup \bar{S}}$; i.e., G is untwisted, and set $\tau = \tau_u$.

For the first part, consider a stub s of G and the effect of $\sigma \circ \tau$ on s. The stub s is contained in an edge incident on a univalent vertex if and only if s is a fixed point of τ by construction, and $\sigma(s) = \sigma(\tau(s))$ in this case is the stub following s in the cyclic ordering at the nonunivalent endpoint of this edge. In the contrary case that s is not a fixed point of τ, the stubs s and $\tau(s)$ are half-edges contained in a common edge of G, and $s, \tau(s), \rho(s) = \sigma(\tau(s))$ is likewise a consecutive triple of stubs occurring in an edge-path of G corresponding to a boundary component of $F(G)$ oriented with $F(G)$ on its left. It follows that a cycle of $\sigma \circ \tau$ is comprised of every other stub traversed by an edge-path in G that corresponds to a boundary component of $F(G)$ oriented in this way, proving the first part.

For the second part, the collection of stubs in X always lies in a single component of G in light of the previous remarks, so if at some stage of the algorithm $X = S$, then G is indeed connected. If at some stage of the algorithm there is no stub s with $\tau(s) \in X$, then X is comprised of all the stubs in some component of G in light of the previous discussion, so $X \neq S$ in this case implies that G has at least two components.

Turning now to the general case, $F(G')$ is the orientation double cover of $F(G)$, and the induced projection map on stubs just erases the bars by Proposition 2.9. The proof in this case is therefore entirely analogous. □

To exemplify these constructions and results for the fatgraphs illustrated in Figure 2.2, we find

$$\sigma(G_1) \circ \tau_u(G_1) = (5,7)(3,4,8)(1,2,9,6),$$
$$\sigma(G_2) \circ \tau_u(G_2) = (1,2,9,5,8,3,4,7,6).$$

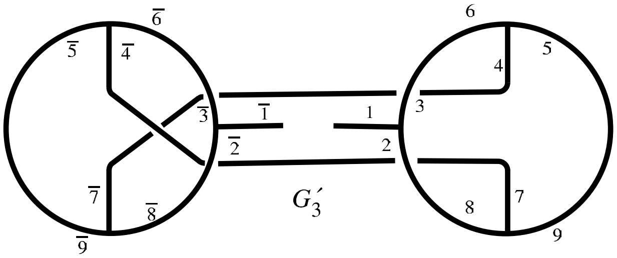

FIGURE 2.3. Example of the orientation double cover.

Thus, $r(G_1) = 3$ and $r(G_2) = 1$, and since $\chi(G_1) = \chi(G_2) = -1$, the (modified) genera are $g^*(G_1) = 0$ and $g^*(G_2) = 1$.

As to G_3, according to Construction 2.8 and Proposition 2.9, the permutations for the orientation double cover are given by

$$\sigma' = (1,2,3)(4,5,6)(7,8,9)(\overline{3},\overline{2},\overline{1})(\overline{6},\overline{5},\overline{4})(\overline{9},\overline{8},\overline{7}),$$
$$\tau' = (2,8)(3,6)(5,9)(\overline{2},\overline{8})(\overline{3},\overline{6})(\overline{5},\overline{9})(4,\overline{7})(\overline{4},7).$$

The untwisted fatgraph G_3' corresponding to $(\sigma', \tau', 1_{S(G_3)\sqcup\overline{S}(G_3)})$ is illustrated in Figure 2.3, and it is connected reflecting the fact that $F(G_3)$ is nonorientable. The cycles of $\rho' = \sigma' \circ \tau'$ are given by

$$(1,2,9,6), (\overline{1},\overline{3},\overline{5},\overline{8}) \quad \text{and} \quad (\overline{2},\overline{7},5,7,\overline{6}), (3,4,\overline{9},\overline{4},8)$$

corresponding to the oriented boundary cycles of G_3', and the reduced cycles of ρ' are therefore

$$(1,2,9,6), (1,3,5,8) \quad \text{and} \quad (2,7,5,7,6), (3,4,9,4,8),$$

each pair corresponding to the two orientations of a single boundary component of $F(G_3)$. It follows that $r(G_3) = 2$ and thus $g^*(G_3) = \frac{1}{2}$ since again $\chi(G_3) = -1$.

2.4 Fatgraphs on the Computer

Given a linear ordering on the vertices of a fatgraph, we may choose an a priori labeling on it that is especially convenient, where the stubs about a fixed vertex are consecutive and the stubs about each vertex precede those of each succeeding vertex as in Figure 2.2. Owing to Proposition 2.7, the strong equivalence class of a fatgraph G with set S of stubs can conveniently be stored on the computer as a triple (σ, τ_u, τ_t) of permutations on the labels $\{1, 2, \ldots, \#S\}$ of stubs. The number of nonunivalent vertices of G is the number of disjoint cycles in σ, the number of edges of G that are not incident on a univalent vertex is the number of disjoint transpositions in $\tau_u \circ \tau_t$, and the Euler characteristic of G or $F(G)$ is given by the former minus the latter. Construction 2.8 provides an algorithm, which is easily implemented on the computer, to produce a triple $(\sigma', \tau', 1_{S\sqcup\overline{S}})$ from (σ, τ_u, τ_t) that determines an untwisted fatgraph G' whose corresponding surface

$F(G')$ is the orientation double cover of $F(G)$ according to Proposition 2.9. Proposition 2.10(i) provides an algorithm to determine the compatibly oriented boundary components of $F(G')$ and hence the boundary components of $F(G)$ itself, and Proposition 2.10(ii) then gives an algorithm to determine whether G' is connected from this data, where both of these methods are again easily implemented on the computer.

In our applications of these techniques, the fatgraph G will typically be connected as we now assume. The orientability of $F(G)$ can thus be ascertained from the connectivity of $F(G')$. The boundary components of $F(G)$, and their number r in particular, can be determined, as above, and hence the modified genus $g^* = (2 - r - \chi)/2$ is likewise easily computed. Thus, the topological type of $F(G)$ can be determined algorithmically on the computer from the triple (σ, τ_u, τ_t) of permutations for a connected fatgraph G, and the particular edge-paths in G corresponding to boundary components of $F(G)$ can be ascertained from the cycles of $\sigma' \circ \tau'$.

3 The Model

We take as input to the method the specification for a polypeptide at equilibrium in some environment the following data:

Input (i): the primary structure given as a sequence R_i of letters in the 20-letter alphabet of amino acids for $i = 1, 2, \dots, L$,

Input (ii): the specification of hydrogen bonding among the various nitrogen and oxygen atoms $\{\mathrm{N}_i, \mathrm{O}_i : i = 1, 2, \dots, L\}$ described as a collection $\mathcal{B}$ of pairs (i, j) indicating that $\mathrm{N}_i - \mathrm{H}_i$ is hydrogen bonded to O_j, where $i, j \in \{1, 2, \dots, L\}$,

Input (iii): the displacement vectors $\vec{x}_i$ from C_i to N_{i+1}, $\vec{y}_i$ from C_i^α to C_i, and $\vec{z}_i$ from N_{i+1} to C_{i+1}^α in each peptide unit for $i = 1, 2, \dots, L - 1$.

These data, which we shall term a *polypeptide structure* P, are either immediately given in or readily derived from the PDB and DSSP files for a folded protein. Practical and other details concerning the determination of these inputs will be discussed in Section 5.1.

A fatgraph is constructed from a polypeptide structure in two basic steps: modeling the backbone using the planarity of the peptide units and the conformational geometry along the backbone based on input (iii), and then adding edges to the model of the backbone for the hydrogen bonds based on inputs (ii) and (iii); finally, certain edges or vertices of the constructed fatgraph may be labeled by residues or their constituent atoms using input (i).

Roughly, inputs (i)–(iii) correspond to the primary, secondary, and tertiary structure of the polypeptide. We must emphasize that the basic fatgraph we construct actually depends only on inputs (ii)–(iii), and input (i) is used only to label it. In a more refined all-atom version of our construction discussed later, the primary structure plays a more fundamental role and does affect the construction of the fatgraph. From a more philosophical point of view, one could argue that even the refined fatgraph structure is determined by primary sequence, so a hidden empirical dependence on primary structure is already manifest in our basic fatgraph model.

We shall assume that input (ii) is consistently based upon fixed energy thresholds with each nitrogen or oxygen atom involved in at most one hydrogen bond (so-called "simple" hydrogen bonding) and relegate the discussion of more general models (with so-called bifurcated hydrogen bonding) to Section 3.4. The assumption thereby imposed on $\mathcal{B}$ in input (ii) is that if $(i, j), (i', j') \in \mathcal{B}$, then $i = i'$ if and only if $j = j'$.

To each peptide unit is associated a fatgraph building block as illustrated in Figure 3.1. These building blocks are concatenated to produce a model of the backbone as illustrated in Figure 3.2, where the determination of whether the edge connecting the two building blocks is twisted is based on input (iii). Specifically, we shall associate to each peptide unit a positively oriented orthonormal 3-frame determined from input (iii). A pair of consecutive peptide units thus gives a pair of such 3-frames, and there is a unique element of the Lie group SO(3) mapping one to the other. Using this, we may assign an element of SO(3) to each oriented edge of the graph underlying the fatgraph model and thereby determine an "SO(3) graph connection" (cf. the next section) on the underlying graph, which is a fundamental and independently interesting aspect of our constructions. This assignment is discretized using the bi-invariant metric on SO(3) to determine twisting and define the fatgraph model of the backbone, where there are special considerations to handle the case of the cis-conformation, which can be detected using input (iii).

Edges are finally added to this model of the backbone in the natural way, one edge for each hydrogen bond in input (ii); see Figure 3.4. These added edges for hydrogen bonds may be twisted or untwisted, and this determination is again made by considering the SO(3) graph connection.

Section 3.1 discusses generalities about 3-frames and SO(3) graph connections. Section 3.2 details the concatenation of fatgraph building blocks to construct the model of the backbone, and Section 3.3 explains the addition of edges corresponding to hydrogen bonds, thus completing the description of the basic model. The final Section 3.4 discusses the general model with bifurcated hydrogen bonds plus other innovations and extensions of the method. An alternative to the basic model, which gives an equivalent but not strongly equivalent fatgraph that is arguably more natural than the basic model, is discussed in Appendix A, and the standard motifs of polypeptide secondary structure are described in the alternative model.

3.1 SO(3) Graph Connections and 3-Frames

The Lie group SO(3) is the group of 3×3 matrices A whose entries are real numbers satisfying $AA^{\mathsf{T}} = I$ and $\det(A) = 1$, where A^{T} denotes the transpose of A and I denotes the identity matrix. A metric $d : \mathrm{SO}(3) \times \mathrm{SO}(3) \to \mathbb{R}$ on SO(3) is said to be *bi-invariant* provided $d(CAD, CBD) = d(A, B)$ for any $A, B, C, D \in$

SO(3). The Lie group SO(3) supports the unique (up to scale) bi-invariant metric

$$d(A,B) = -\tfrac{1}{2}\,\mathrm{trace}(\log(AB^{\mathsf{T}}))^2,$$

where the trace of a matrix is the sum of its diagonal entries and the logarithm is the matrix logarithm [6].

PROPOSITION 3.1 *For any $A_1, A_2 \in \mathrm{SO}(3)$, we have $d(A_1, I) < d(A_2, I)$ if and only if* $\mathrm{trace}(A_2) < \mathrm{trace}(A_1)$*, where d is the unique bi-invariant metric on* SO(3).

PROOF: For any $A \in \mathrm{SO}(3)$, there is $B \in \mathrm{SO}(3)$ so that

$$BAB^{\mathsf{T}} = \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

for some angle $0 \le \theta \le \pi$; cf. [6]. It follows from bi-invariance that

$$d(A,I) = d(BAB^{\mathsf{T}}, BIB^{\mathsf{T}}) = d(BAB^{\mathsf{T}}, I) = d(BAB^{-1}, I),$$

i.e., distance to I is a conjugacy invariant, and from the formula for d that $d(A,I)$ is a monotone increasing function of θ. On the other hand,

$$\mathrm{trace}(A) = \mathrm{trace}(BAB^{-1}) = \mathrm{trace}(BAB^{\mathsf{T}}) = 1 + 2\cos\theta$$

is a monotone decreasing function of θ that is also a conjugacy invariant, and the result follows. □

A (positively oriented) *3-frame* is an ordered triple $\mathcal{F} = (\vec{u}_1, \vec{u}_2, \vec{u}_3)$ of three mutually perpendicular unit vectors in $\mathbb{R}^3$ so that $\vec{u}_3 = \vec{u}_1 \times \vec{u}_2$. For example, the standard unit basis vectors $(\vec{i}, \vec{j}, \vec{k})$ give a standard 3-frame.

PROPOSITION 3.2 *An ordered pair $\mathcal{F} = (\vec{u}_1, \vec{u}_2, \vec{u}_3)$ and $\mathcal{G} = (\vec{v}_1, \vec{v}_2, \vec{v}_3)$ of 3-frames uniquely determines an element $D \in$* SO(3)*, where $D\vec{u}_i = \vec{v}_i$ for $i = 1, 2, 3$. Furthermore, the trace of D is given by $\vec{u}_1 \cdot \vec{v}_1 + \vec{u}_2 \cdot \vec{v}_2 + \vec{u}_3 \cdot \vec{v}_3$, where "$\cdot$" is the usual dot product of vectors in $\mathbb{R}^3$.*

PROOF: Express

$$\vec{u}_i = a_{1i}\vec{i} + a_{2i}\vec{j} + a_{3i}\vec{k}, \qquad \vec{v}_i = b_{1i}\vec{i} + b_{2i}\vec{j} + b_{3i}\vec{k},$$

for $i = 1, 2, 3$, as linear combinations of $\vec{i}$, $\vec{j}$, and $\vec{k}$. The matrices $A = (a_{ij})$ and $B = (b_{ij})$ thus map $\vec{i}, \vec{j}, \vec{k}$ to $\vec{u}_1, \vec{u}_2, \vec{u}_3$ and $\vec{v}_1, \vec{v}_2, \vec{v}_3$, respectively. It follows that the matrix $D = BA^{-1}$ indeed has the desired properties. If D' is another such matrix, then $D^{-1}D'$ must fix each vector $\vec{u}_1, \vec{u}_2, \vec{u}_3$, and hence must agree with the identity proving the first part. For the second part since trace is a conjugacy invariant, we have

$$\mathrm{trace}(BA^{-1}) = \mathrm{trace}(A^{-1}B) = \mathrm{trace}(A^{\mathsf{T}}B) = \sum_{i=1}^{3} \vec{u}_i \cdot \vec{v}_i$$

as was claimed. □

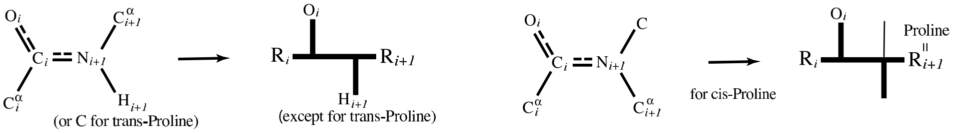

FIGURE 3.1. Fatgraph building block.

Suppose that Γ is a graph. An SO(3) *graph connection* on Γ is the assignment of an element $A_f \in$ SO(3) to each oriented edge f of Γ so that the matrix associated to the reverse of f is the transpose of A_f. Two such assignments A_f and B_f are regarded as equivalent if there is an assignment $C_u \in$ SO(3) to each vertex u of Γ so that $A_f = C_u B_f C_w^{-1}$ for each oriented edge f of Γ with initial point u and terminal point w. An SO(3) graph connection on Γ determines an isomorphism class of flat principal SO(3) bundles over Γ; cf. [9]. Given an oriented edge-path γ in Γ described by consecutive oriented edges $f_0 - f_1 - \cdots - f_{k+1}$, where the terminal point of f_i is the initial point of f_{i+1} for $i = 0, 1, \ldots, k$, the *parallel transport operator* of the SO(3) graph connection along γ is given by the matrix product $\rho(\gamma) = A_{f_0} A_{f_1} \cdots A_{f_k} \in$ SO(3). In particular, if the terminal point of f_k agrees with the initial point of f_0 so that γ is a closed oriented edge-path, then trace$(\rho(\gamma))$ is the *holonomy* of the graph connection along γ and is well-defined on the equivalence class of graph connections.

3.2 Modeling the Backbone

In this section, we shall define our model $T(P)$ for the backbone of a polypeptide structure P. To this end, consider the fatgraph building block depicted in Figure 3.1, which consists of a horizontal segment and two vertical segments joined to distinct interior points of the horizontal segment, the vertical segment on the left lying above and on the right below the horizontal segment. Each such building block represents a peptide unit. This is also indicated in the figure, where the left and right endpoints of the horizontal segment represent C$^\alpha_i$ and C$^\alpha_{i+1}$ and are labeled by the corresponding residue R$_i$ and R$_{i+1}$, respectively, the left and right trivalent vertices represent C$_i$ and N$_{i+1}$, respectively, and the endpoints of the vertical segments above and below the horizontal segment represent O$_i$ and H$_{i+1}$, respectively. In the case that R$_{i+1}$ is proline, the endpoint of the vertical segment below the horizontal segment instead represents the non-alpha carbon atom bonded to N$_{i+1}$ in the proline ring. In the case of cis-proline as depicted in Figure 3.1—or indeed any other peptide unit in the cis-conformation—a more geometrically accurate building block would have the vertical segment on the right also lying above the horizontal segment as indicated by the skinny line in the figure, but we nevertheless use a single building block in all cases for convenience.

Fix a polypeptide structure P and start by defining a fatgraph $T'(P)$ as the concatenation of $L - 1$ copies of this fatgraph building block, where the two univalent

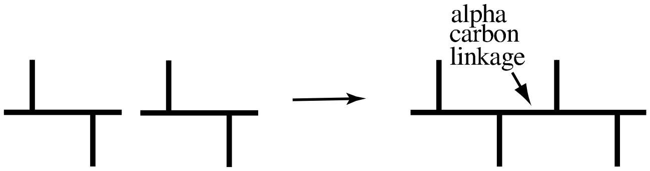

FIGURE 3.2. Concatenating fatgraph building blocks.

vertices representing C^{α}_{i+1} are identified so that the two incident edges are combined to form a single horizontal edge of T' called the $(i+1)^{\text{st}}$ *alpha carbon linkage* for $i = 1, 2, \ldots, L-2$, as illustrated in Figure 3.2. Let us also refer to the horizontal edges incident on the vertex corresponding to C^{α}_1 and C^{α}_L as the first and L^{th} *alpha carbon linkages*, respectively, so the i^{th} alpha carbon linkage is naturally labeled by the amino acid R_i for $i = 1, 2, \ldots, L$. Thus, $T'(P)$ consists of a long horizontal segment composed of $2L-1$ horizontal edges, L of which are alpha carbon linkages and $L-1$ of which correspond to peptide bonds, with $2L-2$ short vertical edges attached to it alternately lying above and below the long horizontal segment. We shall define the fatgraph $T(P)$ by specifying twisting on the alpha carbon linkages of $T'(P)$.

CONSTRUCTION 3.3 Associate a 3-frame $\mathcal{F}_i = (\vec{u}_i, \vec{v}_i, \vec{w}_i)$ to each peptide unit using input (iii) by setting

$$\begin{aligned}
\vec{u}_i &= \frac{1}{|\vec{x}_i|}\,\vec{x}_i,\\
\vec{v}_i &= \frac{1}{|\vec{y}_i - (\vec{u}_i \cdot \vec{y}_i)\,\vec{u}_i|}(\vec{y}_i - (\vec{u}_i \cdot \vec{y}_i)\,\vec{u}_i),\\
\vec{w}_i &= \vec{u}_i \times \vec{v}_i,
\end{aligned}$$

for $i = 1, 2, \ldots, L-1$, where $|\vec{t}|$ denotes the norm of the vector $\vec{t}$.

Thus, $\vec{u}_i$ is the unit displacement vector from C_i to N_{i+1}, $\vec{v}_i$ is the projection of $\vec{y}_i$ onto the specified perpendicular of $\vec{u}_i$ in the plane of the peptide unit, and $\vec{w}_i$ is the specified normal vector to this plane.

According to Proposition 3.2, there is a unique element $A_i \in \mathrm{SO}(3)$ mapping $\mathcal{F}_i$ to $\mathcal{F}_{i+1}$ for $i = 1, 2, \ldots, L-2$. Define the *backbone graph connection* on the graph underlying $T'(P)$ to take value I on all oriented edges except on the i^{th} alpha carbon linkage oriented from its endpoint representing N_i to its endpoint representing C_i, where it takes value A_{i-1} for $i = 2, 3, \ldots, L-1$.

We shall discretize the backbone graph connection to finally define the backbone fatgraph model $T(P)$. To this end, in addition to the 3-frames $\mathcal{F}_i = (\vec{u}_i, \vec{v}_i, \vec{w}_i)$ of Construction 3.3, we consider also the 3-frames $\mathcal{G}_i = (\vec{u}_i, -\vec{v}_i, -\vec{w}_i)$, which correspond to simply turning $\mathcal{F}_i$ upside down by rotating through 180° in 3-space about the line containing C_i and N_{i+1} for $i = 1, 2, \ldots, L-1$. Again, by the first part of Proposition 3.2, there is a unique element $B_i \in \mathrm{SO}(3)$ taking $\mathcal{F}_i$ to $\mathcal{G}_{i+1}$. By construction, A_i also takes $\mathcal{G}_i$ to $\mathcal{G}_{i+1}$, and B_i takes $\mathcal{G}_i$ to $\mathcal{F}_{i+1}$.

CONSTRUCTION 3.4 For any polypeptide structure P, define the fatgraph $T(P)$ derived from $T'(P)$ by taking twisting only on certain of the alpha carbon linkages, where the $(i+1)^{\mathrm{st}}$ alpha carbon linkage is twisted if and only if

$$\begin{cases} d(I, B_i) \le d(I, A_i) \\ \quad \text{if the peptide unit before } \mathrm{R}_{i+1} \text{ is not in the cis-conformation}, \\ d(I, B_i) \ge d(I, A_i) \\ \quad \text{if the peptide unit before } \mathrm{R}_{i+1} \text{ is in the cis-conformation}, \end{cases}$$

for $i = 1, 2, \ldots, L-2$, where d is the unique bi-invariant metric on SO(3).

COROLLARY 3.5 *The* $(i+1)^{st}$ *alpha carbon linkage of the backbone model* $T(P)$ *is twisted if and only if*

$$\begin{cases} \vec{v}_i \cdot \vec{v}_{i+1} + \vec{w}_i \cdot \vec{w}_{i+1} \le 0 & \text{if } \vec{y}_i \cdot \vec{z}_i \ge 0, \\ \vec{v}_i \cdot \vec{v}_{i+1} + \vec{w}_i \cdot \vec{w}_{i+1} \ge 0 & \text{if } \vec{y}_i \cdot \vec{z}_i < 0, \end{cases}$$

for $i = 1, 2, \ldots, L-2$.

PROOF: According to Proposition 3.1, $d(A_i, I) \le d(B_i, I)$ if and only if $\mathrm{trace}(B_i) \le \mathrm{trace}(A_i)$. According to the second part of Proposition 3.2, we have

$$\mathrm{trace}(A_i) = \vec{u}_i \cdot \vec{u}_{i+1} + \vec{v}_i \cdot \vec{v}_{i+1} + \vec{w}_i \cdot \vec{w}_{i+1},$$
$$\mathrm{trace}(B_i) = \vec{u}_i \cdot \vec{u}_{i+1} - \vec{v}_i \cdot \vec{v}_{i+1} - \vec{w}_i \cdot \vec{w}_{i+1},$$

so that $\mathrm{trace}(A_i) - \mathrm{trace}(B_i) = 2(\vec{v}_i \cdot \vec{v}_{i+1} + \vec{w}_i \cdot \vec{w}_{i+1})$.

Thus, if R_{i+1} is in the trans-conformation, then we twist the $(i+1)^{\mathrm{st}}$ alpha carbon linkage if and only if $\mathcal{F}_i$ is closer to $\mathcal{G}_{i+1}$ than it is to $\mathcal{F}_{i+1}$ in the sense that $d(I, A_i) \ge d(I, B_i)$, and this is our natural discretization of the backbone graph connection in Construction 3.4 in this case. Clearly, R_{i+1} is in the cis-conformation if and only if $\vec{y}_i \cdot \vec{z}_i < 0$ as determined by input (iii) so we twist the $(i+1)^{\mathrm{st}}$ alpha carbon linkage only if $d(I, A_i) \le d(I, B_i)$. To see that this is the natural discretization of the backbone graph connection in this case, notice that the 3-frame $\mathcal{F}_i$ in Construction 3.3 is determined using the displacement vectors $\vec{x}_i$ from C_i to N_{i+1} and $\vec{y}_i$ from C_i^{α} to C_i, which are insensitive to whether R_{i+1} is in the cis-conformation. It is therefore only upon exiting a cis-peptide unit along the backbone that the earlier determination should be modified since the latter displacement vector should be replaced by its antipode. □

Define the *flip sequence* of $G(P)$ to be the word in the alphabet {F, N} whose i^{th} letter is N if and only if the $(i+1)^{\mathrm{st}}$ alpha carbon linkage is untwisted for $i = 1, 2, \ldots, L(G) - 2$. The flip sequence thus gives a discrete invariant assigned to each alpha carbon linkage derived from the conformational geometry along the backbone. The flip sequence can be determined directly from the conformational angles along the backbone using the following result:

PROPOSITION 3.6 *Under the idealized geometric assumptions of tetrahedral angles among bonds at each alpha carbon atom and* 120° *angles between bonds within a peptide unit, the matrix $A = A_i$ in Construction* 3.4 *can be calculated in terms of the conformational angles $\varphi = \varphi_i$ and $\psi = \psi_i$ as follows*:

$$A = B_3(\varphi) B_2(\varphi + \psi) \begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} & 0 \\ \frac{\sqrt{3}}{2} & \frac{1}{2} & 0 \\ 0 & 0 & -1 \end{pmatrix},$$

where

$$B_3(\varphi) = \begin{pmatrix} \frac{2}{3} - \frac{C^2}{3} + \frac{S^2}{6} & -2[\frac{\sqrt{2}C}{3} + \frac{S^2}{4\sqrt{3}}] & 2[\frac{CS}{2\sqrt{3}} - \frac{S}{3\sqrt{2}}] \\ 2[\frac{\sqrt{2}C}{3} - \frac{S^2}{4\sqrt{3}}] & \frac{2}{3} - \frac{C^2}{3} - \frac{S^2}{6} & -2[\frac{CS}{6} + \frac{S}{\sqrt{6}}] \\ 2[\frac{CS}{2\sqrt{3}} + \frac{S}{3\sqrt{2}}] & 2[\frac{S}{\sqrt{6}} - \frac{CS}{6}] & \frac{2}{3} + \frac{C^2}{3} - \frac{S^2}{3} \end{pmatrix}$$

for $C = \cos\varphi$, $S = \sin\varphi$, and

$$B_2(\varphi + \psi) = \begin{pmatrix} 1 - \frac{3}{2}S^2 & \frac{\sqrt{3}}{2}S^2 & \sqrt{3}CS \\ \frac{\sqrt{3}}{2}S^2 & 1 - \frac{1}{2}S^2 & -CS \\ -\sqrt{3}CS & CS & 1 - 2S^2 \end{pmatrix}$$

for $C = \cos\frac{\varphi+\psi}{2}$, $S = \sin\frac{\varphi+\psi}{2}$.

Explicitly, this is the representative $A = A_i$ in its conjugacy class for which the 3-frame vectors $\vec{u}_i = \vec{i}$, $\vec{v}_i = \vec{j}$, and $\vec{w}_i = \vec{k}$ in Construction 3.3 *are given by the standard unit basis vectors*; *i.e., this is the choice of so-called gauge fixing.*

PROOF: Let ξ be an angle and $\vec{v}$ be a nonzero vector in $\mathbb{R}^3$. We denote by $(\xi, \vec{v})$ the linear transformation $\mathbb{R}^3 \to \mathbb{R}^3$ that rotates $\mathbb{R}^3$ through the angle ξ around the line spanned by $\vec{v}$ in the right-handed sense in the direction of $\vec{v}$. By following the standard 3-frame along the backbone in the natural way one bond at a time, we find

$$A = B_6(\varphi, \psi) B_5(\varphi, \psi) B_4(\varphi, \psi) B_3(\varphi) B_2(\varphi) B_1(\tfrac{\pi}{3})$$

where

$$B_1(\xi) = (\xi, \vec{k}), \quad B_2(\varphi) = (\varphi, B_1(\tfrac{\pi}{3})\vec{i}), \quad B_3(\phi) = (\pi - \theta, B_2(\phi)\vec{k}),$$

$$B_4(\varphi, \psi) = (\psi, B_3(\varphi) B_1(\tfrac{\pi}{3})\vec{i}), \quad B_5(\varphi, \psi) = (\tfrac{2\pi}{3}, -B_4(\varphi, \psi) B_3(\varphi) B_2(\varphi)\vec{k}),$$

$$B_6(\varphi, \psi) = (\pi, B_5(\varphi, \psi) B_4(\varphi, \psi) B_3(\varphi) B_2(\varphi) B_1(\tfrac{\pi}{3})\vec{j}),$$

and where $\theta = 2\arctan(\sqrt{2})$ is the tetrahedral angle $\approx 109.5°$ for which $\cos\theta = -\frac{1}{3}$.

We observe that

$$B_4(\varphi, \psi) B_3(\varphi) = B_3(\varphi) B_2(\psi)$$

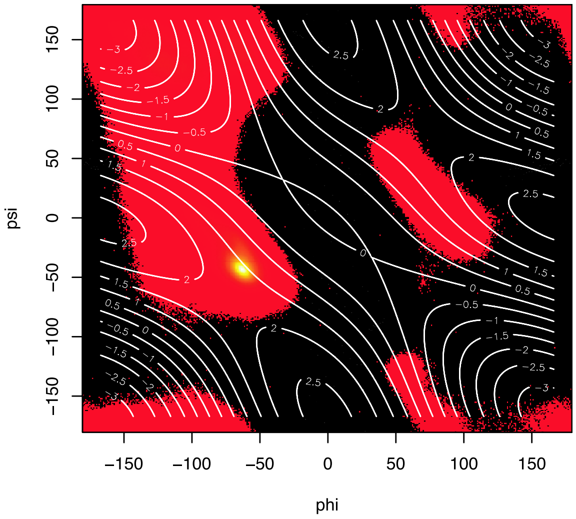

FIGURE 3.3. Level sets of trace(A) − trace(B) on a Ramachandran plot.

whence

$$B_4(\varphi,\psi)B_3(\varphi)B_2(\varphi) = B_3(\varphi)B_2(\varphi+\psi),$$

and therefore

$$A = B_6(\varphi,\psi)B_3(\phi)B_2(\varphi+\psi)B_1(-\tfrac{\pi}{3}).$$

Setting $B_0 = (\pi, \vec{j})$, we conclude

$$A = B_3(\varphi)B_2(\varphi+\psi)B_1(-\tfrac{\pi}{3})B_0,$$

which devolves after some computation to the given expression. □

Remark 3.7. It is interesting to graph the level sets of trace(A) − trace(B) on the Ramachandran plot; i.e., the plot of pairs of conformational angles (φ_i, ψ_i) for the entire CATH database [25] using Proposition 3.6 as depicted in Figure 3.3, where the matrix $B = B_i$ of Construction 3.3 is obtained from $A = A_i$ in Proposition 3.6 by precomposing it with rotation by π about $\vec{i}$. In particular, the zero level set fairly well avoids highly populated regions, so the case of near equality in Construction 3.4 is a relatively rare phenomenon for proteins.[6]

3.3 Modeling Hydrogen Bonds

The fatgraph model $T(P)$ of the backbone of a polypeptide structure P defined in the previous section is here completed to our fatgraph model $G(P)$. Just as in the previous section, we shall first define another fatgraph $G'(P)$ from which

[6] Indeed, further scrutiny of details not depicted in Figure 3.3 shows that the zero level set does penetrate into conformations of "beta turns of types II and VI"; cf. the discussion of Figure A.3. This could be further documented empirically, but we have not done so.

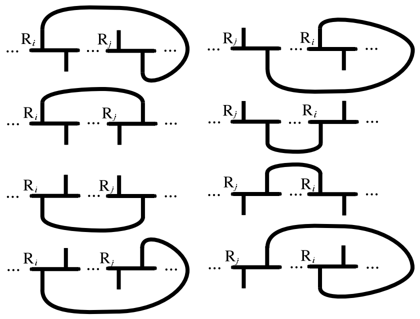

FIGURE 3.4. Adding edges to $T(P)$ for hydrogen bonds.

$G(P)$ is derived by further twisting certain of its edges. As described in the previous section, $T(P)$ consists of a long horizontal segment, certain of whose alpha carbon linkages are twisted, together with small vertical segments alternately lying above and below the long horizontal segment, where the $(i+1)^{\mathrm{st}}$ alpha carbon linkage is labeled by its corresponding amino acid R_{i+1} for $i = 1, 2, \ldots, L$. The endpoints of the vertical segments above and below the horizontal segment, respectively, represent the atoms O_i and H_{i+1} except for the vertical segments below the horizontal segment preceding an alpha carbon linkage labeled by proline, whose endpoint represents the non-alpha carbon atom bonded to N_{i+1} in the corresponding proline ring for $i = 1, 2, \ldots, L-1$.

CONSTRUCTION 3.8 For each $(i, j) \in \mathcal{B}$ in input (ii), adjoin an edge to $T(P)$ without introducing new vertices connecting the endpoints of short vertical segments corresponding to H_i and O_j to produce a fatgraph denoted $G'(P)$.

See Figure 3.4. It is important to emphasize that the relative positions of these added edges corresponding to hydrogen bonds other than their endpoints are completely immaterial to the strong equivalence class of $G'(P)$. The edges of $T(P)$ corresponding to the non-alpha carbon atoms in a proline rings are never hydrogen bonded in our model. In the remaining extremely rare cases of nonproline cis-conformations, the model is slightly inaccurate.

To complete the construction of $G(P)$, it remains only to determine which edges of the fatgraph $G'(P)$ are twisted. To this end, suppose that $(i, j) \in \mathcal{B}$ in input (ii). According to our enumeration of peptide units, H_i occurs in peptide unit $i-1$ and O_j occurs in peptide unit j, and there are corresponding 3-frames

$$\begin{aligned}\mathcal{F}_{i-1} &= (\vec{u}_{i-1}, \vec{v}_{i-1}, \vec{w}_{i-1}),\\ \mathcal{F}_j &= (\vec{u}_j, \vec{v}_j, \vec{w}_j),\\ \mathcal{G}_j &= (\vec{u}_j, -\vec{v}_j, -\vec{w}_j),\end{aligned}$$

from Construction 3.3.

CONSTRUCTION 3.9 As before by the first part of Proposition 3.2, there are unique $D_{i,j}, E_{i,j} \in \mathrm{SO}(3)$ taking $\mathcal{F}_{i-1}$ to $\mathcal{F}_j, \mathcal{G}_j$, respectively. An edge of $G'(P)$ corresponding to the hydrogen bond $(i, j) \in \mathcal{B}$ is twisted in $G(P)$ if and only if

$$d(I, E_{i,j}) \leq d(I, D_{i,j}),$$

where d is the unique bi-invariant metric on SO(3).

As before, a short computation gives the following:

COROLLARY 3.10 *The edge of $G(P)$ corresponding to the hydrogen bond $(i, j) \in \mathcal{B}$ is twisted if and only if $\vec{v}_{i-1} \cdot \vec{v}_i + \vec{w}_{i-1} \cdot \vec{w}_j \leq 0$.*

Remark 3.11. The backbone graph connection on the graph that underlies $T(P)$ clearly has trivial holonomy since $T(P)$ is contractible. It extends naturally to an SO(3) graph connection on the graph underlying $G(P)$, where to the oriented edge corresponding to the hydrogen bond connecting $\mathrm{N}_i - \mathrm{H}_i$ and O_j, we assign the unique element of SO(3), whose existence is guaranteed by Proposition 3.2, which maps $\mathcal{F}_{i-1}$ to $\mathcal{F}_j$ for $i = 2, 3, \ldots, L-2$. This graph connection on $G(P)$ also has trivial holonomy by construction. Our fatgraph model $G(P)$ arises from a discretization of this SO(3) graph connection giving a $\mathbb{Z}/2$ graph connection, rotated so that the oriented edges with nontrivial holonomy are the twisted ones, and this $\mathbb{Z}/2$ graph connection on the graph underlying $G(P)$ typically does not have trivial holonomy.

3.4 Basic Model and Its Extensions

The previous section completed the definition of our basic fatgraph model $G(P)$ of a polypeptide structure P. Notice that hydrogen bonds and alpha carbon linkages are treated in precisely the same manner in this construction.

A crucial point in practice is that the polypeptide structure itself depends upon data that must be considered as idealized for various reasons: proteins actually occur in several closely related conformations, varying under thermal fluctuations, for example, whose sampling is corrupted by experimental uncertainties as well as errors. The fatgraph $G(P)$ must therefore not be taken as defined absolutely, but rather as defined only in some statistical sense as a family of fatgraphs $\{G(P) : P \in \mathcal{P}\}$ based on a collection $\mathcal{P}$ of polypeptide structures that differ from one another by a small number of such idealizations, uncertainties, or errors. Properties of the fatgraph $G(P)$ that we can meaningfully assign to the polypeptide structure P must be nearly constant on $\mathcal{P}$ and lead to the notion of "robustness" of invariants of $G(P)$ as descriptors of P, which is discussed in Section 4. Nevertheless, the construction of our model has been given based on the inputs above regarded as exact and error free.

In particular, there is the tacit assumption that there is never equality in the determination of whether to twist in Constructions 3.4. In practice, $\vec{v}_i \cdot \vec{v}_{i+1} + \vec{w}_i \cdot \vec{w}_{i+1} = 0$ never occurs exactly, but there is the real possibility that this condition *nearly holds*; that is, we cannot reliably determine whether to twist if $|\vec{v}_i \cdot \vec{v}_{i+1} +$

$\vec{w}_i \cdot \vec{w}_{i+1}|$ is below some small threshold because of experimental uncertainty; cf. Remark 3.7. There are similar issues in the specification of which hydrogen bonds exist in input (ii) based upon the possibly problematic exact atomic locations from which the electrostatic potentials are inferred as well as whether to twist in Construction 3.9.

However, there is the following control over the topological type of $F(G(P))$, which will be the basis for several of the robust invariants of fatgraphs and resulting meaningful descriptors of polypeptides studied in Section 4.

COROLLARY 3.12 *Let P and P' be polypeptide structures with the same inputs* (i) *but differing in inputs* (ii)–(iii) *in the determinations of the existence of m hydrogen bonds and of the twisting of n alpha carbon linkages or hydrogen bonds. Then $|r(G(P)) - r(G(P'))| \leq m + n$.*

PROOF: This is an immediate consequence of Proposition 2.2. □

There are several generalizations of the basic fatgraph model $G(P)$ of a polypeptide structure. As already mentioned, we might specify energy thresholds $E_- < E_+ < 0$ and demand that the potential energy of a hydrogen bond lie in the range between E_- and E_+ in order that it be regarded as a hydrogen bond to include in input (ii) so as to produce a fatgraph denoted $G_{E_-,E_+}(P)$. We shall describe in Section 5 certain experiments with proteins using various such energy thresholds.

One may also model bifurcated hydrogen bonds and allow hydrogen or oxygen atoms in the peptide units to participate in at most $\beta \geq 1$ hydrogen bonds by altering the fatgraph building block in Figure 3.1 by replacing the univalent vertices representing hydrogen and oxygen atoms by vertices of valence $\beta + 1$. Different valencies less than $\beta + 1$ for oxygen and hydrogen can be implemented with this single building block by appropriately imposing different constraints in input (ii). Natural fattenings on these new vertices representing hydrogen or oxygen atoms are determined as follows: project centers of partners in bonding into the plane of the peptide unit with origin at the center of the corresponding nitrogen or carbon atom, respectively, where the positive x-axis contains the bond axis of the incident peptide bond, and take these projections in order of increasing argument.

Our definition of polypeptide structure assumes that there are no atoms missing along the backbone, and this is actually somewhat problematic in practice. A useful aspect of the methods in Section 3.2 is that such gaps present no essential difficulty since an edge connecting fatgraph building blocks can just as well be taken to represent a gap between peptide units as to represent an alpha carbon linkage as in our model articulated before. The determination of twisting on these new gap edges is just as in Construction 3.4, but now the 3-frames in this construction do not correspond to consecutive peptide units.

A more profound extension of the method is to use the bi-invariant metric on SO(3) to give finer discretizations of the SO(3) graph connection on $G(P)$ discussed in Remark 3.11. For example, rather than our $\mathbb{Z}/2$ graph connection modeled by fatgraphs, one can easily implement the analogous construction of a $\mathbb{Z}/n$

graph connection based on the natural extensions of Constructions 3.4 and 3.9 modeled by graphs with fattenings and $\mathbb{Z}/n$ colorings. These "rotamer fatgraphs" capture the "protein rotamers," which are highly studied in the biophysics literature.

A still more profound innovation rests on the observation that our techniques are of greater utility and can be adapted to model essentially any molecule since 3-frames can analogously be associated to any bond axis. One might thus model entire amino acids themselves as rotamer fatgraphs to give a truly realistic model of a polypeptide. Such an all-atom fatgraph model thus explicitly includes the primary structure of the protein in keeping with current methods. As argued earlier, if structure is indeed fully determined by sequence, then attributes of the all-atom model lie hidden even in our basic model in its empirical consequences.

Furthermore, the discussion thus far has concentrated on molecules at equilibrium, and one might instead regard the fatgraph or rotamer fatgraph as a dynamic model by taking time- or temperature-dependent inputs (ii)–(iii).

4 Robust Polypeptide Descriptors

We have described in the previous sections the fatgraph $G(P)$ of a polypeptide structure P with simple hydrogen bonding determined by inputs (i)–(iii) based upon specified energy thresholds. With the understanding that the input data can be problematic due to errors and experimental indeterminacies, we must consider the fatgraph as defined only in a statistical sense, where a family of fatgraphs arises from a collection $\mathcal{P} \ni P$ of polypeptide structures that differ from P by a small number of such errors or indeterminacies. As such, only certain properties of the fatgraph $G(P)$ can meaningfully be assigned as descriptors of P, namely, those properties that do not vary significantly over the various polypeptide structures in $\mathcal{P}$. In this section, we shall first formalize this notion of meaningful properties of fatgraphs and then describe and discuss a myriad of such polypeptide descriptors.

Let $\mathcal{G}$ denote the collection of all strong equivalence classes of fatgraphs $G(P)$ arising from nonempty polypeptide structures P. We may perform the following modifications to any $G \in \mathcal{G}$, leaving all other data unchanged:

Mutation (i): change the color of one alpha carbon linkage of G,

Mutation (ii): change the color of one edge of G corresponding to a hydrogen bond,

Mutation (iii): add or delete an untwisted edge of G corresponding to a hydrogen bond,

Mutation (iv): replace a fatgraph building block of G by two building blocks connected by an untwisted alpha carbon linkage, where any edges corresponding to hydrogen bonds incident on the original building block are connected to the replacement building block that occurs first along the backbone from N to C termini, and the reverse of this operation.

Suppose that X is some set with metric ρ. We say that a function $\nu : \mathcal{G} \to X$ is *κ-robust of radius Q on $\mathcal{H} \subseteq \mathcal{G}$*, where $\kappa \geq 0$ is real and $Q \geq 0$ is an integer, if $\rho(\nu(G), \nu(G')) \leq q\kappa$ whenever G' arises from $G \in \mathcal{H}$ by a sequence

$$G = G_0 - G_1 - \cdots - G_q = G' \quad \text{with } q \leq Q,$$

where G_{j+1} arises from G_j by a single mutation of type (i)–(iv) for $j = 0, 1, 2, \ldots, q-1$. If ν is κ-robust of infinite radius on all of $\mathcal{G}$, then we say simply that ν is κ-robust.

By definition if X supports operations of addition and scalar multiplication and if ν is κ-robust of radius Q on $\mathcal{H}$, then for any $\alpha \in \mathbb{R}$, $\alpha\nu$ is $\alpha\kappa$-robust of radius Q on $\mathcal{H}$, and furthermore, if ν' is κ'-robust of radius Q' on $\mathcal{H}'$, then $\nu \pm \nu'$ is $(\kappa + \kappa')$-robust of radius $\min(Q, Q')$ on $\mathcal{H} \cap \mathcal{H}'$.

It is only the κ-robust functions ν of reasonably large radius Q and sufficiently small value of κ on $\mathcal{H} \subseteq \mathcal{G}$ that are significant characteristics of polypeptide structures whose fatgraphs G lie in $\mathcal{H}$. This is because a combination of mutations arising from $q \leq Q$ errors or indeterminacies of the input data then affects the value of $\nu(G)$ by an amount bounded by $q\kappa$, which must be small compared to the value of $\nu(G)$.

It is clear that any two fatgraphs arising from a nonempty polypeptide structure are related by a finite sequence of mutations (i)–(iv). By assigning a penalty of some nonzero magnitude to each type of mutation, the *mutation distance* between two such fatgraphs can be defined as the minimum sum of penalties corresponding to sequences of mutations relating them. This gives a metric, albeit seemingly difficult to compute, on $\mathcal{G}$ itself, and we may regard two polypeptide structures as being similar if the mutation distance between their corresponding fatgraphs is small. The assignment of fatgraph $G(P)$ to polypeptide structure P is κ-robust by definition with this metric, where the parameter κ is the largest penalty.

For several obvious numerical examples, the numbers $L(G)$ of residues and $B(G)$ of hydrogen bonds of G are 1-robust, and the Euler characteristic $\chi(G)$ of G or $F(G)$ is likewise 1-robust since $\chi(G) = 1 - B(G)$. The numbers $v(G) = 2L(G) - 2$ of vertices and $e(G) = B(G) + 2L(G) - 3$ of edges of G are therefore 2- and 3-robust, respectively. The number of twisted edges corresponding to hydrogen bonds and the number of twisted alpha carbon linkages of G are each also clearly 1-robust.

With X the set of all words of finite length in the alphabet {F, N} given the edit distance with unit operation cost [13], the flip sequence of G is 1-robust by definition. In contrast, the plus/minus sequence of the alternative model $K(P)$ in Appendix A as a word in the alphabet $\{+, -\}$ with the same metric is not κ-robust of radius greater than 0 on $\mathcal{G}$ for any κ since a single modification of type (i) to G can change all the entries of the plus/minus sequence.

For another negative example with $X = \mathbb{R}$, the genus $g(G)$ of $F(G)$ is not κ-robust of any radius greater than 0 for any κ on $\mathcal{G}$ since a single modification of type (ii) on an untwisted G can produce a fatgraph G' with $F(G')$ nonorientable, and $|g(G) - g(G')| = [1 + B(G) - r(G)]/2$. In contrast, the modified genus is robust of infinite radius according to the following result.

PROPOSITION 4.1 *The number $r(G)$ of boundary components and the modified genus $g^*(G)$ of $F(G)$ are* 1*-robust. Moreover, the number of appearances in the flip sequence of G of any fixed word of length k in the alphabet* $\{0, 1\}$ *is k-robust.*

PROOF: The function r satisfies the required properties by Corollary 3.12, hence so too does $g^* = (1+B-r)/2$. The remaining assertion follows essentially by definition. □

Given a closed edge-path γ on $G \in \mathcal{G}$, define the *peptide length* of γ to be the number of pairs of distinct peptide units visited by γ and define the *edge length* of γ to be the number of edges of G traversed by γ, each counted with multiplicity. For example, the dotted boundary components in Figure A.3 that are characteristic of alpha helices and beta strands all have peptide length 4 and various edge lengths 4, 6, and 8. Define the *peptide length spectrum* $\mathbb{P}(G)$ and the *edge length spectrum* $\mathbb{E}(G)$ of $G \in \mathcal{G}$, respectively, to be the unordered set of peptide lengths and edge lengths of boundary components of $F(G)$. Let $\overline{\mathbb{P}}(G)$ and $\overline{\mathbb{E}}(G)$ denote their respective means. It is worth pointing out that the preponderance of alpha helices and beta strands in practice heavily biases $\overline{\mathbb{P}}(G)$ towards 4.

Let X denote the collection of all finite unordered collections of natural numbers. The elements of a member of X may be ordered by increasing magnitude. The distance between two such ordered finite collections of natural numbers may then be defined by standard methods [13], and this induces a metric on X itself. We may thus regard $\mathbb{P}$ and $\mathbb{E}$ as functions on $\mathcal{G}$ with values in the metric space X. As in the proof of Corollary 3.12, these functions are κ-robust where the parameter κ depends on the choice of metric.

LEMMA 4.2 *Suppose that $\mu : \mathcal{G} \to \mathbb{Z}$ is k-robust of radius at least Q on $\mathcal{G}$ and that $\nu : \mathcal{G} \to \mathbb{R}$ is κ-robust of radius Q on*

$$\mathcal{H} = \{G \in \mathcal{G} : \mu(G) > kQ \text{ and } \nu(G) + Q\kappa \le [\mu(G) - kQ]^2\}.$$

Then $\nu(G)/\mu(G) : \mathcal{G} \to \mathbb{R}$ is $(\kappa + k)$-robust of radius Q on $\mathcal{H}$.

PROOF: Suppose that $G \in \mathcal{H}$ and that $G = G_0 - G_1 - \cdots - G_q = G'$ is a sequence as before, with $q \le Q$. First note that

$$\nu(G_{i+1}) \le \nu(G_0) + i\kappa \quad \text{and} \quad \mu(G_{i+1}) \ge \mu(G_0) - ki$$

by hypothesis, and so

$$\frac{\nu(G_{i+1})}{[\mu(G_{i+1})]^2} \le \frac{\nu(G_0) + i\kappa}{[\mu(G_0) - ki]^2} \le \frac{\nu(G_0) + Q\kappa}{[\mu(G_0) - kQ]^2} \le 1$$

since $G_0 \in \mathcal{H}$ for $i = 0, 1, 2, \ldots, p$. Furthermore, we have that $|\nu(G_i) - \nu(G_{i+1})| \leq \kappa$ and $|\mu(G_i) - \mu(G_{i+1})| \leq k$ for each $i = 0, 1, 2, \ldots, q-1$, and hence

$$\begin{aligned}\left|\frac{\nu(G_i)}{\mu(G_i)} - \frac{\nu(G_{i+1})}{\mu(G_{i+1})}\right| &= \left|\frac{\mu(G_{i+1})\nu(G_i) - \mu(G_i)\nu(G_{i+1})}{\mu(G_i)\mu(G_{i+1})}\right| \\ &\leq \begin{cases} \frac{\kappa}{|\mu(G_i)|} & \text{if } \mu(G_{i+1}) = \mu(G_i), \\ \frac{\kappa}{|\mu(G_i)|} + k\frac{|\nu(G_{i+1})|}{[\mu(G_{i+1})]^2} & \text{if } \mu(G_{i+1}) < \mu(G_i), \\ \frac{\kappa}{|\mu(G_i)|} + k\frac{|\nu(G_i)|}{[\mu(G_i)]^2} & \text{if } \mu(G_{i+1}) > \mu(G_i), \end{cases} \\ &\leq \kappa + k.\end{aligned}$$

The triangle inequality then gives

$$\left|\frac{\nu(G)}{\mu(G)} - \frac{\nu(\widetilde{G})}{\mu(\widetilde{G})}\right| \leq q(\kappa + k)$$

as required. □

PROPOSITION 4.3 *The mean $\overline{\mathbb{P}}(G)$ of the peptide length spectrum is* 3*-robust of radius Q on*

$$\{G \in \mathcal{G} : r(G) > Q \text{ and } L(G) + Q - 1 \leq \tfrac{1}{2}[r(G) - Q]^2\},$$

and the mean $\overline{\mathbb{E}}(G)$ of the edge length spectrum is 7*-robust of radius Q on*

$$\{G \in \mathcal{G} : r(G) > Q \text{ and } B(G) + 2L(G) - 3 + 6Q \leq [r(G) - Q]^2\}.$$

PROOF: Since each peptide unit occurs exactly twice in the union of all the boundary components, the sum of all the elements in $\mathbb{P}(G)$ is constant equal to $2[L(G) - 1]$, which is 2-robust according to earlier comments. Since $\overline{\mathbb{P}}(G) = 2[L(G) - 1]/r(G)$ and $r(G)$ is 1-robust by Lemma 4.1, the first assertion follows from Lemma 4.2. Similarly, each edge occurs exactly twice in the union of all boundary components, so the sum of all the elements in $\mathbb{E}(G)$ is equal to $2e(G) = 2[B(G)+2L(G)-3]$, which is 6-robust according to earlier comments. The second assertion therefore likewise follows from Lemma 4.2. □

Other notions of lengths of closed edge-paths in G may also be useful. For example, for each amino acid type, each boundary component of $F(G)$ visits a certain number of alpha carbon linkages labeled by amino acids of this type, and alternative notions of length arise by assigning weights to the various amino acids and taking the weighted sum over amino acids visited. The robustness of these sorts of invariants seems difficult to analyze.

It is also worth pointing out that the underlying graph of the fatgraph $G(P)$ has its own related characteristics for any polypeptide structure P. For example, there is an associated notion of length spectrum, namely, one or another of the notions of generalized length discussed before of the closed edge-paths or simple closed edge-paths on the graph. Invariants of this type, which can be derived from the graph underlying the fatgraph, may also be of importance in practice, and their robustness is based on the invariance of the underlying graph under the modifications (i)–(ii).

The fatgraph G is of a special type in that it has a "spine" arising from the backbone, namely, the long horizontal segment arising from the concatenation of horizontal segments in the fatgraph building blocks that was discussed in Section 3.2. This "spined fatgraph" admits a canonical "reduction" by serially removing each edge incident on a univalent vertex and amalgamating the pair of edges incident on the resulting bivalent vertex into a single edge. The graph underlying this reduced spined fatgraph is a "chord diagram," and there are countless "finite-type invariants associated with weight systems" [24], which could provide useful protein invariants whose robustness depends upon the choice of weight system. See Section 6 for a further discussion of related quantum invariants.

5 First Results

5.1 Aspects of Implementation

In this section, we shall first make several practical remarks about the implementation in this paper of our methods for a protein from its PDB and DSSP files (cf. Section 1), where we shall consider here only the model with simple hydrogen bonds, i.e., $\beta = 1$, which depends upon energy thresholds $E_- < E_+ < 0$ as follows. In effect, we employ the standard methods of DSSP described in Section 1 to estimate electrostatic potentials of possible hydrogen bonds, and we tabulate to hundredths of kcal/mole the two strongest such potentials in which each hydrogen or oxygen atom in a polypeptide unit participates. Any such energies beyond our energy thresholds are then discarded. Displacements of corresponding backbone atoms are used to discriminate between equal tabulated electrostatic potentials in order to derive a strict linear ordering on them: a hydrogen bond with energy E between atoms at distance δ precedes a hydrogen bond with energy E' between atoms at distance δ' if $E < E'$ or if $E = E'$ and $\delta \leq \delta'$, where $E = E'$ to hundredths of kcal/mole and $\delta = \delta'$ to thousandths of angstroms never occurs in practice. We finally greedily add to $\mathcal{B}$ in input (ii) the hydrogen bonds in this linear ordering provided they do not violate the a priori simple hydrogen bond assumption $\beta = 1$.

Minor technical comments are that we only implemented the flipped conformation of fatgraph building blocks for cis-conformations in the case of cis-proline and not for other residues. In any case, other cis-conformations are so rare as to be inconsequential for the empirical work we report here. Furthermore, unspecified or missing residue types are assumed not to be proline for input (i), atomic locations in the PDB with highest occupancy numbers are those used for determining input (iii), and we take only the first model in case there are several models in a PDB file.

Whenever there is a missing datum, for example the atomic location of a backbone atom in a PDB file, that is required for the algorithmic construction of the 3-frame corresponding to its peptide unit, we concatenate an associated fatgraph building block without twisting the alpha carbon linkage, and we prohibit any hydrogen bonding to its constituent edges. Such "gap frames" are included for each problematic peptide unit. A number of such gap frames may occur between two fatgraph building blocks that *can* consistently be assigned 3-frames, and the last alpha carbon linkage connecting a gap frame to a nongap frame is twisted or untwisted based upon the usual criteria for the two adjacent well-defined nongap frames. In particular, the fatgraph constructed is always connected. Other examples of gap frames arise from breaks along the backbone as detected by a separation of more than 2.0 Å between atoms C_i and N_{i+1} for any i.

5.2 Injectivity Results

The database CATH version 3.2.0 [25] is a collection $\mathcal{P}_{\text{CATH}}$ of 114,215 protein domains, which are uniquely catalogued by a nine-tuple of natural numbers; this is a hierarchical classification with a "standard" representative domain chosen in each class. Our methods have been applied to the associated PDB and DSSP files so as to produce corresponding connected fatgraphs $G_{-\infty,E}(P)$ for each $P \in \mathcal{P}_{\text{CATH}}$ and various energy thresholds $E < 0$. We have concentrated here just on the question of finding tuples of robust invariants that uniquely determine the domain P among all the domains in $\mathcal{P}_{\text{CATH}}$, or the standard representatives of all the classes at some level; this section simply presents these empirical "injectivity" results.

Our first results rely only on the most basic of robust invariants, which depend only on the topological type of the surface, namely, the modified genus $g^*_E(P)$ and the number $r_E(P)$ of boundary components of $F(G_{-\infty,E}(P))$.

RESULT 5.1 *The 14 numbers* $(g^*_E(P), r_E(P))$, *with* $E = -0.5(1+t)$ *for integral* $0 \le t \le 6$, *uniquely determine the primary structure of each* $P \in \mathcal{P}_{\text{CATH}}$ *except for the special cases given in Table* B.1. *In particular, these 14 numbers uniquely determine the depth 7 classes* (CATHSOL) *except for the four following special cases*:

- 3.40.50.720.63.1.1.1.1 *and* 3.40.50.720.63.1.2.1.1,
- 3.30.70.270.7.1.2.1.1 *and* 3.30.70.270.2.1.5.5.2,
- 2.10.210.10.1.1.1.1.1 *and* 1.10.8.10.13.1.1.1.2,
- 2.10.69.10.3.2.2.1.1 *and* 2.10.69.10.3.2.5.1.1.

The next injectivity result relies upon several robust invariants of the fatgraph.

RESULT 5.2 *For any polypeptide structure* P *and energy threshold* $E < 0$, *consider the 10 numbers given by*

- *the number of residues of* P,
- *the number of hydrogen bonds of* P *with energy at most* E, $r_E(P)$, *and* $g^*_E(P)$,
- *the mean of the peptide length spectrum to one significant digit,*
- *the number of twisted alpha carbon linkages of* $G_{-\infty,E}(P)$,
- *the number of twisted edges of* $G_{-\infty,E}(P)$ *corresponding to hydrogen bonds,*

- *the respective number of pairs* FF, FN, *and* NN *occurring in the flip sequence.*

These numbers for the single energy level $E = -0.5$ uniquely determine the standard representatives of $\mathcal{P}_{\text{CATH}}$ classes at depth 4 (CATH) *except for the 19 exceptions enumerated in Table* B.2.

Our final injectivity result relies only on the model of the backbone, namely, on the flip sequence.

RESULT 5.3 *The flip sequence nearly uniquely determines elements of $\mathcal{P}_{\text{CATH}}$ with the 45 exceptions enumerated in Table* B.3.

We regard Results 5.1 to 5.3 as topological classifications of protein domains in the spirit of topology determining geometry as is familiar from rigidity results for three-dimensional manifolds, for example. We have intentionally taken so complete a collection of robust invariants as to obtain complete invariants of globules, yet subsets of these complete collections provide new tools for classification.

5.3 Classification and Prediction Results

To illustrate the use of fatgraph invariants for classification, we focus on two specific CATH domain topologies in version 3.2.0, *pectate lyase C-like* (2.160.20) and *glycosyltransferase* (1.50.10), both comprising five homologous superfamilies (H-level). Pairwise scatterplots (Figures B.1 and B.2) of the three robust invariants, the modified genus, the number of boundary components, and the number of twisted alpha carbon linkages clearly indicate separation of the homologous superfamilies in both cases.

We have implemented a machine-learning approach for domain classification, at the H-level as well as at the refined S-level, where domains are further grouped according to sequence similarity. To this end, we use the machine-learning algorithm "random forests" [7], which is a probabilistic approach and depends on the specific run of the algorithm; hence we repeated the training step 100 times. Two-thirds of each H-level (or S-level) family are used for training, and the remaining one-third is used for testing/prediction. Each domain is represented by the three robust invariants mentioned above in addition to the number of residues.

In 2.160.20 there are five homologous superfamilies, 2.160.20.10, 2.160.20.20, 2.160.20.50, 2.160.20.60, and 2.160.20.70 with 102, 14, 3, 9, and 14 members, respectively. The overall performance (percentage of correctly predicted domains in the testing set) was almost identical for all runs with an avarage of 99.3%. The domains in the three homologous superfamilies 2.160.20.20, 2.160.20.50, and 2.160.20.60 are all correctly predicted in each run; the remaining two are all predicted correctly in 72 and 99 cases, respectively. At the S-level in the CATH hierarchy, we still observe clear separation of families (Figure B.1 and B.2). There are a total of 16 different (nonsingleton) S-families in 2.160.20 ranging in size from 2 to 21 domains. The average performance in the 100 runs is 97.1%, and 10 of the S-families are all correctly classified in each run. The remaining S-families are predicted correctly between 54 and 99 times out of the 100 runs.

The topology 1.50.10 comprises five homologous superfamilies of sizes 2, 21, 41, 96, and 210, respectively, and in 100 tests the mean performance of the classifier is 95.6%. The S-level comprises 27 nonsingleton classes ranging in sizes from 2 to 60, and the average performance in 100 runs is 94.4%. Ten of these are 100% correctly classified in all of 100 independent runs with the remaining predicted correctly between 7 and 99 times. Note that the lowest scoring S-class is 1.50.10.10.2, which contains only three domains.

To illustrate how the modified genus and the number of boundary components vary at different levels of CATH, we have taken as an example the domain 1o88A00 with CATHSOLID classification 2.160.20.10.11.2.1.1.1 belonging to the "pectate lyase C-like" topology; see Figure B.3. The deepest level D contains a unique identifier (hence only one domain with a given CATHSOLID classification), whereas higher levels are potentially populated by more domains. Levels S, O, L, and I are defined based on sequence similarity, e.g., domains having the same I-levels are substrings of each other sharing at least 80% sequence overlap; the variables g and r alone are unable to differentiate domains at this I-level; cf. Figure B.3.

More detailed and systematic statistical analyses across the entire CATH database will be taken up elsewhere.

6 Closing Remarks

The fatgraph corresponding to a polypeptide structure defined here and its generalizations discussed in Section 3.4 are based on the intrinsic geometry of a protein at equilibrium. We believe that we have just scratched the surface of defining meaningful protein descriptors derived from robust invariants of these fatgraphs in this paper, whose primary intent is simply to introduce these methods. Further applications are either ongoing or anticipated, and we briefly discuss aspects of these various projects in this closing section.

Recall from Section 3.4 that rotamer fatgraphs arise from our basic fatgraph model of a polypeptide structure by refining the simplest discretization of the backbone graph connection. Such a rotamer fatgraph or invariants of it may be assigned to the subsequence of a protein corresponding to a turn or coil in order to give a new classification of these structural elements. Construction 3.9 associates matrices to hydrogen bonds, thus providing new tools for their analysis, for example, discretizations likewise providing new classifications of hydrogen bonds.

More generally, the fatgraph or rotamer fatgraph of a protein or protein domain and robust invariants of it provide new descriptors that can be used to refine existing structural classifications. A key attribute of these new descriptors, as exemplified by the injectivity results in Section 5.2, is that they are automatically computable from PDB files without the need for human interpretation into the usual architectural motifs. In a similar vein, [28] associates protein descriptors inspired by quantum invariants of links, which are different from the quantum invariants proposed

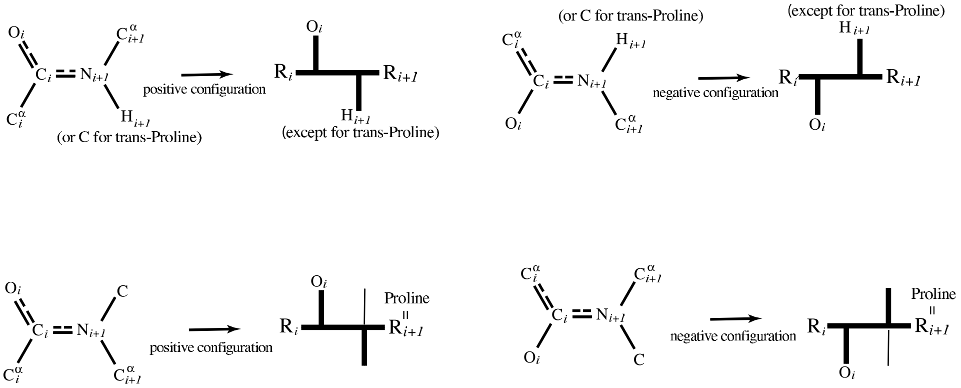

FIGURE A.1. Fatgraph building blocks for the alternative model.

in Section 3.4, and proves injectivity results analogous to those in Section 5.2. In contrast to [28], where the geometric or topological meaning of the descriptors is unclear, the significance of our descriptors such as those considered in Section 5.2 is manifest.

The recent paper [5] studies probability densities on the space of conformational angles with applications to structure prediction, and densities on the Lie group SO(3) can be computed and applied to structure prediction in an analogous manner. Furthermore, the prediction of corresponding discretizations such as the flip sequence and its rotamer analogues from protein primary structure has already proved interesting.

Appendix A: Alternative Description of the Model

There is another representative $K(P)$ of the equivalence class of the fatgraph $G(P)$ associated to a polypeptide structure P, which we shall describe in this appendix. In some ways, the alternative description is more natural, though Corollary 3.12 is true but not obvious in this formulation.

The backbone is modeled as the concatenation of fatgraph building blocks, one such building block for each peptide unit. The two possible building blocks for the i^{th} peptide unit are illustrated in Figure A.1 and are called the *positive* and *negative configurations* corresponding to whether the oxygen atom O_i lies to the left or right of the backbone, respectively, when traversed from N to C termini. The model of the backbone is determined by the sequence of configurations, positive or negative, assigned to the consecutive peptide units and is thus described by a word of length $L - 1$ in the alphabet $\{+, -\}$, which is called the *plus/minus sequence* of the polypeptide structure. The untwisted fatgraph $Y(P)$, which is an alternative model of the backbone, is constructed from this data by identifying endpoints of the consecutive horizontal segments of the fatgraph building blocks in the natural

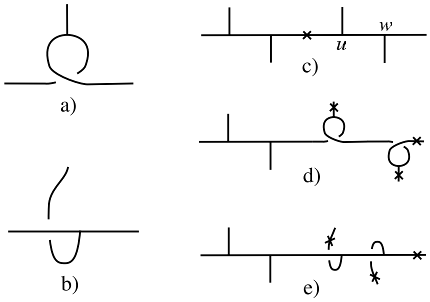

FIGURE A.2. Elementary equivalences of fatgraphs.

way as before. There is an arbitrary choice of configuration $c_1 = +$ for the first building block as positive.

Suppose recursively that configurations $c_\ell \in \{+, -\}$ have been determined for $\ell < i < L$. The configuration c_i is calculated from the configuration c_{i-1} as follows:

$$c_i = \begin{cases} +c_{i-1} & \text{if } \vec{v}_{i-1} \cdot \vec{v}_i + \vec{w}_{i-1} \cdot \vec{w}_i > 0 \text{ and } \mathrm{R}_i \text{ is not in the cis-conformation}, \\ -c_{i-1} & \text{if } \vec{v}_{i-1} \cdot \vec{v}_i + \vec{w}_{i-1} \cdot \vec{w}_i \le 0 \text{ and } \mathrm{R}_i \text{ is not in the cis-conformation}, \\ -c_{i-1} & \text{if } \vec{v}_{i-1} \cdot \vec{v}_i + \vec{w}_{i-1} \cdot \vec{w}_i \ge 0 \text{ and } \mathrm{R}_i \text{ is in the cis-conformation}, \\ +c_{i-1} & \text{if } \vec{v}_{i-1} \cdot \vec{v}_i + \vec{w}_{i-1} \cdot \vec{w}_i < 0 \text{ and } \mathrm{R}_i \text{ is in the cis-conformation}, \end{cases}$$

completing the construction of the alternative backbone model $Y(P)$. Notice that the flip sequence uniquely determines the plus/minus sequence and conversely.

As in Construction 3.8, if $(i, j) \in \mathcal{B}$ in input (ii), then we add an edge to $Y(P)$ connecting the short vertical segments corresponding to the atoms H_i and O_j. To complete the construction of $K(P)$, it remains only to specify which edges of the resulting fatgraph are twisted. To this end, suppose that $(i, j) \in \mathcal{B}$ in input (ii). There are corresponding 3-frames

$$\mathcal{F}_{i-1} = (\vec{u}_{i-1}, \vec{v}_{i-1}, \vec{w}_{i-1}),$$
$$\mathcal{F}_j = (\vec{u}_j, \vec{v}_j, \vec{w}_j),$$

from Construction 3.2 and corresponding configurations c_{i-1} and c_j defined above. An edge corresponding to the hydrogen bond $(i, j) \in \mathcal{B}$ is taken to be twisted in $K(P)$ if and only if $c_{i-1} c_j \operatorname{sign}(\vec{v}_{i-1} \cdot \vec{v}_j + \vec{w}_{i-1} \cdot \vec{w}_j)$ is negative.

The proof that $K(P)$ and $G(P)$ are equivalent depends upon the following simple diagrammatic result.

LEMMA A.1 *The fatgraphs that are depicted in Figures* A.2a) *and* A.2b) *are strongly equivalent, and the fatgraphs depicted in Figures* A.2c), A.2d)*, and* A.2e) *are pairwise equivalent.*

PROOF: The strong equivalence of A.2a) and A.2b) is proved directly. Perform vertex flips on the vertices labeled u, w in A.2c) and erase pairs of icons × on

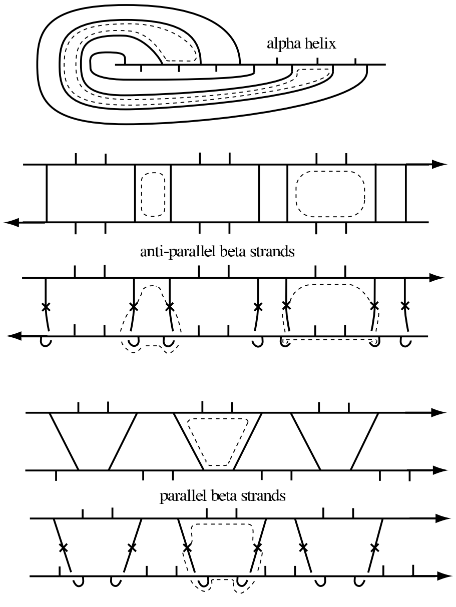

FIGURE A.3. Alpha helices and beta strands.

common edges to produce A.2d), which is strongly equivalent to A.2e) according to the first assertion. □

PROPOSITION A.2 *The fatgraphs $G(P)$ and $K(P)$ are equivalent.*

PROOF: The underlying graphs of $G(P)$ and $K(P)$ are isomorphic by construction. Furthermore, recursive application of Lemma A.1 shows that there is a sequence of vertex flips starting at $T(P)$ and ending at $Y(P)$, so the two backbone models are equivalent by Proposition 2.4. We claim that an edge of $G(P)$ representing a hydrogen bond is twisted if and only if the corresponding edge of $K(P)$ is twisted, and there are two cases depending upon the parity of the number of twisted alpha carbon linkages of $G(P)$ between the endpoints of such an edge. This number is even, and hence so too is the number of icons × on the edge, if and only if the configurations of fatgraph building blocks in $K(P)$ at these endpoints agree, and the claim therefore follows by the definition of twisting in $K(P)$. □

We finally consider how the standard motifs of protein secondary structure are manifest in our alternative model $K(P)$. The illustration on the top of Figure A.3 depicts our fatgraph model of an *alpha helix*, which is defined by the indicated pattern of hydrogen bonding. It is well-known for proteins [10] that the plus/minus sequence of an alpha helix is given by a constant[7] plus/minus sequence $+ + + + +$ or $- - - - -$. Indeed, this is the standard graphical depiction of an alpha helix

[7] This can be seen, for example, from the Ramachandran plot Figure 3.3 or from the direct consideration of 3-frames according to Construction 3.4.

in the protein literature, but for us, there is the deeper meaning of the figure as a fatgraph rather than simply as a graph in its usual interpretation. The dotted line indicates a typical boundary component of the corresponding surface.

The second and fourth illustrations from the top in Figure A.3 depict our fatgraph models of an *antiparallel beta strand* and a *parallel beta strand*, respectively, which are again defined by the indicated pattern of hydrogen bonding and the orientations along the backbone from the N to C termini indicated by the arrows in the figure. Again, it is well-known for proteins [10] that a beta strand, whether parallel or antiparallel, has an alternating (see footnote 7) plus/minus sequence $+ - + - +$ or $- + - + -$. Again, these are the standard graphical depictions of beta strands but now with our enhanced fatgraph interpretation, and the dotted lines indicate typical boundary components of the corresponding surface.

Consider the effect of a change of single configuration type in the plus/minus sequence, from $+$ to $-$ or $-$ to $+$, on the backbone between these two backbone snippets as depicted in the third and fifth illustrations from the top in Figure A.3. It follows from the definition of twisting in $K(P)$ that the vertical edges corresponding to hydrogen bonds will now be twisted. The boundary components in the second and fourth illustrations from the top persist in the third and fifth illustrations, respectively, in accordance with Corollary 3.12. Indeed, an odd number of changes of configuration types in the backbone between the two backbone snippets will produce an analogous result, and an even number leaves the figure unchanged.

Let us also clarify a point about antiparallel beta strands. It is *not necessarily the case* that the second and third illustrations from the top in Figure A.3 accurately depict our fatgraph model of an antiparallel beta strand: it may happen that our model produces the second figure but with twisted edges representing the hydrogen bonds in the strand or the third figure without such twisting. This is because the determination of twisting in $K(P)$ depends upon the sign of $cc'(\vec{v} \cdot \vec{v}' + \vec{w} \cdot \vec{w}')$, where $(\vec{u}, \vec{v}, \vec{w})$ and $(\vec{u}', \vec{v}', \vec{w}')$ are the 3-frames of the peptide units with configurations c and c' corresponding to the endpoints of the edge. Though the oxygen and hydrogen atoms involved in the hydrogen bond are within a few angstroms, the configurations c and c' may not reflect this, and furthermore, the sign of $cc'(\vec{v} \cdot \vec{v}' + \vec{w} \cdot \vec{w}')$ depends not only on c and c', but also on *both* of $\vec{v} \cdot \vec{v}'$ and $\vec{w} \cdot \vec{w}'$. This leads naturally to the notion of "untwisted antiparallel beta strands," namely, those for which Figure A.3 is accurate, and "twisted antiparallel beta strands," those for which it is not. In contrast, alpha helices and parallel beta strands *are always* represented as in Figure A.3.

In short, the passage from graph to fatgraph enhances the usual graphical depiction of alpha helices and beta strands. Changes of configuration type away from the alpha helices and beta strands leaves undisturbed the boundary components of the surface associated to the fatgraphs that model them. Furthermore, the distinction between twisted and untwisted antiparallel beta strands is new and depends upon modeling the backbone as a fatgraph rather than merely as a graph.

Appendix B: Tables and Figures

TABLE B.1: Exceptions to injectivity in Result 5.1.

Invariants	CATH domains
(26.5,80,23.5,66,21.5,58,16.5,44,11.5,18,5.0,4,3.0,2)	2.60.120.20.4.3.1.2.2 and 2.60.120.20.4.3.1.1.n for $2 \leq n \leq 24$ and $n \neq 3, 4, 10, 11, 12, 14$
(36.5,81,32.5,72,31.5,66,29.0,56,23.5,34,14.0,12,2.0,2)	2.70.98.10.2.1.1.n.1 for $3 \leq n \leq 17$ and $n \neq 9$
(34.5,84,32.5,71,31.5,69,29.5,56,22.5,41,14.0,20,5.0,9)	2.70.98.10.2.1.1.n.1 for $19 \leq n \leq 33$ and $n \neq 26, 30$
(20.5,89,17.5,82,14.0,66,8.5,48,6.5,25,3.0,12,1.0,3)	3.20.20.70.69.3.1.n.1 for $4 \leq n \leq 10$ or $n = 12, 15, 17$
(41.0,99,30.5,76,25.5,51,14.0,31,8.0,19,5.5,9,0.5,3)	3.75.10.10.1.2.2.n.1 for $1 \leq n \leq 6$ or $n = 8, 11$
(20.5,89,17.5,82,14.0,67,8.5,48,6.5,25,3.0,12,1.0,3)	3.20.20.70.69.3.1.n.1 for $n = 1, 2, 3, 13, 14, 16$
(8.0,71,6.0,63,5.5,55,5.0,43,3.0,17,0.0,4,0.0,1)	3.40.50.510.1.1.1.1.$m.n$ for $m.n = 1.1, 1.3, 2.3, 3.1$
(19.5,68,16.5,54,12.5,48,12.5,28,7.5,18,1.0,11,1.0,4)	3.90.70.10.3.2.1.$m.n$ for $m.n = 2.15, 4.1, 5.1, 8.1, 9.1$
(4.0,96,4.0,91,2.5,86,1.0,64,0.0,18,0.0,1,0.0,1)	1.10.490.10.5.1.1.$m.n$ for $m.n = 1.52, 1.53, 28.1, 28.2$
(4.0,102,3.0,93,2.5,84,1.0,58,0.0,22,0.0,2,0.0,1)	1.10.490.10.4.1.1.$m.n$ for $m.n = 1.54, 1.55, 2.17, 2.18$
(7.5,38,7.0,33,5.0,32,2.5,20,1.5,10,1.0,5,0.5,4)	2.60.40.10.2.1.1.$m.n$ for $m.n = 1.258, 1.259, 7.23, 7.24$
(1.0,29,0.5,29,0.5,27,0.0,20,0.0,11,0.0,5,0.0,2)	4.10.220.20.1.1.2.n.1 for $n = 1, 2, 3$
(4.5,169,4.0,157,2.5,145,2.0,113,1.0,59,0.5,10,0.5,1)	1.20.1070.10.1.1.1.$m.n$ for $m.n = 1.12, 1.21, 9.1$
(34.0,83,32.0,76,31.5,71,29.0,55,20.0,41,16.0,26,5.0,7)	2.70.98.10.2.1.1.n.1 for $n = 42, 44, 46$
(36.0,136,34.0,123,32.0,114,23.5,85,8.0,40,2.0,15,0.0,5)	3.20.20.140.22.1.1.n.1 for $n = 2, 3, 4$
(0.0,11,0.0,6,0.0,4,0.0,2,0.0,2,0.0,1,0.0,1)	2.10.210.10.1.1.1.1.1 and 1.10.8.10.13.1.1.1.2
(0.5,32,0.0,30,0.0,29,0.0,20,0.0,8,0.0,2,0.0,1)	4.10.220.20.1.1.1.n.1 for $n = 13, 15$
(0.5,97,0.5,94,0.5,85,0.5,65,0.5,25,0.5,5,0.0,1)	1.20.1500.10.3.1.1.n.1 for $n = 1, 2$
(1.0,21,1.0,17,0.5,15,0.5,14,0.5,9,0.0,3,0.0,2)	2.10.69.10.3.2.2.1.1 and 2.10.69.10.3.2.5.1.1
(1.5,42,1.5,42,1.5,39,0.5,32,0.0,16,0.0,5,0.0,1)	1.20.1280.10.1.1.1.$m.n$ for $m.n = 1.1, 2.47$
(1.5,43,1.5,42,1.5,38,0.5,32,0.0,17,0.0,5,0.0,1)	1.20.1280.10.1.1.1.$m.n$ for $m.n = 1.2, 2.48$
(3.0,21,3.0,18,3.0,15,3.0,13,1.5,9,0.5,3,0.5,1)	4.10.410.10.1.1.3.n.2 for $n = 4, 7$
(3.0,21,3.0,18,3.0,16,3.0,13,2.0,8,0.5,3,0.0,1)	4.10.410.10.1.1.3.n.1 for $n = 5, 8$
(4.5,8,4.5,8,3.5,7,2.5,5,1.0,4,0.0,2,0.0,1)	2.10.25.10.20.2.1.n.1 for $n = 1, 2$
(4.5,35,3.5,34,3.0,29,1.5,23,1.0,14,0.0,6,0.0,1)	1.10.1200.30.1.1.2.$m.n$ for $m.n = 1.3, 4.1$
(4.5,51,4.5,42,4.5,32,4.0,20,3.5,15,2.5,5,1.0,4)	3.30.70.270.4.1.1.$m.n$ for $m.n = 1.185, 2.1$
(5.5,48,4.5,40,4.0,32,3.0,18,1.0,12,0.0,10,0.0,4)	1.10.238.10.3.1.2.n.1 for $n = 5, 6$
(6.0,42,5.5,36,5.5,31,5.0,29,4.5,15,2.0,1,0.0,1)	2.40.70.10.3.1.1.$m.n$ for $m.n = 5.6, 6.10$
(6.0,44,5.5,39,4.5,30,4.5,23,3.5,14,1.0,6,1.0,2)	1.10.760.10.6.1.1.n.1 for $n = 1, 25$
(6.5, 32,6.0,30,5.5,27,3.5,28,2.5,16,1.5,4,0.0,1)	2.30.30.140.3.1.1.$m.n$ for $m.n = 1.3, 2.1$
(6.5,44,4.5,41,4.5,35,4.0,25,3.5,13,2.5,7,0.5,4)	3.30.70.270.7.1.2.1.1 and 3.30.70.270.2.1.5.5.2
(6.5,57,6.0,52,6.0,52,5.5,42,3.5,25,2.5,7,0.5,1)	3.30.365.10.4.1.1.$m.n$ for $m.n = 1.1, 2.2$
(7.0,65,7.0,64,6.5,60,3.0,54,2.0,28,0.5,5,0.0,1)	1.10.1040.10.4.1.1.n.1 for $n = 1, 2$
(7.5,71,6.5,63,4.5,57,4.5,41,2.0,19,2.0,6,0.0,3)	3.30.1330.10.1.1.1.n.1 for $n = 2, 4$
(7.5,72,5.5,64,5.0,56,5.0,43,3.0,17,0.0,4,0.0,1)	3.40.50.510.1.1.1.$m.n$ for $m.n = 2.4, 3.2$
(8.0,65,8.0,57,7.5,50,6.0,35,3.5,24,1.0,8,0.0,1)	3.30.1330.10.1.1.1.n.1 for $n = 3, 5$
(8.5,35,8.0,33,7.5,31,6.0,26,4.0,17,3.5,4,0.5,2)	2.30.30.140.3.1.1.$m.n$ for $m.n = 1.4, 2.2$
(8.5,69,7.5,62,6.5,56,5.5,45,5.5,24,3.5,3,0.0,1)	3.40.47.10.8.1.1.n.4 for $n = 2, 6$
(8.5,70,7.5,62,7.0,56,6.0,40,4.5,20,2.5,2,0.0,1)	3.40.47.10.8.1.1.n.1 for $n = 2, 6$
(9.0,68,8.0,60,6.5,53,6.0,40,5.0,12,1.5,1,0.5,1)	3.40.47.10.8.1.1.n.8 for $n = 2, 6$
(9.0,69,7.5,63,6.5,54,5.5,43,4.5,14,1.0,2,0.0,1)	3.40.47.10.8.1.1.n.6 for $n = 2, 6$
(9.0,70,7.5,63,6.5,55,6.0,43,5.0,19,2.0,1,0.0,1)	3.40.47.10.8.1.1.n.2 for $n = 2, 6$
(9.5,67,7.5,60,5.5,52,5.0,41,4.0,12,2.0,3,0.0,1)	3.40.47.10.8.1.1.n.7 for $n = 2, 6$
(9.5,67,8.0,61,6.0,54,5.0,43,5.0,19,2.0,3,0.0,1)	3.40.47.10.8.1.1.n.3 for $n = 2, 6$
(9.5,68,8.0,62,7.5,52,6.0,37,4.0,16,1.5,2,0.0,1)	3.40.47.10.8.1.1.n.5 for $n = 2, 6$
(9.5,71,6.5,62,6.0,52,3.5,43,2.5,27,2.0,8,1.5,5)	3.40.420.10.2.2.4.n.1 for $n = 1, 2$
(10.5,36,10.5,32,9.0,28,7.5,24,4.5,14,0.0,7,0.0,3)	3.10.20.30.6.1.1.n.1 for $n = 2, 4$
(10.5,58,10.0,49,10.0,47,8.5,33,7.0,15,3.5,4,0.5,2)	3.10.310.10.6.1.2.n.1 for $n = 1, 2$
(13.5,73,13.5,65,11.5,60,10.5,45,7.5,22,2.5,7,0.0,2)	3.40.50.720.82.1.1.n.1 for $n = 4, 9$
(13.5,74,13.5,67,11.5,64,10.5,37,6.5,14,1.5,7,0.0,2)	3.40.50.720.82.1.1.n.1 for $n = 2, 6$
(14.0,49,14.0,44,13.0,43,13.0,39,9.5,17,3.0,5,0.0,1)	3.30.1330.40.2.1.1.n.1 for $n = 1, 3$
(14.0,58,12.0,52,11.5,47,10.5,33,7.5,14,3.0,8,0.0,5)	3.10.310.10.8.1.1.n.1 for $n = 6, 7$
(17.5,79,15.0,64,12.5,54,9.5,38,6.5,21,4.0,6,1.5,2)	2.60.120.20.9.3.1.$m.n$ for $m.n = 1.15, 6.1$
(18.5,81,14.5,65,13.5,55,10.0,40,7.0,24,3.5,8,1.5,2)	2.60.120.20.9.3.1.$m.n$ for $m.n = 1.18, 6.2$
(19.0,20,18.0,21,16.0,17,9.5,18,7.0,10,2.0,4,0.5,2)	2.60.30.10.2.1.1.n.1 for $n = 7, 9$
(19.0,55,18.0,50,17.0,45,14.0,34,8.0,18,2.0,5,0.0,1)	3.40.50.720.63.1.n.1.1 for $n = 1, 2$
(19.5,149,19.5,137,18.0,124,12.5,97,7.5,50,1.5,7,0.0,1)	3.20.20.110.1.1.3.n.1 for $n = 11, 13$
(19.5,180,18.5,161,16.0,135,14.0,77,10.0,28,1.0,8,0.0,2)	3.20.20.70.55.2.1.$m.n$ for $m.n = 5.8, 7.4$
(19.5,185,15.5,163,11.5,130,11.5,82,6.0,42,3.5,10,0.0,1)	3.20.20.70.55.2.1.$m.n$ for $m.n = 5.5, 7.1$
(20.0,43,18.5,38,15.5,31,13.5,22,9.5,14,7.0,6,2.0,4)	3.90.650.10.1.1.1.n.1 for $n = 3, 5$

Data continue on the next page

TABLE B.1: *Continued*

Invariants	CATH domains
(20.5,61,18.5,51,16.5,47,15.5,31,10.5,20,5.5,5,0.0,4)	2.60.90.10.1.3.1.n.1 for $n = 1, 3$
(21.5,46,17.0,38,15.0,33,13.5,23,9.5,14,4.5,5,2.0,2)	3.90.650.10.1.1.1.n.1 for $n = 2, 4$
(22.0,178,19.0,157,18.0,129,15.0,86,9.5,30,2.0,6,0.0,1)	3.20.20.70.55.2.1.$m.n$ for $m.n = 5.6, 7.2$
(23.0,178,20.0,160,18.0,134,14.5,82,11.0,34,2.0,9,0.0,2)	3.20.20.70.55.2.1.$m.n$ for $m.n = 5.7, 7.3$
(24.0,274,19.5,257,16.0,228,13.0,176,10.0,90,1.0,22,0.0,2)	1.10.620.20.6.1.1.$m.n$ for $m.n = 1.2, 2.48$
(26.5,171,24.0,151,20.5,134,16.5,105,12.5,52,3.0,16,1.0,1)	3.40.718.10.4.6.1.$m.n$ for $m.n = 1.4, 3.2$
(27.5,180,22.0,160,19.5,141,16.5,105,10.5,51,6.0,12,0.5,3)	3.40.718.10.4.6.1.$m.n$ for $m.n = 1.3, 3.1$
(36.0,102,28.5,94,26.0,81,20.0,58,12.5,27,6.5,9,2.0,2)	3.50.50.60.55.1.1.n.1 for $n = 7, 9$
(36.5,81,32.5,72,31.5,66,29.0,56,24.5,33,14.0,12,2.0,2)	2.70.98.10.2.1.1.n.1 for $n = 9, 18$
(36.5,145,34.0,130,27.5,124,25.0,92,15.5,37,3.5,6,0.5,1)	3.20.20.70.72.1.1.$m.n$ for $m.n = 3.8, 5.4$
(36.5,145,34.0,131,28.5,123,25.5,96,17.0,41,5.0,6,0.5,1)	3.20.20.70.72.1.1.$m.n$ for $m.n = 3.6, 5.2$
(38.5,141,36.0,126,30.5,117,27.0,90,19.0,39,4.5,6,0.5,1)	3.20.20.70.72.1.1.$m.n$ for $m.n = 3.7, 5.3$
(39.0,142,35.5,127,30.0,119,26.5,92,16.5,37,5.5,5,1.0,1)	3.20.20.70.72.1.1.$m.n$ for $m.n = 3.5, 5.1$
(41.0,99,30.5,76,25.5,51,14.0,30,8.0,19,5.5,9,0.5,3)	3.75.10.10.1.2.2.n.1 for $n = 7, 10$

TABLE B.2. Exceptions to injectivity in Result 5.2.

Invariants	CATH domains
(49,45,46,0.0,4.0,0,0,0,0,46)	1.20.5.190.1.1.2.1.4, 1.20.5.530.1.1.1.1.2, 1.20.5.170.1.1.2.1.1
(56,51,52,0.0,4.0,0,0,0,0,53)	1.20.5.190.1.1.3.1.1, 1.20.5.500.1.1.1.1.3, 1.20.5.170.9.1.1.1.1
(42,38,39,0.0,4.0,0,0,0,0,39)	1.20.5.190.1.1.3.2.1, 1.20.5.170.3.1.1.1.12
(46,31,30,1.0,5.0,2,3,0,2,39)	1.10.60.10.3.1.1.1.2, 1.10.287.680.1.1.1.1.16
(49,43,44,0.0,4.1,0,0,0,0,46)	1.20.5.300.2.1.1.1.7, 1.20.5.170.2.2.1.1.6
(49,25,24,1.0,6.0,6,3,1,5,35)	1.10.10.60.32.1.1.1.42, 4.10.51.10.1.1.1.1.25
(50,45,46,0.0,4.0,0,0,0,0,47)	1.20.5.80.2.1.1.2.2, 1.20.5.170.2.2.1.1.2
(52,48,49,0.0,4.0,0,0,0,0,49)	1.20.5.530.1.1.1.1.1, 1.20.5.170.2.1.1.1.2
(52,32,33,0.0,5.0,6,1,2,3,40)	4.10.220.20.1.1.1.1.1, 1.20.5.810.3.1.1.7.1
(53,30,27,2.0,6.0,5,6,1,4,41)	1.10.1220.10.3.1.3.1.3, 1.10.890.20.1.1.1.1.3
(59,55,56,0.0,4.0,0,0,0,0,56)	1.20.5.500.1.1.1.1.2, 1.20.5.170.10.1.1.3.1
(60,56,57,0.0,4.0,0,0,0,0,57)	1.20.5.500.1.1.1.1.1, 1.20.5.170.10.1.1.3.2
(62,58,59,0.0,4.0,0,0,0,0,59)	1.20.5.170.6.1.1.2.1, 1.20.5.110.6.1.1.2.3
(64,58,59,0.0,4.1,0,0,0,0,61)	1.20.5.300.1.1.1.1.2, 1.20.5.170.6.1.1.1.8
(65,37,35,1.5,5.7,9,5,2,7,46)	1.10.8.200.1.1.1.2.1, 1.10.2030.10.1.1.1.1.8
(72,48,46,1.5,5.1,7,3,2,5,57)	1.10.40.30.1.1.2.1.6, 1.10.220.10.8.1.1.1.2
(79,75,76,0.0,4.0,0,0,0,0,76)	1.20.5.170.16.1.1.1.5, 1.20.5.110.7.1.1.2.1
(88,60,53,4.0,5.5,10,11,4,6,69)	1.10.238.10.9.2.1.1.10, 1.10.288.10.2.1.1.1.1
(95,54,42,6.5,7.0,38,23,26,11,43)	3.30.1050.10.5.1.1.1.6, 3.30.1490.70.4.1.1.1.2

TABLE B.3. Exceptions to injectivity in Result 5.3, where N^k denotes $k \geq 1$ consecutive N.

Flip Sequence	CATH domains
N^{19}	1.20.5.460.1.1.1.6.1, 1.20.5.110.15.1.1.1.1
N^{27}	1.20.5.800.1.1.2.1.1, 1.10.10.380.1.1.1.1.1
N^{29}	1.20.5.140.3.1.1.1.1, 1.20.5.420.5.1.1.1.1, 1.20.5.170.18.1.1.1.1
N^{30}	1.20.5.700.1.1.1.1.1, 1.20.5.100.2.1.1.1.1
N^{32}	1.20.5.770.1.1.1.1.1, 1.20.5.700.1.1.1.1.3
N^{37}	1.20.5.40.1.1.2.1.6, 1.20.5.80.2.1.1.2.5
N^{38}	1.20.5.440.1.1.1.1.1, 4.10.810.10.1.1.1.1.1, 1.20.5.170.8.1.1.1.5
N^{40}	1.20.5.190.1.1.3.2.1, 1.20.5.170.3.1.1.1.12
N^{42}	1.20.5.430.1.1.2.1.3, 1.20.5.80.2.1.1.1.3, 1.20.5.490.1.1.1.1.1
N^{43}	1.20.5.240.1.2.1.1.1, 1.10.930.10.1.1.2.1.2, 1.20.5.170.3.1.1.1.1
N^{44}	1.20.5.230.1.1.1.1.1, 1.20.5.80.1.1.1.1.2
N^{45}	1.20.5.190.1.1.2.1.5, 1.20.5.300.2.1.1.1.12, 1.20.5.170.14.1.1.1.1
N^{46}	1.20.5.300.2.1.1.1.9, 1.10.287.300.1.1.1.1.1
N^{47}	1.20.5.190.1.1.2.1.4, 1.20.5.530.1.1.1.1.2, 1.20.5.300.2.1.1.1.7, 1.20.5.170.1.1.2.1.1
N^{48}	1.20.5.190.1.1.1.1.2, 1.20.5.80.2.1.1.2.1, 1.20.5.300.2.1.1.1.1, 1.20.5.170.2.2.1.1.1
N^{49}	1.20.5.190.1.1.2.1.1, 1.20.5.170.2.2.1.1.11, 1.20.5.110.2.1.1.1.3
N^{50}	1.20.5.290.1.1.1.1.1, 1.20.5.530.1.1.1.1.1, 1.20.5.170.2.1.1.1.2, 1.20.5.110.14.1.1.1.1
N^{51}	1.20.5.190.1.1.5.1.1, 1.20.5.370.2.1.2.1.1, 1.20.5.170.10.1.1.1.1
N^{52}	1.10.287.750.1.1.8.1.1, 1.20.5.170.2.2.1.2.2, 1.20.5.110.11.1.1.1.1
N^{53}	1.20.5.170.2.2.1.2.1, 1.20.5.110.10.1.1.1.1
N^{54}	1.20.5.190.1.1.3.1.1, 1.20.5.500.1.1.1.1.3, 1.20.5.170.4.1.1.1.1
N^{56}	1.20.5.300.1.2.1.1.2, 1.20.5.110.5.1.1.1.2
N^{57}	1.20.5.500.1.1.1.1.2, 1.20.5.170.10.1.1.3.1, 1.10.287.130.2.1.1.1.6
N^{58}	1.20.5.390.1.1.1.1.1, 1.20.5.500.1.1.1.1.1, 1.20.5.170.10.1.1.3.2, 1.20.5.110.8.1.1.1.1
N^{59}	1.20.5.620.1.1.1.1.1, 1.10.287.230.1.1.1.1.2, 1.20.5.170.4.2.1.1.1, 1.20.5.110.5.1.1.1.1
N^{60}	1.20.5.300.1.1.1.1.1, 1.20.5.170.4.1.1.2.2, 1.20.5.110.6.1.1.2.3
N^{61}	1.10.287.210.2.2.1.8.1, 1.20.5.170.6.1.1.1.11, 1.20.5.110.3.1.1.1.1
N^{62}	1.20.5.300.1.1.1.1.2, 1.20.5.170.6.1.1.1.8, 1.20.5.110.4.1.1.1.1
N^{63}	1.20.5.500.1.1.1.1.4, 1.20.5.170.5.1.1.1.1
N^{65}	1.10.1440.10.1.1.1.1.1, 1.20.5.170.5.1.1.1.2, 1.2.5.110.6.1.1.1.1
N^{66}	1.20.5.730.1.1.1.1.1, 1.20.5.170.6.1.1.1.3, 1.20.5.110.2.1.1.1.1
N^{71}	1.20.5.400.1.1.1.1.1, 1.10.287.210.2.2.1.4.4, 1.20.5.110.6.1.2.2.2
N^{72}	1.10.287.210.2.2.1.4.3, 1.20.5.170.16.1.1.1.3, 1.20.5.110.6.1.2.2.3
N^{75}	1.20.5.340.1.1.1.1.4, 1.20.5.110.7.1.1.4.3
N^{76}	1.10.287.210.7.1.1.1.1, 1.20.5.170.16.1.1.1.4
N^{77}	1.20.20.10.1.1.1.1.3, 1.20.5.340.1.1.1.1.3, 1.20.5.170.16.1.1.1.5, 1.20.5.110.7.1.1.2.1
$\mathrm{N}^{28}\mathrm{FN}^{25}$	1.10.287.660.1.1.1.2.1, 1.10.287.230.1.1.2.1.5, 1.10.287.750.1.1.6.1.1
$\mathrm{N}^{2}\mathrm{FN}^{61}$	1.20.5.170.5.1.1.2.1, 1.20.5.110.6.1.1.2.1
$\mathrm{N}^{27}\mathrm{FN}^{26}$	1.10.287.230.1.1.1.4.1, 1.10.287.210.2.1.2.1.3
$\mathrm{N}^{29}\mathrm{FN}^{24}$	1.10.287.230.1.1.2.1.4, 1.10.287.750.1.1.5.1.1
$\mathrm{N}^{31}\mathrm{FN}^{26}$	1.10.287.750.1.1.3.1.1, 1.10.287.210.2.2.1.7.1
$\mathrm{N}^{34}\mathrm{FNF}^{2}\mathrm{N}$	4.10.81.10.2.1.1.1.1, 1.20.5.50.9.1.1.1.8
$\mathrm{N}^{41}\mathrm{F}$	1.20.5.490.1.1.1.1.3, 1.20.1070.10.7.1.1.1.2
$\mathrm{N}^{43}\mathrm{F}$	1.10.10.200.2.2.1.1.1, 1.20.5.170.15.1.1.1.1
$\mathrm{N}^{50}\mathrm{F}$	1.20.5.170.10.1.1.2.1, 1.10.287.190.1.1.1.1.2

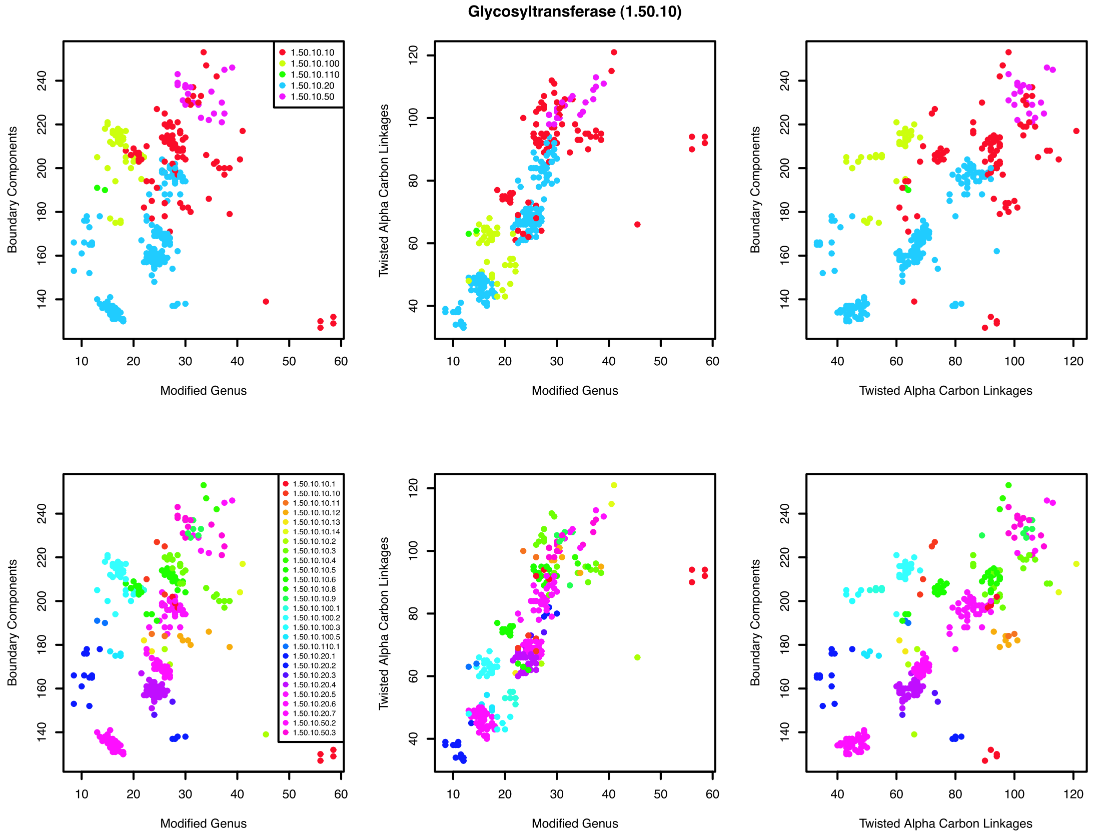

FIGURE B.1. Top row: Pairwise scatterplots for the five H-level families in the topology glycosyltransferase. Bottom row: Pairwise scatterplots for the S-level families.

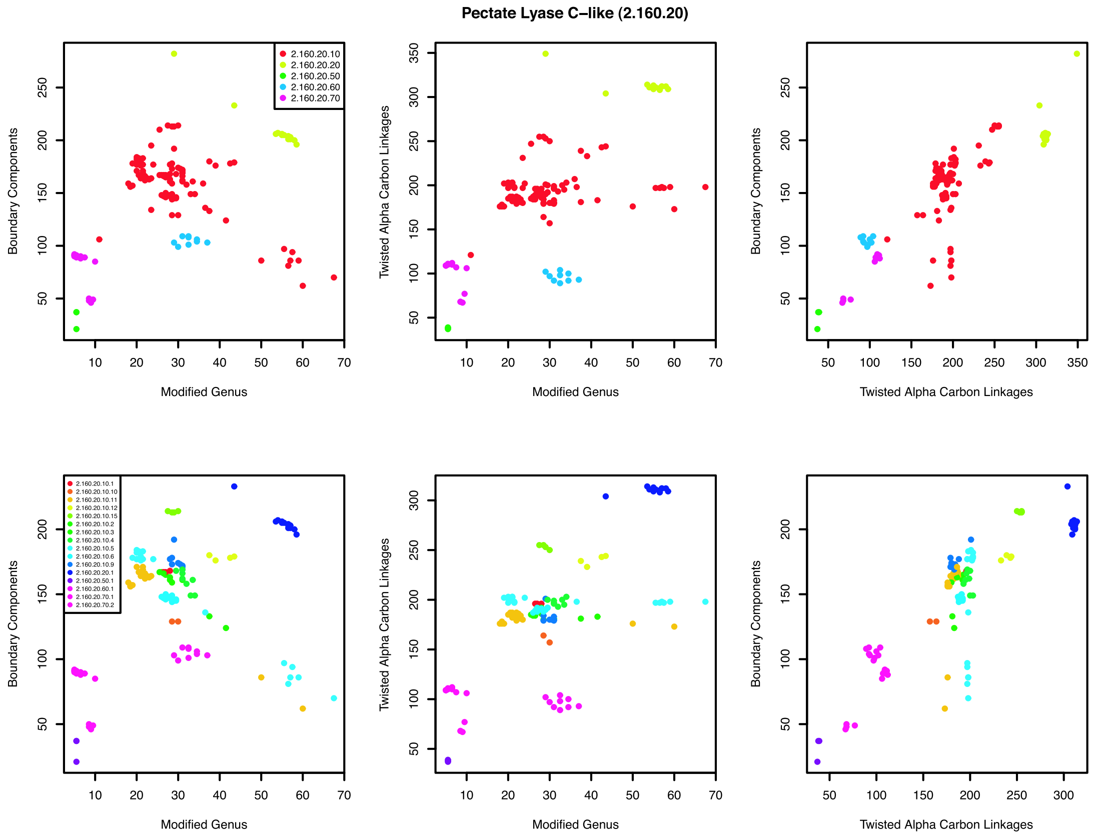

FIGURE B.2. Top row: Pairwise scatterplots for the five H-level families in the topology pectate lyase C-like. Bottom row: Pairwise scatterplots for the S-level families.

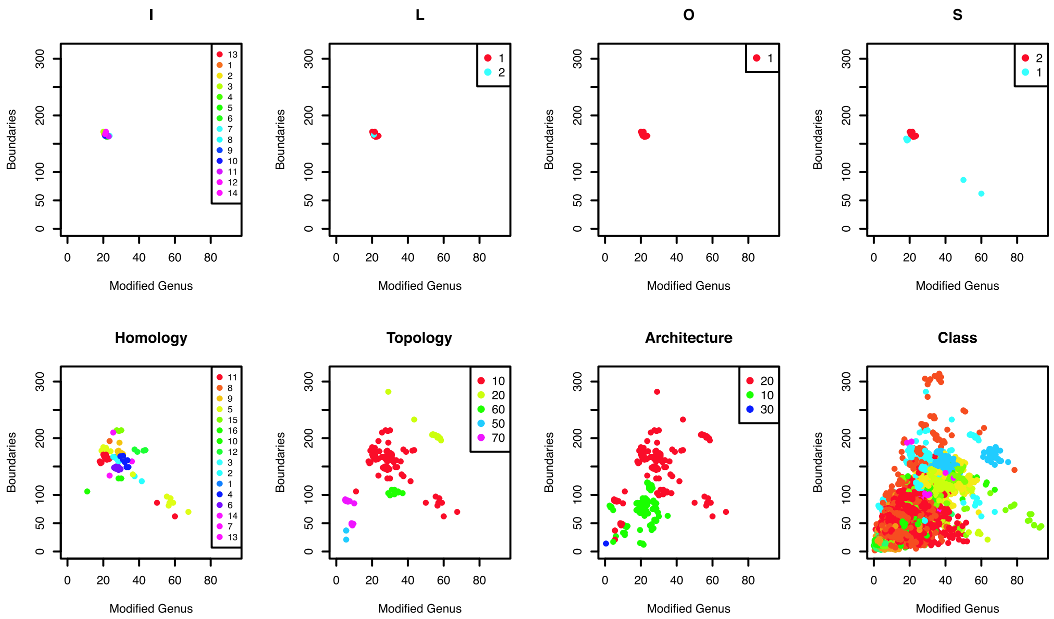

FIGURE B.3. Shown are scatterplots of the modified genus versus the number of boundary components at various levels of CATH. We start with the domain 1o88A00 having CATHSOLID classification 2.160.20.10.11.2.1.1.1, and at each CATH level, plot all domains sharing classification with 1o88A00; e.g., at the C-level, we plot all domains with classification 2 colored according to their A-level. Similarly, all domains with CA-classification 2.160 are shown with distinct colors for all topologies, and we continue all the way down to the CATHSOLI level. Note that by definition the D-level is used to distinguish individual CATH entries, so all domains with same CATHSOLI level are assigned to different D-levels.

Acknowledgments. It is a pleasure for RCP to thank the Preuss Foundation for funding an exploratory conference in 2005 at the University of Southern California, immediately after which a precursor of the model described here was derived, and to thank the participants of that conference, Ken Dill, Soren Istrail, Hubert Saleur, Arieh Warshel, Michael Waterman, and especially Alexei Finkelstein, for stimulating and provocative discussion. Excellent computer programming assistance was provided by Piotr Karasinski and funded by the Center for Theory in Natural Sciences, Aarhus University, and the Danish Ministry of Science, Technology, and Innovation.

The methods disclosed in this paper are protected by U.S. provisional patent filing 61/077,277 (July 1, 2008) and the Danish priority application PA 2008 01009 (July 17, 2008).

MK is supported by the Center for Theory in the Natural Sciences at Aarhus University. CW is partially supported by the Danish Research Councils.

Bibliography

[1] Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walters, P. *The shape and structure of proteins. Molecular biology of the cell*, 4th ed. Garland Science, New York–London, 2002. Available at: `http://www.ncbi.nlm.nih.gov/books/bv.fcgi?highlight=structure,shape,proteins&rid=mboc4.section.388`

[2] Bairoch, A. Serendipity in bioinformatics, the tribulations of a Swiss bioinformatician through exciting times! *Bioinformatics* **16** (2000), no. 1, 48–64. Available at: `http://bioinformatics.oxfordjournals.org/cgi/reprint/16/1/48`.

[3] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucl. Acids Res.* **28** (2000), no. 1, 235–242. Available at: `http://www.ncbi.nlm.nih.gov/pubmed/10592235`.

[4] Bessis D.; Itzykson, C.; Zuber, J. B. Quantum field theory techniques in graphical enumeration. *Adv. in Appl. Math.* **1** (1980), no. 2, 109–157.

[5] Boomsma, W.; Mardia, K. V.; Taylor, C. C.; Ferkinghoff-Borg, J.; Krogh, A.; Hamelryck, T. A generative, probabilistic model of local protein structure. *Proc. Nat. Acad. Sci. U.S.A.* **105** (2008), no. 26, 8932–8937.

[6] Bourbaki, N. *Elements of mathematics: Lie groups and Lie algebras*. Addison-Wesley, Reading, Mass., 1975.

[7] Breiman, L. Random forests. *Machine Learning* **45** (2001), 5–32.

[8] Brézin, E.; Kazakov, V.; Serban, D.; Wiegmann, P.; Zabrodin, A., eds. *Applications of random matrices in physics*. Proceedings of the NATO Advanced Study Institute held in Les Houches, June 6–25, 2004. NATO Science Series II: Mathematics, Physics and Chemistry, 221. Springer, Dordrecht, 2006.

[9] Darling, R. W. R. *Differential forms and connections*. Cambridge University Press, Cambridge, 1994.

[10] Finkelstein, A. V.; Ptitsyn, O. B. *Protein physics: a course of lectures (soft condensed matter, complex fluids and biomaterials)*. Academic, London–San Diego, 2002.

[11] Finkelstein, A. V. Private communication, 2008.

[12] Finn, R. D.; Tate, J.; Mistry, J.; Coggill, P. C.; Sammut, J. S.; Hotz, H. R.; Ceric, G.; Forslund, K.; Eddy, S. R.; Sonnhammer, E. L.; Bateman, A. The Pfam protein families database. *Nucl. Acids Res.* **36** (2008), D281–D288. Available at: `http://nar.oxfordjournals.org/cgi/content/full/36/suppl_1/D281?maxtoshow=&`

hits=10&RESULTFORMAT=&fulltext=the+pfam+protein+families&searchid=1&FIRSTINDEX=0&resourcetype=HWCIT

[13] Gusfield, D. *Algorithms on strings, trees, and sequences. Computer science and computational biology*. Cambridge University Press, Cambridge, 1997.

[14] Harer, J. L.; Zagier, D. The Euler characteristic of the moduli space of curves. *Invent. Math.* **85** (1986), no. 3, 457–485.

[15] Holm, L.; Kääriäinen, S.; Rosenström, P.; Schenkel, A. Searching protein structure databases with DaliLite v.3. *Bioinformatics* **24** (2008), no. 23, 2780–2781.

[16] Igusa, K. Combinatorial Miller-Morita-Mumford classes and Witten cycles. *Algebr. Geom. Topol.* **4** (2004), 473–520.

[17] Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **12** (1983), no. 12, 2577–637.

[18] Kontsevich, M. Intersection theory on the moduli space of curves and the matrix Airy function. *Comm. Math. Phys.* **147** (1992), no. 1, 1–23.

[19] Kortemme, T.; Morozov, A. V.; Baker, D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J. Mol. Biol.* **326** (2003), no. 4, 1239–1259.

[20] Lindauer, K.; Bendic, C.; Sühnel, J. HBexplore–a new tool for identifying hydrogen bonding patterns in biological macromolecules. *Comput. Appl. Biosci.* **12** (1996), no. 4, 281–289.

[21] Massey, W. S. *Algebraic topology: an introduction*. Reprint of the 1967 ed. Graduate Texts in Mathematics, 56. Springer, New York–Heidelberg 1977.

[22] Mondello, G. Combinatorial classes on $\overline{\mathscr{M}}_{g,n}$ are tautological. *Int. Math. Res. Not.* (2004), no. 44, 2329–2390.

[23] Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247** (1995), no. 4, 536–540.

[24] Ohtsuki, T. *Quantum invariants. A study of knots, 3-manifolds, and their sets*. Series on Knots and Everything, 29. World Scientific, River Edge, N.J., 2001.

[25] Orengo, C. A.; Michie, A. D.; Jones, D. T.; Swindells, M. B.; Thornton, J. M. CATH–a hierarchic classification of protein domain structures. *Structure* **5** (1997), no. 8, 1093–1108.

[26] Penner, R. C. Perturbative series and the moduli space of Riemann surfaces. *J. Differential Geom.* **27** (1988), no. 1, 35–53.

[27] Penner, R. C.; Waterman, M. S. Spaces of RNA secondary structures. *Adv. Math.* **101** (1993), no. 1, 31–49.

[28] Røgen, P.; Fain, B. Automatic classification of protein structure by using Gauss integrals. *Proc. Nat. Acad. Sci. U.S.A.* **100** (2003), no. 1, 119–124.

[29] Strebel, K. *Quadratic differentials*. Ergebnisse der Mathematik und ihrer Grenzgebiete (3), 5. Springer, Berlin, 1984.

[30] ’t Hooft, G. A planar diagram theory for strong interactions. *Nucl. Phys. B* **72** (1974), 461–473.

[31] Wu, C. H.; Apweiler, R.; Bairoch, A.; Natale, D. A.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Mazumder, R.; O’Donovan, C.; Redaschi, N.; Suzek, B. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucl. Acids Res.* **34** (2006), D187–D191.

R. C. PENNER
University of Southern California
Departments of Mathematics
and Physics/Astronomy
Los Angeles, CA 90089
E-mail: rpenner@math.usc.edu
and
Aarhus University
Department of Mathematics
Center for the Topology
and Quantization of Moduli Spaces
DK-8000 Aarhus C
DENMARK

MICHAEL KNUDSEN
Aarhus University
Bioinformatics Research Center
DK-8000 Aarhus C
DENMARK
E-mail: micknudsen@gmail.com

CARSTEN WIUF
Aarhus University
Bioinformatics Research Center
and
Danish National Research Foundation
Centre for Membrane Pumps
in Cells and Disease–PUMPKIN
DK-8000 Aarhus C
DENMARK
E-mail: wiuf@birc.au.dk

JØRGEN ELLEGAARD ANDERSEN
Aarhus University
Department of Mathematics
Center for the Topology
and Quantization of Moduli Spaces
DK-8000 Aarhus C
DENMARK
E-mail: andersen@imf.au.dk

Received July 2009.