OPEN ACCESS Freely available online

PLoS one

An Algebro-Topological Description of Protein Domain Structure

Robert Clark Penner[1,2], Michael Knudsen[3], Carsten Wiuf[3,4]*, Jørgen Ellegaard Andersen[1]

1 Center for the Topology and Quantization of Moduli Spaces, Department of Mathematical Sciences, Aarhus University, Aarhus, Denmark, **2** Departments of Mathematics and Physics/Astronomy, University of Southern California, Los Angeles, California, United States of America, **3** Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark, **4** Centre for Membrane Pumps in Cells and Disease, Aarhus University, Aarhus, Denmark

Abstract

The space of possible protein structures appears vast and continuous, and the relationship between primary, secondary and tertiary structure levels is complex. Protein structure comparison and classification is therefore a difficult but important task since structure is a determinant for molecular interaction and function. We introduce a novel mathematical abstraction based on geometric topology to describe protein domain structure. Using the locations of the backbone atoms and the hydrogen bonds, we build a combinatorial object – a so-called *fatgraph*. The description is discrete yet gives rise to a 2-dimensional mathematical surface. Thus, each protein domain corresponds to a particular mathematical surface with characteristic *topological invariants*, such as the genus (number of holes) and the number of boundary components. Both invariants are global fatgraph features reflecting the interconnectivity of the domain by hydrogen bonds. We introduce the notion of robust variables, that is variables that are robust towards minor changes in the structure/fatgraph, and show that the genus and the number of boundary components are robust. Further, we invesigate the distribution of different fatgraph variables and show how only four variables are capable of distinguishing different folds. We use local (secondary) and global (tertiary) fatgraph features to describe domain structures and illustrate that they are useful for classification of domains in CATH. In addition, we combine our method with two other methods thereby using primary, secondary, and tertiary structure information, and show that we can identify a large percentage of new and unclassified structures in CATH.

Citation: Penner RC, Knudsen M, Wiuf C, Andersen JE (2011) An Algebro-Topological Description of Protein Domain Structure. PLoS ONE 6(5): e19670. doi:10.1371/journal.pone.0019670

Editor: Christopher L. Douglas, University of California, Berkeley, United States of America

Received January 5, 2011; **Accepted** April 3, 2011; **Published** May 24, 2011

Copyright: © 2011 Penner et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: MK is supported by the Centre for Theory in Natural Sciences; CW is supported by the Danish Research Councils and by the Humboldt Foundation, Germany. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: 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). This does not alter the authors' adherence to all the PLoS ONE policies on sharing data and materials, as detailed online in the guide for authors.

* E-mail: wiuf@birc.au.dk

Introduction

Protein domains are protein subsequences that may fold and function independently of the rest of the protein [1,2]. Experimentally determined protein structures deposited in PDB [3] have been classified according to their fold and function in hierarchical databases of which CATH [4] and SCOP [5] are the most widely used. These databases involve manual steps, assisted by computational methods, for fold characterization and classification [4–6]. The database DALI [7,8], on the other hand, uses a fully automated procedure to classify domains non-hierarchically based on structural similarities only. Other methods have been proposed to reduce the description of a domain fold to a vector of numerical attributes that are characteristic for the fold [9,10]; recent methods are, for instance, based on geometric characteristics [11–14], secondary structure information [15–18], sequence information [19], and physical properties derived from the primary sequence [20]. These methods might be useful, not only for classification, but also for annotation and understanding features of protein folding.

Using techniques from geometric topology, we propose a novel mathematical abstraction for studying protein domain structures [21]. In particular, we conceive the structure as a fatgraph [21,22], which is a graph in the ordinary sense extended in a particular way to be explained below. Fatgraphs have been used for studying various problems in mathematical physics; here we investigate their use for studying complex molecular structures.

The construction of a fatgraph corresponding to a protein domain is illustrated in Fig. 1. The peptide unit is the basic unit of description in our model disregarding the amino acid residue. In Fig. 1a, the i-th and $(i+1)$-st peptide units of a protein domain are shown. Each peptide unit is a planar region [23] and is represented as a building block with two stubs corresponding to the oxygen and hydrogen atoms (Fig. 1b). The domain backbone is thus depicted as a series of concatenated building blocks. Fig. 1c shows four such building blocks with one hydrogen bond between two peptide units indicated by an edge connecting the H-stub of the first building block with the O-stub of the last. Subsequently, each edge (hydrogen bond and link between building blocks, termed *alpha carbon linkage*) is considered as twisted or not twisted (untwisted) depending on the relative orientations of the peptide units in physical space (Fig. 1d; *Materials and Methods*, section 1). Finally, each edge is widened to become a strip (Fig. 1e), hence the term *fat*graph. The strip is twisted whenever a twisted bond is encountered, similar to how a piece of paper is twisted when forming a Möbius band. In Fig. 1f, the surface is shown without

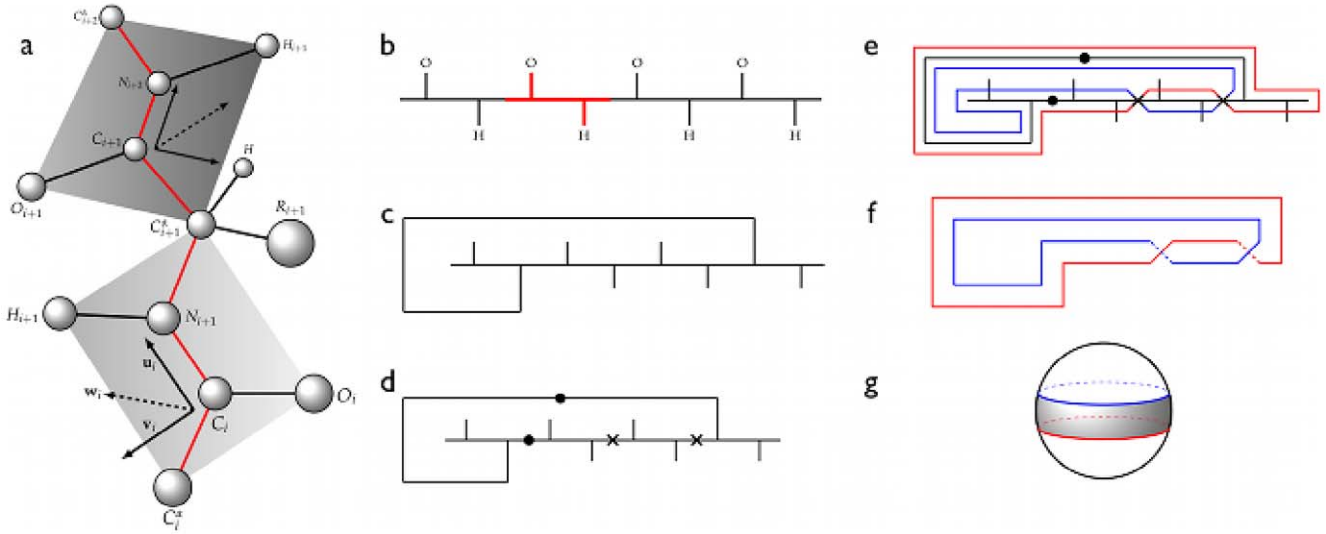

Figure 1. The fatgraph construction. (a) Two neighbouring peptide units of a protein domain labelled as the i-th and $(i+1)$-st unit and the vector triple $(\mathbf{u}_i,\mathbf{v}_i,\mathbf{w}_i)$ defining the associated coordinate system of unit i. The vector $\mathbf{u}_i$ follows the direction of the bond C_i–N_{i+1}, $\mathbf{v}_i$ is perpindicular to $\mathbf{u}_i$ and points towards the same side as H_{i+1}, and $\mathbf{w}_i$ is constructed to form a right-handed coordinate system (b) Concatenated building blocks (one block shown in red), representing a backbone of four peptide units. Vertical stubs correspond to O_i and H_{i+1}. (c) Same as (b) but with one hydrogen bond attached. (d) Hydrogen bonds and alpha carbon linkages are labelled according to the relative orientations of the coordinate frames in (a) such that little change in the orientation results in an untwisted edge (*Materials and Methods*, section 1); • = untwisted and x = twisted. (e) The fattening of the graph is illustrated by colored lines depicting the margins (boundaries) of the strip. The strip is twisted whenever a twisted bond is encountered. (f) The band in (e) with the underlying graph removed. (g) The two adjacent twists cancel out, resulting in band similar to a sphere with two discs removed. It has two boundary components (blue and red). In (c) the hydrogen bond may also be drawn around the right end of the backbone or even cross over or under. The fatgraph is not sensitive to how the hydrogen bonds are drawn and all possibilities yield the same surface.
doi:10.1371/journal.pone.0019670.g001

the underlying fatgraph. The two twists in Fig. 1f cancel each other, and the resulting strip is equivalent, *homeomorphic* in the parlance of topology, to a sphere with the regions around the North and South poles removed (Fig. 1g and *Materials and Methods*, section 2). The concept of a homeomorphism is exemplified in Fig. 2, where a cube is continuously transformed into a sphere. This also illustrates why geometric topology is often referred to as *rubber-sheet geometry*.

In general, the surface corresponding to a fatgraph is homeomorphic to a particular 2-dimensional surface, implying that a fatgraph can be studied by algebraic topological methods and categorized using concepts going back to the work of Leonhard Euler in the 18th century [21,22,24]. For example, when allowing self-intersections during deformation as well as insertions and deletions of full twists, the resulting class of surfaces is uniquely determined by its genus g^*, number of boundary components r, and whether it is *orientable* or not [24]. A surface is called orientable if it is possible to define a consistent orientation (e.g. defined by the right-hand rule) on the entire surface. The strip in Fig. 1 is orientable, whereas the Möbius band is not: One may

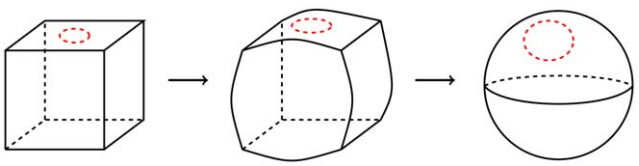

Figure 2. Inflation of a cube with one boundary component constitutes an example of a homeomorphism. The inflation happens without breaking the surface, and only bending and stretching are used.
doi:10.1371/journal.pone.0019670.g002

start a walk from any point and come back again upside-down. The variables g^* and r are examples of *topological invariants* which are quantities that do not change when the surface is bent or stretched (i.e. changed under homeomorphic transformations). The Euler characteristic, defined for any surface as $\chi=2-2g^*-r$, is another invariant summarizing the overall shape of a surface in a single number. In the special case of surfaces arising from fatgraphs of proteins, the Euler characteristic can be computed directly from the fatgraph. In fact, one may show that $\chi=1-b$, where b is the number of hydrogen bonds [21]. As demonstrated in Fig. 1e, the number of boundary components is easily counted, whereas the modified genus is less transparent. However, by using the two alternative descriptions of the Euler characteristic, we may express g^* in terms of simpler quantities, $g^*=(b-r+1)/2$. Thus, χ has direct biologial interpretation (in terms of number of hydrogen bonds) whereas g^* and r are quantities derived through the fatgraph abstraction.

The surface in Fig. 1 is a sphere ($g^*=0$) with two boundary components ($r=2$), one for each of the removed discs (that is the North and South poles), and the structure has only one hydrogen bond. In particular, the alternative expression $\chi=1-b=0$ agrees with $\chi=2-2g^*-r=0$. The fatgraph abstraction thus opens an entirely new perspective on protein structure by replacing complex structures by much simpler constructs.

An example of a surface corresponding to a domain fatgraph is shown in Fig. 3. The CATH protein domain 1ptoF00 is a mixed alpha-beta domain classified as *OB fold* and has $g^*=3$ and $r=48$. The corresponding surface is non-orientable and thus has only one side (up and down are the same), just like a Möbius band. Furthermore, the surface has 48 discs cut out, each giving rise to a boundary component.

We demonstrate that the *global* variables g^* and r capture structural differences in domains. We show this by example and

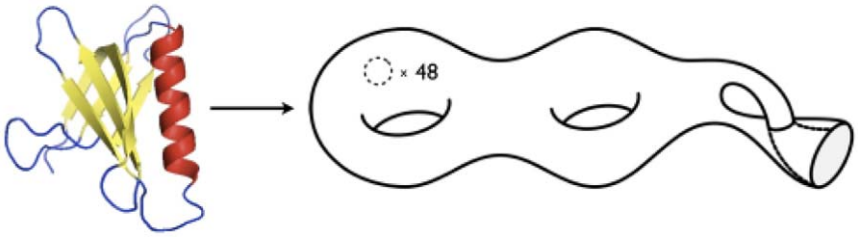

Figure 3. The protein domain 1ptoF00 is an alpha-beta domain classified as *OB fold* (Dihydrolipoamide Acetyltransferase, E2P; 2.40.50). The corresponding surface is not orientable and has genus 3 and 48 boundary components. The surface is homeomorphic to two tori connected to a Klein bottle with 48 discs removed. The surface is difficult to visualize in 3D; in the figure the handle crosses through the bottle with no physical contact.
doi:10.1371/journal.pone.0019670.g003

also by analysis of the distributions of g^* and r values over domains. Further, we illustrate how g^*, r and other fatgraph variables can be utilized for classification. In addition to the global variables, the fatgraph abstraction allows us to introduce a simple local or secondary structure annotation, namely the backbone as a sequence of twisted and untwisted edges. Using machine learning techniques, the usefulness of the fatgraph abstraction is illustrated by classification of domains in the CATH database. We compare to an alternative geometric approach [11] and to an approach based on sequence information only. We show that combining information from all three makes a very strong classifier. Further, we investigate the causes of false predictions and show that our methods are able to detect domains in v3.3.0 with classifications that are non-existing in the previous version, v3.2.0. The classification scheme devises a method for flagging domains as possible new or problematic folds.

Results

Robust variables

For each domain we compute the corresponding fatgraph and calculate four robust variables derived from it (*Materials and Methods*, section 3 and Fig. 4). Robust variables are defined such that they are relatively insensitive to noisy and imprecise experimental data; that is, noise in data that may result in errors in the fatgraph.

We represent a domain by four robust variables, among these the genus g^* and the number of boundary components r of the corresponding surface. These variables are global in the sense that they cannot be related to any particular region of the domain. Furthermore, we consider the domain length L, measured as the number of amino acid residues, and the number of twisted alpha carbon linkages F that measures how often the backbone twists in the orientation of the planar peptide units (Fig. 1d). Recall that insertions and deletions of full twists are not captured in g^* and r, but this is compensated for in F (Fig. 1f).

The CATH database classifies domains in a hierarchical scheme with four main levels (listed from the top and down) called class (C), architecture (A), topology (T), and homologous superfamily (H), hence the name CATH [4,25]. At the C-level domains are grouped according to their secondary structure content into four categories with the three main ones being *mainly alpha*, *mainly beta*, and *mixed alpha-beta*. The last category contains domains with only very few secondary structures. The A-level groups domains according to the general orientations of their secondary structures, and at the T-level the connectivity (the order) of the secondary structures is taken into account. The grouping of domains at the H-level is based on a combination of both sequence similarity and a measure of structural similarity. Below the four main lavels, CATH has an additional five layers called S, O, L, I, and D. The first four group domains according to increasing sequence overlap and similarity, and the D-level assigns a unique identifier to every domain thus ensuring that no two domains have the exact same CATHSOLID classification.

Fig. 5 shows an example of how g^* and r separate domains at different CATHSOLID levels. It transpires that the best separation is obtained at T-, H-, and S-levels. The grouping at the A-level is often very broad, and an architecture may comprise domains of very different sizes. Furthermore, since the order of the secondary structure elements is not taken into account at the A-level, a single architecture may contain domains with very different connectivities [4,25]. This is likely the explanation for the lack of separation of A-levels observed in Fig. 5. On the other hand, because the fatgraph approach is based on structural

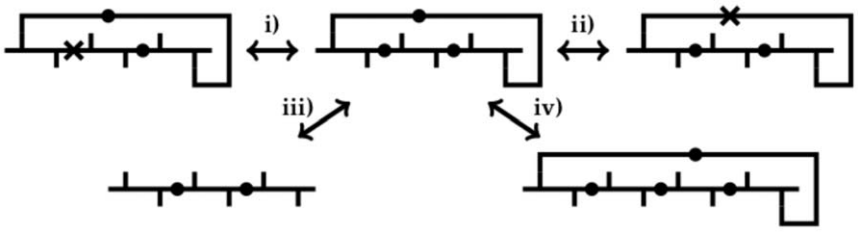

Figure 4. The four basic modifications of fatgraphs. Variables defined on a fatgraph that change at most linearly with the number of basic modifications are called robust. i) Change the color of a bond between peptide units, ii) Change the color of a hydrogen bond, iii) Add or remove an untwisted hydrogen bond, and iv) Replace a fatgraph building block 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. Any fatgraph can be created from an arbitrary starting fatgraph by repeated application of i)–iv).
doi:10.1371/journal.pone.0019670.g004

features, we do not expect to see a clear separation at the SOLID levels, since these are defined in terms of sequence overlap and similarity. Fig. 6 shows that g^* and r separate the H-level families in the CATH topology *Pectate Lyase C-like* (CATH classification 2.160.20) with one family (red in Fig. 6) being larger than the others. To test the empirical robustness of the variables, we generated 25 modified structures for each domain using the CONCOORD algorithm [26] and calculated g^* and r from the resulting structures (Fig. 6 and *Materials and Methods*, section 4). The figure indicates that even after modifications, the variables are able to separate domains at the H-level. Furthermore, for individual domains, the variables did not in general deviate significantly from the original values (illustrated in Fig. S2).

Distribution of fatgraph variables

Fig. 7 shows a scatter plot of g^*, r, and F for the three main classes (C-level) in v.3.3.0. Generally, the mainly alpha domains have lower g^* and higher r than the mainly beta domains with the mixed alpha-beta domains falling in between. For example, mainly alphas have many domains with $g^*=0$ corresponding to a sphere with r discs cut out. For small values of g^* and r, almost all combinations are found, but for higher values, only a small fraction of all possible combinations are observed. More details are shown in Figs. S3–S4, with pairwise scatterplots of the variables g^*, r, F, L, and the Euler characteristic $\chi=2-2g^*-r$. Empirically, fairly sharp boundaries appear for possible values, and longer domains tend to have higher values of both g^*, r, and F than shorter domains. A total of 127,491 domains are non-orientable, and only 1,141 domains are orientable. We expect this because a single twisted hydrogen bond may introduce a Möbius band and potentially alter the orientability of the corresponding surface: For example, in Fig. 1 the two adjacent twists cancel out, and the surface becomes orientable, but removing one of the existing twist or adding an extra twist along the backbone results in a Möbius band. Similarly, moving the right-most end of the hydrogen bond one stub to the left, thus separating the two twists, yields two Möbius bands which do not cancel out.

Structural divergence may be caused by only modest modifications at the amino acid sequence level, and we compared how differences in sequences are reflected in the topological invariants. Fig. 8 shows scatter plots of normalized alignment scores (*Materials and Methods*, section 7) versus normalized differences in g^* and r, respectively, for all pairs of S95-domains in the topology Pectate Lyase C-like (2.60.120). In general, low sequence similarity implies relatively large differences in g^* and r with only a few outliers. For example, three domains have sequences very similar to that of 2iq7A00 (alignment score >0.6), but still the normalized differences in g^* (resp. r) are almost 0.5 (resp. 0.3). This may be explained by a lower number of hydrogen bonds in 2iq7A00 compared with the three other domains – a feature captured by the topological invariants but not by sequences alone.

To further assess the ability of the four fatgraph variables to distinguish different folds, we performed pairwise Wilcoxon tests comparing the distributions of each variable using the 1,161 H-level families in v3.3.0 containing ten or more domains (in total 124,372 domains or 96.7% of all domains). The results are summarized in Fig. S5, and the plot indicates that in general the four variables are sufficient to distinguish most H-levels. In fact, at significance level $\alpha=10^{-3}$, almost all pairs of H-levels (91.4%) are distinguishable by at least three of the four variables.

Secondary structure elements

The secondary structure is a particularly rigid part of a protein structure, and this is reflected in the corresponding fatgraph. Fig. 9 depicts idealized fatgraphs arising from the three most common secondary structure motifs: (a) alpha helices, (b) parallel beta

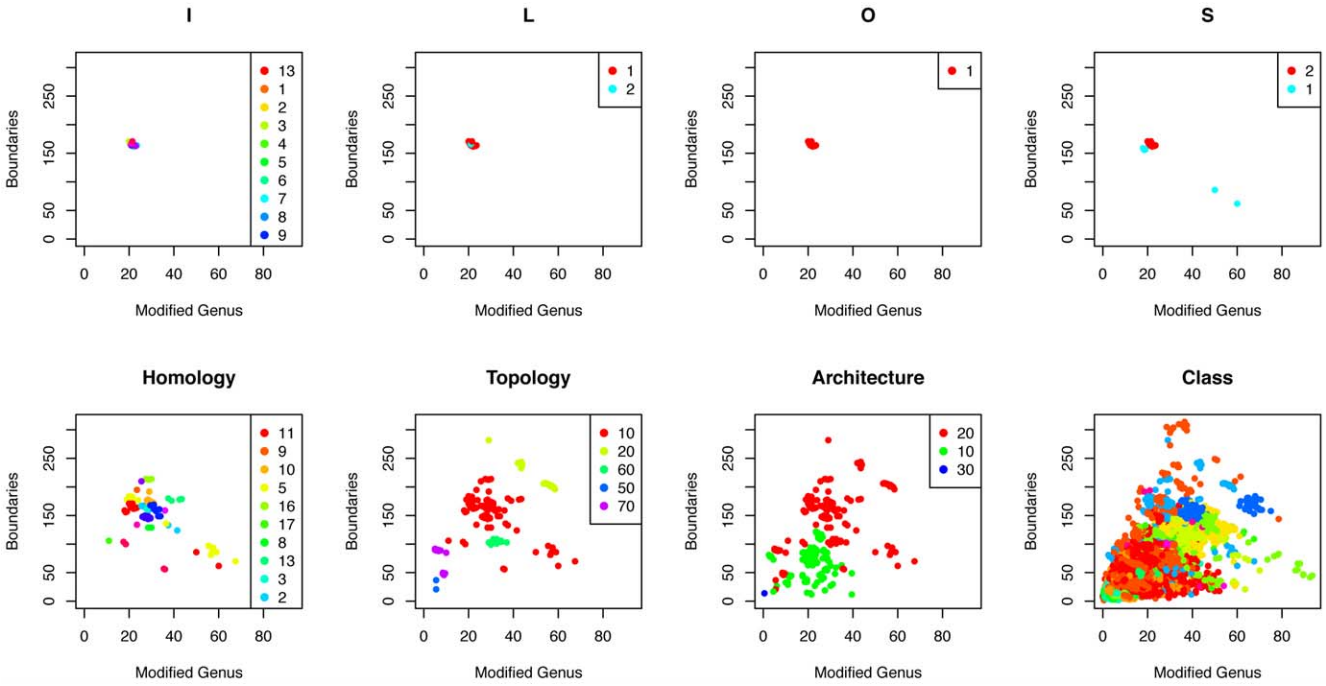

Figure 5. The domain 1o88A00 is classified as Pectate Lyase C-like (2.160.20) with complete CATHSOLID classification 2.160.20.10.11.2.1.1.1. The Class plot shows (g^*,r) for all domains with C=2 (colored according to A-level), and the Architecture plot shows (g^*,r) for all domains with (C,A)=(2.160) (colored according to the three T-levels). This continues all the ways down to the last plot where (g^*,r) are shown for for (C,A,T,H,S,O,L,I)=(2.160.20.10.11.2.1.1).
doi:10.1371/journal.pone.0019670.g005

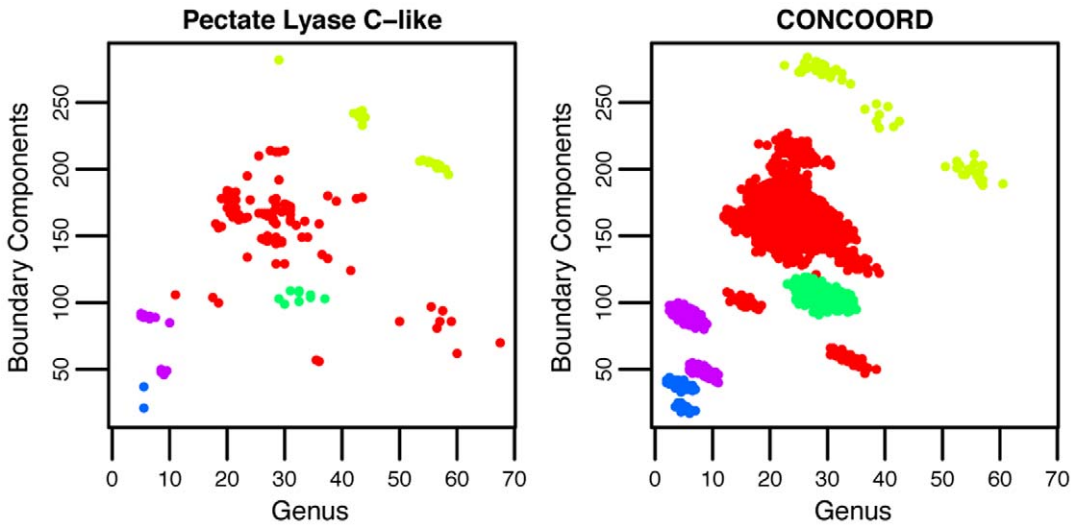

Figure 6. Robustness of topological invariants to noise. Left: Scatter plots of g^* and r for domains in CATH topology *Pectate Lyase C-like* (2.160.20). The five H-levels are indicated by different colors (3 to 102 domains in size), and a clear separation of H-levels is observed. Right: g^* and r calculated from CONCOORD modified structures. Even with noise, separation at H-levels is still clearly visible. Note that in the bottom right corner in the left figure there are eight red dots without counterparts in the right figure. The number of hydrogen bonds typically increases when a domain is modified using CONCOORD, but the eight domains corresponding to the missing dots each showed a decrease in the number of hydrogen bonds (Fig S1).
doi:10.1371/journal.pone.0019670.g006

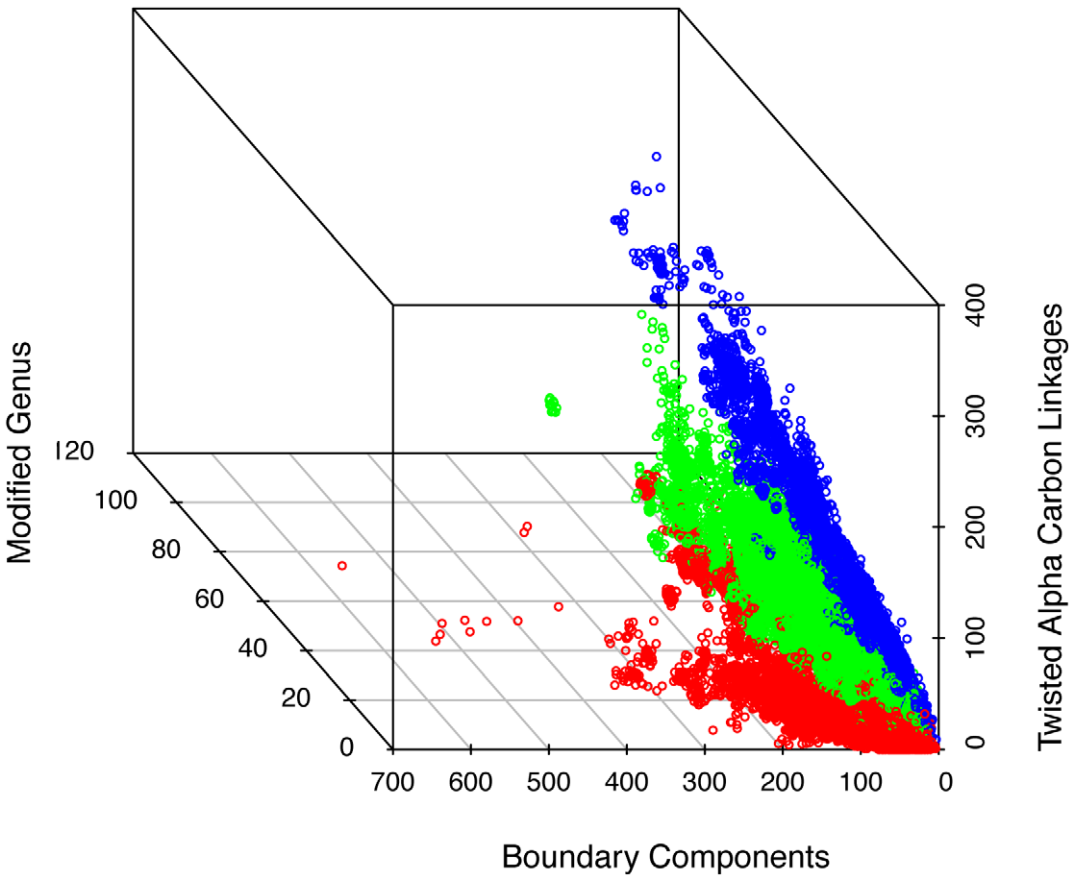

Figure 7. Scatter plot of (g^*,r,F) for all domains in the three main classes mainly alpha (red), mainly beta (blue), and mixed alpha-beta (green). Visually, the mainly alpha and mainly beta domains are separated with the mixed alpha-beta domains residing in between.
doi:10.1371/journal.pone.0019670.g007

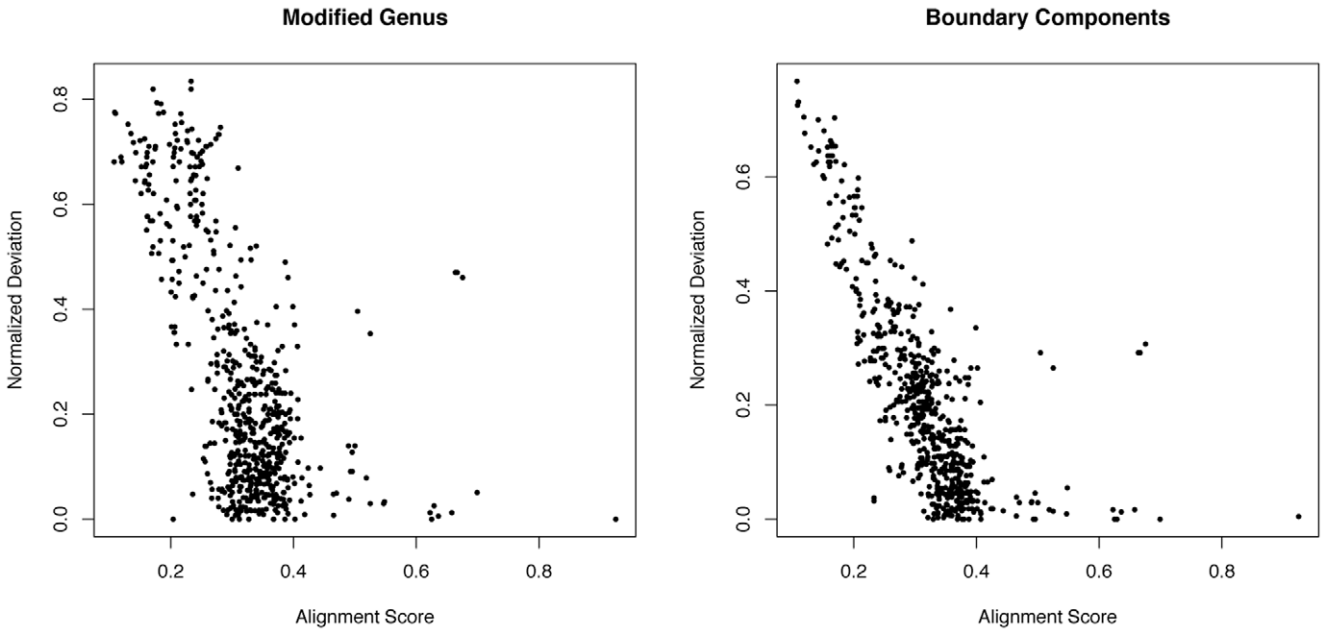

Figure 8. Alignment scores versus differences in g^* and r for all pairs of S95-domains in the Pectate Lyase C-like topology (2.160.20). We use the normalized difference $|(g_i^*-g_j^*)/(g_i^*+g_j^*)|$ between modified genera (and similarly for boundary components) to take discrepancies in domain length into account. A high alignment score indicates high sequence similarity and the plot illustrates that similarity is at the primary and tertiary levels are correlated.
doi:10.1371/journal.pone.0019670.g008

sheets, and (c) anti-parallel beta sheets with typical boundary components indicated by dashed red lines. All boundary components pass through exactly four different peptide units, and the backbone of an alpha helix consists of untwisted edges, whereas the backbone of a beta sheet consists of twisted edges.

A domain consisting of one long alpha helix has genus zero, and the number of boundary components is proportional to the length of the domain. Likewise, a beta sheet contributes to the number of boundary components proportionally to the sheet size and only marginally to the genus. The apparently abstract topological quantities thus exhibit direct relationships to the secondary structures of the domains (Fig. S6). This observation agrees with the empirical result above that the main CATH classes show differences in the distribution of g^* and r.

Non-additivity of fatgraph variables

The topological quantities corresponding to an entire domain cannot be obtained directly by adding quantities from individual secondary structure components alone; most domains have stabilizing hydrogen bonds between secondary structure elements, and these contribute in a non-linear fashion to the fatgraph. This non-additivity is e.g. reflected in the mainly alpha class (Fig. S6), where the genus increases with increasing number of alpha helices (despite each has $g^*=0$) because the helices are stabilized by bonds between them.

The lack of additivity is perhaps even clearer when considering entire proteins comprising multiple domains. As a concrete example, consider the protein 1DAR (Fig. 10) with five CATH domains. Considered as one contiguous structure, 1DAR has $g^*=90.5$ and $r=181$, but the genera and boundary components corresponding to the individual domains add up to 52 and 247, respectively. Examples where the genus (resp. the number of boundary components) of the entire protein is smaller (resp. larger) than the sum of those corresponding to the individual domains also exist.

Despite the relatively high deviation of g^* and r from the sums obtained from the individual constituents of a structure, the Euler characteristic $\chi=2-2g^*-r=1-b$ is generally more consistent. In the example 1DAR, the sum of the Euler characteristics is -349 compared to -360 for the entire protein. If there are k domains, then

$$\sum_{i=1}^{k}\chi_i=k-\sum_{i=1}^{k}b_i,$$

where index i refers to the ith domain, $i=1,\ldots,k$. That is, the difference between χ of the entire protein and the sum is $(k-1)+B_k$, where B_k denotes the number of bonds *between* domains. Since, in general there are fewer hydrogen bonds between domains than within, the sum is close to χ of the whole protein.

Classification using robust variables

We attempted to reproduce the CATH classification using only the four robust variables, g^*, r, F, and L. We applied different classification techniques to the data and found that the method Random Forests [27] generally performed well.

In v3.2.0 (SAll, see *Materials and Methods*, sections 5), we selected the 500 largest H-levels (86% of all domains) and randomly sampled $2/3$ of the domains for training, while keeping $1/3$ for testing (*Materials and Methods*, sections 5 and 6). In addition, we tested the classifier on the new domains in v3.3.0 that are not already in v3.2.0. Fig. 11 shows the results. We assigned 74.9% of the domains in v3.2.0 into their correct H-level, whereas 78.4%, 84.6% and 96.1% are correctly assigned at the T-, A-, and C-level, respectively. For the new domains in v3.3.0, the percentages are smaller: 55.3% (H), 63.0% (T), 74.1% (A), and 92.8% (C). When the classifier makes a correct prediction, it does so with high confidence whereas

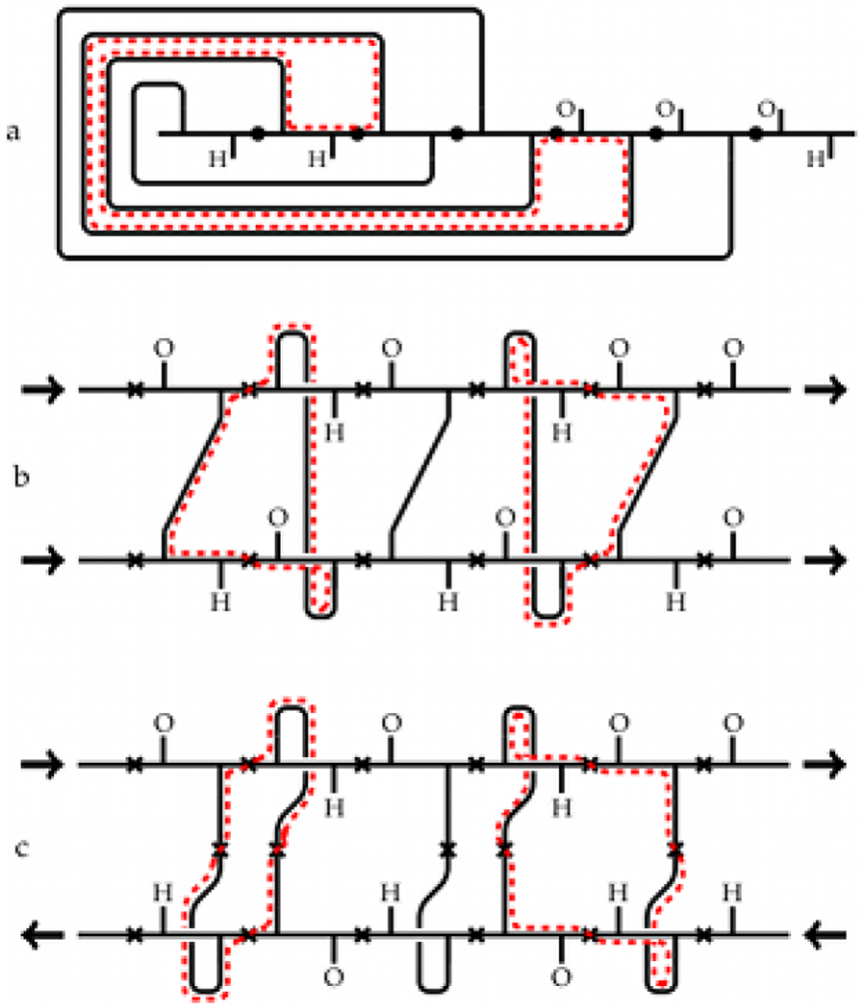

Figure 9. Fatgraphs corresponding to the common secondary structures. (a) alpha helix, (b) parallel beta sheet, and (c) anti-parallel beta sheet. Each boundary component (dashed red line) passes through four peptide units; the alpha helix is local in that it connects only closely situated peptide units whereas the beta sheet also connects peptide units potentially far away from each other. The number of components depends on the length of the structure. The less frequently occurring 3_{10}-helices and π-helices give rise to similar pictures as (a), the only difference is that hydrogen bonds connect stubs three and five peptide units apart instead of four. The backbone of an α-helix is a string of untwisted alpha carbon linkages, whereas for the β-sheets these are twisted.
doi:10.1371/journal.pone.0019670.g009

it is less certain when making a false prediction, and a similar lack of confidence is observed when the classifier is applied to domains which have not been assigned a classification by CATH or domains with H-levels that are new in v3.3.0 (Fig. 12).

Performances for other selections of variables at the H-level are shown in Fig. S7. We found that the domain length is an important variable for correct prediction (an observation also made in [11]). We also varied the energy cut-off used to infer hydrogen bonds (*Materials and Methods*, section 1). The resulting values of the robust variables are strongly correlated with those calculated using the default cut-off, and the classification results did not change significantly (Fig. S7).

Some of the largest families show a remarkable homogeneity in g^*- and r-values across domains (Fig. S8), which to some extent stands in contrast to reports indicating structural diversity within H-levels based on RMSD measures [28]. Our approach may therefore provide an important complement to existing classification schemes.

The performance is generally lower on the new domains in v3.3.0 than on the domains in v3.2.0. The lower performance could in principle be due to skewness in family sizes in the two data sets, but this is not observed (Fig. S9). To further explore this discrepancy, we used the S-level immediately below the H-level as a proxy for the complexity of the H-levels. Fig. 11 shows that the classifier performs much better on domains with known S-level (i.e., S-levels that are associated with domains in the H-level training set) than on domains with unknown S-level (i.e., S-levels not found for any domain in the training set). Furthermore, 21.8% of the new domains have unknown S-levels while this was only the case for 1.8% of the domains in the v3.2.0 testing set. In the training set, this must be due to sampling whereas in the new set, the difference is mainly caused by genuinely new S-levels introduced in v3.3.0. This finding indicates that the known S-levels and the S-levels new in v3.3.0, despite being defined based on sequence similarity, also differ in their fatgraph characteristics.

Classification using flip sequences

The domain example in Fig. 1d comprises three alpha carbon linkages, one untwisted and two twisted. Reading from left to right, the conformations may be represented as a string UTT with U (T) meaning untwisted (twisted). In this way, each CATH

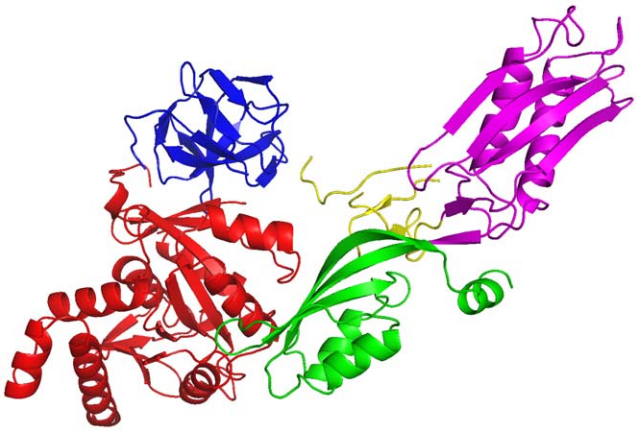

Figure 10. The protein 1DAR comprises five CATH domains with individual genera 5,8,4.5,9, and 25.5, and 43,54,5,39, and 106 boundary components. The entire protein considered as one contiguous structure has $g^*=90.5$ and $r=181$, but the sums of the individual g^* and r are 52 and 181. The robust variables g^* and r are thus not additive. The figure is made using PyMOL (www.pymol.org).
doi:10.1371/journal.pone.0019670.g010

domain has an associated sequence of these letters of length one less than the number of peptide units, and we refer to this sequence as the *flip sequence*. Alpha helices and beta sheets have particularly simple flip sequences, namely, UUU… and TTT…, respectively (Fig. 9).

The alignment score computed from an alignment of two flip sequences gives a measure of similarity between the corresponding domains, and we applied this score to build an alternative CATH classifier using flip sequences (secondary level) rather than robust variables (tertiary level). To do so, we randomly selected 2/3 of all domains in v3.2.0 (S95 or SAll; *Materials and Methods*, section 5) for training, keeping 1/3 for testing. In addition, the domains that are new in v3.3.0 or unclassified in v3.3.0 were also kept for testing.

For all domain pairs $(d,\tilde{d})$ in the training set, we calculated the pairwise, normalized alignment score $S(d,\tilde{d})$ (*Materials and Methods*, section 7). Subsequently, we defined the similarity between a domain d and an H-level h (similarly for C-, A-, and T-levels) as

$$S_d(h)=\max\{S(d,\tilde{d})|\tilde{d}\neq d,\tilde{d}\ has\ H-level\ h\}. \qquad (1)$$

For each d, we identified the two H-levels with the highest scores (Fig. 13). In the figure a green (red) dot indicates that the H-level with the highest score is the same (not the same) as that of d. The relationship between the scores is clearly indicative of the H-level of d. The scores for the two nearest H-levels were combined into one variable (z-score) and a test was designed to facilitate comparison between methods (*Materials and Methods*, section 7). The same procedure was adapted to amino acid sequences (primary structure). Additionally, we compared to a geometric method for classification that uses tertiary structure information [11] in the following way: A domain is represented as a 30-component vector comprising the number of residues and 29 quantities derived from a geometric description of the backbone curve. The domain backbone is viewed as a piecewise linear curve in three dimensions with each piece corresponding to a bond, and the average number of crossings when the curve is viewed from all possible angles is computed. This *average crossing number* is one of the 30 variables used in the classification, and except the number of residues, the remaining 28 variables are all generalizations (known as *Gauss integrals*) of the average crossing number.

Overall the results are comparable (Table 1 for S95); further results for S95 and SAll are shown in Table S1 with amino acid sequences generally performing better than Gauss integrals and flip sequences. All methods show a decrease in sensitivity and specificity for the domains that are new in v3.3.0 (Table S1), which is similar to that observed for the robust variables though less pronounced and attributed to differences between the old and new domains, such as S-levels, indicating that the new domains are evolutionary more diverse than the old ones.

We combined all three tests into one to achieve higher performance (*Materials and Methods*, section 7, and Fig. 14). For S95, the AUC (area under curve) increases to 96% for the combined test from 86% (flips), 90% (Gauss), and 91% (amino acids) for the individual methods. In particular, for a sensitivity of

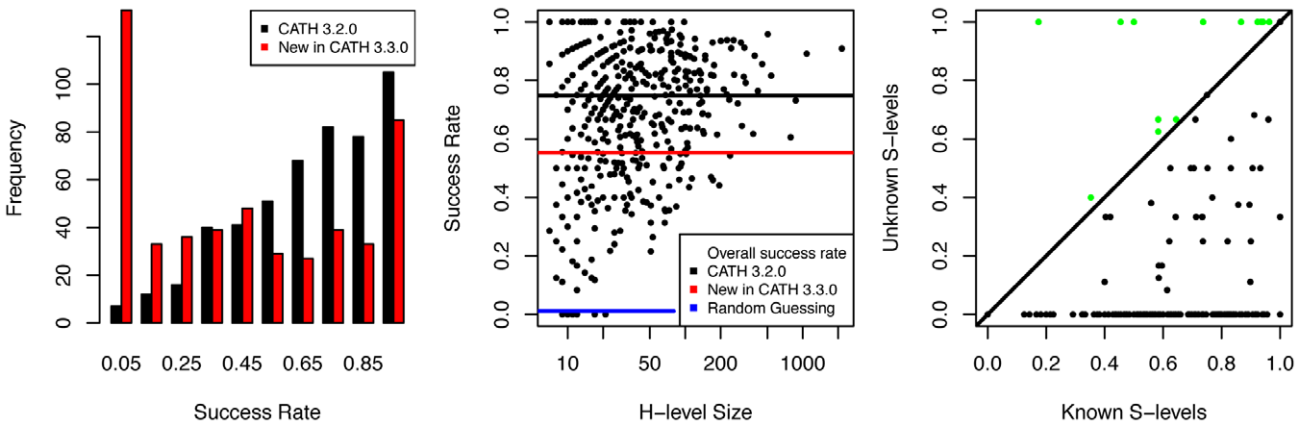

Figure 11. Classification at H-level using g^*, r, F, and L. Left: Distributions of success rates for the 500 largest H-levels in the CATH 3.2.0 test set (black) and the new domains in CATH 3.3.0 (red). Middle: Success rates for the CATH 3.2.0 training set plotted against H-level sizes. Average success rates are indicated by lines. Right: In the CATH 3.2.0 test set, 182 H-levels contain domains with S-levels not present in the training set. In all but 13 cases (green), the classifier performs better on the domains with known S-level.
doi:10.1371/journal.pone.0019670.g011

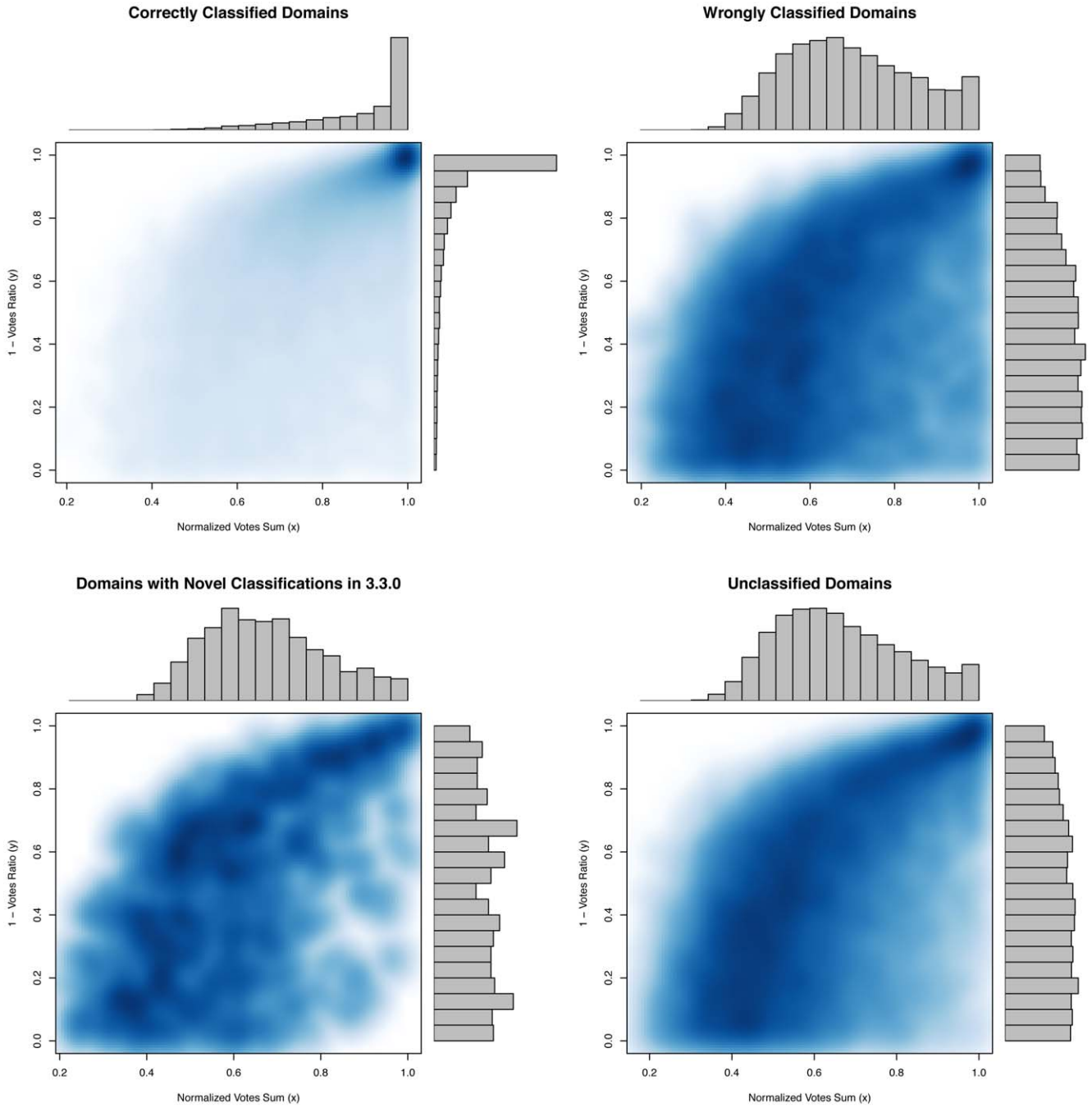

Figure 12. The classifier consists of a collection (forest) of classification trees. For each domain in the test set, each classification tree votes, and a consensus is reached. We used 500 trees, and for each domain we calculated the percentages, v_1 and v_2, of the most frequent and the second most frequent votes occuring. The top plots show the distributions of $x = 1 - v_2 / v_1$ and $y = v_1 + v_2$ for the correctly and wrongly classified domains at the H-level in v3.2.0. If y is large, the majority of votes are cast for the top two candidates, and if furthermore x is large, many more votes have been cast for the winner compared to the runner-up. Therefore, if both x and y are large, this indicates that the classifier makes a confident prediction. The distributions of x and y for correctly predicted domains show that the confidence in correct predictions is generally high. On the other hand, confidence in wrong predictions is much more uniform. The bottom plots show the distributions corresponding to the 908 newly added domains in v3.3.0 with non-existent classification in v3.2.0 and the 17,918 domains for which CATH has not yet provided a classification. Both show the same kind of uncertainty as observed for the wrongly classified domains.
doi:10.1371/journal.pone.0019670.g012

95%, the specificity is at least 50% higher for the combined test than on any of the individual tests individually. Likewise, the combined test is able to identify 78.4% of all domains with unknown H-level in v3.2.0, and 46.6% of all domains that are unclassified in v3.3.0. For SAll, the latter percentages raise to 100% and 92.8%, respectively, making the classifier very capable of detecting domains with structures potentially not in CATH v3.3.0.

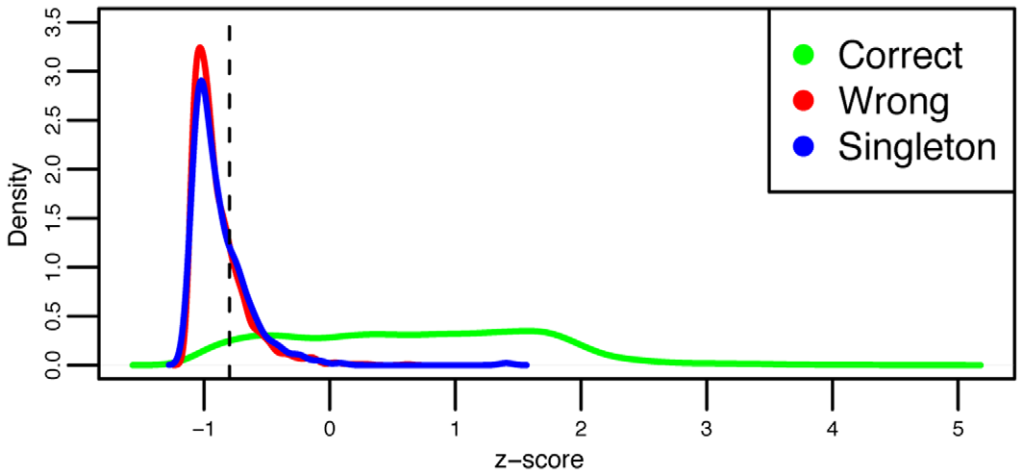

Figure 13. The plot shows the z-score l for each domain in the CATH 3.2.0 training set (S95) for H-level, based on flip sequences. The vertical dashed black line indicates the 95% sensitivity decision threshold; if the z-score of a domain is above the line it is assigned the classification of the nearest H-level, otherwise it is left unassigned. H-levels with only one domain in the test set cannot be classified in this way. Green = Closest H-level is correct, Red = Closest H-level is wrong, Blue = Unclassifiable.
doi:10.1371/journal.pone.0019670.g013

Discussion

We discuss a new representation of protein structure and show how local and global variables vary across domains and structural classification. The description makes use of concepts from geometric topology and represents a domain structure by a fatgraph that in turn can be interpreted as a 2-dimensional surface. The structure of the fatgraph relies on atomic coordinates of the protein and its hydrogen bonds but not on primary sequence information.

The representation provides a complementary and alternative view to structures as purely 3D geometric objects [7,8,11–14] and shows several strengths. The domain structure is conceptualized in terms of entities that are amenable to further manipulation and characterization; for example, we compute the genus and the number of boundary components that are both topological invariants. Even though several unrelated domains may share the same invariants, we are able to classify the majority of the domains correctly. Further, we have introduced the idea of a robust variable, namely, a variable defined on the fatgraph that is robust towards noise and errors in the experimental determination of the structure, and we have formulated robustness in terms of operations on the fatgraph and investigated the robustness empirically.

The invariants g^* and r of a domain cannot in general be computed from the secondary structures alone. Due to stabilizing bonds connecting secondary structure elements, g^* and r may differ significantly from what is obtained by summing the genera and the number of boundary components of individual secondary structure elements. The invariants thus capture tertiary structure information.

We showed that using secondary structure information we could classify a large percentage of domains in S95 (and SAll) correctly. However, correct (and better) classification could also be achieved using other methods using primary or tertiary structure information. We combined all three methods and achieved a method with higher performance as well as sensitivity and specificity. The combined method is also able to identify the majority of unknown or unclassified domains.

We used the CATH database as a gold standard, but using appropriate clustering algorithms it would be possible to make a *de novo* classification and compare this to existing classification. Given the observation that classification based on primary structures is best at reproducing the CATH database, it is conceivable that a *de novo* classification based on structural properties alone would lead to a different hierarchy. However, a *de novo* classification would require further investigations to e.g. determine the number of classes needed, and this is beyond the scope of this paper. The lack of agreement between the two most widely used databases, CATH and SCOP, certainly indicates that more analyses are needed in order to fully comprehend the universe of domain structures [29].

Table 1. Comparison of methods.

Method	C	A	T	H
Gauss Integrals	95.3/95.3/42.0/NA	89.0/95.5/36.1/NA	85.2/95.0/43.5/47.1	80.8/95.7/30.9/34.3
Flip Sequences	94.0/95.3/33.0/NA	82.4/95.1/15.7/NA	75.4/94.6/21.0/22.6	73.0/94.9/19.0/24.3
Amino Acid Sequences	86.2/95.4/17.4/NA	79.8/94.9/22.1/NA	79.2/95.3/28.9/31.3	80.6/95.1/38.8/35.0
Combined	95.8/95.2/69.6/NA	90.3/95.3/62.7/NA	87.9/95.4/69.6/67.8	87.2/95.4/70.4/70.7

Each cell contains performance, sensitivity, specificity, and the number of *unknown* domains flagged correctly (in percent) in that order on the CATH v3.2.0 test set at each CATH level for all three methods. Decision thresholds were calibrated on the CATH 3.2.0 training set to achieve 95% sensitivity. Levels comprising only a single domain in CATH 3.2.0 are never included in the training set and thus account for the numbers in the unknown column; for C- and A-levels, there are none of these.
doi:10.1371/journal.pone.0019670.t001

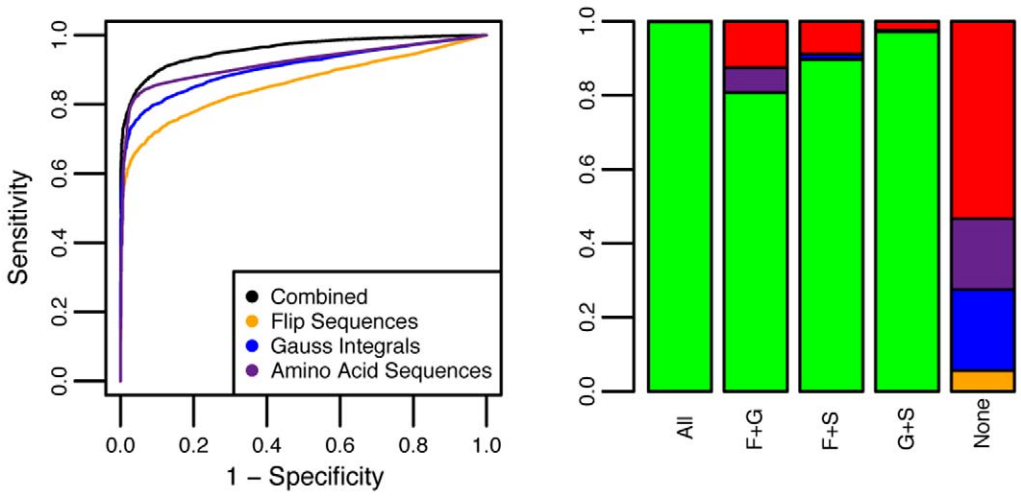

Figure 14. Analysis of sensitivity and specificity. Left: ROC curves for S95 for the three tests individually and the combined test. Combining information from all three clearly improves sensitivity and specificity considerably. Right: The columns show how often agreeing methods (S = amino acids, F = flips, G = Gauss) are correct (green), and how often the non-agreeing method is correct; S (purple), G (blue), F (orange). Red indicates the number of times none of them are correct. The numbers in each column are 6057 (F+G+S), 476 (F+G), 348 (F+S), 697 (G+S), and 2084 (None).
doi:10.1371/journal.pone.0019670.g014

Automated approaches, including ours, benefit from not relying on human judgment. However, a key difficulty is that structural and evolutionary homology does not always go hand in hand, and different protein families show a wide spectrum in sequence, structural and functional similarities [3,7,30]. Currently, the level of classification achieved by manually assisted methods, such as CATH or SCOP, might not be realizable by automated means only, but we believe that advances in mathematical modeling of protein structure will change this in the future and that our model should be a step in this direction. For example, it would be interesting to put our method into a probabilistic setting. In recent papers, it has been shown that probabilistic models might have large potential [31,32].

Various extension of the fatgraph model are conceivable. For example, one approach could be to include bifurcating hydrogen bonds in the model, as CO and NH groups of the protein backbone engage in two or more hydrogen bonds [23]. However, this phenomenon poses a mathematical question that must be addressed in a biologically meaningful way: The fatgraph model depends on an ordering of the edges around a vertex, and with bifurcating bonds there is no *a priori* way of choosing such orientations. Other extensions could be the inclusion of sulphur bridges, or extending the two types of edges (backbone and hydrogen bonds) to multiple types reflecting more accurately the twisting of the backbone.

Materials and Methods

1. Fatgraph construction

The fatgraph is constructed as explained in the main text, Fig. 1. To each peptide unit we associate a coordinate system (a frame). If the frames of two linked peptide units are similar, the edge connecting them are untwisted and otherwise they are twisted. 'Similar' is here measured via a metric on frames, see [21] for details. The procedure is identical to twisting if the sum of scalar products $s_{ij}=\mathbf{v}_i\cdot\mathbf{v}_j+\mathbf{w}_i\cdot\mathbf{w}_j$ is negative [21], a sum combining the change in the normal vector $\mathbf{w}$ with the change in the orientation of the peptide plane. For example, if the two frames are identical $s_{ij}=2$ and the edge is not twisted; whereas if frame j is frame i up-side-down, $s_{ij}=-2$ and the edge is twisted.

We infer hydrogen bonds using the DSSP algorithm [33]. The algorithm depends on an energy cut-off, and by default a hydrogen bond is inferred if the electrostatic interaction energy E is lower than -0.5 kcal/mol.

2. Classification of surfaces and fatgraphs

For integers $g,r\geq 0$, the following families of surfaces are particularly interesting: 1) a sphere with r discs cut out, 2) the sum of g tori with r discs cut out, and 3) the sum of g real projective planes with r discs cut out. The first two types can be visualized in 3D, whereas the last cannot; 1) is straightforward, and 2) is a series of g doughnuts glued together with r discs removed. Two surfaces are called homeomorphic if one can be transformed into the other by stretching and bending but no tearing. A classical result in algebraic topology [24] states that any closed connected surface is homeomorphic to exactly one of the surfaces above. For example any deformation of a ballon is homeomorphic to a sphere and thus has $g=0$. The number g is called the genus, and r is the number of discs or boundary components of the surface. The surfaces in 1) and 2) are orientable, whereas the surfaces in 3) are not. It follows that a surface is uniquely determined by its g, r, and whether it is orientable. We define the modified genus g^* as g if the surface is orientable and as $g/2$ if it is non-orientable [21]. With this definition, the Euler characteristic (a term combining g and r) is $\chi=2-2g^*-r$ in either case. The number of hydrogen bonds b relates to χ through $\chi=1-b$.

The invariants, including whether the fatgraph is orientable, can be calculated computationally efficiently and quickly [21]. For example, b can readily be found from the fatgraph, whereas r requires more work. For small fatgraphs as the one in Fig. 1, r is easily counted, but a more systematic approach must be applied when dealing with larger fatgraphs. This may be accomplished by a purely algebraic approach which is easily implemented in a computer program [21]. Even using a naive and straight-forward implementation, parsing a domain and calculating the corresponding topological invariants takes less than a second on a standard laptop computer. Finally, g^* follows from $\chi=2-2g^*-r=1-b$. Note that by construction (Fig. 1e), a fatgraph corresponding to a

protein always has a least one boundary component, that is $r \geq 1$. A general surface has $r \geq 0$ and $g^* \geq 0$.

3. Robust variables

A function $v(G)$ defined on fatgraphs is called κ-robust for an integer κ, if $|v(G)-v(G_q)| \leq q\kappa$ whenever $G=G_0 \rightarrow G_1 \rightarrow \cdots \rightarrow G_q$ is a sequence of $q \geq 0$ fatgraphs, and G_{i+1} is obtained from G_i by one of four basic modifications: i) change the label of a bond between peptide units (i.e., remove a twist or introduce a twist); ii) change the label of a hydrogen bond; iii) add or remove an untwisted hydrogen bond; and iv) insert a fatgraph building block B_1 next to an existing building block B_0 (in the direction from N to C termini), such that the alpha carbon linkage between the two blocks is untwisted, the alpha carbon linkage to the right of B_1 is (un)twisted according to whether the linkage to the right of B_0 in the original fatgraph is (un)twisted, and B_1 has no hydrogen bonds attached. Or the reverse operation, i.e. removing a block with no hydrogen bonds which has an untwisted linkage to its left neighbour (illustrated in Fig. 4). It can be shown [21] that the functions g^*, r, F, and L all are 1-robust. This implies that if G is changed by a sequence of at most q basic modifications, the functions g^*, r, F and L change at most q.

4. CONCOORD modified structures

All modified structures were generated using the CONCOORD algorithm [26] which generates conformations from a known structure based on restrictions on interatomic distances. We used default parameter values.

5. CATH domain data

The newest version of CATH (3.3.0) contains 128,688 domains whereas the previous version (3.2.0) only comprises 114,215 domains. We obtained the classifications of domains in both versions as well as the raw data consisting of chopped PDB-files and corresponding DSSP-files from the CATH homepage (www.cathdb.info). CATH is hierarchical; we focus on the class (C), architecture (A), topology (T) and homologous superfamily (H), and sequence (S) level.

We operate with the following basic data sets: A) CATH 3.2.0, B) the domains in v3.3.0, but not in v3.2.0 (called new in v3.3.0), and C) CATH 3.3.0. In addition, CATH lists 17,918 domains that have not yet received CATH classification, either because the domain annotation is questionable or because there is uncertainty about the structural similarity to other domains. We denote this set unclassified in v3.3.0. Among the new in v3.3.0, there are 908 domains with novel H-level, that is, levels that are not already in v3.2.0, but added to v3.3.0. Using CATH terminology, the full data sets are denoted by SAll, whereas data sets *only* including sequences with less than 95% similarity are denoted by S95 (this does not apply to the unclassified set).

There are 2,386 H-level families in v3.3.0, and 322 of these are singletons. Almost half of the families contain less than ten domains whereas a handful of families contain more than 1,000 domains (the largest contains 7,674 domains). Moreover, the distributions of family sizes are highly skewed and resemble power-laws (Fig. S10).

6. Classification using robust variables

The algorithm Random Forests [27] is used for classification. It is a probabilistic approach, that is, rerunning the algorithm might produce a (slightly) different result. A random forest builds $NTREE = 500$ (user specified) classification trees based on a training set (where we always take the training set to be $2/3$ of the full data set) and uses majority voting for predicting the classes of the testing set.

Classification by random guessing is done by assigning family levels to domains according to family sizes, that is, the average success rate is the sum over $(f_i)^2$, where f_i is the frequency of family i.

7. Classification using flip sequences, amino acid sequences and Gauss integrals

For flip sequences, alignments were made using the Smith-Waterman algorithm with mismatch and gap penalties set to -1 and match score to 2. For amino acid sequences, BLOSUM40 was used. Let $S_0(s_1,s_2)$ denote the score of the alignment of sequences s_1 and s_2. To facilitate a pairwise comparison of all sequences in CATH, regardless of lengths, we use the normalized score given by $S(s_1,s_2)=S_0(s_1,s_2)/\max(S_0(s_1,s_2),S_0(s_2,s_2))$, such that all scores are between 0 and 1.

For a given a domain, denote by z_1 and z_2 the normalized alignment scores of the domain to the nearest and second nearest H-levels in the v3.3.0 training set. A univariate measure was used to compare methods, $z=(\delta-\mathrm{mean}(\delta))/\mathrm{sd}(\delta)$, where $\delta=z_1-z_2$ and the mean and sd are over all domains in the training set. For Gauss integrals, we used $z=(r-\mathrm{mean}(r))/\mathrm{sd}(r)$ with $r=\log(d_2/d_1)$, and d_1 (d_2) the distance to the (second) nearest H-levels [11]. Fig. S11 shows the distributions of the z-values for the three methods. Domains were classified according to their nearest H-level. A decision threshold was calibrated to achieve 95% sensitivity; all domains with z above the threshold are flagged as correctly identified if the nearest H-level corresponds to the true level. All domains with z below the threshold are likewise correctly identifed as problematic if the nearest H-level is not the true level. Same procedure for C, A, and T-levels.

The three methods were combined into one method. Domains were classified according to the majority rule. If none of the methods agree, the method with the highest z-score decides the classification. A combined z-score was calculated as the maximum z-score of the agreeing methods. If all disagree, then the maximum z-score is used. Fig. S10 shows the distribution of the combined z-score.

Supporting Information

Figure S1 Each domain in the Pectate Lyase C-like topology was subjected to 25 **independent modifications using the CONCOORD algorithm.** In the figure each column is a domain and the distribution of the normalized number of hydrogen in the modified structures is shown. The number is normalized relatively to the number observed in the original (unmodified) domain. The values corresponding to the eight outliers in Fig. 6 are highlighted in red, and all show a conspicuous decrease in the number of hydrogen bonds in the modified structures compared to the general trend of the remaining domains.
(TIF)

Figure S2 The deviation of g^* and r from the observed values for eight randomly selected domains subjected to 1,000 **modifications using the CONCOORD algorithm.** In general the modified values are centered around the observed values, though in some cases the distribution is biased to the left or right.
(TIF)

Figure S3 Distributions of the three quantities genus (g^*), number of boundary components (r), and number of twisted alpha carbon linkages (F) for all domains in

v3.3.0. Mainly beta and mixed alpha-beta have very similar distributions of g^* whereas mainly beta and mixed alpha-beta have very similar distributions of r.
(TIF)

Figure S4 Pairwise scatter plots of the five variables: the genus g^*, the number of boundary components r, the number of twisted alpha carbon linkages F, the number of residues L and the Euler characteristic χ for all domains in v3.3.0. The variables g^* and F are positive or zero, r and L are strictly positive, and the Euler characteristic is at most 1. Further, the relationship $\chi = 2 - 2g^* - r$ provides bounds, e.g. $\chi \leq 2 - r$. The plots indicate that the variables are capable of distinguishing CATH at the Class (C) level. For example, the (g^*,r), (g^*,F), and (r,F) plots all show separation of the mainly alpha and the mainly beta classes with the mixed alpha-beta class falling somewhere between.
(TIF)

Figure S5 Wilcoxon plot corresponding to pairwise comparisons of the 1,161 H-levels comprising 10 or more domains with significance level 10^{-3} (above diagonal) and 10^{-6} (below diagonal). Each row and column correspond to a H-level, and these are ordered by size in decreasing order. Colors indicate the number of variables (g^*, r, L, F) separating a pair of families at the given significance level: 0 (black), 1 (red), 2 (yellow), 3 (green), and 4 (white). Only every fifth H-level is used in the plot.
(TIF)

Figure S6 Plots of the variables g^*, r, and F, versus the number of alpha helices and beta sheets, respectively, for all domains in v3.3.0. Separation of the mainly alpha and mainly beta classes with the mixed alpha-beta class falling somewhere between is observed. The higher genera observed in the mainly beta and mixed alpha-beta classes are mainly caused by beta sheets. Separation between classes is harder to spot on the plots with beta sheet counts.
(TIF)

Figure S7 Boxplots summarizing the success rates obtained on v3.3.0 using different subsets of variables for classification. For all plots, an energy cut-off at $E = -0.5 kcal/mol$ is used to determine hydrogen bonds. The last plot in the middle row is identical to Fig. 5. The last row shows success rates for (g^*,r,L,F) with alternative energy cut-offs used for determining hydrogen bonds.
(TIF)

Figure S8 Standard deviations of the genus and the number of boundary components for each H-level in v3.3.0 (SAll). The standard deviations are generally not increasing with increasing H-level size, indicating that even large families are homogeneous. There is, however, more variation in the number of boundary components than in the genus.
(TIF)

Figure S9 Correlation plots illustrating the difference between v3.2.0 (SAll) and the newly added domains in v3.3.0. The left plot shows the sizes of the families in v3.2.0 test set versus the family sizes among the newly added domains and the right plot shows the corresponding performance rates. The new domains in v3.3.0 evidently have lower performance while family sizes roughly are proportional to those in v3.2.0.
(TIF)

Figure S10 The distribution of H-level sizes in CATH 3.3.0 exhibits power-law behavior with many small levels and a few very large levels.
(TIF)

Figure S11 The distributions of normalized votes for all methods on the S95 training set.
(TIF)

Table S1 Comparison of all three classifiers on the non-redundant S95 subset of CATH (first three tables) as well as the entire CATH (SAll, last three tables). At each level (C, A, T, and H) we split CATH v3.2.0 into two sets: For a level with N members, we used $\lfloor 2/3 \cdot N \rfloor$ domains for training and the remaining N ï¿½ $\lfloor 2/3 \cdot N \rfloor$ domains for testing. Note that for $N = 1$ and $N = 2$, no domains are used for training. Therefore, in the CATH v3.2.0 test set as well as in the set of new domains in CATH v3.3.0, some domains do not have a classification present in the training set (despite the fact that the classification does exist in CATH v3.2.0). We call such domains unknown. All classifiers were trained to provide a 95% sensitivity on the training sets. For each set (S95 and SAll), the three tables show the following: **Top**: Performance, sensitivity, specificity, and unknown domains flagged as novel/problematic (in percent and in that order) on the CATH v3.2.0 training set. **Middle**: Similarly on the set of new domains in CATH v3.3.0 with classifications existing in CATH v3.2.0. **Bottom**: Some domains in CATH v3.3.0 have novel classifications not existing in CATH v3.2.0. This table summarized how many of these are flagged as novel/problematic by the three classifiers. Finally, the percentage of unclassified domains flagged by each method is shown. Note that there is only one set of unclassified domains, and this is used in both the S95 and the SAll case.
(TIF)

Acknowledgments

Piotr Karasinski is thanked for programming assistance. Lars Madsen is thanked for providing part of Fig. 3. Mikael Christensen, Lars Nørvang Andersen, Palle Villesen Fredsted, and Freddy Bugge Christiansen are thanked for commenting on the manuscript.

Author Contributions

Conceived and designed the experiments: MK CW. Performed the experiments: MK CW. Analyzed the data: MK CW. Wrote the paper: MK CW. Developed the fatgraph model: RCP JEA.

References

1. Koehl P, Levitt M (1999) Structure-based conformational preferences of amino acids. Proceedings of the National Academy of Sciences of the United States of America 96: 12524–12529.
2. Lindorff-Larsen K, Rogen P, Paci E, Vendruscolo M, Dobson CM (2005) Protein folding and the organization of the protein topology universe. Trends in Biochemical Sciences 30: 13–19.
3. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Research 28: 235–242.
4. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, et al. (1997) CATH – a hierarchic classification of protein domain structures. Structure (London, England: 1993) 5: 1093–1108.
5. Murzin A, Brenner S, Hubbard T, Chothia C (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology 247: 536–540.
6. Cuff AL, Sillitoe I, Lewis T, Redfern OC, Garratt R, et al. (2009) The CATH classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. Nucleic Acids Research 37: D310–4.
7. Dietmann S, Holm L (2001) Identification of homology in protein structure classification. Nature Structural Biology 8: 953–957.
8. Dietmann S, Park J, Notredame C, Heger A, Lappe M, et al. (2001) A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. Nucleic Acids Research 29: 55–57.

9. Rackovsky S (2006) Classification of protein sequences and structures. In: Walker JM, ed. The Proteomics Protocol Handbook. Totowa, NJ: Humana Press. pp 861–874.
10. Taylor WR, May ACW, Brown NP, Aszodi A (2001) Protein structure: geometry, topology and classification. Reports on Progress in Physics 64: 517–590.
11. Rogen P, Fain B (2003) Automatic classification of protein structure by using Gauss integrals. Proceedings of the National Academy of Sciences of the United States of America 100: 119–124.
12. Rogen P, Karlsson PW (2008) Parabolic section and distance excess of space curves applied to protein structure classification. Geometriae Dedicata 134: 91–107.
13. Choi IG, Kwon J, Kim SH (2004) Local feature frequency profile: a method to measure structural similarity in proteins. Proceedings of the National Academy of Sciences of the United States of America 101: 3797–3802.
14. Gramada A, Bourne PE (2006) Multipolar representation of protein structure. BMC bioinformatics 7: 242.
15. Kim YJ, Patel JM (2006) A framework for protein structure classification and identification of novel protein structures. BMC bioinformatics 7: 456.
16. Zotenko E, O'Leary DP, Przytycka TM (2006) Secondary structure spatial conformation footprint: a novel method for fast protein structure comparison and classification. BMC Structural Biology 6: 12.
17. Aung Z, Tan KL (2005) Automatic 3D protein structure classification without structural alignment. Journal of computational biology: a journal of computational molecular cell biology 12: 1221–1241.
18. Yang JM, Tung CH (2006) Protein structure database search and evolutionary classification. Nucleic Acids Research 34: 3646–3659.
19. Getz G, Vendruscolo M, Sachs D, Domany E (2002) Automated assignment of SCOP and CATH protein structure classifications from FSSP scores. Proteins 46: 405–415.
20. Rackovsky S (2009) Sequence physical properties encode the global organization of protein structure space. Proceedings of the National Academy of Sciences of the United States of America 106: 14345–14348.
21. Penner RC, Knudsen M, Wiuf C, Andersen JE (2010) Fatgraph models of proteins. Communications on Pure and Applied Mathematics 63: 1249–1297.
22. Penner RC (1988) Perturbative series and the moduli space of Riemann surfaces. J Differential Geom 27: 35–53.
23. Finkelstein AV, Ptitsyn OB (2002) Protein Physics – a course of lectures. New York: Academic Press.
24. Massey WS (1997) Algebraic Topology: An Introduction. New York: Springer-Verlag.
25. Knudsen M, Wiuf C (2010) The CATH database. Human genomics 4: 207–212.
26. de Groot BL, van Aalten DM, Scheek RM, Amadei A, Vriend G, et al. (1997) Prediction of protein conformational freedom from distance constraints. Proteins 29: 240–251.
27. Breiman L (2001) Random forests. Machine learning 45: 5–32.
28. Cuff A, Redfern OC, Greene L, Sillitoe I, Lewis T, et al. (2009) The CATH hierarchy revisitedstructural divergence in domain superfamilies and the continuity of fold space. Structure (London, England: 1993) 17: 1051–1062.
29. Csaba G, Birzele F, Zimmer R (2009) Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis. BMC Structural Biology 9: 23.
30. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J (2006) On the origin and highly likely completeness of single-domain protein structures. Proceedings of the National Academy of Sciences of the United States of America 103: 2605–2610.
31. Hamelryck T, Kent JT, Krogh A (2006) Sampling realistic protein conformations using local structural bias. PLoS Comput Biol 2: e131.
32. Boomsma W, Mardia KV, Taylor CC, Ferkinghoff-Borg J, Krogh A, et al. (2008) A generative, probabilistic model of local protein structure. Proceedings of the National Academy of Sciences of the United States of America 105: 8932–8937.
33. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577–2637.