Japan J. Indust. Appl. Math. (2015) 32:1–17
DOI 10.1007/s13160-014-0153-5

ORIGINAL PAPER **Area 1**

A topological measurement of protein compressibility

Marcio Gameiro · Yasuaki Hiraoka · Shunsuke Izumi · Miroslav Kramar · Konstantin Mischaikow · Vidit Nanda

Received: 6 January 2014 / Revised: 7 July 2014 / Published online: 11 October 2014

Abstract In this paper we partially clarify the relation between the compressibility of a protein and its molecular geometric structure. To identify and understand the relevant topological features within a given protein, we model its molecule as an alpha filtration and hence obtain multi-scale insight into the structure of its tunnels

M. Gameiro
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Caixa Postal 668, São Carlos, SP 13560-970, Brazil
e-mail: gameiro@icmc.usp.br

Y. Hiraoka (✉)
Institute of Mathematics for Industry, Kyushu University, 744, Motooka, Nishi-ku, Fukuoka 819-0395, Japan
e-mail: hiraoka@imi.kyushu-u.ac.jp

S. Izumi
Department of Mathematical and Life Sciences, Hiroshima University, 1-3-1, Kagamiyama, Higashi-Hiroshima 739-8526, Japan
e-mail: sizumi@sci.hiroshima-u.ac.jp

M. Kramar
Department of Mathematics, Hill Center-Busch Campus Rutgers, The State University of New Jersey, 110 Frelinghusen Rd, Piscataway, NJ 08854-8019, USA
e-mail: miroslav@math.rutgers.edu

K. Mischaikow
Department of Mathematics and BioMaPS, Hill Center-Busch Campus Rutgers, The State University of New Jersey, 110 Frelinghusen Rd, Piscataway, NJ 08854-8019, USA
e-mail: mischaik@math.rutgers.edu

V. Nanda
Department of Mathematics, The University of Pennsylvania, DRL, 209 South 33rd Street, Philadelphia, PA 19104, USA
e-mail: vnanda@sas.upenn.edu

and cavities. The persistence diagrams of this alpha filtration capture the sizes and robustness of such tunnels and cavities in a compact and meaningful manner. From these persistence diagrams, we extract a measure of compressibility derived from those topological features whose relevance is suggested by physical and chemical properties. Due to recent advances in combinatorial topology, this measure is efficiently and directly computable from information found in the Protein Data Bank (PDB). Our main result establishes a clear linear correlation between the topological measure and the experimentally-determined compressibility of most proteins for which both PDB information and experimental compressibility data are available. Finally, we establish that both the topological measurement and the linear correlation are stable with respect to small perturbations in the input data, such as those arising from experimental errors in compressibility and X-ray crystallography experiments.

Keywords Protein compressibility · Computational topology · Persistence diagram · Stability

Mathematics Subject Classification 55N99

1 Introduction

The softness of a protein is known to be closely related to its geometric structure and biological function [20,23–25]. One of the quantities which partially characterizes protein softness is *compressibility* [1,10,11,13,14,17,18,20]. The accurate evaluation of protein compressibility is essential to elucidating the physical mechanism responsible for the structure-function relationship of proteins. It has been hypothesized [10,11] that softness depends on the size of the individual molecular cavities. Thus, proteins whose molecules contain larger cavities are predicted to be softer than proteins with smaller cavities even if the total volume of the cavities is the same. This suggests that concentrating on the total volume or the number of cavities separately cannot provide a satisfactory prediction of compressibility. It is possible that the ratio of total volume to number of cavities would provide a better indicator; however, this ratio does not offer any insight into sizes and geometric features of the individual cavities.

In this paper, we propose a new topological predictor for compressibility of proteins based on a relatively new mathematical tool known as persistent homology [3,9,26]. Persistent homology has several features which suggest that it is a potentially powerful tool for analyzing geometric characteristics of molecules. First, it provides topological information; that is, it counts cavities and tunnels (a precise definition is provided in Sect. 2), but in addition it provides information about their sizes. Second, the information it provides is robust to perturbations such as measurement error. In other words, small changes in the input data lead to small changes in the output of the persistent homology computations. Finally, persistent homology can be computed efficiently using freely available open-source software [30] starting with standard molecular datasets such as those of the Protein Data Bank (PDB) [31].

A standard model for the geometry of a molecule is obtained by representing each atom by a solid ball in three-dimensional space with the van der Waal radius of the corresponding atom. While this model is easy to generate, it has two obvious limitations.

First, the van der Waal radius of a given atom depends on its chemical environment and is not a universal constant. Second, the model depends on precise knowledge of the locations of all the atoms. For PDB data, the typical error bounds on atom locations often exceed 1 Å. If one assumes a fixed radius, then the aforementioned experimental variability can easily lead to the creation of false cavities or the destruction of real ones. Given the hypothesis that cavities are correlated with compressibility, this becomes a serious issue. To circumvent this problem we make use of *alpha filtrations* [8] weighted by the van der Waal radii. An exact description is given in Sect. 3, for the moment it is sufficient to view it as providing an increasing family of geometric models indexed by a parameter $\alpha \geq 0$, where at $\alpha = 0$ one obtains the standard model described above.

As α increases, various topological features such as cavities and tunnels are created and destroyed. Persistent homology provides a coherent means by which topological features at different scales can be uniquely identified. Thus, it allows us to obtain for each topological feature z a unique α-value b_z at which this feature first appears and another α-value d_z at which this feature disappears. The collection of all such pairs (b_z, d_z) forms a finite subset of the plane called the *persistence diagram* of our molecular alpha filtration. As is discussed in Sect. 2.3, there is a natural notion of distance between persistence diagrams along with theorems which guarantee that slight changes in the assumptions concerning the location of the atoms or the particular choice of van der Waal radii result in small changes to the persistence diagram.

From each point (b_z, d_z) in the persistence diagram of a molecular alpha filtration, we can make inferences about the size and structure of the associated feature z which it identifies. This leads us to define a *topological compressibility measure*, denoted Σ, as the ratio of the number of cavities to tunnels which lie in a specific portion of the persistence diagram. Our main result—illustrated in Fig. 6—is that Σ exhibits a remarkable linear correlation with most experimental compressibility data present in [11]. Moreover, to test a different hypothesis requires only a modification of the measure Σ based on the persistence diagrams of the proteins which can be found at [28].

It should be noted that while there are alternative methods (e.g., Naccess [29], RosettaHoles [21], etc) which identify cavities in a protein at a fixed scale, our approach using persistent homology has several advantages. As indicated above, it is because we use a 1-parameter family of geometric models, that we can be guaranteed of the stability of the results with respect to small experimental errors or variants in the chemical environment. Furthermore, persistent homology provides a consistent unique identification of the desired topological features over the family of models. Finally, because we make use of alpha filtrations to perform these computations we can be sure of capturing the appearance and disappearance of all cavities and tunnels that are used to define Σ.

2 Background on topological methods

We provide a brief and heuristic introduction to simplicial complexes and their homology groups. The reader is encouraged to consult [16] for a complete presentation. A detailed account of filtered simplicial complexes and their persistent homology groups may be found in [3,9] and the references therein.

2.1 Simplicial complexes and homology groups

Let V be a finite set. A *simplicial complex* K with vertex set V is a collection of non-empty subsets of V so that the following two conditions hold. First, for each vertex v in V, the singleton $\{v\}$ lies in K, and second, K is closed under the containment relation. That is, if $\sigma \subset V$ is in K and $\tau \subset \sigma$, then τ is also in K. Given a simplicial complex K with vertex set V, an element σ in K is called a *simplex* of K and its *dimension*, denoted $\dim \sigma$, equals its cardinality minus 1. Note that the collection of 0-dimensional simplices in K is naturally identified with the underlying vertex set V.

If we assume that the vertices in V are *ordered*, then we can impose an algebraic structure on K in the following manner. For each dimension m, one defines a vector space $C_m(\mathsf{K})$ of *m-chains* spanned by the m-dimensional simplices as an orthonormal basis. The *boundary operators* $\partial_m : C_m(\mathsf{K}) \to C_{m-1}(\mathsf{K})$ are linear maps whose action on a basis simplex $\sigma = (v_0, \ldots, v_m)$ with ordered vertices is given by

$$\partial_m(\sigma) = \sum_{j=0}^{m} (-1)^j \sigma_j,$$

where σ_j denotes the $(m-1)$-dimensional simplex containing all the same vertices as σ but with the vertex v_j removed.

Thus, we obtain a sequence of vector spaces connected by linear maps

$$\cdots \to C_{m+1}(\mathsf{K}) \xrightarrow{\partial_{m+1}} C_m(\mathsf{K}) \xrightarrow{\partial_m} C_{m-1}(\mathsf{K}) \to \cdots$$

The kernel of ∂_m in $C_m(\mathsf{K})$ is called the subspace of *m-cycles* and denoted by $Z_m(\mathsf{K})$. Similarly, the image of ∂_{m+1} in $C_m(\mathsf{K})$ is known as the subspace of *m-boundaries* and written as $B_m(\mathsf{K})$. A routine calculation shows that $\partial_m \circ \partial_{m+1}$ is the zero map on $C_{m+1}(\mathsf{K})$ for each dimension, so we have an inclusion of the vector spaces $B_m(\mathsf{K}) \subset Z_m(\mathsf{K})$. The m-dimensional *homology group* of K is defined as the quotient space

$$H_m(\mathsf{K}) = \frac{Z_m(\mathsf{K})}{B_m(\mathsf{K})}.$$

Two m-cycles are called *homologous* in K if their algebraic difference is a m-boundary, and in this case they represent the same element of the m-th homology group. The m-th *Betti number* of K, written $\beta_m(\mathsf{K})$, is the dimension of $H_m(\mathsf{K})$ as a vector space. We say that K is *acyclic* if $\beta_0(\mathsf{K}) = 1$ and $\beta_m(\mathsf{K}) = 0$ for all $m > 0$.

In practice, one can compute Betti numbers and presentations of homology groups by putting the boundary operators ∂_m in *Smith normal form*. Various algorithms [7,15] have been implemented [28] to perform such computations efficiently.

2.2 Geometric realizations and nerves

To each simplicial complex K, one can associate a geometric object $|\mathsf{K}|$ embedded in an Euclidean space. On the other hand, from each finite collection $\mathcal{U}$ of subsets of an

Euclidean space, one can create a simplicial complex $N(\mathcal{U})$. In this section we briefly describe both constructions.

Let K be a simplicial complex with vertices $v_1, \ldots, v_n$ and associate the j-th vertex to the j-th *coordinate point* p_j of $\mathbb{R}^n$ whose coordinate expansion contains a 1 in the j-th position and zeros elsewhere. Then, each simplex σ of K is associated with the convex hull $|\sigma|$ of the coordinate points corresponding to its vertices. For instance, when $n = 3$ and $\sigma = (v_1, v_2)$, then $|\sigma|$ is precisely the line segment stretching between $(1, 0, 0)$ and $(0, 1, 0)$ in $\mathbb{R}^3$. The union of all such $|\sigma|$ as σ ranges over the simplices of K is called the *geometric realization* of K and is typically denoted by $|\mathsf{K}|$. We call a geometric object *triangulable* if it is homeomorphic to the geometric realization of some simplicial complex, and remark that the homology groups of triangulable spaces are defined to equal those of their homeomorphic simplicial counterparts.

Consider a finite collection $\mathcal{U}$ of non-empty triangulable subsets of $\mathbb{R}^n$ and let U be their union. The *nerve* of $\mathcal{U}$ is a simplicial complex $N(\mathcal{U})$ whose vertex set is $\mathcal{U}$, and whose d-dimensional simplices for $d > 0$ correspond precisely to those sub-collections of $\mathcal{U}$ which have a non-empty intersection in $\mathbb{R}^n$. The *nerve theorem* from [2] states that if each non-empty intersection of sets from $\mathcal{U}$ is acyclic, then the homology groups of their union U are the same as those of the nerve $N(\mathcal{U})$.

The Betti numbers of a triangulable space $\mathbf{T}$ embedded in $\mathbb{R}^3$ capture coarse properties of $\mathbf{T}$. The zeroth Betti number $\beta_0(\mathbf{T})$ counts the number of *connected components* of $\mathbf{T}$. The first and second Betti numbers count the number of *tunnels* and *cavities* in $\mathbf{T}$ respectively. For the purposes of this paper, a tunnel refers to a cylindrical structure within $\mathbf{T}$ whereas a cavity corresponds to a region completely enclosed by $\mathbf{T}$. For instance, if $\mathbf{T}$ is a solid ball with a thickened diameter removed, then it has a single tunnel and no cavities. On the other hand, if the entire interior of that solid ball is removed, then the hollow shell which remains has a single cavity but no tunnels.

It follows from the nerve theorem that if a triangulable space $\mathbf{T} \subset \mathbb{R}^3$ is decomposed into a finite collection $\mathcal{T}$ of triangulable subsets whose non-empty intersections are acyclic, then the geometric realization $|N(\mathcal{T})|$ has precisely the same number of connected components, tunnels and cavities as the original space $\mathbf{T}$.

2.3 Filtrations, persistent homology and stability

Let K be a simplicial complex. A *subcomplex* K' of K is a sub-collection of simplices of K which is a simplicial complex in its own right. That is, if some simplex σ of K is contained in K', then any subset of σ is also contained in K'. We denote this subcomplex relation by $\mathsf{K}' \hookrightarrow \mathsf{K}$. It is easy to check that cycles and boundaries of K' remain cycles and boundaries when viewed as chains of K. A *filtration* $\mathcal{F}$ with length $A \in \mathbb{R}$ of a simplicial complex K assigns to each index a in $[0, A]$ a subcomplex $\mathcal{F}_a\mathsf{K}$ of K so that $\mathcal{F}_\alpha\mathsf{K} \hookrightarrow \mathcal{F}_{\alpha'}\mathsf{K}$ whenever $\alpha \leq \alpha'$.

It is customary to define the subcomplexes which constitute a filtration only on a finite subset $S_\mathcal{F} \subset [0, A]$ of indices[1] with the understanding that $\mathcal{F}_\alpha\mathsf{K} = \mathcal{F}_s\mathsf{K}$ where s is the largest element of $S_\mathcal{F}$ smaller than α.

[1] From the finiteness of K, we have a natural choice of $S_\mathcal{F}$ for any filtration $\mathcal{F}$ of K since there are only finitely many indices in $[0, A]$ where new simplices get introduced.

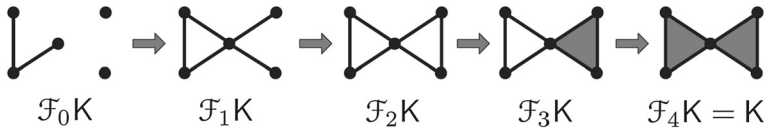

Fig. 1 A length 4 filtration $\mathcal{F}$ of the simplicial complex K consisting of two 2-dimensional simplices joined at a common vertex. Since $S_{\mathcal{F}} = \{0, 1, 2, 3, 4\}$, it is assumed for instance that $\mathcal{F}_\alpha\mathsf{K} = \mathcal{F}_1\mathsf{K}$ whenever $1 \leq \alpha < 2$

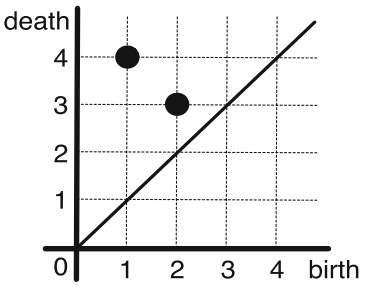

Fig. 2 The 1-dimensional persistence diagram $\mathrm{dgm}_1(\mathcal{F})$ of the filtration $\mathcal{F}$ from Fig. 1. The point (1, 4) corresponds to the hollow triangle on the *left* whereas the point (2, 3) corresponds to that on the *right*

Persistent homology is to filtrations what homology is to simplicial complexes. Fix a dimension m and choose an index s in the finite set $S_{\mathcal{F}}$. To each vector z in the homology group $H_m(\mathcal{F}_s\mathsf{K})$, we associate an interval (b_z, d_z) as follows. The *birth* index $b_z \leq s$ is the smallest $r \in S_{\mathcal{F}}$ for which there is some cycle x in $Z_m(\mathcal{F}_r\mathsf{K})$ homologous to a representative cycle of z in $Z_m(\mathcal{F}_s\mathsf{K})$. Similarly, the *death* index $d_z > s$ is the smallest index $t \in S_{\mathcal{F}}$ so that any representative cycle of z is a boundary in $B_m(\mathcal{F}_t\mathsf{K})$ – with the understanding that $d_z = \infty$ if this never happens.

Definition 1 The m-dimensional *persistence diagram* $\mathrm{dgm}_m(\mathcal{F})$ of the filtration $\mathcal{F}$ is the multi-set of points (b_z, d_z) where z ranges over homologically independent cycles in $Z_m(\mathcal{F}_s\mathsf{K})$ for $s \in S_{\mathcal{F}}$.

The m-th Betti number $\beta_m(\mathcal{F}_s\mathsf{K})$ of any intermediate subcomplex in the filtration $\mathcal{F}$ can be recovered simply by counting all those intervals in $\mathrm{dgm}_m(\mathcal{F})$ which contain s. One often depicts a persistence diagram as a set of points in the plane (see Fig. 2) where the vertical distance $d_z - b_z$ of each point (b_z, d_z) from the *diagonal* measures the *lifespan* of the feature z. Algorithms [26] and highly optimized implementations [30] for computing persistence diagrams of filtrations are freely available.

We now turn to the issue of comparing persistence diagrams. Recall that the ℓ_∞-distance between any pair of points $u = (x, y)$ and $u' = (x', y')$ in the plane is given by

$$\|u - u'\|_\infty = \max\left\{|x - x'|, |y - y'|\right\}.$$

Let dgm and dgm′ be a pair of persistence diagrams and let $\epsilon > 0$ be a positive number. An ϵ-*matching* between dgm and dgm′ is a subset Φ of the product dgm $\times$ dgm′ for which the following three conditions hold:

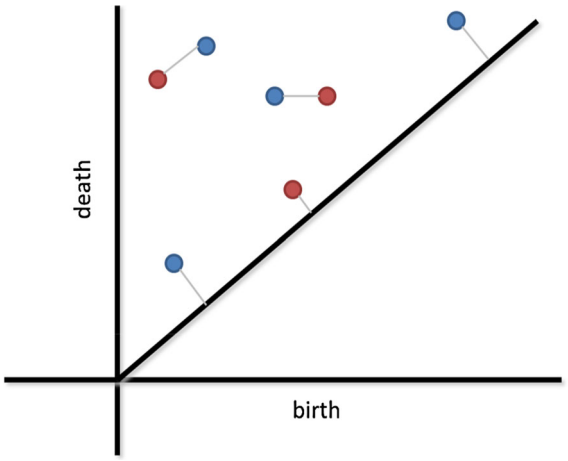

Fig. 3 Two overlaid persistence diagrams whose points are shown in *red* and *blue*. The optimal matching between these diagrams is depicted by *gray line* segments which either match points across the two diagrams or with the diagonal

- each point $u = (b, d) \in \mathrm{dgm}$ appears as the first component of at most one element in Φ, and each point $u' = (b', d') \in \mathrm{dgm}'$ appears as the second component of at most one element in Φ,
- if $(u, u') \in \Phi$ then $\|u - u'\|_\infty < \epsilon$,
- if a point in dgm or dgm$'$ does not appear in Φ at all, then it is within ℓ^∞ distance ϵ of a point on the diagonal.

The *bottleneck distance* $\mathbf{d}_{\mathrm{bot}}(\mathrm{dgm}, \mathrm{dgm}')$ between dgm and dgm$'$ is defined to be the smallest ϵ for which there exists an ϵ-matching between dgm and dgm$'$. Figure 3 illustrates a matching which realizes the bottleneck distance between two simple persistence diagrams.

Let $\epsilon > 0$ be a real number, and assume that $\mathcal{F}$ and $\mathcal{G}$ are length-A filtrations of the same simplicial complex K. We say that $\mathcal{F}$ and $\mathcal{G}$ are *ϵ-interleaved* if there are subcomplex relations $\mathcal{F}_\alpha \mathsf{K} \hookrightarrow \mathcal{G}_{\alpha+\epsilon} \mathsf{K}$ and $\mathcal{G}_\alpha \mathsf{K} \hookrightarrow \mathcal{F}_{\alpha+\epsilon} \mathsf{K}$ for each α in $[0, A - \epsilon]$. The following result is a consequence of the *stability theorem* [4,5] for persistence diagrams.

Theorem 1 *If two filtrations $\mathcal{F}$ and $\mathcal{G}$ are ϵ-interleaved, then the bottleneck distance $\mathbf{d}_{\mathrm{bot}}(\mathrm{dgm}_m(\mathcal{F}), \mathrm{dgm}_m(\mathcal{G}))$ is smaller than ϵ for each m.*

It follows from this theorem that the lifespan of a given point in a persistence diagram measures the robustness of the corresponding feature to changes in the filtration.

3 Persistence diagrams of protein molecules

A simple geometric representation of an atom is the three-dimensional solid ball

$$\mathbf{B}(c; \mathbf{w}) = \left\{ x \in \mathbb{R}^3 : \|x - c\| < \mathbf{w} \right\},$$

where c is the center of that atom, $\mathbf{w}$ is its van der Waals radius (see Table 1) and $\|\cdot\|$ is the usual Euclidean distance in three dimensions. A molecule, then, may be

Table 1 van der Waals radii of atoms commonly found in protein molecules

Atom	C	N	O	P	S
Radius (Å)	1.70	1.55	1.52	1.80	1.80

modeled as the union of such balls corresponding to its constituent atoms. In [19] this *static* model was employed to estimate surface areas and the volumes of various protein molecules. The key tool used in their analysis is the *alpha filtration* [8], which we briefly describe below.

Let $\mathbb{P} = (P, \mathbf{w})$ be a pair where P is a finite set of points in $\mathbb{R}^3$ and $\mathbf{w}$ assigns to each p in P a non-negative *weight* $\mathbf{w}_p$. The *weighted distance* from x in $\mathbb{R}^3$ to p is given by

$$\mathbf{d}_p(x) = \|x - p\| - \mathbf{w}_p,$$

and the *Voronoi cell* $\mathbf{V}_p$ associated to p is the collection of those points in $\mathbb{R}^3$ which are nearer to p than to any other point of P under this weighted distance. That is,

$$\mathbf{V}_p = \left\{ x \in \mathbb{R}^3 \mid \mathbf{d}_p(x) \leq \mathbf{d}_{p'}(x) \quad \text{for each } p' \neq p \text{ in } P \right\}.$$

For each *scale* $\alpha \geq 0$ and point p in P, define

$$w_p(\alpha) = \sqrt{\alpha + \mathbf{w}_p^2},$$

noting that $w_p(0)$ coincides with the assigned weight $\mathbf{w}_p$ of p, and that $w_p(\alpha)$ is a strictly increasing function of the scale α. Consequently, the α-indexed family of solid balls $\mathbf{B}(p; w_p(\alpha))$ centered at p has radii which increase with α. Define the intersections

$$\mathbf{U}_p(\alpha) = \mathbf{B}(p; w_p(\alpha)) \cap \mathbf{V}_p,$$

of these balls with the Voronoi cell corresponding to p to produce a new increasing family of convex sets around each p in P. At each scale α, the union over all points p of $\mathbf{U}_p(\alpha)$ equals the union of balls $\mathbf{B}(p; w_p(\alpha))$ because the Voronoi cells partition the latter union into the former (Fig. 4).

Let K_P be the *complete* simplicial complex with vertex set P, i.e., all possible subsets of P constitute simplices of K_P.

Definition 2 The *alpha filtration* $\mathcal{G}$ around $\mathbb{P}$ is a filtration of K_P defined as follows: the subcomplex $\mathcal{G}_\alpha \mathsf{K}_P \hookrightarrow \mathsf{K}_P$ is the nerve of the sets $\mathbf{U}_p(\alpha)$ where p ranges over the points in P.

Whenever $\alpha \leq \alpha'$, we have the desired subcomplex relation $\mathcal{G}_\alpha \mathsf{K}_P \hookrightarrow \mathcal{G}_{\alpha'} \mathsf{K}_P$ because $\mathbf{U}_p(\alpha)$ is a subset of $\mathbf{U}_p(\alpha')$ for each point p in P. Moreover, it follows from the finiteness of P that there are only finitely many values of the scale α at which new simplices get introduced into the filtration $\mathcal{G}$.

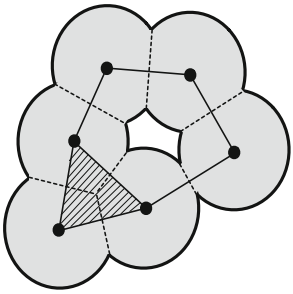

Fig. 4 A union of scale α balls in the plane. The *dashed lines* indicate partitions by Voronoi cells and the associated subcomplex of the alpha filtration is overlaid

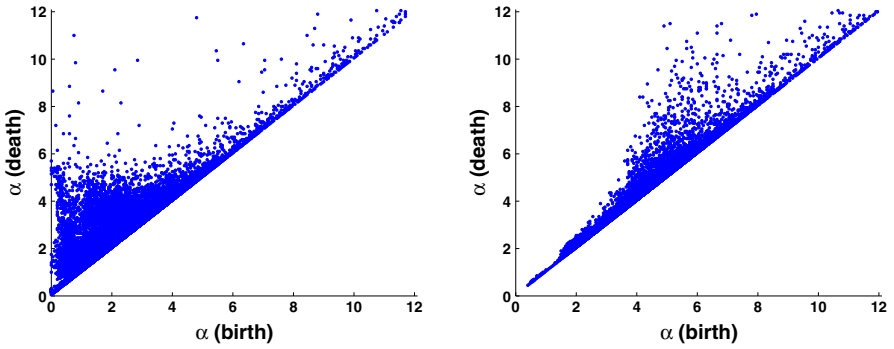

Fig. 5 dgm_1 (*left*) and dgm_2 (*right*) of 1OVA

It is an immediate consequence of the nerve theorem that at each scale $\alpha \geq 0$, the homology groups of the subcomplex $\mathcal{G}_\alpha \mathsf{K}_P$ are the same as those of the union of balls $\mathbf{B}(p, w_p(\alpha))$ where p ranges over the points in P.

Definition 3 Let $\mathbb{P} = (P, \mathbf{w})$ be a pair consisting of finitely many points P in $\mathbb{R}^3$ which represent positions of atom-centers and weights $\mathbf{w}_p$ equaling the van der Wall radii of the corresponding atoms. The *m-th persistence diagram of* $\mathbb{P}$—denoted $\mathrm{dgm}_m \mathbb{P}$—is the m-dimensional persistence diagram associated to the alpha filtration $\mathcal{G}$ around $\mathbb{P}$.

Figure 5 shows the 1 and 2-dimensional persistence diagrams of ovalbumin, which is identified in the PDB as 1OVA. For the computations of alpha filtrations, we use CGAL [27]. In practice, we do not build the filtration past $\alpha = 12$ Å because no interesting topological features are observed in protein molecules at higher scale values.

4 Compressibility from persistence diagrams

Experimentally, the compressibility of a protein molecule is determined from measurement of the ultrasonic wave velocities in both its solution and solvent [12]. A pressure wave in a fluid causes alternating compressions and rarefactions. Because the period is short compared with the time required for thermal equilibrium of the solution, the process is reversible and adiabatic. In general, the experimentally determined adiabatic compressibility of a protein would be mainly due to the contributions

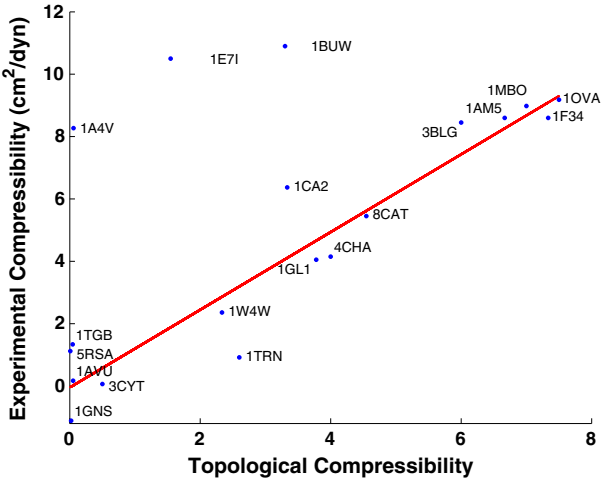

Fig. 6 Topological compressibility Σ plotted against experimental compressibility for several proteins. Every point represents a protein and is labeled with its corresponding PDB ID

of cavities and hydration [11]. Since the hydration of globular proteins is proportional to the volume of the protein, it can be concluded that the compressibility (per unit volume) of globular proteins is determined by the cavities in the protein.

In this section, we associate to each protein a topological quantity which is measurable directly from crystallography data, and linearly correlated with experimental compressibility values of proteins from [11].

Let dgm be a persistence diagram. Given an open interval $\mathcal{I} = (x, y)$ and a real number $\epsilon > 0$, we denote by $|\mathrm{dgm}(\mathcal{I}; \epsilon)|$ the number of points of dgm whose birth b lies in $\mathcal{I}$ and whose lifespan exceeds ϵ. More precisely, $|\mathrm{dgm}(\mathcal{I}; \epsilon)|$ counts those (b, d) in dgm which satisfy the inequalities $x < b < \min\{y, d - \epsilon\}$.

Definition 4 The *topological compressibility* of a protein $\mathbb{P}$ is given by

$$\Sigma_{\mathbb{P}} := \frac{|\mathrm{dgm}_2 \mathbb{P}(\mathcal{I}_2; \delta)|}{|\mathrm{dgm}_1 \mathbb{P}(\mathcal{I}_1; \delta)|}, \tag{1}$$

where $\delta = 1.25$, $\mathcal{I}_1 = (4.8, 7.6)$ and $\mathcal{I}_2 = (4.6, 7.6)$.

Figure 6 shows a clear linear correlation between Σ and the experimental compressibility of most proteins for which both X-ray crystallography data and experimental compressibility data are available. The rest of this section is dedicated to the consideration of physical and geometric factors which motivate Definition 4 along with providing an experimental justification for the explicit choices of parameters $\delta, \mathcal{I}_1$ and $\mathcal{I}_2$.

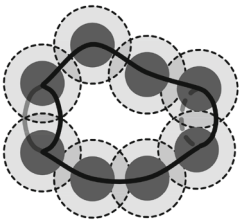

Fig. 7 Longitudinal section of a short tunnel whose thickening forms a single cavity

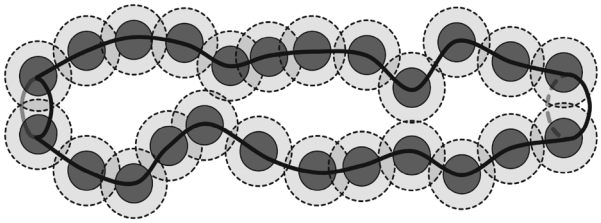

Fig. 8 Longitudinal section of a long tunnel whose thickening forms three cavities

4.1 Compressibility as a ratio

At its core, the topological compressibility $\Sigma_{\mathbb{P}}$ of a protein $\mathbb{P}$ is a ratio of the number of certain types of cavities to the number of certain types of tunnels present in the alpha filtration around $\mathbb{P}$. Before investigating specifically which cavities and tunnels contribute to this ratio, we provide a geometric justification for using a ratio in the first place.

We propose that a fundamental geometric quantity that determines compressibility is the *presence of long tunnels* in the alpha filtration around $\mathbb{P}$. Although the length of any given tunnel is not directly encoded by its birth and death coordinates in $\mathrm{dgm}_1\mathbb{P}$, we can extract useful metric information by focusing on how the tunnels evolve on average as the scale α increases. Increasing α increases the radius of balls whose union forms the walls of each tunnel, and eventually leads to the formation of cavities as the expanded walls get pinched together. In general, longer tunnels correspond to more undulant surface regions, and hence generate a larger number of cavities upon thickening. This phenomenon is illustrated in Figs. 7 and 8.

4.2 The parameters δ, $\mathcal{I}_1$ and $\mathcal{I}_2$

In this section we provide a justification for the chosen values of various parameters involved in Definition 4.

It follows from Theorem 1 that points in a persistence diagram which happen to be near the diagonal are unstable to changes in the underlying filtration. Therefore, we introduce a parameter $\delta > 0$ and restrict our attention to only those points in the persistence diagrams of a given protein which are at least δ away from the diagonal.

Consider the two cavities whose cross sections have been shown in Fig. 9. We conjecture that the sparse cavity on the right is deformable to a much larger extent

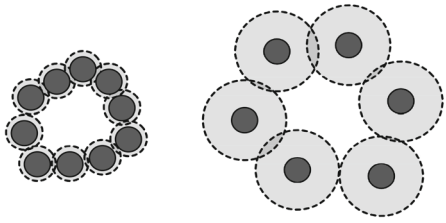

Fig. 9 Dense hole (*left*) and sparse hole (*right*). *Solid balls* correspond to the van der Waals radii and *dashed balls* have radii slightly larger than the birth scale

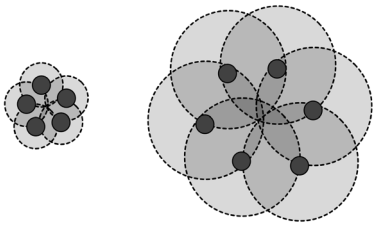

Fig. 10 Death scale and size. Note that the larger cavity to the *right* has a larger death scale than the smaller one to the *left*

than the dense one on the left, and hence assume that a larger number of sparse holes leads to greater compressibility. This distinction between dense and sparse topological features is readily captured by persistence diagrams: the denser the cavity, the smaller its birth scale. This heuristic argument is our justification for counting only those topological features whose birth scales are sufficiently large. To this end, we introduce thresholds $x_m > 0$ for $m = 1, 2$ and restrict attention to those points (b, d) in $\mathrm{dgm}_m \mathbb{P}$ for which $b > x_m$.

We do not introduce similar thresholds to control the death scales of features. The death scale of a hole is closely related to the *size* of that hole as shown in Fig. 10. Introducing such bounds on the size would remove large holes from consideration and compromise the analysis of compressibility.

Performing a simple least-squares computation reveals that there are no values of δ, x_1 and x_2 for which the ratio of corresponding points in $\mathrm{dgm}_2 \mathbb{P}$ to those in $\mathrm{dgm}_1 \mathbb{P}$ yields a good approximation to the experimental compressibility of $\mathbb{P}$. This is perhaps not surprising. One can imagine that tunnels or cavities are created on a scale exceeding those that are relevant to compressibility, e.g. cavities in polymers created by the monomer subunits. For this reason we introduce new parameters y_m for $m = 1, 2$ with $y_1 \leq y_2$ and further restrict the birth scales of points in $\mathrm{dgm}_m \mathbb{P}$ to quantities smaller than y_m. Performing the least-squares computation with these five parameters instead of the first three, the linear correlation from Fig. 6 emerges for the following optimal values:

$$\begin{aligned} \delta &= 1.25, \\ \mathcal{I}_1 &= (x_1, y_1) = (4.8, 7.6), \text{ and} \\ \mathcal{I}_2 &= (x_2, y_2) = (4.6, 7.6). \end{aligned}$$

Thus, the parameter values in Definition 4 arise from a mixture of hypothesis and experiment. These parameters carve out those regions of 1 and 2-dimensional per-

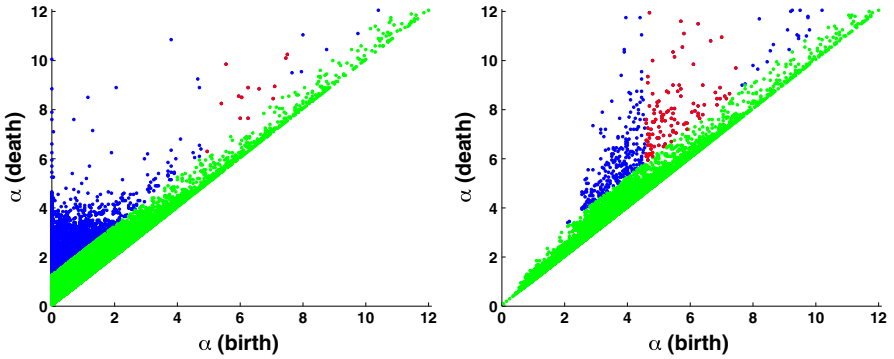

Fig. 11 The topological compressibility $\Sigma_{1\text{OVA}}$ is a ratio of number of *red points* in dgm_2 (*left*) to the number of *red points* in dgm_1 (*right*). The *green points* are excluded because of δ while the *blue points* are excluded because of $\mathcal{I}_1$ (*left*) and $\mathcal{I}_2$ (*right*)

sistence diagrams whose points are relevant to compressibility calculations. These regions are shown in Fig. 11 for the protein 1OVA.

At the time of writing, there are only a handful of proteins for which experimental compressibility data is available. We expect that the availability of such data for more proteins will provide opportunities to further refine the parameters δ, $\mathcal{I}_1$ and $\mathcal{I}_2$.

5 Stability of topological compressibility

In this section we use Theorem 1 to show that $\Sigma_{\mathbb{P}}$ is invariant to small errors in measurement of atom positions $\mathbb{P}$ as well as in the values of van der Waal radii.

Let $\mathbb{P} = (P, \mathbf{w})$ be the usual pair consisting of atom positions and van der Waal radii. Consider another pair $\mathbb{Q} = (Q, \mathbf{v})$ which is to be understood as a perturbation of $\mathbb{P}$ in the following sense. There exist some distances $\lambda, \mu > 0$ so that each p in P corresponds bijectively to some q in Q with the Euclidean distance $\|p - q\|$ smaller than λ; moreover, the maximum difference $|\mathbf{w}_p - \mathbf{v}_q|$ of van der Waal radii over all points p in P is smaller than μ.

We prove the following stability theorem, establishing that topological compressibility remains unchanged when we replace $\mathbb{P}$ by a sufficiently small perturbation $\mathbb{Q}$.

Theorem 2 $\Sigma_{\mathbb{P}} = \Sigma_{\mathbb{Q}}$ *if* λ *and* μ *are small enough.*

In the course of proving this theorem, we derive an explicit inequality which constrains how small λ and μ must be in order for the conclusion to hold. For now, we treat these distances as free parameters.

Define $\overline{\alpha}$ to be the largest value of α encountered in the alpha filtration built around $\mathbb{P}$ and note that for all our diagrams we have $\overline{\alpha} = 12\,\text{Å}$. Also define

$$\overline{\omega} = \max_{p \in P} \{\mathbf{w}_p\} + \mu,$$

and note that this quantity is an upper bound on the van der Waal radii of atoms across both P and Q. The next lemma uses the notation introduced in Sect. 3

Lemma 1 *For each p in P, we have the containment of open balls*

$$\mathbf{B}(p; w_p(\alpha)) \subset \mathbf{B}(q; v_q(\alpha + \epsilon))$$

for any $\epsilon \geq \lambda^2 + 2\lambda\sqrt{\overline{\alpha} + \overline{\omega}^2} + 2\overline{\omega}\mu$.

Proof The desired containment relation holds whenever $v_q(\alpha+\epsilon)$ exceeds $\|p-q\| + w_p(\alpha)$. Using the formulas for w_p and v_q along with the assumption that $\|p-q\| < \lambda$, the following inequality gives a sufficient condition:

$$\sqrt{(\alpha + \epsilon) + \mathbf{v}_q^2} \geq \lambda + \sqrt{\alpha + \mathbf{w}_p^2}.$$

Squaring both sides, we require

$$\epsilon \geq \lambda^2 + 2\lambda\sqrt{\alpha + \mathbf{w}_p^2} + \left(\mathbf{w}_p^2 - \mathbf{v}_q^2\right).$$

It is easy to see that the right side of this inequality is bounded above by $\lambda^2 + 2\lambda\sqrt{\overline{\alpha} + \overline{\omega}^2} + 2\overline{\omega}\mu$ as desired. □

In light of this lemma, it is convenient to define the function

$$\Psi(\lambda, \mu) = \lambda^2 + 2\lambda\sqrt{\overline{\alpha} + \overline{\omega}^2} + \overline{\omega}\mu.$$

By interchanging the roles of $\mathbb{P}$ and $\mathbb{Q}$ in the lemma, we see that if $\epsilon \geq \Psi(\lambda, \mu)$ then the containment of $\mathbf{B}(q; v_q(\alpha))$ in $\mathbf{B}(p; w_p(\alpha+\epsilon))$ is also guaranteed. Consequently, the alpha filtration around $\mathbb{P}$ is $\Psi(\lambda, \mu)$-interleaved with the alpha filtration around $\mathbb{Q}$. By Theorem 1, we have the following result.

Proposition 1 *The bottleneck distance between* $\mathrm{dgm}_m\mathbb{P}$ *and* $\mathrm{dgm}_m\mathbb{Q}$ *is smaller than* $\Psi(\lambda, \mu)$ *for* $m = 1, 2$.

It is clear that $\Psi(\lambda, \mu)$ can be made as small as desired by shrinking the distances λ and μ which separate $\mathbb{P}$ from its perturbation $\mathbb{Q}$. We therefore turn our attention to proving the conclusion of Theorem 2 when the bottleneck distances $\mathbf{d}_{\mathrm{bot}}(\mathrm{dgm}_m\mathbb{P}, \mathrm{dgm}_m\mathbb{Q})$ for $m = 1, 2$ are sufficiently small.

Proof of Theorem 2 Recall from Definition 4 that the parameters δ and $\mathcal{I}_m = (x_m, y_m)$ identify those regions of $\mathrm{dgm}_m\mathbb{P}$ which are relevant to the computation of $\Sigma_{\mathbb{P}}$. The *confidence* $\nu_{\mathbb{P}}$ in the measurement of $\Sigma_{\mathbb{P}}$ is the minimum ℓ_∞ distance by which a point in $\mathrm{dgm}_m\mathbb{P}$ must be moved in order to change the value of $\Sigma_{\mathbb{P}}$. More precisely, for each point $u = (b, d)$ in $\mathrm{dgm}_m\mathbb{P}$ define

$$\nu_u = \min\left\{|b - x_m|, |b - y_m|, |(d - b) - \delta|\right\},$$

so $\nu_{\mathbb{P}}$ equals the minimum ν_u encountered as u ranges over the points in dgm_m for $m = 1, 2$.

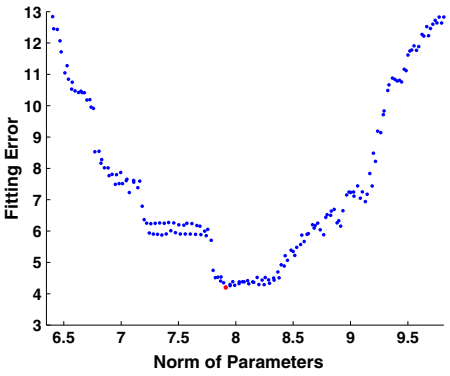

Fig. 12 Least-square errors plotted against the $||\cdot||_2$-norm of parameters $\mathcal{I}_1, \mathcal{I}_2$, and δ. The *red point* corresponds to the optimal parameters

By construction, $\mathbf{d}_{\text{bot}}(\text{dgm}_m\mathbb{P}, \text{dgm}_m\mathbb{Q}) < \nu_{\mathbb{P}}$ implies $\Sigma_{\mathbb{P}} = \Sigma_{\mathbb{Q}}$ for any perturbation $\mathbb{Q}$ of $\mathbb{P}$. Combining this fact with Proposition 1 shows that whenever the distances λ and μ associated to the perturbation $\mathbb{Q}$ of $\mathbb{P}$ are small enough to guarantee $\Psi(\lambda, \mu) < \nu_{\mathbb{P}}$, we have $\Sigma_{\mathbb{P}} = \Sigma_{\mathbb{Q}}$. This concludes the proof of Theorem 2. □

We note that the confidence $\nu_{\mathbb{P}}$ introduced in the proof provides us with a rigorous bound which guarantees that $\Sigma_{\mathbb{P}}$ remains unchanged for the optimal parameters shown in Definition 4. The stability with respect to changes of the parameters $\mathcal{I}_1, \mathcal{I}_2$, and δ follows from Theorem 2, and the explicit bound to satisfy this stability from the optimal parameters is given by $\min_{\mathbb{P}}\{\nu_{\mathbb{P}}\}$.

The fitting errors of the least-square computations obtained by changing the parameters $\mathcal{I}_1, \mathcal{I}_2$, and δ beyond the above bound are shown in Fig. 12, where the red point corresponds to the optimal parameters.

Finally, we provide evidence that no overfitting has occurred in our least-squares analysis. Since available experimental compressibility data is limited to only 16 proteins, we apply the delete-2 jackknife method to experimentally validate our parameter fitting. Let $\mathcal{P}$ be the set of the proteins listed in Fig. 6, and consider the 120 subsets $\mathcal{R}_i \subset \mathcal{P}$ which contain 14 proteins. For each $\mathcal{R}_i$, denote the line of best fit by $y = a_i x + b_i$. Figure 13 shows the plots of means (blue) and variances (red) of a_i (filled) and b_i (empty) across i values. From this figure we note that overfitting is unlikely since the means are almost stationary while the variances are almost zero throughout.

6 Conclusions and further analysis

Figure 6 clearly indicates that the topological measurement successfully extracts some of the essential structural features which determine protein compressibility. On the other hand, it is unclear why the three exceptional proteins 1A4V, 1E7I, and 1BUW deviate from the main correlation line. One possible source of this discrepancy is that there were significant differences in the experimental conditions under which their crystallography and compressibility were measured. Regardless of the causes, we would like to stress that our analysis in this paper is a first approximation and that there is considerable room for improvement. For instance, our geometric models of

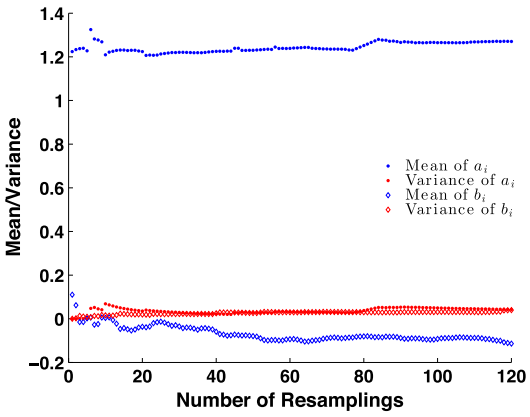

Fig. 13 Plots of means (*blue*) and variances (*red*) of a_i (*filled*) and b_i (*empty*) against the numbers of resamplings based on the delete-2 jackknife method

the proteins are derived entirely from the atom locations and do not take into account chemical aspects such as the types of chemical bonds involved. Also, the parameters which define significant regions of persistence diagrams will be amenable to further refinements as more experimental compressibility data becomes available. We expect that modifications of Σ obtained by aggregating geometric, chemical and experimental factors, will yield much better fits to experimental compressibility.

The points in persistence diagrams $\mathrm{dgm}_m\mathbb{P}$ which contribute towards the topological compressibility $\Sigma_{\mathbb{P}}$ may be regarded as tunnels and cavities having significant impact on the compressibility of the protein represented by $\mathbb{P}$. Recall that each point in $\mathrm{dgm}_m\mathbb{P}$ is given by a vector z in $H_m(\mathcal{G}_\alpha K_P) = Z_m(\mathcal{G}_\alpha K_P)/B_m(\mathcal{G}_\alpha K_P)$ which is represented by some cycle $x \in Z_m(\mathcal{G}_\alpha K_P)$. Let $||x||_0$ be the ℓ_0-norm, i.e., the number of the nonzero elements in the vector x. Then, a solution of the following optimization problem

$$\text{Minimize } ||\bar{x}||_0, \quad \text{subject to} \quad \bar{x} = x + b,\ b \in B_m(\mathcal{G}_\alpha K_P)$$

is a minimum representative of z and gives us geometric locations of tunnels or cavities (e.g., [6,22]). Hence, by solving the optimizations on the points in $\mathrm{dgm}_m\mathbb{P}$ used for the computations of $\Sigma_{\mathbb{P}}$, we can specify regions in the protein inducing high compressibility, and may obtain further insights of the relationship between geometry and compressibility.

Acknowledgments The authors thank Fumihide Nouno for valuable discussions. M. G. was partially supported by FAPESP Grants 2013/07460-7 and 2010/00875-9 and by CNPq Grant 306453/2009-6. Y. H. and S. I. were partially supported by JSPS Grant-in-Aid for Challenging Exploratory Research. M. K., K. M., and V. N. were partially supported by NSF Grants DMS-0915019, DMS-1125174, and CBI-0835621 and by contracts from DARPA and AFOSR.

References

1. Balog, E., Perahia, D., Smith, J., Merzel, F.: Vibrational softening of a protein on ligand binding. J. Phys. Chem. B **115**(21), 6811–6817 (2011)

2. Borsuk, K.: On the imbedding of systems of compacta in simplicial complexes. Fund. Math. **35**, 217–234 (1948)
3. Carlsson, G.: Topology and data. Bull. Am. Math. Soc. **46**(2), 255–308 (2009)
4. Chazal, F., Cohen-Steiner, D., Glisse, M., Guibas, L., Oudot S. : Proximity of persistence modules and their diagrams. In: Proceedings of the Twenty-fifth Annual Symposium on Computational Geometry, pp. 237–246 (2009)
5. Cohen-Steiner, D., Edelsbrunner, H., Harer, J.: Stability of persistence diagrams. Discrete Comput. Geom. **37**, 103–120 (2007)
6. Dey, T., Hirani, A., Krishnamoorthy B.: Optimal homologous cycles, total unimodularity and linear programming. SIAM J. Comput. **40**, 1026–1044 (2011)
7. Dumas, J.-G., Heckenbach, F., Saunders, B.D., Welker, V.: Computing simplicial homology based on efficient Smith normal form algorithms. Algebra, Geometry and Software Systems, pp. 177–206 (2003)
8. Edelsbrunner, H.: The union of balls and Its dual shape. Discrete Comput. Geom. **13**, 415–440 (1995)
9. Edelsbrunner, H., Harer, J.: Persistent homology—a survey. In: Surveys on Discrete and Computational Geometry, vol. 453, pp. 257–282. American Mathematical Society, Providence (2008)
10. Gekko, K., Araga, M., Kamiyama, T., Ohmae, E., Akasaka, K.: Nonneutral evolution of volume fluctuations in lysozymes revealed by normal-mode analysis of compressibility. Biophys. Chem. **144**(1–2), 67–71 (2009)
11. Gekko, K., Hasegawa, Y.: Compressibility-structure relationship of globular proteins. Biochemistry **25**, 6563–6571 (1986)
12. Gekko, K., Noguchi, H.: Compressibility of globular proteins in water at 25 °C. J. Phys. Chem. **83**(21), 2706–2714 (1979)
13. Gekko, K., Tamura, Y., Ohmae, E., Hayashi, H., Kagamiyama, H., Ueno, H.: A large compressibility change of protein induced by a single amino acid substitution. Protein Sci. **5**(3), 542–545 (1996)
14. Gromiha, M., Ponnuswamy, P.K.: Relationship between amino acid properties and protein compressibility. J. Theor. Biol. **165**, 87–100 (1993)
15. Harker, S., Mischaikow, K., Mrozek, M., Nanda, V.: Discrete Morse theoretic algorithms for computing homology of complexes and maps. Found. Comput. Math. (2012). doi:10.1007/s10208-013-9145-0
16. Hatcher, A.: Algebraic Topology. Cambridge University Press (2002)
17. Kharakoz, D.: Protein compressibility, dynamics, and pressure. Biophys. J. **79**, 511–525 (2000)
18. Leu, B., Alatas, A., Sinn, H., Alp, E., Said, A., Yavaş, H., Zhao, J., Sage, J., Sturhahn, W.: Protein elasticity probed with two synchrotron-based techniques. J. Chem. Phys. **132**, 085103 (2010)
19. Liang, J., Edelsbrunner, H., Fu, P., Sudharkar, P.V., Subramaniam, S.: Analytic shape computation of macromolecules I: molecular area and volume through alpha shape. Proteins Struct. Funct. Genet. **33**, 1–17 (1998)
20. Sanchez-Ruiz, J.M.: Protein kinetic stability. Biophys. Chem. **148**(1–3), 1–15 (2010)
21. Sheffler, W., Baker, D.: RosettaHoles: rapid assessment of protein core packing for structure prediction, refinement, design, and validation. Protein Sci. **18**, 229–239 (2009)
22. Tahbaz-Salehi, A., Jadbabaie, A.: Distributed coverage verification in sensor networks without location information. IEEE Trans. Auto. Control **55**, 1837–1849 (2010)
23. Yamamoto, T., Izumi, S., Gekko, K.: Mass spectrometry on hydrogen/deuterium exchange of dihydrofolate reductase: effects of ligand binding. J. Biochem. **135**(6), 663–671 (2004)
24. Uversky, V.: Natively unfolded proteins: a point where biology waits for physics. Protein Sci. **11**(4), 739–756 (2002)
25. Zaccai, G.: How soft is a protein? A protein dynamics force constant measured by neutron scattering. Science **288**, 1604 (2000)
26. Zomorodian, A., Carlsson, G.: Computing persistent homology. Discrete Comput. Geom. **33**, 249–274 (2005)
27. CGAL webpage. http://www.cgal.org/
28. CHomP webpage. http://chomp.rutgers.edu/
29. Naccess. http://www.bioinf.manchester.ac.uk/naccess/
30. Perseus webpage. http://www.math.rutgers.edu/~vidit/perseus.html
31. PDB. http://www.rcsb.org/