Published in final edited form as:
Int J Numer Method Biomed Eng. 2014 August ; 30(8): 814–844. doi:10.1002/cnm.2655.

Persistent homology analysis of protein structure, flexibility and folding

Kelin Xia[1,2] and **Guo-Wei Wei**[1,2,3,4,*]

[1]Department of Mathematics, Michigan State University, MI 48824, USA

[2]Center for Mathematical Molecular Biosciences, Michigan State University, MI 48824, USA

[3]Department of Electrical and Computer Engineering, Michigan State University, MI 48824, USA

[4]Department of Biochemistry and Molecular Biology, Michigan State University, MI 48824, USA

Abstract

Proteins are the most important biomolecules for living organisms. The understanding of protein structure, function, dynamics and transport is one of most challenging tasks in biological science. In the present work, persistent homology is, for the first time, introduced for extracting molecular topological fingerprints (MTFs) based on the persistence of molecular topological invariants. MTFs are utilized for protein characterization, identification and classification. The method of slicing is proposed to track the geometric origin of protein topological invariants. Both all-atom and coarse-grained representations of MTFs are constructed. A new cutoff-like filtration is proposed to shed light on the optimal cutoff distance in elastic network models. Based on the correlation between protein compactness, rigidity and connectivity, we propose an accumulated bar length generated from persistent topological invariants for the quantitative modeling of protein flexibility. To this end, a correlation matrix based filtration is developed. This approach gives rise to an accurate prediction of the optimal characteristic distance used in protein B-factor analysis. Finally, MTFs are employed to characterize protein topological evolution during protein folding and quantitatively predict the protein folding stability. An excellent consistence between our persistent homology prediction and molecular dynamics simulation is found. This work reveals the topology-function relationship of proteins.

Keywords

persistent homology; molecular topological fingerprint; protein topology-function relationship; protein topological evolution; computational topology

1 Introduction

Proteins are of paramount importance to living systems not only because of their role in providing the structural stiffness and rigidity to define the distinct shape of each living being, but also due to their functions in catalyzing cellular chemical reactions, immune

*Address correspondences to Guo-Wei Wei. wei@math.msu.edu.

systems, signaling and signal transduction. It is commonly believed that protein functions are determined by protein structures, including primary amino acid sequences, secondary alpha helices and beta sheets, and the associated tertiary structures. Rigidity and flexibility are part of protein functions. Since 65–90% of human cell mass is water, structural proteins, such as keratin, elastin and collagen, provide stiffness and rigidity to prevent biological material from flowing around. The prediction of protein flexibility and rigidity is not only crucial to structural proteins, but also important to membrane and globular proteins due to the correlation of flexibility to many other protein functions. Protein functions are well known to correlate with protein folding, a process in which random coiled polypeptides assume their three-dimensional structures. Although Anfinsen's dogm[1] has been challenged due to the existence of prions and amyloids, most functional proteins are well folded. The folding funnel hypothesis associates each folded protein structure with the global minimum of the Gibbs free energy. Unfolded conformations have higher energies and are thermodynamically unstable, albeit they can be kinetically favored.

The understanding of the structure-function relationship of proteins is a central issue in experimental biology and is regarded by many to be the holy grail of computational biophysics. Numerous attempts have been made to unveil the structure-function relationship in proteins. One approach to this problem is to design experiments from the evolutionary point of view to understand how evolutionary processes have led to various protein functions that strengthen the sustainability of live beings. The past decade has witnessed a rapid growth in gene sequencing. Vast sequence databases are readily available for entire genomes of many bacteria, archaea and eukaryotes. New genomes are updated on a daily basis. The Protein Data Bank (PDB) has also accumulated near one hundred thousand tertiary structures. The availability of these structural data enables the comparative study of evolutionary processes, which has a potential to decrypt the structure-function relationship. Another approach is to utilize abundant protein sequence and structural information at hand to set up theoretical models for relationships between protein structures and functions. An ultimate goal is to predict protein functions from known protein structures, which is one of the most challenging tasks in biological sciences.

Theoretical study of the structure-functions relationship of proteins is usually based on fundamental laws of physics, i.e., quantum mechanics (QM), molecular mechanism (MM), statistical mechanics, thermodynamics, etc. QM methods are indispensable for chemical reactions and protein degradations. MM approaches are able to elucidate the conformational landscapes and flexibility patterns of proteins.[2] However, the all-electron or all-atom representations and long-time integrations lead to such an excessively large number of degrees of freedom that their application to real-time scale protein dynamics becomes prohibitively expensive. One way to reduce the number of degrees of freedom is to pursue time-independent formulations, such as normal mode analysis (NMA),[3–6] elastic network model (ENM),[7] including Gaussian network model (GNM)[8–10] and anisotropic network model (ANM).[11] Multiscale methods are some of the most popular approaches for studying protein structure, function, dynamics and transport.[12–15] Recently, we have introduced differential geometry based multiscale models for biomolecular structure, solvation, and transport.[16–19] A new interaction free approach, called flexibility-rigidity index (FRI), has

also been proposed for the estimation of the shear modulus in the theory of continuum elasticity with atomic rigidity (CEWAR) for biomolecules.[20]

A common feature of the above mentioned models for the study of structure-functions relationship of proteins is that they are structure or geometry based approaches.[21, 22] Mathematically, these approaches make use of local geometric information, i.e., coordinates, distances, angles, surfaces[22–24] and sometimes curvatures[25–27] for the physical modeling of biomolecular systems. Indeed, the importance of geometric modeling for structural biology,[21] biophysics[28, 29] and bioengineering[30–36] cannot be overemphasized. However, geometry based models are often inundated with too much structural detail and computationally extremely expensive. In many biological problems, such as the open or close of ion channels, the association or disassociation of ligands, and the assembly or disassembly of proteins, there exists an obvious topology-function relationship. In fact, just qualitative topological information, rather than quantitative geometric information is needed to understand many physical and biological functions. To state it differently, there is a *topology-function relationship* in many biomolecular systems. Topology is exactly the branch of mathematics that deals with the connectivity of different components in a space and is able to classify independent entities, rings and higher dimensional faces within the space. Topology captures geometric properties that are independent of metrics or coordinates. Topological methodologies, such as homology and persistent homology, offer new strategies for analyzing biological functions from biomolecular data, particularly the point clouds of atoms in macromolecules.

In the past decade, persistent homology has been developed as a new multiscale representation of topological features.[37–39] In general, persistent homology characterizes the geometric features with persistent topological invariants by defining a scale parameter relevant to topological events. Through filtration and persistence, persistent homology can capture topological structures continuously over a range of spatial scales. Unlike commonly used computational homology which results in truly metric free or coordinate free representations, persistent homology is able to embed geometric information to topological invariants so that “birth” and “death” of isolated components, circles, rings, loops, pockets, voids and cavities at all geometric scales can be monitored by topological measurements. The basic concept was introduced by Frosini and Landi,[40] and in a general form by Robins,[41] Edelsbrunner et al.,[37] and Zomorodian and Carlsson,[38] independently. Efficient computational algorithms have been proposed to track topological variations during the filtration process.[42–46] Usually, the persistent diagram is visualized through barcodes,[47] in which various horizontal line segments or bars are the homology generators lasted over filtration scales. It has been applied to a variety of domains, including image analysis,[48–51] image retrieval,[52] chaotic dynamics verification,[53, 54] sensor network,[55] complex network,[56, 57] data analysis,[58–62] computer vision,[50] shape recognition[63] and computational biology.[64–66] Compared with computational topology[67, 68] and/or computational homology, persistent homology inherently has an additional dimension, the filtration parameter, which can be utilized to embed some crucial geometric or quantitative information into the topological invariants. The importance of retaining geometric information in topological analysis has been recognized in a survey.[69] However, most successful applications of

persistent homology have been reported for qualitative characterization or classification. To our best knowledge, persistent homology has hardly been employed for quantitative analysis, mathematical modeling, and physical prediction. In general, topological tools often incur too much reduction of the original geometric/data information, while geometric tools frequently get lost in the geometric detail or are computationally too expensive to be practical in many situations. Persistent homology is able to bridge between geometry and topology. Given the big data challenge in biological science, persistent homology ought to be more efficient for many biological problems.

The objective of the present work is to explore the utility of persistent homology for protein structure characterization, protein flexibility quantification and protein folding stability prediction. We introduce the molecular topological fingerprint (MTF) as a unique topological feature for protein characterization, identification and classification, and for the understanding of the topology-function relationship of biomolecules. We also introduce all-atom and coarse-grained representations of protein topological fingerprints so as to utilize them for appropriate modeling. To analyze the topological fingerprints of alpha helices and beta sheets in detail, we propose the method of slicing, which allows a clear tracking of geometric origins contributing to topological invariants. Additionally, to understand the optimal cutoff distance in the GNM, we introduce a new distance based filtration matrix to recreate the cutoff effect in persistent homology. Our findings shed light on the topological interpretation of the optimal cutoff distance in GNM. Moreover, based on the protein topological fingerprints, we propose accumulated bar lengths to characterize protein topological evolution and quantitatively model protein rigidity based on protein topological connectivity. This approach gives rise to an accurate prediction of optimal characteristic distance used in the FRI method for protein flexibility analysis. Finally the proposed accumulated bar lengths are also employed to predict the total energies of a series of protein folding configurations generated by steered molecular dynamics.

The rest of this paper is organized as follows. Section 2 is devoted to fundamental concepts and algorithms for persistent homology, including simplicial complex, homology, persistence, Čech complex, Rips complex, filtration process, reduction algorithm, Euler characteristic, etc. To offer a pedagogic description, we discuss and illustrate the definition, generation and calculation of simplicial homology in detail. The persistent homology analysis of protein structure, flexibility and folding is developed in Section 3. Extensive examples, including alpha helices, beta sheets and beta barrel, are used to demonstrate the generation and analysis of protein topological fingerprints. Additionally, we utilize MTFs to explore the topology-function relationship of proteins. Protein flexibility and rigidity is quantitative modeled by MTFs. We further explore protein topological evolution by analyzing the trajectory of protein topological invariants during the protein unfolding. The quantitative prediction of protein folding stability is carried out over a series of protein configurations. This paper ends with some concluding remarks.

2 Theory and algorithm

In general, homology utilizes a topological space with an algebraic group representation to characterize topological features, such as isolated components, circles, holes and void. For a

given topological space $\mathbb{T}$ a *p*-dimensional hole in $\mathbb{T}$induces the corresponding homology group $H_p(\mathbb{T}$. For a point set of data, such as atoms in a protein, one wishes to extracted the original topological invariants in its continuous description. Persistent homology plays an important role in resolving this problem. By associating each point with an ever-increasing radius, a multi-scale representation can be systematically generated. The corresponding series of homology groups is capable of characterizing the intrinsic topology in the point set. Additionally, efficient computational algorithms have been proposed. The resulting persistent diagrams provide detailed information of the birth and death of topological features, namely, different dimensional circles or holes. In order to facilitate the detailed analysis of biomolecular systems, we briefly review basic concepts and algorithms relevant to persistent homology, including simplicial complex, Čech complex, Rips complex, filtration process, reduction algorithm, paring algorithm, etc. in this section. We illustrate several aspects including the definition, the generation and the computation of the simplicial homology with many simple examples.

2.1 Simplicial homology and persistent homology

Simplicial complex is a topological space consisting of vertices (points), edges (line segments), triangles, and their high dimensional counterparts. Based on simplicial complex, simplicial homology can be defined and further used to analyze topological invariants.

Simplicial complex—The essential component of simplicial complex *K* is a *k*-simplex, σ^k, which can be defined as the convex hall of *k* + 1 affine independent points in $\mathbb{R}^N$ (*N* > *k*). If we let υ_0, υ_1, υ_2, ⋯, υ_k be *k* + 1 affine independent points, a *k*-simplex $\sigma^k = \{\upsilon_0, \upsilon_1, \upsilon_2, \cdots, \upsilon_k\}$ can be expressed as

$$\sigma^k=\left\{\lambda_0\upsilon_0+\lambda_1\upsilon_1+\cdots+\lambda_k\upsilon_k|\sum_{i=0}^{k}\lambda_i=1;0\leq\lambda_i\leq 1,i=0,1,\cdots,k\right\}. \quad (1)$$

Moreover, an *i*-dimensional face of σ^k is defined as the convex hall formed by the nonempty subset of *i* + 1 vertices from σ^k (*k* > *i*). Geometrically, a 0-simplex is a vertex, a 1-simplex is an edge, a 2-simplex is a triangle, and a 3-simplex represents a tetrahedron. We can also define the empty set as a (−1)-simplex.

To combine these geometric components, including vertices, edges, triangles, and tetrahedrons together under certain rules, a simplicial complex is constructed. More specifically, a simplicial complex *K* is a finite set of simplicies that satisfy two conditions. The first is that any face of a simplex from *K* is also in *K*. The second is that the intersection of any two simplices in *K* is either empty or shared faces. The dimension of a simplicial complex is defined as the maximal dimension of its simplicies. The underlying space |*K* | is a union of all the simplices of *K*, i.e., $|K| = \cup_{\sigma^k \in K}\sigma^k$. In order to associate the topological space with algebra groups, we need to introduce the concept of chain.

Homology—A *k*-chain $[\sigma^k]$ is a linear combination $\sum_i^k \alpha_i\sigma_i^k$ of *k*-simplex σ_i^k. The coefficients α_i can be chosen from different fields such as, rational field $\mathbb{Q}$, integer field $\mathbb{Z}$, and prime integer field $\mathbb{Z}_p$ with prime number *p*. For simplicity, in this work the coefficients

α_i is chosen in the field of $\mathbb{Z}_2$, for which the addition operation between two chains is the modulo 2 addition for the coefficients of their corresponding simplices. The set of all k-chains of simplicial complex K together with addition operation forms an Abelian group $C_k(K, \mathbb{Z}_2)$. The homology of a topological space is represented by a series of Abelian groups.

Let us define the boundary operation ∂_k as $\partial_k : C_k \to C_{k-1}$. With no consideration of the orientation, the boundary of a k-simplex $\sigma^k = \{\upsilon_0, \upsilon_1, \upsilon_2, \cdots, \upsilon_k\}$ can be denoted as,

$$\partial_k \sigma^k = \sum_{i=0}^{k} \{\upsilon_0, \upsilon_1, \upsilon_2, \cdots, \hat{\upsilon}_i, \cdots, \upsilon_k\}. \quad (2)$$

Here $\{\upsilon_0, \upsilon_1, \upsilon_2, \cdots, \hat{\upsilon_i}, \cdots, \upsilon_k\}$ means that the $(k-1)$-simplex is generated by the elimination of vertex υ_i from the sequence. A key property of the boundary operator is that applying the boundary operation twice, any k-chain will be mapped to a zero element as $\partial_{k-1}\partial_k = \emptyset$. Also we have $\partial_0 = \emptyset$. With the boundary operator, one can define the cycle group and boundary group. Basically, the k-th cycle group Z_k and the k-th boundary group B_k are the subgroups of C_k and can be defined as,

$$Z_k = \mathrm{Ker}\partial_k = \{c \in C_k | \partial_k c = \varnothing\}, \quad (3)$$

$$B_k = \mathrm{Im}\partial_{k+1} = \{c \in C_k | \exists d \in C_{k+1} : c = \partial_{k+1} d\}. \quad (4)$$

Element in the k-th cycle group Z_k or the k-th boundary group B_k is called the k-th cycle or the k-th boundary. As the boundary of a boundary is always empty $\partial_{k-1}\partial_k = \emptyset$, one has $B_k \subseteq Z_k \subseteq C_k$. Topologically, the k-th cycle is a k dimensional loop or hole.

With all the above definitions, one can introduce the homology group. Specifically, the k-th homology group H_k is the quotient group generated by the k-th cycle group Z_k and k-th boundary group B_k: $H_k = Z_k/B_k$. Two k-th cycle elements are then called homologous if they are different by a k-th boundary element. From the fundamental theorem of finitely generated abelian groups, the k-th homology group H_k can be expressed as a direct sum,

$$H_k = Z \oplus \cdots \oplus Z \oplus Z_{p_1} \oplus \cdots \oplus Z_{p_n} = Z^{\beta_k} \oplus Z_{p_1} \oplus \cdots \oplus Z_{p_n}, \quad (5)$$

where β_k, the rank of the free subgroup, is the k-th Betti number. Here Z_{p_i} is torsion subgroup with torsion coefficients $\{p_i | i = 1, 2, \ldots, p_n\}$, the power of prime number. Therefore, whenever H_k is torsion free. The Betti number can be simply calculated by

$$\beta_k = \mathrm{rank} H_k = \mathrm{rank} Z_k - \mathrm{rank} B_k. \quad (6)$$

Topologically, cycle element in H_k forms a k-dimensional loop or hole that is not from the boundary of a higher dimensional chain element. The geometric meanings of Betti numbers in $\mathbb{R}^3$ are the follows: β_0 represents the number of isolated components, β_1 is the number of one-dimensional loop or circle, and β_2 describes the number of two-dimensional voids or

holes. Together, the Betti number sequence $\{\beta_0, \beta_1, \beta_2, \cdots\}$ describes the intrinsic topological property of the system.

Persistent homology—For a simplicial complex K, the filtration is defined as a nested sub-sequence of its subcomplexes,

$$\emptyset = K^0 \subseteq K^1 \subseteq \cdots \subseteq K^m = K. \quad (7)$$

The introduction of filtration is of essential importance and directly leads to the invention of persistent homology. Generally speaking, abstract simplicial complexes generated from a filtration give a multiscale representation of the corresponding topological space, from which related homology groups can be evaluated to reveal topological features. Furthermore, the concept of persistence is introduced for long-lasting topological features. The p-persistent k-th homology group K^i is

$$H_k^{i,p} = Z_k^i / (B_k^{i+p} \cap Z_k^i). \quad (8)$$

Through the study of the persistent pattern of these topological features, the so called persistent homology is capable of capturing the intrinsic properties of the underlying space solely from the discrete point set.

2.2 Simplicial complex construction and filtration

Čech complex, Rips complex and alpha complex—The concept of nerve is essential to the construction of simplicial complex from a given topological space. Basically, given an index set I and open set $\mathbf{U} = \{U_i\}_{i \in I}$ which is a cover of a point set $X \in \mathbb{R}^N$, i.e., $X \subseteq \{U_i\}_{i \in I}$, the nerve $\mathbf{N}$ of $\mathbf{U}$ satisfies two basic conditions. One of the conditions is that $\emptyset \in \mathbf{N}$. The other states that if $\cap_{j \in J} U_j \neq \emptyset$ for $J \subseteq I$, then one has $J \in \mathbf{N}$. In general, for a set of point cloud data, the simplest way to construct a cover is to assign a ball of certain radius around each point. If the set of point data is dense enough, then the union of all the balls has the capability to recover the underlying space.

The nerve of a cover constructed from the union of balls is a Čech complex. More specifically, for a point set $X \in \mathbb{R}^N$, one defines a cover of closed balls $\mathbf{B} = \{B(x,\varepsilon) \mid x \in X\}$ with radius ε and centered at x. The Čech complex of X with parameter ε is denoted as $\mathscr{C}(X, \varepsilon)$, which is the nerve of the closed ball set $\mathbf{B}$,

$$\mathscr{C}(X, \varepsilon) = \{\sigma \mid \cap_{x \in \sigma} B(x, \varepsilon) \neq \emptyset\}. \quad (9)$$

The condition for a Čech complex can be relaxed to generate a Vietoris-Rips complex, in which, a simplex σ is generated if the largest distance between any of its vertices is at most 2ε. Denote $\mathscr{R}(X, \varepsilon)$ the Vietoris-Rips complex, or Rips complex.[70] These two abstract complexes satisfy the relation,

$$\mathscr{C}(X, \varepsilon) \subset \mathscr{R}(X, \varepsilon) \subset \mathscr{C}(X, \sqrt{2}\varepsilon). \quad (10)$$

In practice, Rips complex is much more preferred, due to the above sandwich relation and its computational efficiency.

Čech complex and Rips complex are abstract complexes. Derived from computational geometry, alpha complex is also an important geometric concept. To facilitate the introduction, we review some basic definitions. Let X be a point set in Euclidean space $\mathbb{R}^d$. The Voronoi cell of a point $x \in X$ is defined as

$$V_x=\{u \in R^d | |u-x| \leq |u-x'|, \forall x' \in X\}. \quad (11)$$

The collection of all Voronoi cells forms a Voronoi diagram. Further, the nerve of the Voronoi diagram generates a Delaunay complex.

We define $R(x,\varepsilon)$ as the intersection of Voronoi cell V_x with ball $B(x,\varepsilon)$, i.e., $R(x,\varepsilon) = V_x \cap B(x,\varepsilon)$. The alpha complex $\mathscr{A}(X,\varepsilon)$ of point set X is defined as the nerve of cover $\cup_{x\in X} R(x,\varepsilon)$,

$$\mathscr{A}(X,\varepsilon)=\{\sigma | \cap_{x\in\sigma} R(x,\varepsilon) \neq \varnothing\}. \quad (12)$$

It can be seen that an alpha complex is a subset of a Delaunay complex.

General filtration processes—To construct a simplicial homology from a set of point cloud data, a filtration process is required.[42–46] For a specific system, the manner in which a suitable filtration is generated is key to the persistent homology analysis. In practice, two filtration algorithms, Euclidean-distance based and the correlation matrix based ones, are commonly used. These nitrations can be modified in many different ways to address physical needs as shown in the application part of this paper.

The basic Euclidean-distance based filtration is straightforward. One associates each point with an ever-increasing radius to form an ever-growing ball for each point. When these balls gradually overlap with each other, complexes can be identified from various complex construction algorithms described above. In this manner, the previously formed simplicial complex is an inclusion of latter ones and naturally, a filtration process is created. One can formalize this process by the use of a distance matrix $\{d_{ij}\}$. Here the matrix element d_{ij} represents the distance between atom i and atom j. For diagonal terms, it is nature to assume $d_{ij} = 0$. Let us denote the filtration threshold as a parameter ε. A 1-simplex is generated between vertices i and j if $d_{ij} \leq \varepsilon$. Similarly higher dimensional complexes can also be defined. Figure 1 demonstrates an Euclidean-distance based filtration process of an icosahedron.

Sometimes, in order to explore the topology-function relationship or illustrate a physical concept, such as cutoff distance, a modification to the distance based filtration is preferred. See Section 3.2.2 for an example.

Usually, the physical properties are associated with geometric or topological features. However, these features, more often than not, cannot be used directly to predict the physical characteristics. Instead, correlation functions, either form fundamental laws or experimental observation, should be employed. Based on the correlation matrix generated by these functions, one can build another more abstract type of filtration. The resulting persistent homology groups can be simply understood as the topological intrinsic properties of the object. In this manner, one has a powerful tool to explore and reveal more interesting

topology-function relationship and essential physical properties. Figure 2 demonstrates a correlation matrix based filtration process for fullerene C_{70}. The correlation matrix is generated from the geometry to topology mapping discussed in Section 3.2.1.

2.3 Computational algorithms for homology

For a given simplicial complex, we are interested in the related Betti numbers, which are topological invariants. In computational homology, the reduction algorithm is a standard method for Betti number evaluation. In this algorithm, the boundary operator is represented as a special matrix. Using invertible elementary row and column operations, this matrix can be further reduced to the Smith normal form. Finally, the Betti number can be expressed in terms of the rank of the matrix.

The matrix representation is essential to the reduction algorithm. For a boundary operator $\partial_i : C_i \rightarrow C_{i-1}$, under the chain group basis, it can be represented by an integer matrix M_i, which has i columns and $i - 1$ rows. Here entries in M_i are directly related to the field chosen. Elementary row and column operation, such as exchanging two rows (columns), multiplying a row (a column) with an invertible number, and replacing two rows (columns), can be employed to diagonalize the matrix M_i to the standard Smith normal form M_i = diag(a_1, a_2, ⋯, a_{i_n}). With this reduction algorithm, the rank M_i equals parameter i_n. From the definition of cycle group and boundary group, one has rank Z_i = rank C_i − rank M_i and rank B_i = rank M_{i+1}. Therefore the Betti number can be calculated as

$$\beta_i = \mathrm{rank} C_i - \mathrm{rank} M_i - \mathrm{rank} M_{i+1}. \quad (13)$$

For a large simplicial complex, the constructed matrix may seem to be cumbersome. Sometimes, the topological invariant called Euler characteristic χ can be helpful in the evaluation of Betti numbers. More specifically, for the k-th simplicial complex K, χ(K) is defined as

$$\chi(K) = \sum_{i=0}^{k} (-1)^i \mathrm{rank} C_i(K). \quad (14)$$

By using Eq. (13), the Euler characteristic can be also represented as

$$\chi(K) = \sum_{i=0}^{k} (-1)^i \mathrm{rank} \beta_i(K). \quad (15)$$

Since K is the k-th simplicial complex, one has Im $\partial_{k+1} = \emptyset$ and rank rank M_{k+1}= 0.As rank $\partial_0 = \emptyset$, one has rank $M_0 = 0$.

Proteins are often visualized by their surfaces of various definitions, such as van der Waals surfaces, solvent excluded surfaces, solvent accessible surfaces, minimal molecular surfaces[23] and Gaussian surfaces[22] at some given van der Waals radii. Therefore, another topological invariant, the genus number, is useful too. However, it is beyond the scope of the present work to elaborate this aspect.

To illustrate Euler characteristic, Betti number and reduction algorithm in detail, we have designed two toy models as shown in Fig. 3. Let us discuss in detail of the first three charts

of the figure, which are about three tetrahedron-like geometries. The first object is made of only points (0-simplex) and edge (1-simplex). Then face information (2-simplex) is added in the second chart. Further, a tetrahedron (3-simplex) is included in the third chart. This process resembles a typical filtration as the former one is the inclusion of the latter one. The same procedure is also used in the last three charts of Fig. 3 for a cube. Table 1 lists the basic properties of two simplicial complexes in terms of numbers of vertices, edges, faces and cells. Both Betti numbers and Euler characteristic are calculated for these examples.

For the filtration process of complicated point cloud data originated from a practical application, the calculation of the persistence of the Betti numbers is nontrivial. It is out of the scope of the present paper to discuss the Betti number calculation in detail. The interested reader is referred to the literature.[37, 58] In the past decade, many software packages have been developed based on various algorithms, such as Javaplex, Perseus, Dionysus etc. In this work, all the computations are carried out by using Javaplex[72] and the persistent diagram is visualized through barcodes.[47]

Figure 4 illustrates the persistent homology analysis of icosahedron (left chart) and fullerene C_{70} (right chart). Both distance based filtration as illustrated in Fig. 1 and correlation matrix based filtration as depicted in Fig. 2 are employed. For the icosahedron chart, there exist three panels corresponding to β_0, β_1 and β_2 from top to bottom. For β_0 number, originally 12 bars coexist, indicating 12 isolated vertices. As the filtration continues, 11 of them disappear simultaneously with only one survived and persisting to the end. Geometrically, due to the high symmetry, these 12 vertices connect with each other simultaneously at 2Å, i.e., the designed “bond length”. The positions where the bars terminate are exactly the corresponding bond lengths. Therefore, barcode representation is able to incorporate certain geometric information. As no one-dimensional circle has ever formed, no β_1 bar is generated. Finally, in the β_2 panel, there is a single bar, which represents a two-dimensional void enclosed by the surface of the icosahedron.

In fullerene C_{70} barcodes, there are 70 β_0 bars. Obviously, there are 6 distinct groups in β_0 bars due to the factor that there are 6 types of bond lengths in the C_{70} structure. Due to the formation of rings, β_1 bars emerge. There is a total of 36 β_1 bars corresponding to 12 pentagon rings and 25 hexagon rings. It appears that one ring is not accounted because any individual ring can be represented as the linear combination of all other rings. Note that there are 6 types of rings. Additionally, 25 hexagon rings further evolve to two-dimensional holes, which are represented by 25 bars in the β_2 panel. The central void structure is captured by the persisting β_2 bar.

3 Persistent homology analysis of proteins

In this section, the method of persistent homology is employed to study the topology-function relationship of proteins. Specifically, the intrinsic features of protein structure, flexibility, and folding are investigated by using topological invariants. In protein structure analysis, we compare an all-atom representation with a coarse-grained (CG) model. Two most important protein structural components, namely, alpha helices and beta sheets, are analyzed to reveal their unique topological features, which can be recognized as their

topological “ID”s or fingerprints. A beta barrel is also employed as an example to further demonstrate the potential of persistent homology in protein structure analysis.

In protein flexibility and rigidity analysis, many elegant methods, such as normal mode analysis (NMA), Gaussian network model (GNM), elastic network model (ENM), anisotropic network model (ANM), our molecular nonlinear dynamic (MND) and flexibility and rigidity index (FRI), have been proposed. Although they differ in terms of theoretical foundations and computational realization, they share a similar parameter called cut off distance or characteristic distance. The physical meaning of this parameter is the relative influence domain of certain atoms. Usually, for the CG model with each amino acid represented by its C_α atom, the optimized cut off distance is about 7 to 8 Å,[73] based on fitting with a large number of experimental B-factors. To provide a different perspective and unveil the topological significance of the cutoff distance or characteristic distance, simplicial complexes and related filtration processes are built up. Different parameter values induce dramatically distinguished topological patterns. Optimal characteristic distances are revealed from our persistent homology analysis.

To study topological features of protein folding, a simulated unfolding process is considered. We use a constant velocity pulling algorithm in our steered molecule dynamics to generate a family of configurations. Through the analysis of their topological invariants, we found that the accumulated bar length, which represents the total connectivity, continuously decreases in the protein unfolding process. As the relative stability of a protein is proportional to its topological connectivity, which is a topology-function relationship, the negative accumulated bar length can be used to describe the stability of a protein.

3.1 Topological fingerprints of proteins

Protein molecules often consist of one or more coiled peptide chains and have highly complicated 3D structures. Protein topological features include isolated entities, rings and cavities. However, each protein structure is unique. Our goal is to unveil protein intrinsic topologies and identify their molecular fingerprints. Alpha helices and beta sheets are basic structural components of proteins. Biologically, alpha helix is a spiral conformation, stabilized through hydrogen bond formed between the backbone N-H group and the backbone C=O group of four residues earlier in the protein polymer. Usually, the backbone of an alpha helix is right-handedly coiled. Each amino acid residue contributes about a 100 degree rotation in the helix structure. State differently, each spiral in the backbone is made of 3.6 amino acid residues. In contrast, beta sheet is a stretched polypeptide chain with about 3 to 10 amino acids. Usually, beta strands connect laterally with each other through the backbone hydrogen bonds to form a pleated sheet. Two adjacent beta strands can be parallel or anti-parallel to each other with a slightly different pattern due to the relative position between N-H group and C=O group. Many amyloidosis related diseases, such as mad cow disease and Alzheimer’s disease, are due to the insoluble protein aggregates and fibrils made of beta sheets.

In this section, two basic protein structure representations, i.e., an all-atom model and a CG model, are considered. For the all-atom model, various types of atoms, including H, O, C, N, S, P, etc., are all included and regarded as equally important in our computation. The all-

atom model gives an atomic description of the protein and is widely used in molecular dynamic simulation. In contrast, the CG model describes the protein molecule with a reduced number of degrees of freedom and is able to highlight important protein structure features. A standard coarse-grained representation of proteins is to represent each amino acid by the corresponding C_α atom. The CG model is efficient for describing large proteins and protein complexes.

Topological fingerprints of alpha helix and beta sheet

Protein structure data are available through the Protein Data Bank (PDB). To analyze an alpha helix structure, we download a protein of PDB ID: 1C26 and adopt an alpha helix chain with 19 residues. In our all-atom model, we do not distinguish between different types of atoms. Instead, each atom is associated with the same radius in the distance based filtration. The persistent diagram is represented by the barcode as depicted in right chart of Fig. 5. As discussed in Section 2.3, the β_0 bars can reveal the bond length information. Physically, for protein molecule, the bond length is between 1 to 2 Å, which is reflected in the distance based filtration. The occurrences of β_1 and β_2 bars are due to the loop, hole and void type of structures. Because the filtration process generates a large number of bars, it is difficult to directly decipher this high dimensional topological information. From the left chart of Fig. 5, it is seen that the alpha helix backbone has a regular spiral structure. However, residual atoms are quite crowd around the main chain and bury the spiral loop. To extract more geometric and topological details of the helix structure, we utilize the CG with each amino acid represented by its C_α atom. The results are demonstrated in the left chart of Fig. 6. As there are 19 residues in the alpha helix structure, only 19 atoms are used in the CG model and the corresponding barcode is dramatically simplified. From the right chart of Fig. 6, it is seen that there are 19 bars in β_0 panel and the bar length is around 3.8 Å, which is the average length between two C_α atoms. Additionally there are 16 bars in the β_1 panel. With similar birth time and persist length, these bars form a striking pattern. To reveal the topological meaning of these bars, we make use of a technique called slicing. Basically, we slice a piece of 4 C_α atoms from the back bone and study its persistent homology behavior. Then, one more C_α atom is added at a time. We repeat this process and generate the corresponding barcodes. The results are depicted in Fig. 7. It can be seen clearly that each four C_α atoms in the alpha helix form a one-dimensional loop, corresponding to a β_1 bar. By adding more C_α atoms, more loops are created and more β_1 bars are obtained. Finally, 19 residues in the alpha helix produce exactly 16 loops as seen in Fig. 6.

To explore the topological fingerprints of beta sheet structures, we extract two parallel beta strands from protein 2JOX. Figure 8 and Fig. 9 demonstrate the persistent homology analysis of all-atom model and CG model with the distance based filtration. Similar to the alpha helix case, in the all-atom representation, the generated barcode has a complicated pattern due to excessively many residual atoms. The barcode of the CG model, on the other hand, is much simpler with only 7 individual β_1 bars. Each of these bars is formed by two adjacent residue pairs. From the β_0 panel we can see that the lengths of most bars are still around 3.8 Å, i.e., the average length between two adjacent C_α atoms as discussed above. These bars end when the corresponding atoms are connected. However, there exists a unique β_0 bar which has a length about 4.1 Å. This bar reflects the shortest distance between two

closest adjacent two C_α atoms from two individual beta strands. With these geometric information, we can explain the mechanism of the birth and death of 7 individual β_1 bars. First, as the filtration begins, adjoined C_α atoms in the same strand form 1-simplex. After that, adjacent C_α atoms in two different strands connect with each other as the filtration continues, which leads to one-dimensional circles and β_1 bars. The further filtration terminates all the β_1 bars. There is no β_1 bar in the CG representation of beta sheet structures.

The persistent homology analysis of alpha helices and beta sheets reveals their topological features which are useful for deciphering protein fingerprints as demonstrated in the above example. Typically, CG model based filtration provides more global scale properties because local characteristics from amino acids is ignored. However, all-atom model based filtration can preserve more local scale topological information. For instance, there are two isolated bars around 2Å in the β_1 panel of Fig. 5. Meanwhile, there are also two individual bars around 2.5Å in the β_2 panel. These bars are the fingerprints of “PHE” or “TYR” types of amino acid residues. It turns out that there are two “PHE” amino acid residues in the alpha-helix structure. Similar fingerprints are of paramount importance for the identification of protein structural motifs, topological modeling of biomolecules and prediction of protein functions. However, these aspects are beyond the scope of the present paper.

Topological fingerprints of beta barrels

Having analyzed the topological fingerprints of alpha helix and beta sheet, we are interested in revealing the topological patterns of protein structures. As an example, a beta barrel molecule (PDB ID: 2GR8) is used. Figures 10**a** and **b** depict its basic structure viewed from two different perspectives. The sheet and helix structures in the protein are then extracted and demonstrated in magenta and bule colors in Figs. 10**c** and **d**, respectively. The coarse grain model is used. We show the persistent homology analysis of alpha helices in Fig. 11. It can been seen from the β_0 panel that nearly all the bar lengths are around 3.8Å, except three. Two of them persist to around 4.5Å and the other forever. This pattern reveals that there exist three isolated components when the filtration size is larger than 3.8Å and less than 4.5Å, which corresponds exactly to the number of individual alpha helices in the system. There are also 3 bars in the β_2 panel. Using our slicing technique, it can be found that each β_2 bar represents a void formed by one turn of three helices due to the high symmetry of the structure. Each of symmetric turns also generates 3 β_1 bar. To be more specific, a circle is formed between each two alpha helices at the groove of the turning part. Moreover, it can be noticed that at the left end three alpha helices are more close to each other, with three C_α atoms symmetrically distributed to form a 2-simplex during the filtration. Therefore, no β_1 bar is generated. However, when we slice down the helices, this pattern of three C_α atoms forming a 2-simplex only happens once at the eighth C_α counted from the left end. When this occurs, a one-dimensional cycle is terminated. As it is well known that usually 3.6 residues form a turn in an alpha helix, we have three turning structures, which give rise to 9 circles, i.e., 9 bars in the β_1 panel. Furthermore, a total of 44 (i.e., the number of β_0 bars) atoms in three alpha helices contributes to 35 (i.e., $44 - 3 * 3$) circles. Together with the terminated one-dimensional cycle at eighth C_α atom, we therefore have 43 β_1 bars, which agrees with the persistent homology calculation.

The persistent homology analysis of the beta-sheet structure is shown in Fig. 12. It is seen that in the β_0 panel, we have 12 bars that are longer than 3.8Å. These bars correspond to 12 isolated beta sheets. In the β_1 panel, there is a unique bar that lasts from around 4 to 16 Å. Obviously, this β_1 bar is due to the global hole of the beta barrel.

Except the longest β_1 bar, other β_1 bars are generated from every adjacent 4 C_α atoms as discovered in the earlier analysis of parallel beta sheet structure. There are 12 near parallel beta sheets. Due to the mismatch in the structure as shown Fig. 12(a) in detail, adjacent two sheets only contribute around 8 β_1 bars, which gives rise to $12 \times 8 = 96$ short-lived β_1 bars. Additionally, the global circle gives rise to another β_1 bar. Therefore, we predict 97 bars. The persistent homology calculation shows 98 β_1 bars. However, one of these bars has an extremely short life and is hardly visible. Therefore, there is very good consistency between our analysis and numerical computation.

3.2 Persistent homology analysis of protein flexibility

Rigidity and flexibility are part of protein functions. Theoretically, protein flexibility and rigidity can be studied based on fundamental laws of physics, such as molecular mechanics.[2] However, the atomic representation and long time integration involve an excessively large number of degrees of freedom. To avoid this problem, normal mode based models,[3–6] such as elastic network model (ENM),[7] anisotropic network model (ANM),[11] and Gaussian network model[8–10] have been proposed. Combined with the coarse-grained representation, they are able to access the flexibility of macromolecules or protein complexes.

Recently, we have proposed the molecular nonlinear dynamic (MND) model[74] and flexibility-rigidity index (FRI) theory[20] to analyze protein flexibility. The fundamental assumption of our methods is that, protein structures are uniquely determined by various internal and external interactions, while the protein functions, such as stability and flexibility, are solely determined by the structure. Based on this assumption, we introduce a key element, the geometry to topology mapping to obtain protein topological connectivity from its geometry information. Furthermore, a correlation matrix is built up to transform the geometric information into functional relations.

In this section, we provide a persistent homology analysis of protein flexibility. We present a brief review to a few techniques that are utilized in the present flexibility analysis. Among them, MND and FRI not only offer protein flexibility analysis, but also provide correlation matrix based filtration for the persistent homology analysis of proteins. Our results unveil the topology-function relationship of proteins.

3.2.1 Protein flexibility analysis

Normal mode analysis: Due to the limitation of computation power, MD or even coarse-grained MD sometimes falls short in the simulation of protein's real time dynamics, especially for macro-proteins or protein complexes, which have gigantic size and long-time-scale motions. However, if protein's relative flexibility and structure-encoded collective dynamics are of the major concern, MD simulation can be replaced by normal mode analysis, the related methods includes elastic network model, anisotropic network model,

Gaussian network model, etc. In these methods, pseudo-bonds/pseudo-springs are used to connect atoms within certain cutoff distance. The harmonic potential induced by the pseudo-bond/pseudo-spring network dominates the motion of the protein near its equilibrium state. Through low order eigenmodes obtained from diagonalizations of the Hessian matrix of the interaction potential, structure-encoded collective motion can be predicted along with the relative flexibility of the protein, which is measured experimentally by Debye-Waller factors. A more detailed description is available from review literature.[75–78]

Molecular nonlinear dynamics: The key element in our MND model is the geometry to topology mapping.[20, 74] Specifically, we denote the coordinates of atoms in the molecule studied as $\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_j, \cdots, \mathbf{r}_N$, where $\mathbf{r}_j \in \mathbb{R}^3$ is the position vector of the jth atom. The Euclidean distance between ith and jth atom r_{ij} can be calculated. Based on these distances, topological connectivity matrix can be constructed with monotonically decreasing radial basis functions. The general form is,

$$C_{ij}=w_{ij}\Phi(r_{ij},\eta_{ij}), i \neq j, \quad (16)$$

where w_{ij} is associated with atomic types, parameter η_{ij} is the atom-type related characteristic distance, and $\Phi(r_{ij}, \eta_{ij})$ is a radial basis correlation kernel.

The kernel definition is of great importance to the FRI model. From our previous experience, both exponential type and Lorentz type of kernels are very efficient. A generalized exponential kernel has the form

$$\Phi(r,\eta)=e^{-(r/\eta)^{\kappa}}, \kappa>0 \quad (17)$$

and the Lorentz type of kernels is

$$\Phi(r,\eta)=\frac{1}{1+(r/\eta)^{\upsilon}}. \upsilon>0. \quad (18)$$

The parameters κ, υ, and η are adjustable. We usually search over a certain reasonable range of parameters to find the best fitting result by comparing with experimental B-factors.

Under physiological conditions, proteins experience ever-lasting thermal fluctuations. Although the whole molecule can have certain collective motions, each particle in a protein has its own dynamics. It is speculated that each particle in a protein can be viewed as a nonlinear oscillator and its dynamics can be represented by a nonlinear equation.[74] The interactions between particles are represented by the correlation matrix 16. Therefore, for the whole protein of N particles, we set a nonlinear dynamical system as[74]

$$\frac{d\mathbf{u}}{dt}=\mathbf{F}(\mathbf{u})+\mathbf{E}\mathbf{u}, \quad (19)$$

where $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_j, \cdots, \mathbf{u}_N)^T$ is an array of state functions for N nonlinear oscillators (T denotes the transpose), $\mathbf{u}_j = (u_{j1}, u_{j2}, \cdots, u_{jn})^{\mathrm{T}}$ is an n-dimensional nonlinear function for the jth oscillator, $\mathbf{F}(\mathbf{u}) = (f(\mathbf{u}_1), f(\mathbf{u}_2), \cdots, f(\mathbf{u}_N))^T$ is an array of nonlinear functions of N oscillators, and

$$\mathbf{E}=\varepsilon\mathbf{C}\otimes\Gamma. \quad (20)$$

Here, ε is the overall coupling strength, $C = \{C_{ij}\}_{i,j=1,2,\cdots,N}$ is an $N \times N$ correlation matrix, and Γ is an $n \times n$ linking matrix.

It is found that, the transverse stability of the MND system gradually increases during the protein folding from disorder conformations to their well-defined natural structure. The interaction among particles leads to collective motions in a protein. The stronger the interaction is, the more unified dynamics will be. Eventually, the chaotic system assumes an intrinsically low dimensional manifold (ILDM) when the final folded state is reached, which indicates that protein folding tames chaos. To predict protein Debye-Waller factors, we introduced a transverse perturbation to the dynamics of each particle and record the relaxation time defined as the time used to recover the original state within a factor of 1/*e*, which measures the strength of particle-particle and particle-environment interactions. Therefore, the relaxation time is associated with protein flexibility. This method has also been successfully applied to the prediction of Debye-Waller factors.[20]

Flexibility rigidity index: The FRI theory is very simple. It can be directly derived from the correlation matrix. Basically, we define the atomic rigidity index μ_i as[20]

$$\mu_i=\sum_j^N w_{ij}\Phi(r_{ij},\eta_{ij}),\forall i=1,2,\cdots,N. \quad (21)$$

The physical interpretation is straightforward. The stronger connectivity an atom has, the more rigid it becomes. After summarizing over all atoms, one arrives at the averaged molecular rigidity index (MRI),

$$\bar{\mu}_{\mathrm{MRI}}=\frac{1}{N}\sum_{i=1}^N\mu_i. \quad (22)$$

It has been pointed out that this index is related to molecular thermal stability, compressibility and bulk modulus.[20]

We also defined the atomic flexibility index as the inverse of the atomic rigidity index,

$$f_i=\frac{1}{\mu_i},\forall i=1,2,\cdots,N. \quad (23)$$

An averaged molecular flexibility index (MFI) can be similar defined as the averaged molecular rigidity index

$$\bar{f}_{\mathrm{MFI}}=\frac{1}{N}\sum_{i=1}^N f_i. \quad (24)$$

We set $\eta_{ij} = \eta$ and $w_{ij} = 1$ for a CG model with one type of atoms.

The atomic rigidity index of each particle in a protein is associated with the particle's flexibility. The FRI theory has been intensively validated by comparing with the experimental data, especially the Debye-Waller factors.[20] Although it is very simple, its application to B-factor prediction yields excellent results. The predicted results are proved to be very accurate and this method is highly efficient. FRI can also be used to analyze the protein folding behavior.

3.2.2 Topology-function relationship of protein flexibility—Topological connectivity is employed in the elastic network models, or GNM for protein flexibility analysis. To this end, a cutoff distance r_c is often used to specify the spatial range of the connectivity in the protein elastic network. Each atom is assumed to be connected with all of its neighbor atoms within the designed cutoff distance by pseudo-springs or pseudo-bonds. Whereas atoms beyond the cutoff distance are simply ignored. In contrast, our MND model and FRI theory do not use a cutoff distance, but utilize a characteristic distance η to weight the distance effect in the geometry to topology mapping as shown in Eqs. (17) and (18). Atoms within the characteristic distance are assigned with relatively larger weights in the correlation matrix. It has been noted that the optimal cutoff distance can varies from protein to protein and from method to method. For a given method, an optimal cutoff distance can be obtained by statistically averaging over a large number of proteins. Such optimal cutoff distance is about 7 to 8Å for GNM and around 13 to 15Å for ANM.[73] No rigorous analysis or explanation has been given for optimal cutoff distances.

In this section, we explore the topology-function relationship of proteins. First, we present a persistent homology interpretation of optimal cutoff distances in GNM. Additionally, we also provide a quantitative prediction of optimal characteristic distances in MND and FRI based on persistent homology. To this end, we develop a new cutoff distance based filtration method to transfer a protein elastic network into a simplicial complex at each cutoff distance. The resulting patterns of topological invariants shed light on the existence of optimal cutoff distances. We propose a new persistent homology based physical model to predict optimal characteristic distances in MND and FRI.

Persistent homology analysis of optimal cutoff distance—Protein elastic network models usually employ the coarse-grained representation and do not distinguish between different residues. We assume that the total number of C_α atoms in the protein is N, and the distance between ith. and jth C_α atoms is d_{ij}. To analysis the topological properties of protein elastic networks, we propose a new distance matrix $\mathbf{D} = \{d_{ij} | i = 1,2, \cdots, N; j = 1,2, \cdots, N\}$

$$d_{ij} = \begin{cases} d_{ij}, d_{ij} \leq r_c; \\ d_\infty, d_{ij} > r_c, \end{cases} \quad (25)$$

where d_∞ is a sufficiently large value which is much larger than the final filtration size and r_c is a given cutoff distance. Here d_∞ is chosen to ensure that atoms beyond the cutoff distance r_c will never form any high order simplicial complex during the filtration process. Consequently, the resulting persistent homology shares the same topological connectivity property with the elastic network model. With the barcode representation of topological

invariants, the proposed persistent homology analysis gives rise to an effective visualization of topological connectivity.

To illustrate our persistent homology analysis, we consider two proteins, 1GVD and 3MRE, with 52 and 383 residues respectively, as shown in Fig. 13. With different cutoff distances, both the constructed elastic networks and persistent homology simplicial complexes demonstrate dramatically different properties. The resulting persistent homology analysis using the proposed filtration (25) for protein 1GVD and 3MRE are illustrated in Figs. 14 and 15, respectively. It can be seen that when the cutoff distance is 3Å, all C_α atoms are isolated from each other, and β_0 bars persist forever. This happens because the average distance between two adjacent C_α atoms is about 3.8Å, as discussed earlier. This particular distance also explains the filtration results in Figs. 14b and 15b at r_c = 4 Å. As the adjacent C_α atoms connect with each other, only a single β_0 bar survives. However, no nontrivial complex is formed at r_c = 4 Å. When the cutoff distance increases to 5Å, a large number of β_1 bars is produced and persists beyond the filtration size limit. Topologically, this means that almost all the generated loops never disappear during the filtration. This happens because we artificially isolate atoms with distance larger than r_c = 5Å in our filtration matrix (25). Physically, without the consideration of long distance interactions, local structural effects are over-amplified in this setting. With the increase of the cutoff distance r_c, the number of persistent β_1 bars drops. When r_c is set to about 12 Å, almost no persistent β_1 bar can be found for both 1GVD and 3MRE. It should also note that for a small protein like 1GVD, the number of persistent β_1 bars falls quickly. While a larger protein with a larger number of β_1 bars, the reduction in the number of ring structures is relatively slow. However, the further increase of the r_c value does not change the persistent homology behavior anymore because all significant geometric features (i.e., isolated components, circles, voids, and holes) are already captured by the existing network at about $r_c \approx 12$Å.

To understand the physical impact of the topological connectivity found by persistent homology analysis, we analyze GNM predictions of protein B-factors at various cutoff distances. The accuracy of the GNM predictions is quantitatively accessed by correlation coefficient (CC)

$$\mathrm{CC}=\frac{\sum_{i=1}^{N}(B_i^e-\bar{B}^e)(B_i^t-\bar{B}^t)}{[\sum_{i=1}^{N}(B_i^e-\bar{B}^e)^2\sum_{i=1}^{N}(B_i^t-\bar{B}^t)^2]^{1/2}}, \quad (26)$$

where $\{B_i^t, i=1,2,\cdots,N\}$ are a set of predicted B-factors by using the proposed method and $\{B_i^e, i=1,2,\cdots,N\}$ are a set of experimental B-factors extracted from the PDB file. Here B^t and B^e the statistical averages of theoretical and experimental B-factors, respectively. Expression (26) is used for the correlation analysis of other theoretical predictions as well.

We plot the CC with respect to the cutoff distance in the right charts of Figs. 14 and 15 for proteins 1GVD and 3MRE, respectively. For both cases, when the cutoff distance is small than 3.8Å, no CC is obtained because pseudo-spring/pseudo-bond is not constructed and GNM is not properly set. As r_c increases to around 4Å, we acquire relatively small CC

values. Further increase of r_c will lower CC values until it reaches the bottom at around r_c = 5Å, where, as discussed in our persistent homology analysis, the influence of the local topological connectivity is over-estimated. Once the cutoff distance is larger than 5Å, the CC values begin to increase dramatically until it reach the peak value around $r_c \approx 7$Å, where the impact of local connectivity and global connectivity reaches an optimal balance. The CC values fluctuate as the cutoff distance increases further, due to improper balances in local connectivity and global connectivity.

From the above analysis, it can seen that if the cutoff distance used in elastic network models is smaller than 5Å, the constructed network is over-simplified without any global connectivity contribution. At the same time, if the cutoff distance is larger than 14Å, excessive global connections are included, which leads to a reduction in prediction accuracy, especially for small proteins. To be specific, by excessive connections we mean that a given atom is connected by using elastic springs with too many remote atoms. Even in the range of 6 to 14Å, there is always a tradeoff between excessive connection in certain part of the protein and lack of connection in the other regions. Also the relative size and the intrinsic topological properties of a protein should be considered when choosing the optimal cutoff distance. Although the selection of the optimal cutoff distance is complicated by many issues,[73] one can make a suitable choice by the proposed persistent homology analysis. For instance, in the above two cases, when the cutoff distance is around 7 to 9Å, the major global features in β_1 panel have already emerged and thus the selection of r_c = 7 Å to r_c = 9Å will generate a reasonable prediction. Therefore, our persistent homology analysis explains the optimal range of cutoff distances (i.e., r_c = 7 Å to 8Å) for GNM in the literature[73] very well.

Persistent homology prediction of optimal characteristic distance—Unlike elastic network models which utilize a cutoff distance, the MND method and FRI theory employ a characteristic distance η in their correlation kernel. The characteristic distance has a direct impact in the accuracy of protein B-factor prediction. Similar to the cutoff distance in GNM, the optimal characteristic distance varies from protein to protein, although an optimal value can be found based on a statistical average over hundreds of proteins.[20] In this section, we propose a persistent homology model to predict the optimal characteristic distance.

Appropriate filtration process is of crucial importance to the persistent homology analysis. To accurately predict optimal characteristic distance for MND and FRI models, we introduce a new filtration matrix $\{M_{ij} | i = 1, 2, \cdots, N; j = 1, 2, \cdots, N\}$ based on a modification of the correlation matrix (16) of MND and FRI

$$M_{ij} = \begin{cases} 1 - \Phi(r_{ij}, \eta_{ij}), i \neq j, \\ 0, i = j, \end{cases} \quad (27)$$

where $0 \leq \Phi(r_{ij}, \eta_{ij}) \leq 1$ is defined in Eqs. (17) or (18). To avoid confusion, we simply use the exponential kernel with parameter $\kappa = 2$ in the present work. The visualization of this new correlation matrix for fullerence C_{70} is given in Fig. 2.

As the filtration continues, atoms with shorter distances and thus lower M_{ij} values will form higher complexes first just like situation in the distance based filtration. However, when characteristic distance varies, the formation of simplicial complex or topological connectivity changes too. To illustrate this point, we use protein 1YZM as an example. The related persistent connectivity patterns in term of β_1 are depicted in Fig. 16. From the analysis of these topological invariants, several interesting observations can be made. First, it can be seen that our persistent homology results obtained with η = 2Å, 6Å, and 16 Å in **b, c**, and **d** share certain similarity with the β_1 pattern of the distance based filtration in **a**. This similarity is most obvious when η values are 6Å and 16 Å as shown in **c** and **d**. Second, results differ much in their scales or persistent durations. The β_1 bars in **b**, **c** and **d** are located around regions [0.996, 1], [0.4,0.8] and [0.05,0.2], respectively. Third, the global behavior is captured in all cases and the local connectivity is not over-emphasized. This aspect is different from that of the cutoff distance based filtration discussed previously and implies the robustness of the correlation matrix based filtration. However, some bars are missing for certain η values. Specifically, the β_1 barcode in Fig. 16**b** does not have all the bars appeared in other cases, which leads to the underestimation of certain protein connectivity.

To quantitatively analyze protein connectivity and predict optimal characteristic distance, we propose a physical model based on persistent homology analysis. We define accumulation bar lengths A_j as the summation of lengths of all the bars for β_j,

$$A_j = \sum_{i=1} L_i^{\beta_j}, j = 0, 1, 2, \quad (28)$$

where $L_i^{\beta_j}$ is the length of the *i*th bar of the *j*-th Betti number. To reveal the influence of the characteristic distance, we compute the accumulation bar length A_1 over a wide range of filtration parameter η. We vary the value of η from 1Å to 21Å, for protein 1YZM and compare the accumulated bar length A_1 with the CC values obtained with FRI over the same range of η. The exponential kernel with parameter κ = 2 is used in both the FRI method and our persistent homology model. Results are displayed in Fig. 17. It can be seen from the figure that two approaches share the same general trend in their behavior as η is increased. Specifically, when η is less than 3Å, the correlation coefficient is much lower than other values. Topologically, this is directly related to the absence of the certain bars in the persistent barcode as depicted in Fig. **17b**. With the increase of η, both CC and A_1 reach their maximum around η = 6Å. The further increase of η leads to the decrease of both CC and A_1.

We now analyze the aforementioned behavior from the topological point of view. The role of η in the FRI model is to scale the influence of atoms at various distances. An optimal η in the FRI model balances the contributions from local atoms and nonlocal atoms, and offers the best prediction of protein flexibility. In contrast, parameter η in the correlation matrix based filtration is to impact the birth and death of each given *k*-complex. For example, a pair of 2-complexes that do not coexist at a given cutoff distance in the distance based filtration might coexist at an appropriate η value in the correlation matrix based filtration. A maximum A_1 means the largest amount of coexisting 2-complexes (i.e., ring structures) at an

appropriate η value, which implies protein structural compactness and rigidity. Since the same kernel and the same η are used in the FRI model and the persistent homology model (i.e., accumulation bar length), it is natural for the η corresponding to the maximum A_1 to be the optimal characteristic distance in the FRI prediction.

To further validate our topological analysis and prediction, a set of 30 proteins are chosen and their PDB IDs are listed in Table 2. Two methods, namely FRI and MND, are employed for the flexibility analysis via the B-factor prediction. The persistent homology analysis is carried out via the accumulation bar length A_1. We use the exponential kernel with parameter $\kappa = 2$ for FRI, MND and A_1 calculations. The average correlation coefficients of MND and FRI methods are obtained by averaging over 30 proteins at each given η value. We compare these averaged CC values with the average accumulated length over the same set of 30 proteins. These results are illustrated in Fig. 18. It can be seen that the average CC values obtained from FRI and MND behave similarly as η increases. They dramatically increase when η goes beyond 3Å, and reach the peak before decrease at large characteristic distances. The best correlation coefficient is achieved at about 9Å and 8Å for MND and FRI models, respectively. The average accumulation bar length A_1 behaves in a similar manner. However, its peak value is around 6Å, which is consistent with our earlier finding with protein 1YZM. The deviation between optimal characteristic distance in protein the flexibility analysis and the "optimal filtration parameter η" is about 2Å to 3Å. We believe this deviation is due to several aspects. First, cofactors like metal ions, ligands, and small sugar molecules, play important role in protein stability and flexibility. Without any consideration of cofactors, our models may offer optimal B-factor prediction at wrong characteristic distances. Additionally, due to the limitation of our computational power, only relatively small sized proteins are considered in our test set. This may results in a lack of representation for relatively large proteins. Finally, the present persistent homology model is based only on β_1 numbers and may be improved by including β_2 as well. In spite of these issues, our persistent homology analysis successfully captures the basic correlation coefficient behavior. It provides an explanation for both the dramatic increase of correlation coefficients in small η values and the slow decrease in large η values. The predicted optimal characteristic distance value is also in a reasonable range.

3.3 Persistent homology analysis of protein folding

Protein folding produces characteristic and functional three-dimensional structures from unfolded polypeptides or disordered coils. Although Anfinsen's dogma[1] has been challenged due to the existence of prions and amyloids, most functional proteins are well folded. The folding funnel hypothesis associates each folded protein structure with a global minimum of the Gibbs free energy. Unfolded conformations have higher energies and are thermodynamically unstable, albeit they can be kinetically favored.

The protein folding process poses astonishing challenges to theoretical modeling and computer simulations. Despite the progress at the protein structure determination, the mechanism that how the polypeptide coils into its native conformation remains a puzzling issue, mainly due to the complexity and the stochastic dynamics involved in the process. Currently, experimental tools, such as atomic force microscopy, optical tweezers, and bio-

membrane force probe, have been devised to give information about unfolding force distribution, stable intermediates and transitional nonnative states. However, these approaches have a limited utility for unstable intermediate structures. Using the steered molecular dynamics (SMD), more details such as some possible folding or unfolding pathway can be obtained.

Protein folding and unfolding involve massive changes in its local topology and global topology. Protein topological evolution can be tracked by the trajectory of protein topological invariants. Typically, the folding of an amino acid polymer chain leads to dramatic increase in both 2-complexes and 3-complexes at appropriate filtration parameters. Therefore, persistent homology should provide an efficient tool for both qualitative characterization and quantitative analysis of protein folding or unfolding.

In this section, we simulate the unfolding process with the constant velocity pulling algorithm of SMD. Intermediate configurations are extracted from the trajectory. Then, we employ the persistent homology to reveal the topological features of intermediate configurations. Furthermore, we construct a quantitative model based on the accumulation bar length A_1 to predict the energy and stability of protein configurations, which establishes a solid topology-function relationship of proteins.

3.3.1 Steered molecular dynamics—Usually, the SMD is carried out through one of three ways: high temperature, constant force pulling, and constant velocity pulling.[79–81] The study of the mechanical properties of protein FN-III_{10} provides information of how to carefully design the SMD. It is observed that the original implicit solvent models for SMD tend to miss friction terms. For implicit solvent models, the design of the water environment is still nontrivial. If a water sphere is used, water deformation requires additional artificial force. Further, the unfolding process may extend the protein out of the boundary of the initial sphere. Therefore, it is believed that using a large box which can hold the stretched protein is a more reliable way if one can afford the computational cost.[82]

In the present work, the molecular dynamical simulation tool NAMD is employed to generate the partially folded and unfolded protein conformations. Two processes are involved, the relaxation of the structure and unfolding with constant velocity pulling. At first, the protein structure is downloaded from the PDB. Then, it is solvated with a water box which has an extra 5Å layer, comparing with the minimal one that hold the protein structure. The standard minimization and equilibration process is employed. Basically, a total of 15000 time steps of equilibration iterations is carried out with the periodic boundary condition after 10000 time steps of initial energy minimization. The length of each time step is 2fs in our simulation.

The setting of the constant velocity pulling is more complicated. First, pulling points where the force applied should be chosen. Usually, the first C_α atom is fixed and a constant pulling velocity is employed on the last C_α atom along the direction connecting these two points. The identification of the pulling points can be done by assigning special values to B-factor term and occupancy term in PDB data. Additionally, the pulling parameters should be carefully assigned. As proteins used in our simulation are relatively small with about 80

residues. The spring constant is set as 7 kcal/molÅ^2, with 1 kcal/molÅ^2/ equaling 69.74 pN Å. The constant velocity is 0.005Å per time step. As many as 30000 simulation steps for protein 1I2T and 40000 for 2GI9 are employed for their pulling processes. We extract 31 conformations from the simulation results at an equal time interval. The total energies (kcal/mol) are computed for all configurations. For each pair of configurations, their relative values of total energies determine their relative stability. A few representative conformations of unfolding 1I2T are depicted in Fig. 19. Obviously, topological connectivities, i.e., 2-complexes and 3-complexes, reduce dramatically from conformation Fig. 19a to conformation Fig. 19g.

3.3.2 Persistent homology for protein folding analysis—The steered molecular dynamics (SMD) discussed in Section 3.3.1 is used to generate the protein folding process. Basically, by pulling one end of the protein, the coiled structure is stretched into a straight-line like shape. It is found that during the unfolding process, the hydrogen bonds that support the basic protein configuration are continuously broken. Consequently, the number of high order complexes that may be formed during the filtration decreases because of protein unfolding. To validate our hypothesis, we employ the coarse-grained model in our persistent homology analysis although the all-atom model is used in our SMD calculations. The benefit of using the coarse-grained model is that the number of β_1 and β_2 is zero for a straight-line like peptide generated by SMD. We compute topological invariants via the distance based filtration for all 31 configurations of protein 1I2T. The persistent homology barcodes for frames 1, 10, 20 and 30 are illustrated in Fig. 20. We found that as the protein unfolded, their related β_1 value decreases. For four protein configurations in Fig. 20, their β_1 values are 47, 19, 13 and 5, respectively. Also there is a cavity in configuration 1. A comparison between configurations in Fig. 19 and their corresponding barcodes in Fig. 20 shows an obvious correlation between protein folding/unfolding and its topological trait. Therefore, topology and persistent homology are potential tools for protein folding characterization.

Additionally, as a protein unfolds, its stability decrease. State differently, protein becomes more and more unstable during its unfolding process, and its total energy becomes higher during the SMD simulation. As discussed above, the first Betti number decreases as protein unfolds. Therefore, there is a strong anti-correlation between protein total energy and its first Betti number during the protein unfolding process. However, using the least square fitting, we found that this linear relation is not highly accurate with a correlation coefficient about 0.89 for 31 configurations of 1I2T. A more robust quantitative model is to correlate protein total energy with the negative accumulation bar length of $\beta_1(A_1^- = -A_1)$. Indeed, a striking linear relation between the total energy and A_1^- can be found. We demonstrate our results in the left chart of Fig. 21. Using a linear regression algorithm, a correlation coefficient about 0.947 can be obtained for 31 configurations of protein 1I2T. To further validate the relation between the negative accumulation bar length and total energy, the correlation matrix based filtration process is employed. We choose the exponential kernel with optimized parameter $\kappa = 2$ and $\eta = 7$Å. The results are illustrated in the right chart of Fig. 21. A linear correlation is found with the CC value of 0.944.

To further validate our persistent homology based quantitative model for the stability analysis of protein folding, we consider protein 2GI9. The same procedure described above is utilized to create 31 configurations. We use both distance based filtration and correlation matrix based filtration to compute A_1^- for all the extracted intermediate structures. Our results are depicted in Fig. 22. Again the linear correlation between the energy prediction using negative accumulation bar length of the first Betti number and total energy is confirmed. The CC values are as high as 0.972 and 0.971, for distance based filtration and correlation matrix based filtration, respectively.

4 Concluding remarks

Persistent homology is a relatively new tool for topological characterization of signal, image and data, which are often corrupted by noise. Compared with the commonly used techniques in computational topology and computational homology, persistent homology incorporates a unique filtration process, through which, a sequence of nested simplicial complexes are generated by continuously enlarging a filtration parameter. In this manner, a multi-scaled representation of the underlying topological space can be constructed and further utilized to reveal the intrinsic topological properties against noise. Like other topological approaches, persistent homology is able to dramatically reduce the complexity of the underlying problem and offer topological insight for geometric structures and/or intrinsic features that last over multiple length scales. Additionally, thanks to the introduction of the filtration process, persistent homology is capable of reintroducing the geometric information associated with topological features like isolated components, loops, rings, circles, pockets, holes and cavities. The most successful applications of persistent homology in the literature have been about qualitative characterizations or classifications in the past. Indeed, there is hardly any successful quantitative model based on topology in the literature, because topological invariants preclude geometric description. It is interesting and desirable to develop quantitative models based on persistent homology analysis. This work introduces persistent homology to protein structure characterization, flexibility prediction and folding analysis. Our goal is to develop a mathematical tool that is able to dramatically simplify geometric complexity while incorporate sufficient geometric information in topological invariants for both qualitative characterization, identification and classification and quantitative understanding of the topology-function relationship of proteins.

To establish notation and facilitate our quantitative modeling, we briefly summarize the background of persistent homology. Simplicial complex, homology, persistent homology, filtration and their computational algorithms are briefly reviewed. To be pedagogic, we illustrate persistent homology concepts with a few carefully designed toy models, including a tetrahedron, a cube, an icosahedron and a C_{70} molecule. The relation between topological invariants, namely Betti numbers and Euler characteristic, is discussed using some simple models. For sophisticated biomolecular systems, we utilize both the all-atom model and coarse-grained model to deliver a multi-representational persistent homology analysis. Molecular topological fingerprints (MTFs) based on the persistence of molecular topological invariants are extracted.

Proteins are the most important molecules for living organism. The understanding of the molecular mechanism of protein structure, function, dynamics and transport is one of the most challenging tasks of modern science. In order to understand protein topological properties, we study the topological fingerprints of two most important protein structural components, namely alpha helices and beta sheets. To understand the geometric origin of each topological invariant and its persistence, we develop a method of slicing to systematically divide a biomolecule into small pieces and study their topological traits. The topological fingerprints of alpha helices and beta sheets are further utilized to decipher the topological property and fingerprint of a beta barrel structure. Uncovering the connection between topological invariants and geometric features deepens our understanding of biomolecular topology-function relationship.

Traditionally, long-lived persistent bars in the barcode have been celebrated as intrinsic topological features that persistent homology was invented for, while other short-lived bars are generally regarded as useless noise. However, from our analysis, it is emphasized that all features generated from persistent homology analysis are equally important because they represent the topological fingerprint of a given protein structure. Therefore, the birth and death of each *k*-complex uniquely represent either a local or a global geometric feature. It is all of these topological features that make the MTF unique for each biomolecule. Just like nuclear magnetic resonance (NMR) signals or x-ray crystallography data, topological fingerprints are a new class of biometrics for the identification, characterization and classification of biomolecules.

The rigidity of a protein is essentially determined by its non-covalent interactions and manifests in the compactness of its three-dimensional (3D) structure. Such a compactness can be measured by the topological connectivity of the protein polymer chain elements, i.e., amino acid residues. Consequently, topological invariants, such as the first Betti number, give rise to a natural description of protein rigidity and flexibility. We therefore are able to reveal the topology-function relationship. Elastic network models, such as Gaussian network model, have been widely used for protein flexibility analysis. However, the performance of these methods depends on the cutoff distance which determines whether a given pair of atoms can be connected by an elastic spring in the model. In practical applications, the selection of a cutoff distance is empirical. We propose a new cutoff distance type of filtration matrix. The resulting topological diagrams shed light on the optimal cutoff distance used in the protein B-factor prediction with the Gaussian network model.

The aforementioned persistent homology analyses are descriptive and qualitative. One of major goals of the present work is to exploit the geometric information embedded in topological invariants for quantitative modeling. To this end, we propose a correlation matrix based filtration to incorporate geometric information in topological invariants. We define accumulated bar length by summing over all the bars of the first Betti number. We assume that the accumulated bar length correlates linearly to protein rigidity due to hydrogen bond strength, hydrophobic effects, electrostatic and van der Waals interactions. In particular, we investigate the dependence of the accumulated bar length on the characteristic distance used in our filtration process. It is found that the location of the

maximum accumulated bar length gives an accurate prediction of the optimal characteristic distance for the flexibility and rigidity index (FRI) analysis[20] of protein temperature factors.

To further exploit persistent homology for quantitative modeling, we consider protein folding which is an essential process for proteins to assume well-define structure and function. Our basic observation is that well-folded proteins, especially well-folded globular proteins, have abundant non-covalent bonds due to hydrogen bonds and van der Waals interactions, which translates into higher numbers of topological invariants, particularly, a large measurement of the first Betti number. Additionally, the funnel theory of protein folding states that a well-folded protein, i.e., the native structure, has the lowest free energy. In contrast, unfolded protein structures have less numbers of topological invariants and higher free energies. Based on this analysis, we propose a persistent homology based model to characterize protein topological evolution and predict protein folding stability. We correlate the negative accumulated bar length of the first Betti number to the protein total energy for a series of protein configurations generated by the steered molecular dynamics. As such the evolution of topological invariants in the protein folding/unfolding process is tracked. Our persistent homology based model is found to provide an excellent quantitative prediction of protein total energy during the protein folding/unfolding process.

Acknowledgments

This work was supported in part by NSF grants IIS-1302285 and DMS-1160352, NIH grant R01GM-090208 and MSU Center for Mathematical Molecular Biosciences Initiative. The authors acknowledge the Mathematical Biosciences Institute for hosting valuable workshops.

References

1. Anfinsen CB. Einfluss der configuration auf die wirkung den. Science. 1973; 181:223–230. [PubMed: 4124164]
2. McCammon JA, Gelin BR, Karplus M. Dynamics of of folded proteins. Nature. 1977; 267:585–590. [PubMed: 301613]
3. Go N, Noguti T, Nishikawa T. Dynamics of a small globular protein in terms of low-frequency vibrational modes. Proc. Natl. Acad. Sci. 1983; 80:3696–3700. [PubMed: 6574507]
4. Tasumi M, Takenchi H, Ataka S, Dwidedi AM, Krimm S. Normal vibrations of proteins: Glucagon. Biopolymers. 1982; 21:711–714. [PubMed: 7066480]
5. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem. 1983; 4:187–217.
6. Levitt M, Sander C, Stern PS. Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme. J. Mol. Biol. 1985; 181(3):423–447. [PubMed: 2580101]
7. Tirion MM. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys. Rev. Lett. 1996; 77:1905–1908. [PubMed: 10063201]
8. Flory PJ. Statistical thermodynamics of random networks. Proc. Roy. Soc. Lond. A. 1976; 351:351–378.
9. Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Folding and Design. 1997; 2:173–181. [PubMed: 9218955]
10. Bahar I, Atilgan AR, Demirel MC, Erman B. Vibrational dynamics of proteins: Significance of slow and fast modes in relation to function and stability. Phys. Rev. Lett. 1998; 80:2733–2736.
11. Atilgan AR, Durrell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys. J. 2001; 80:505–515. [PubMed: 11159421]

12. Cui Q. Combining implicit solvation models with hybrid quantum mechanical/molecular mechanical methods: A critical test with glycine. Journal of Chemical Physics. 2002; 117(10): 4720.
13. Zhang Y, Yu H, Qin JH, Lin BC. A microfluidic dna computing processor for gene expression analysis and gene drug synthesisn. Biomicrofluidics. 2009; 3(044105)
14. Tian WF, Zhao Shan. A fast ADI algorithm for geometric flow equations in biomolecular surface generations. International Journal for Numerical Methods in Biomedical Engineering. 2014; 30:490–516. [PubMed: 24574191]
15. Geng W, Wei GW. Multiscale molecular dynamics using the matched interface and boundary method. J Comput. Phys. 2011; 230(2):435–457. [PubMed: 21088761]
16. Wei GW. Differential geometry based multiscale models. Bulletin of Mathematical Biology. 2010; 72:1562–1622. [PubMed: 20169418]
17. Wei, Guo-Wei; Zheng, Qiong; Chen, Zhan; Xia, Kelin. Variational multiscale models for charge transport. SIAM Review. 2012; 54(4):699–754. [PubMed: 23172978]
18. Chen, Duan; Chen, Zhan; Wei, GW. Quantum dynamics in continuum for proton transport II: Variational solvent-solute interface. International Journal for Numerical Methods in Biomedical Engineering. 2012; 28:25–51. [PubMed: 22328970]
19. Wei, Guo-Wei. Multiscale, multiphysics and multidomain models I: Basic theory. Journal of Theoretical and Computational Chemistry. 2013; 12(8):1341006.
20. Xia KL, Opron K, Wei GW. Multiscale multiphysics and multidomain models — Flexibility and rigidity. Journal of Chemical Physics. 2013; 139:194109. [PubMed: 24320318]
21. Feng, Xin; Xia, Kelin; Tong, Yiying; Wei, Guo-Wei. Geometric modeling of subcellular structures, organelles and large multiprotein complexes. International Journal for Numerical Methods in Biomedical Engineering. 2012; 28:1198–1223. [PubMed: 23212797]
22. Zheng Q, Yang SY, Wei GW. Molecular surface generation using PDE transform. International Journal for Numerical Methods in Biomedical Engineering. 2012; 28:291–316. [PubMed: 22582140]
23. Bates PW, Wei GW, Zhao Shan. Minimal molecular surfaces and their applications. Journal of Computational Chemistry. 2008; 29(3):380–391. [PubMed: 17591718]
24. Bates PW, Chen Z, Sun YH, Wei GW, Zhao S. Geometric and potential driving formation and evolution of biomolecular surfaces. J. Math. Biol. 2009; 59:193–231. [PubMed: 18941751]
25. Chen Z, Baker NA, Wei GW. Differential geometry based solvation models I: Eulerian formulation. J. Comput. Phys. 2010; 229:8231–8258. [PubMed: 20938489]
26. Chen Z, Baker NA, Wei GW. Differential geometry based solvation models II: Lagrangian formulation. J. Math. Biol. 2011; 63:1139–1200. [PubMed: 21279359]
27. Chen Z, Zhao Shan, Chun J, Thomas DG, Baker NA, Bates PB, Wei GW. Variational approach for nonpolar solvation analysis. Journal of Chemical Physics. 2012; 137(084101)
28. Feng X, Xia KL, Tong YY, Wei GW. Multiscale geometric modeling of macromolecules II: lagrangian representation. Journal of Computational Chemistry. 2013; 34:2100–2120. [PubMed: 23813599]
29. Xia KL, Feng X, Tong YY, Wei GW. Multiscale geometric modeling of macromolecules. Journal of Computational Physics. 2014; 275:912–936.
30. Boileau E, Bevan RLT, Sazonov I, Rees MI, Nithiarasu P. Flow-induced atp release in patient-specific arterial geometries - a comparative study of computational models. International Journal for Numerical Methods in Engineering. 2013; 29:1038–1056.
31. Sazonov, Igor; Nithiarasu, Perumal. Semi-automatic surface and volume mesh generation for subject-specific biomedical geometries. International Journal for Numerical Methods in Biomedical Engineering. 2012; 28:133–157.
32. Sazonov I, Yeo SY, Bevan RLT, Xie XH, van Loon R, Nithiarasu P. Modelling pipeline for subject-specific arterial blood flow – a review. International Journal for Numerical Methods in Engineering. 2012; 28:1868–1910.
33. Sohn JS, Li SW, Li XF, Lowengrub JS. Axisymmetric multicomponent vesicles: A comparison of hydrodynamic and geometric models. International Journal for Numerical Methods in Engineering. 2012; 28:346–368.

34. Ramalho S, Moura A, Gambaruto AM, Sequeira A. Sensitivity to outflow boundary conditions and level of geometry description for a cerebral aneurysm. International Journal for Numerical Methods in Engineering. 2012; 28:697–713.
35. Manzoni, Andrea; Quarteroni, Alfio; Rozza, Gianluigi. Model reduction techniques for fast blood flow simulation in parametrized geometries. International Journal for Numerical Methods in Engineering. 2012; 28:604–625.
36. Mikhal J, Kroon DJ, Slump CH, Geurts BJ. Flow prediction in cerebral aneurysms based on geometry reconstruction from 3d rotational angiography. International Journal for Numerical Methods in Engineering. 2013; 29:777–805.
37. Edelsbrunner H, Letscher D, Zomorodian A. Topological persistence and simplification. Discrete Comput. Geom. 2002; 28:511–533.
38. Zomorodian A, Carlsson G. Computing persistent homology. Discrete Comput. Geom. 2005; 33:249–274.
39. Zomorodian, Afra; Carlsson, Gunnar. Localized homology. Computational Geometry - Theory and Applications. 2008; 41(3):126–148.
40. Frosini, Patrizio; Landi, Claudia. Size theory as a topological tool for computer vision. Pattern Recognition and Image Analysis. 1999; 9(4):596–603.
41. Robins, Vanessa. Towards computing homology from finite approximations. Topology Proceedings. 1999; 24:503–532.
42. Bubenik, Peter; Kim, Peter T. A statistical approach to persistent homology. Homology, Homotopy and Applications. 2007; 19:337–362.
43. Edelsbrunner, Herbert; Harer, John. Computational topology: an introduction. American Mathematical Soc.; 2010.
44. Dey TK, Li KY, Sun J, David CS. Computing geometry aware handle and tunnel loops in 3d models. ACM Trans. Graph. 2008; 27
45. Dey, Tamal K.; Wang, Yusu. Reeb graphs: Approximation and persistence. Discrete and Computational Geometry. 2013; 49(1):46–73.
46. Mischaikow K, Nanda V. Morse theory for filtrations and efficient computation of persistent homology. Discrete and Computational Geometry. 2013; 50(2):330–353.
47. Ghrist R. Barcodes: The persistent topology of data. Bull. Amer. Math. Soc. 2008; 45:61–75.
48. Carlsson G, Ishkhanov T, Silva V, Zomorodian A. On the local behavior of spaces of natural images. International Journal of Computer Vision. 2008; 76(1):1–12.
49. Pachauri D, Hinrichs C, Chung MK, Johnson SC, Singh V. Topology-based kernels with application to inference problems in alzheimer's disease. Medical Imaging, IEEE Transactions on. 2011 Oct; 30(10):1760–1770.
50. Singh G, Memoli F, Ishkhanov T, Sapiro G, Carlsson G, Ringach DL. Topological analysis of population activity in visual cortex. Journal of Vision. 2008; 8(8)
51. Bendich, Paul; Edelsbrunner, Herbert; Kerber, Michael. Computing robustness and persistence for images. IEEE Transactions on Visualization and Computer Graphics. 2010; 16:1251–1260. [PubMed: 20975165]
52. Frosini, Patrizio; Landi, Claudia. Persistent betti numbers for a noise tolerant shape-based approach to image retrieval. Pattern Recognition Letters. 2013; 34:863–872.
53. Mischaikow K, Mrozek M, Reiss J, Szymczak A. Construction of symbolic dynamics from experimental time series. Physical Review Letters. 1999; 82:1144–1147.
54. Kaczynski, T.; Mischaikow, K.; Mrozek, M. Computational homology. Springer-Verlag; 2004.
55. Silva VD, Ghrist R. Blind swarms for coverage in 2-d. Proceedings of Robotics: Science and Systems. 2005:01.
56. Lee H, Kang H, Chung MK, Kim B, Lee DS. Persistent brain network homology from the perspective of dendrogram. Medical Imaging, IEEE Transactions on. 2012 Dec; 31(12):2267–2277.
57. Horak D, Maletic S, Rajkovic M. Persistent homology of complex networks. Journal of Statistical Mechanics: Theory and Experiment. 2009; 2009(03):P03034.
58. Carlsson G. Topology and data. Am. Math. Soc. 2009; 46(2):255–308.

59. Niyogi P, Smale S, Weinberger S. A topological view of unsupervised learning from noisy data. SIAM Journal on Computing. 2011; 40:646–663.
60. Wang, Bei; Summa, Brian; Pascucci, Valerio; Vejdemo-Johansson, M. Branching and circular features in high dimensional data. IEEE Transactions on Visualization and Computer Graphics. 2011; 17:1902–1911. [PubMed: 22034307]
61. Rieck, Bastian; Mara, Hubert; Leitte, Heike. Multivariate data analysis using persistence-based filtering and topological signatures. IEEE Transactions on Visualization and Computer Graphics. 2012; 18:2382–2391.
62. Liu, Xu; Xie, Zheng; Yi, Dongyun. A fast algorithm for constructing topological structure in large data. Homology, Homotopy and Applications. 2012; 14:221–238.
63. Di Fabio, Barbara; Landi, Claudia. A mayer-vietoris formula for persistent homology with an application to shape recognition in the presence of occlusions. Foundations of Computational Mathematics. 2011; 11:499–527.
64. Kasson PM, Zomorodian A, Park S, Singhal N, Guibas LJ, Pande VS. Persistent voids a new structural metric for membrane fusion. Bioinformatics. 2007; 23:1753–1759. [PubMed: 17488753]
65. Gameiro M, Hiraoka Y, Izumi S, Kramar M, Mischaikow K, Nanda V. Topological measurement of protein compressibility via persistence diagrams. preprint. 2013
66. Dabaghian Y, Memoli F, Frank L, Carlsson G. A topological paradigm for hippocampal spatial map formation using persistent homology. PLoS Comput Biol. 2012; 8(8):e1002581. 08. [PubMed: 22912564]
67. Yao Y, Sun J, Huang XH, Bowman GR, Singh G, Lesnick M, Guibas LJ, Pande VS, Carlsson G. Topological methods for exploring low-density states in biomolecular folding pathways. The Journal of Chemical Physics. 2009; 130:144115. [PubMed: 19368437]
68. Chang HW, Bacallado S, Pande VS, Carlsson GE. Persistent topology and metastable state in conformational dynamics. PLos ONE. 2013; 8(4):e58699. [PubMed: 23565139]
69. Biasotti S, De Floriani L, Falcidieno B, Frosini P, Giorgi D, Landi C, Papaleo L, Spagnuolo M. Describing shapes by geometrical-topological properties of real functions. ACM Computing Surveys. 2008; 40(4):12.
70. Edelsbrunner H, Mucke EP. Three-dimensional alpha shapes. Physical Review Letters. 1994; 13:43–72.
71. Xia KL, Wei GW. A stochastic model for protein flexibility analysis. Physical Review E. 2013; 88:062709.
72. Tausz, Andrew; Vejdemo-Johansson, Mikael; Adams, Henry. Javaplex: A research software package for persistent (co)homology. 2011 Software available at http://code.google.com/p/javaplex.
73. Yang L, Song G, Jernigan RL. Protein elastic network models and the ranges of cooperativity. Proceedings of the National Academy of Sciences. 2009; 106(30):12347–12352.
74. Xia KL, Wei GW. Molecular nonlinear dynamics and protein thermal uncertainty quantification. Chaos. 2014; 24:013103. [PubMed: 24697365]
75. Ma J. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. Structure. 2005; 13:373–180. [PubMed: 15766538]
76. Yang LW, Chng CP. Coarse-grained models reveal functional dynamics–i. elastic network models–theories, comparisons and perspectives. Bioinformatics and Biology Insights. 2008; 2:25–45. [PubMed: 19812764]
77. Skjaerven L, Hollup SM, Reuter N. Normal mode analysis for proteins. Journal of Molecular Structure: Theochem. 2009; 898:42–48.
78. Cui, Q.; Bahar, I. Normal mode analysis: theory and applications to biological and chemical systems. Chapman and Hall/CRC; 2010.
79. Paci E, Karplus M. Unfolding proteins by external forces and temperature: The importance of topology and energetics. Proceedings of the National Academy of Sciences. 2000; 97:6521–6526.
80. Hui L, Isralewitz B, Krammer A, Vogel V, Schulten K. Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation. Biophysical Journal. 1998; 75:662–671. [PubMed: 9675168]

81. Srivastava A, Granek R. Cooperativity in thermal and force-induced protein unfolding: integration of crack propagation and network elasticity models. Phys. Rev. Lett. 2013; 110(138101):1–5.

82. Gao M, Craig D, Vogel V, Schulten K. Identifying unfolding intermediates of $fn - iii_{10}$ by steered molecular dynamics. J. Mol. Biol. 2002; 323:939–950. [PubMed: 12417205]

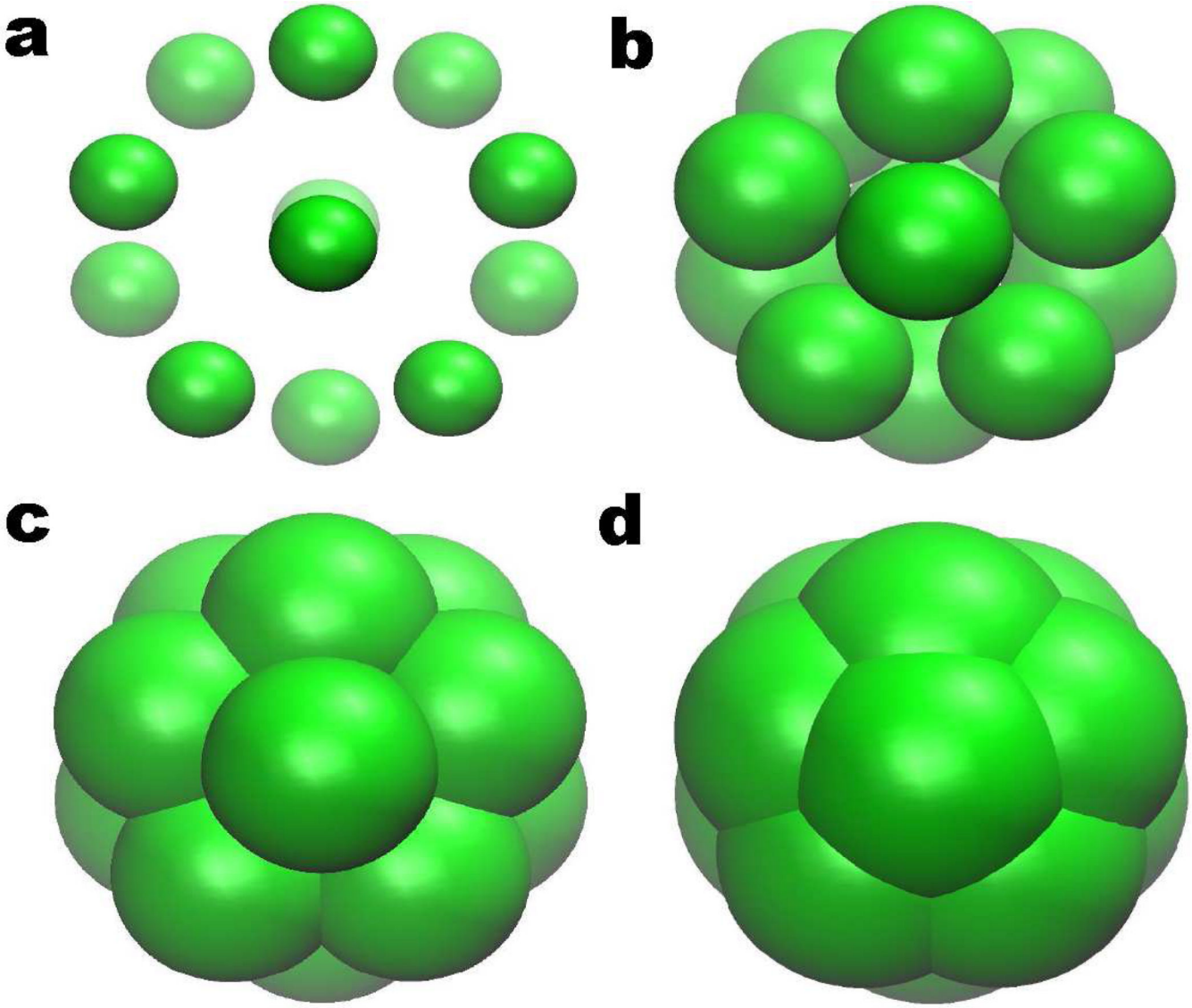

Figure 1.

The distance based filtration process of an icosahedron. Each icosahedron vertex is associated with an ever-increasing radius to form a ball. With the increase of their radii, the balls overlap with each other to form higher simplexes. In this manner, the previously formed simplicial complex is included in the latter ones.

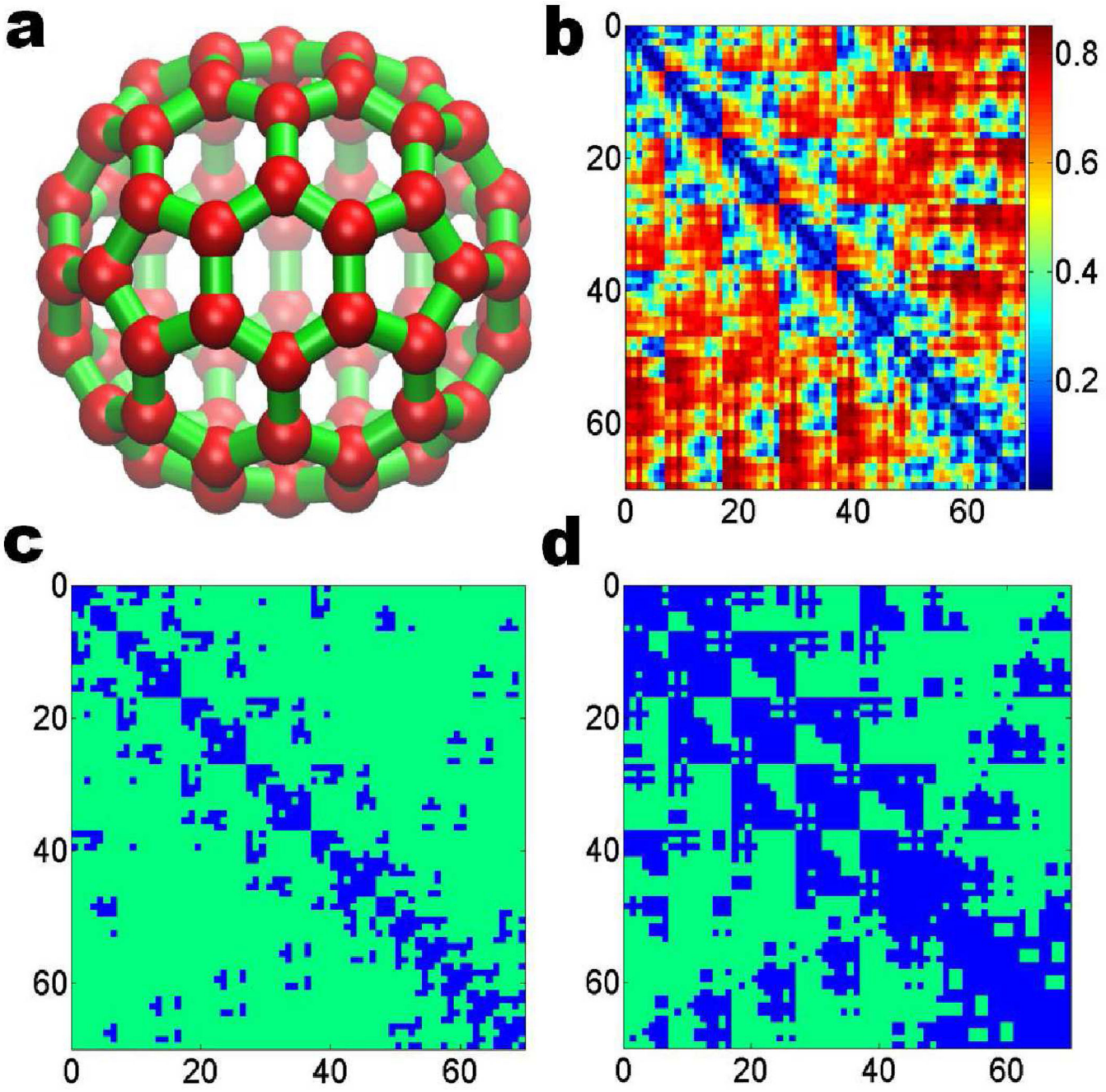

Figure 2.

Correlation matrix based filtration for fullerene C_{70}. The correlation matrix is constructed by using the geometry to topology mapping.[20,71] As the value of the filtration parameter increase, Rips complex formed grows accordingly, **a** is an image of fullerene C_{70}; **b, c** and **d** demonstrate the connectivity among C_{70} atoms at filtration threshold ε = 0.1Å, 0.3Å and 0.5Å, respectively. The blue color represents the atoms already formed simplicies.

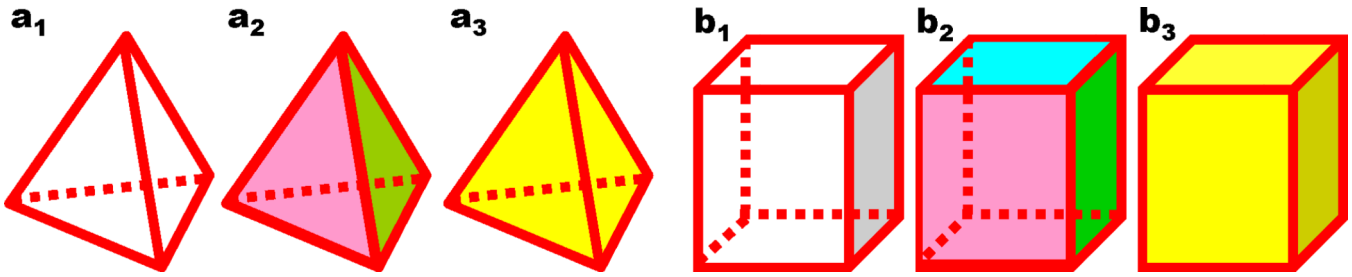

Figure 3.

Illustration of simplicial complexes. The tetrahedron-shaped simplicial complexes are depicted in $\mathbf{a}_1$ to $\mathbf{a}_3$, and the cube-shaped simplicial complexes are demonstrated in $\mathbf{b}_1$ to $\mathbf{b}_3$. In each shape, the leftmost simplicial complex has only 0-simplexes and 1-simplexes. As filtration processes, 2-simplexes emerge in the middle one. In the final stage, a 3-simplex is generated. Table 1 gives a description in terms of vertices, edges, faces and cells.

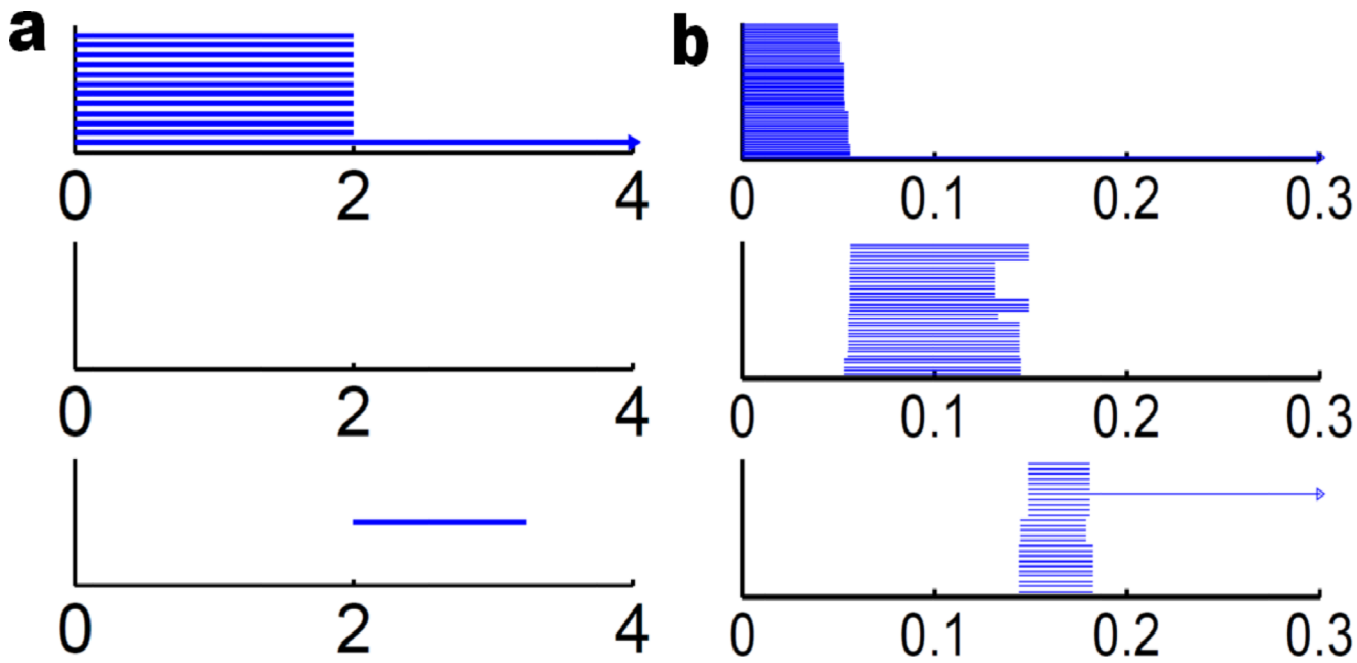

Figure 4.

Persistent homology analysis of the icosahedron (left chart) and fullerene C_{70} (right chart). The horizontal axis is the filtration parameter, which has the unit of angstrom (Å) in the distance based filtration for icosahedron and there is no unit in the correlation matrix based filtration for C_{70}. From the top to the bottom, there are three panels corresponding to β_0, β_1 and β_2 bars, respectively. For the icosahedron barcode, originally there are 12 bars due to 12 vertices in β_0 panel. As the filtration continues, 11 of them terminate simultaneously with only one persisting to the end, indicating that all vertices are connected. Topologically, β_0 bars represent isolated entities. At beginning, there are 12 individual vertices. They connect to each other simultaneously due to their structural symmetry. Nothing occurs at the β_1 panel, which means there is no one-dimensional circle ever formed. Finally, in the β_2 panel, the single bar represents the central void. For fullerene C_{70} barcode, there are 70 β_0 bars and 36 β_1 bars. The β_1 bars are due to 12 pentagon rings and 25 hexagon rings. The hexagon rings further evolve into two-dimensional holes, which are represented by 25 short-lived β_2 bars. The central void structure formed is captured by the persisting β_2 bar.

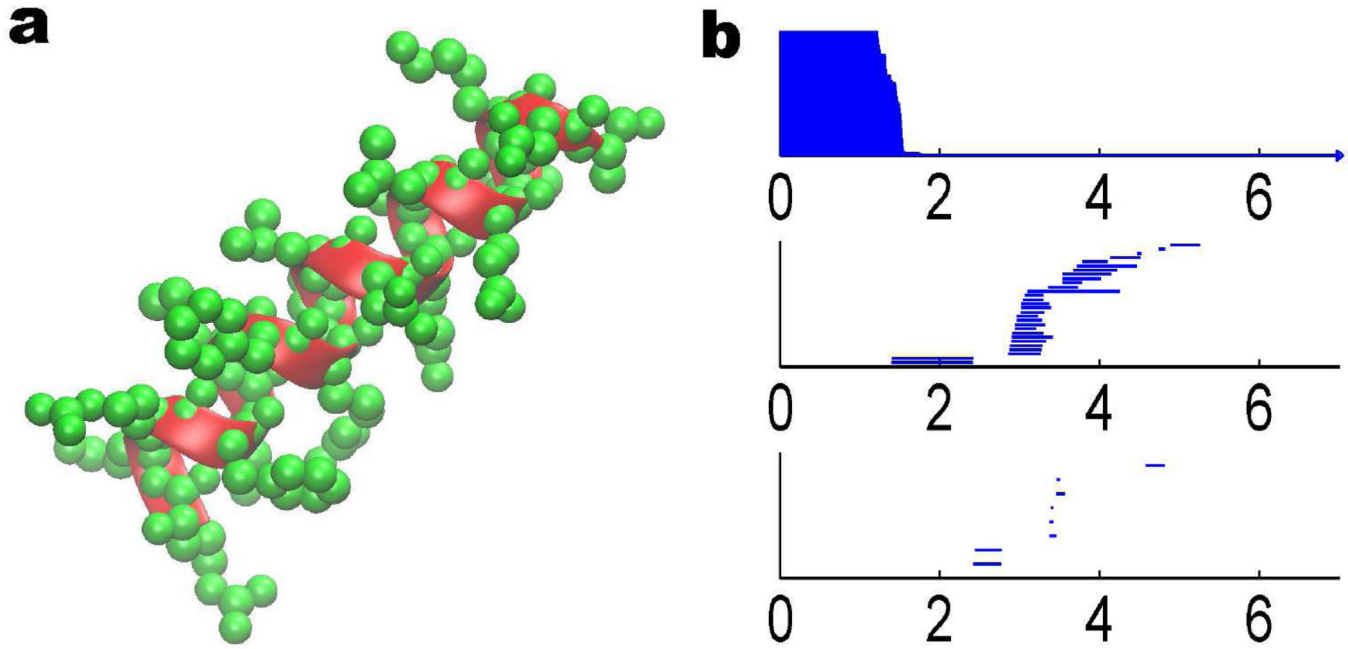

Figure 5.

Illustration of an alpha helix structure (PDB ID: 1C26) and its topological fingerprint obtained by the distance filtration. In the left chart, atoms are demonstrated in green color and the helix structure of the main chain backbone is represent by the cartoon shape in red. The right chart is the corresponding barcode with the all-atom description. The horizontal axis is the filtration size (Å). Although the alpha helix backbone has a loop-type structure, the corresponding barcode does not clearly demonstrate these patterns due to the fact that there are two many atoms around the main chain.

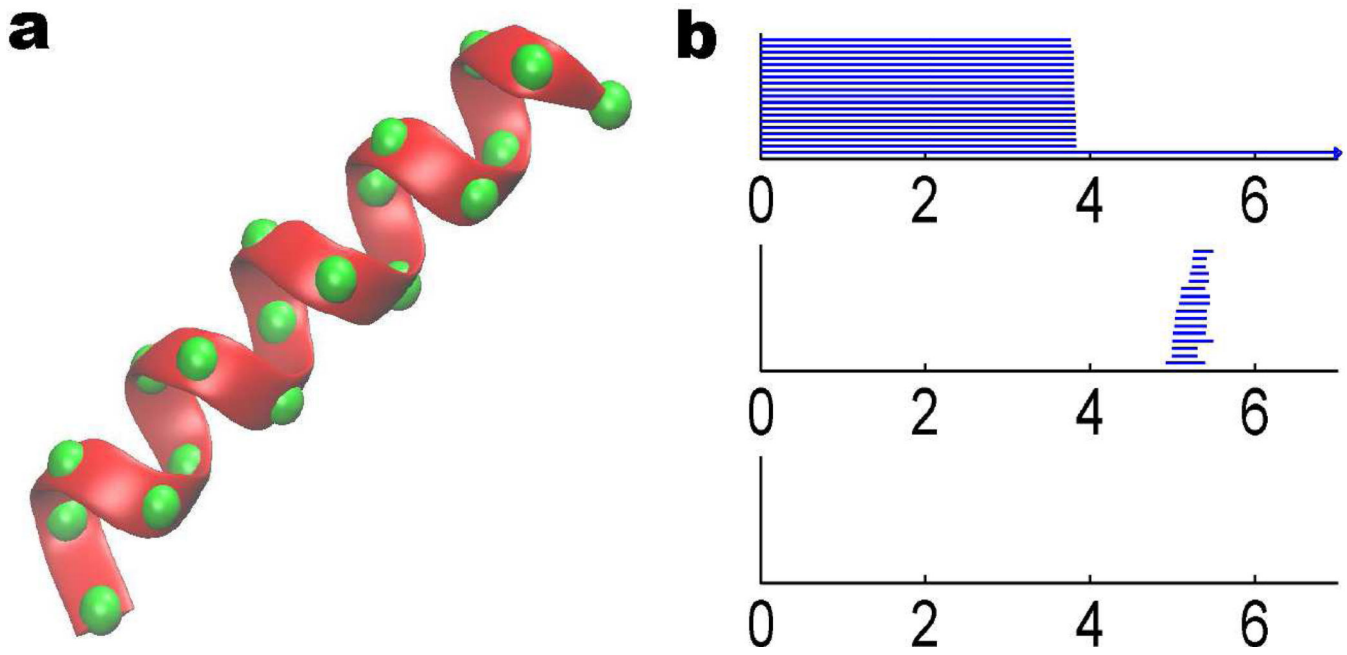

Figure 6.

The coarse-grained representation of an alpha helix structure (left chart) and its topological fingerprint (right chart). The alpha helix structure (PDB ID: 1C26) has 19 residues represented by C_α atoms in green color. Each 4 C_α atoms contribute a β_1 loop and thus there are 16 short-lived bars in the β_1 panel.

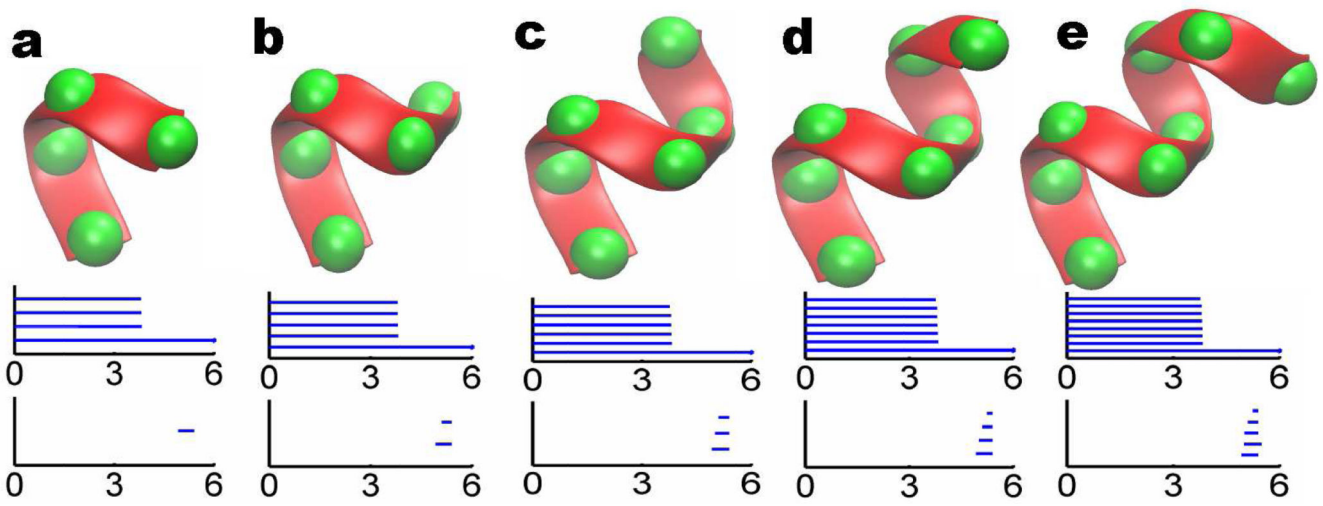

Figure 7.

Method of slicing for the analysis of alpha helix topological fingerprints. In the coarse-grain representation, each residue is represented by a C_α atom. In an alpha helix, each set of four Ca atoms forms a one-dimensional loop in the filtration process as depicted in **a**. By adding one more C_α atom, one more β_1 loop is generated and leads to an additional β_1 bar as shown in **b, c, d**, and **e**. This explains the occurrence of 16 β_1 bars in Fig. 6.

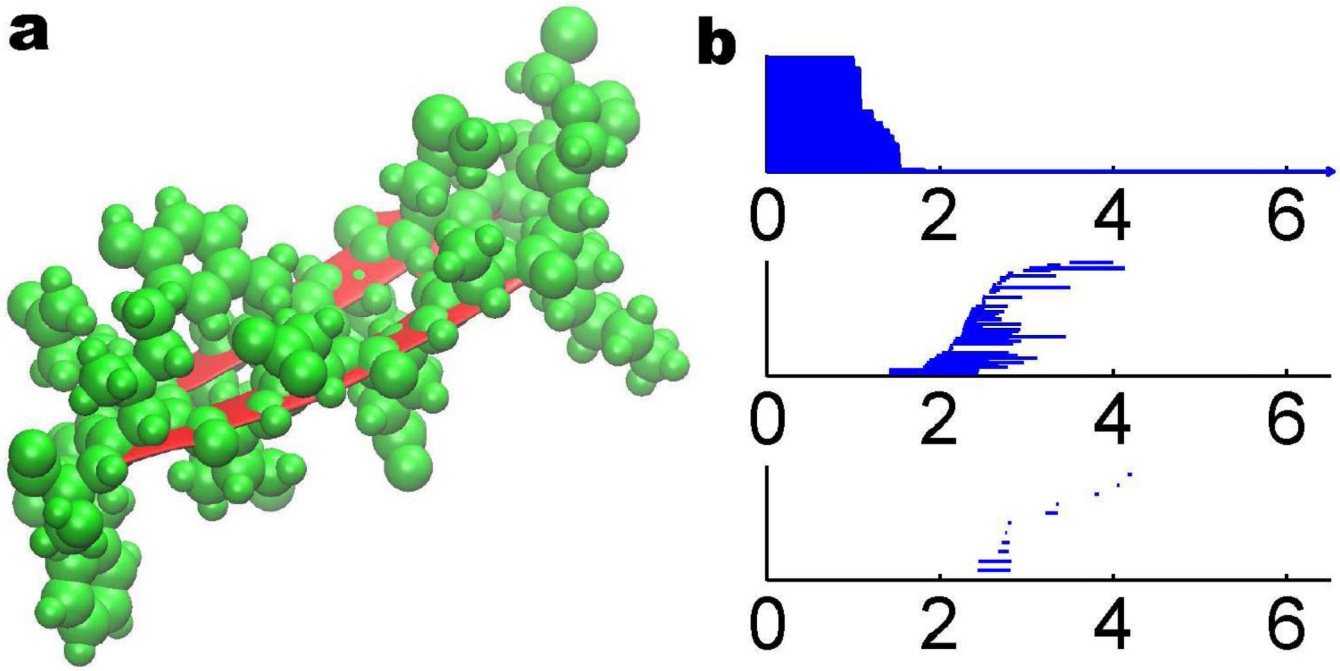

Figure 8.

The all-atom representation of the beta sheet structure generated from PDB 2JOX (left chart) and the related topological fingerprint (right chart). Each beta sheet has 8 residues. The topological fingerprint generated from all-atom based filtration has a complicated pattern due to excessively many residual atoms.

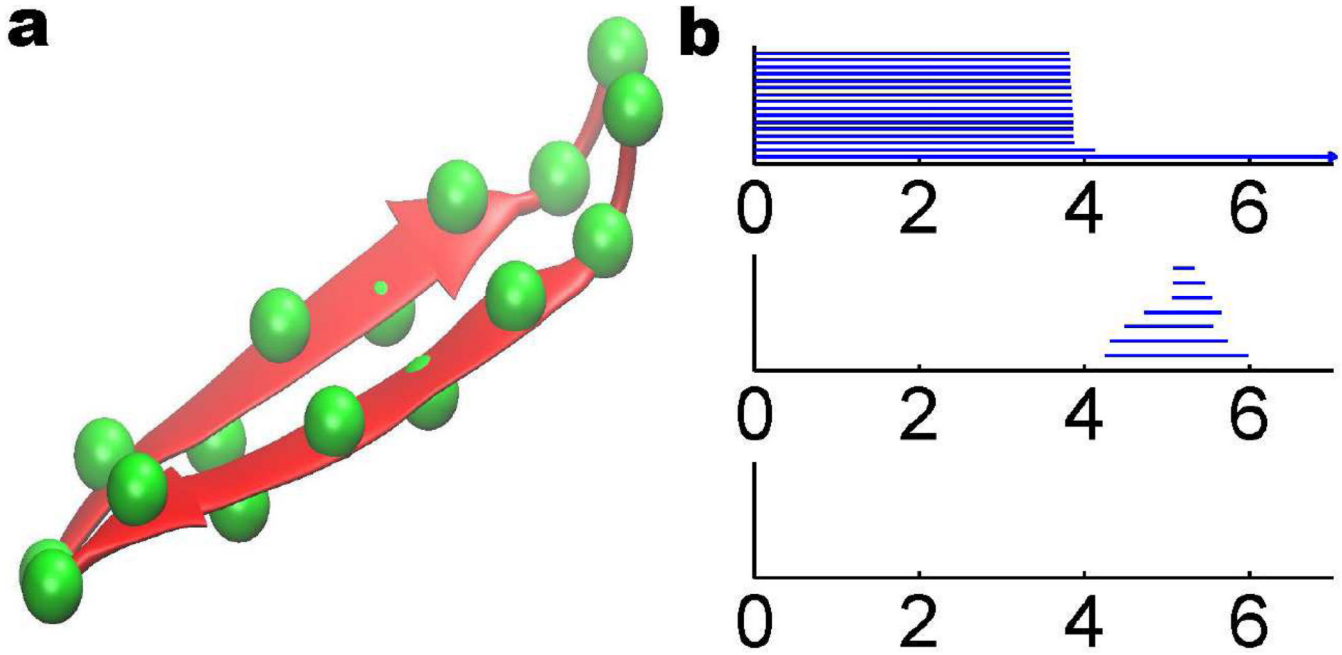

Figure 9.

The coarse-grained representation of two beta sheet strands generated from protein 2JOX (left chart) and the corresponding topological fingerprint (right chart). There are 8 pairs of residues represented by 16 C_α atoms in the β_0 panel. Each 2 pairs of C_α atoms contribute a $\beta_1$loop to make up 7 bars in the β_1 panel.

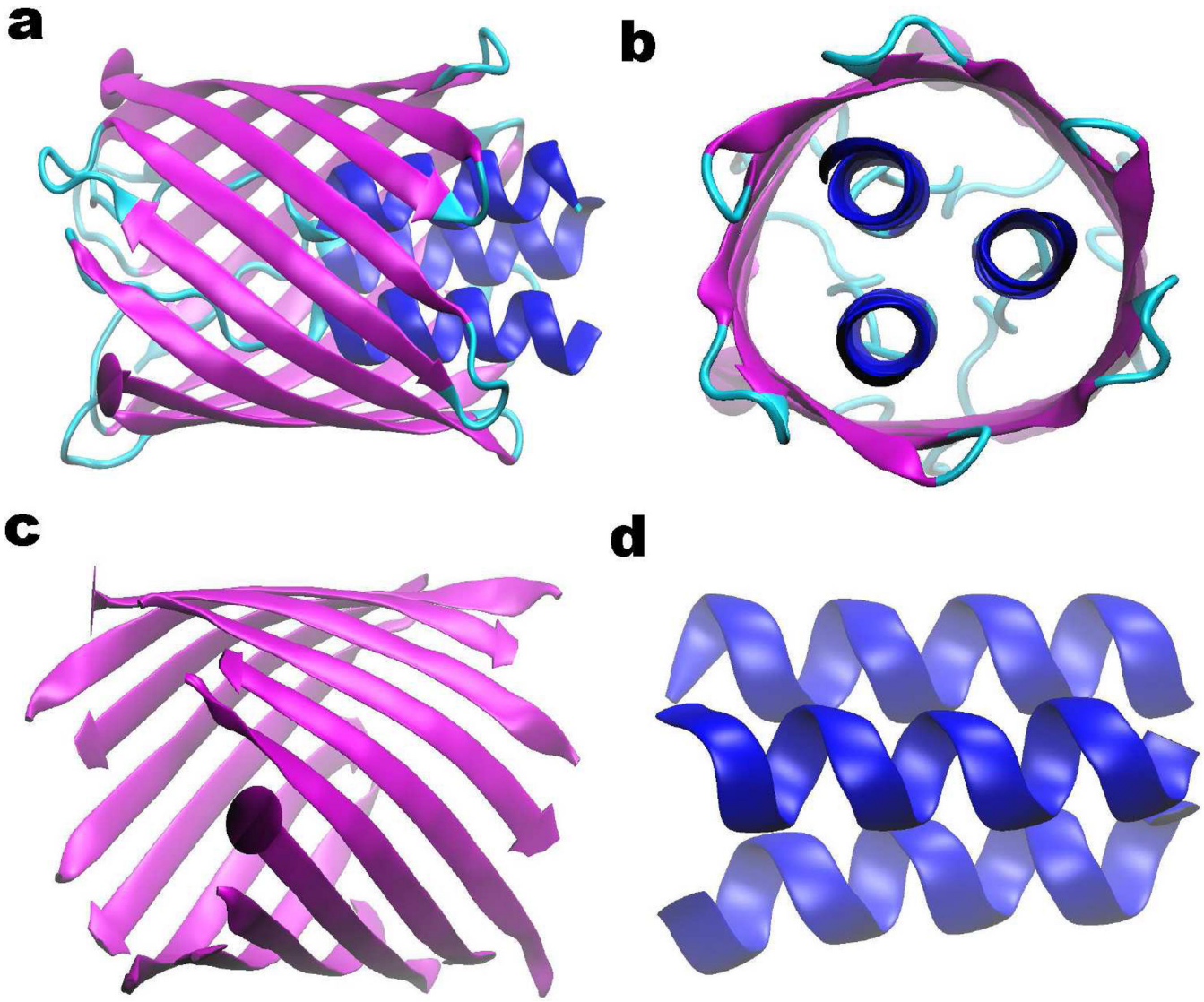

Figure 10.

The basic geometry of the beta barrel generated from protein 20GR8. Here **a** and **b** are cartoon representations from two views. The beta barrel structure is decomposed into beta sheet and alpha helix components for topological pattern recognition in **c** and **d**, respectively.

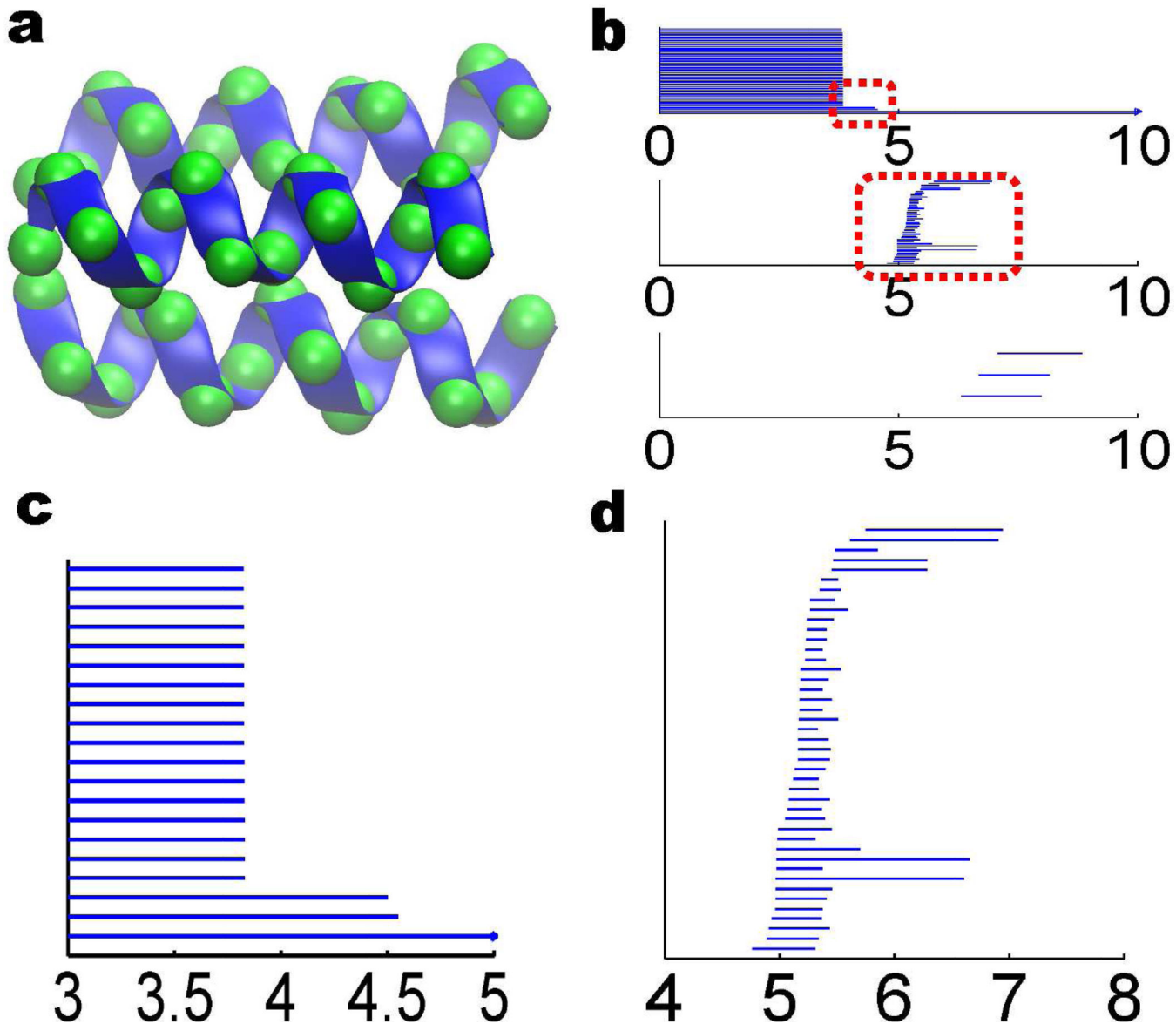

Figure 11.

Topological analysis of the alpha-helix structure from beta barrel 2GR8. The coarse-grained representation is employed with C_α atoms in the green color in **a**. The topological fingerprint of the alpha-helix structure is depicted in **b**. The details of the barcode enclosed in red boxes are demonstrated in **c** and **d**. It is seen from **c** that in the β_0 panel, the length of most bars is around 3.8Å except for those of three bars. Two of these β_0 bars end around 4.5Å, and the other lasts forever. These three bars represent the remaining isolated alpha helices when C_α atoms inside each alpha helix become connected at about 3.8Å. However, when the filtration parameter is increased to 4.5Å, which is exactly the smallest distance between alpha helices, three alpha helices are connected and the total number of independent entities becomes one. There are also three bars in the β_2 panel as shown in **b**. There is a total of 43 β_1 bars as shown in **d**.

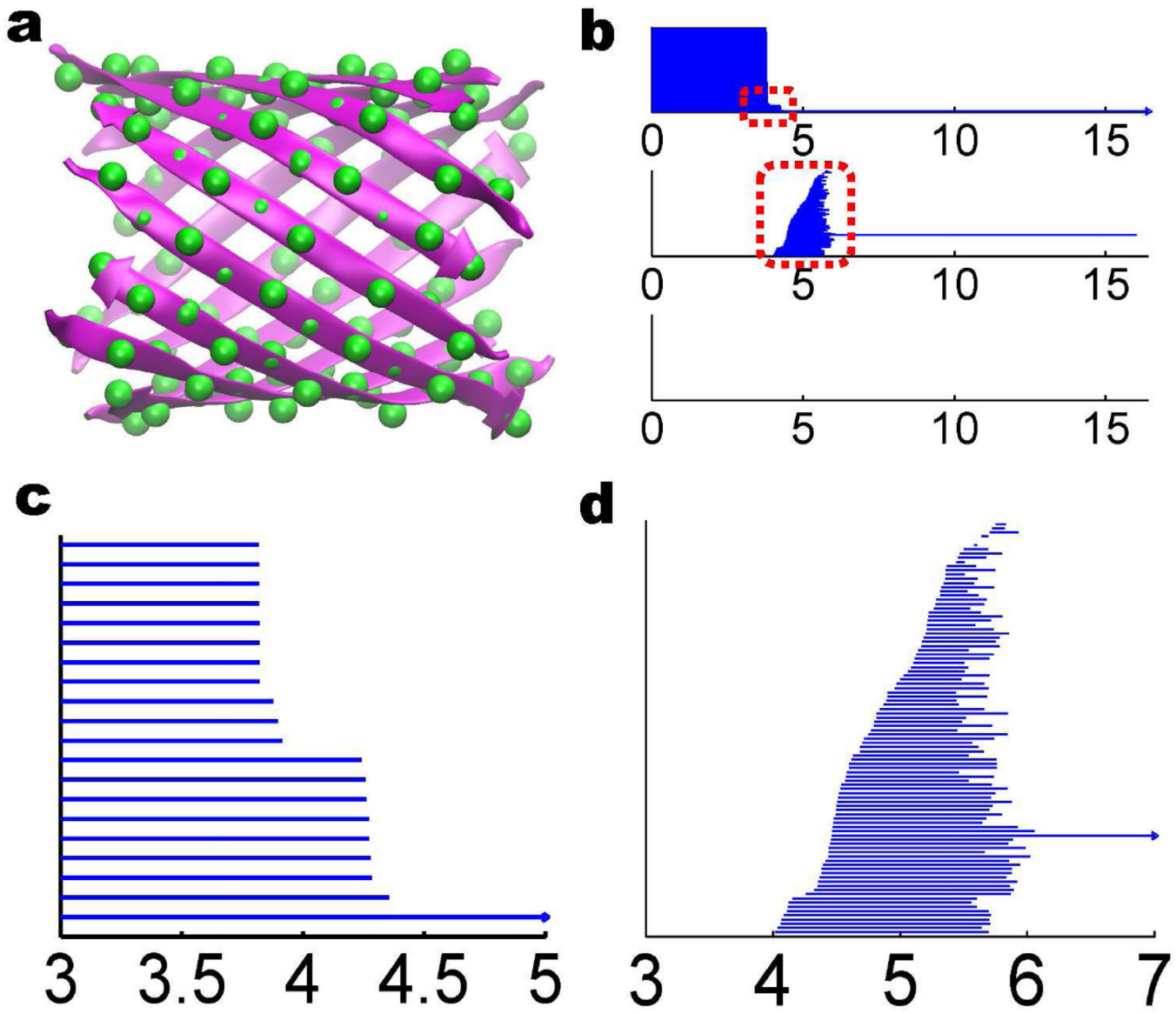

Figure 12.

The topological analysis of the beta sheet structure extracted from beta barrel 2GR8. The coarse-grained representation is employed with C_α atoms in the green color in **a**. There are 128 atoms organized in 12 beta sheets. The topological fingerprint of the beta sheet structure is depicted in **b**. The red boxes are zoomed in and demonstrated in c and d. There are 128 $\beta_0$bars and 98 β_1 bars. It is seen from **b** and **c** that in the β_0 panel, 12 out of 128 bars persist beyond 3.8Å, which corresponds to 12 isolated beta sheets. The longest β_1 bar in **b** is due to the large hole in beta barrel structure. Other β_1 bars are formed from every adjacent 4 C_α atoms as discussed in the earlier analysis of parallel beta sheet structure. Due to the mismatch, adjacent two sheets contribute around 8 β_1 bars, which accounts for $12 \times 8 = 96$ short-lived β_1 bars. Together with the bar for the global intrinsic ring, it is estimated that there are 97 β_1 bars. Although computational result shows 98 β_1 bars, one of them (the fifth from top) is barely visible.

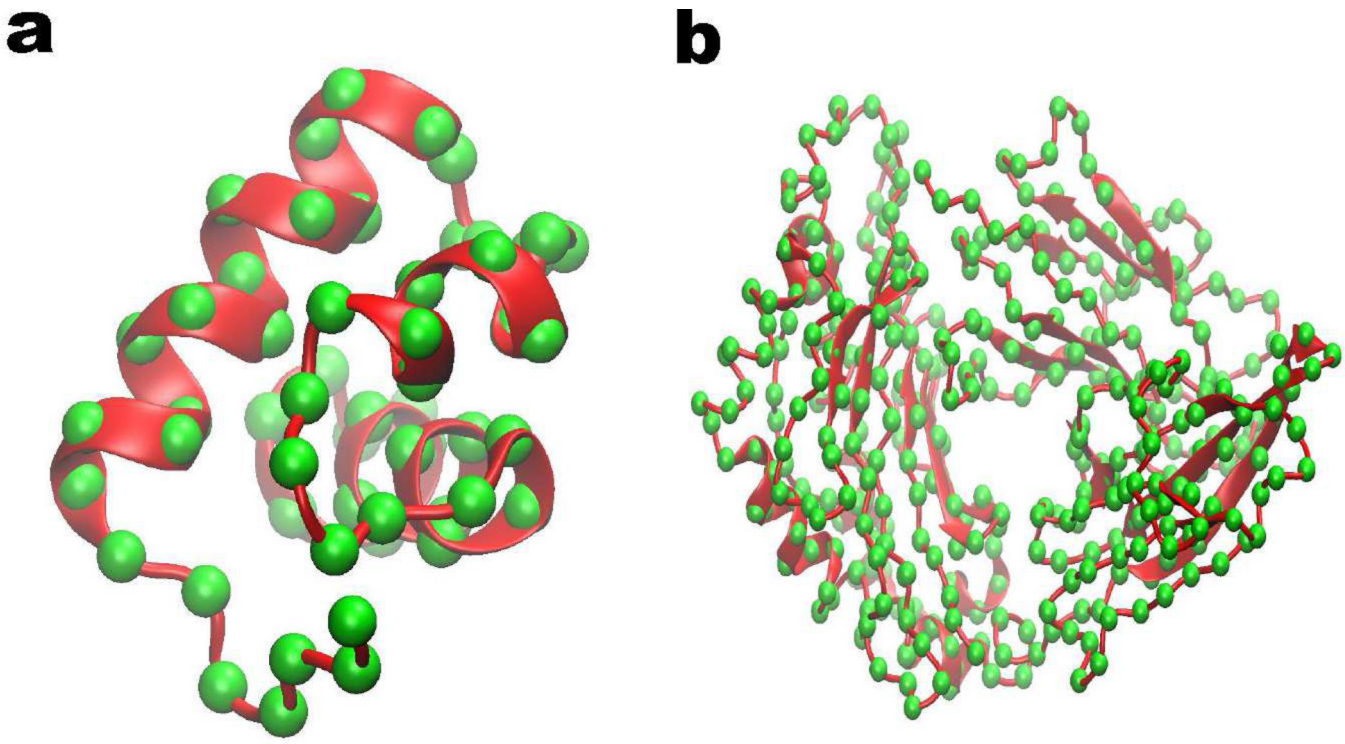

Figure 13.

Illustration of proteins 1GVD (left chart) and 3MRE (right chart) used in analyzing the optimal cutoff distance of the Gaussian network model. The coarse-grained model is employed with residues represented by their C_α atoms and displayed as green balls.

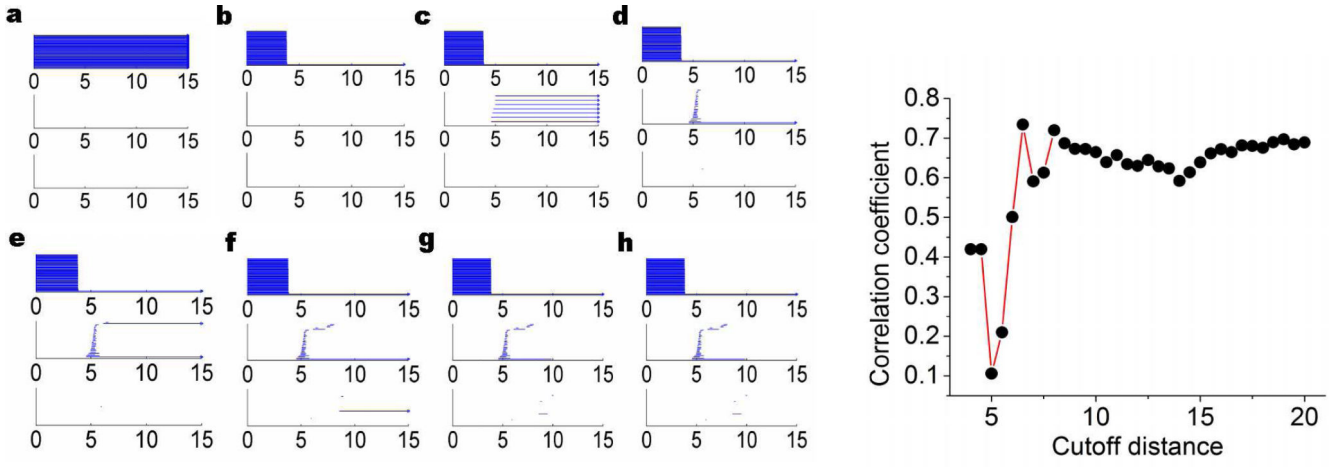

Figure 14.

Visualization of topological connectivity and optimal cutoff distance of Gaussian network model for protein 1GVD. The left charts from **a** to **h** are barcodes generated based on the filtration given in Eq. (25). From **a** to **h**, the cutoff distances used are 3Å, 4Å, 5Å, 6Å, 7Å, 9Å, 11Å, and 13Å, respectively. The right chart is the correlation coefficient obtained from the Gaussian network model under different cutoff distance (Å).

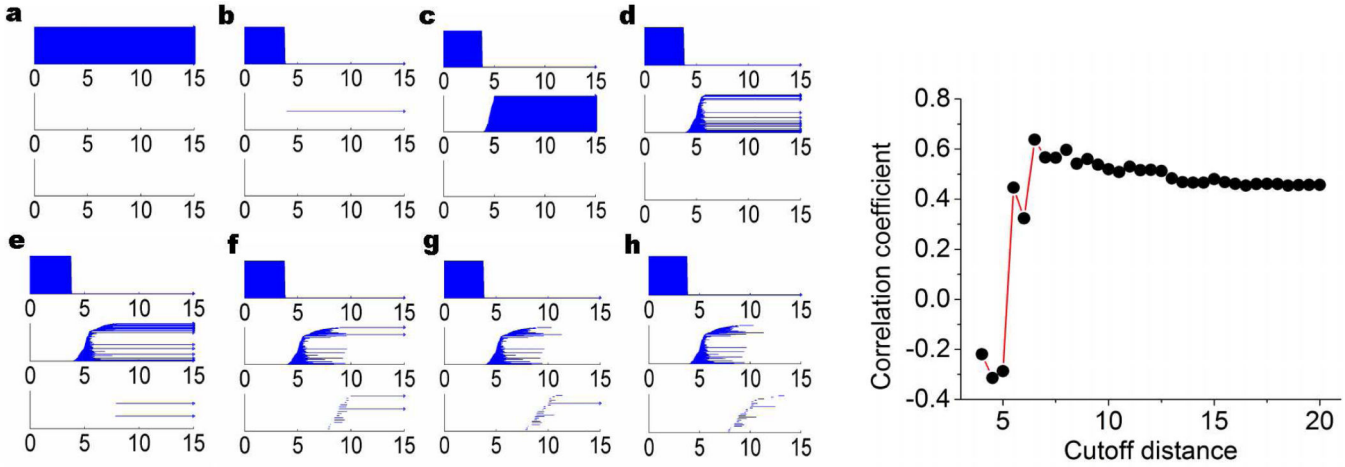

Figure 15.

Visualization of topological connectivity and optimal cutoff distance of Gaussian network model for protein 3MRE. The left charts from **a** to **h** are the barcodes generated based on the filtration given in Eq. (25). From **a** to **h**, the cutoff distances used are 3Å, 4Å, 5Å, 6Å, 8Å, 10Å, 12Å, and 14Å, respectively. The right chart is the correlation coefficient obtained from the Gaussian network model under different cutoff distance (Å).

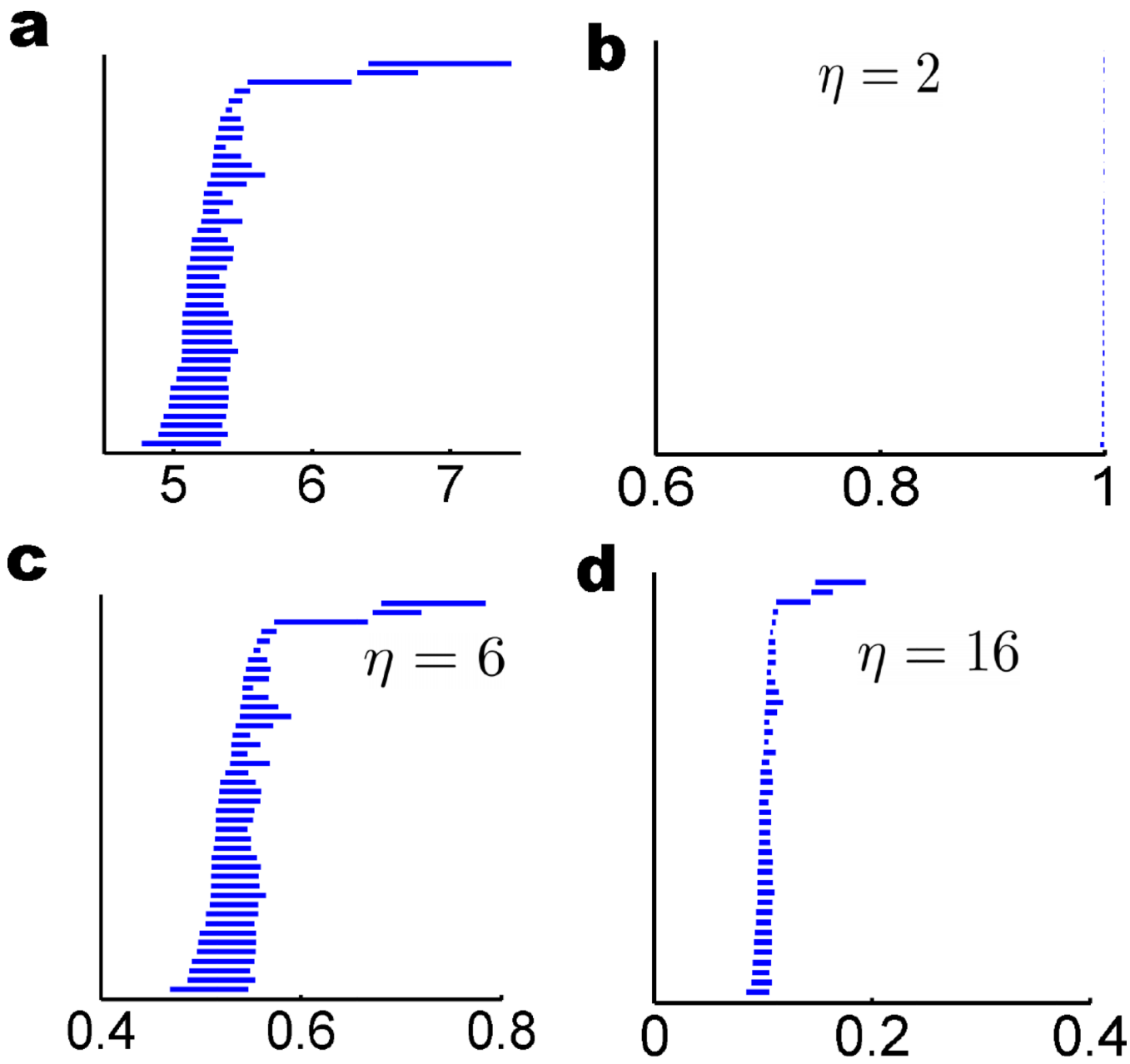

Figure 16.

Comparison of β_1 behaviors in different filtration settings for protein 1YZM C_α point cloud data. Distance based filtration is shown in **a**. The correlation matrix based filtration with exponential kernel ($\kappa = 2$) is used in **b**, **c** and **d**. The η is chosen to be 2Å, 6Å and 16Å in **b, c** and **d**, respectively. The β_1 bar patterns **a, c** and **d** are very similar but have different durations. The β_1 bar pattern in **b** differs much from the rest due to a small characteristic distance $\eta = 2$Å.

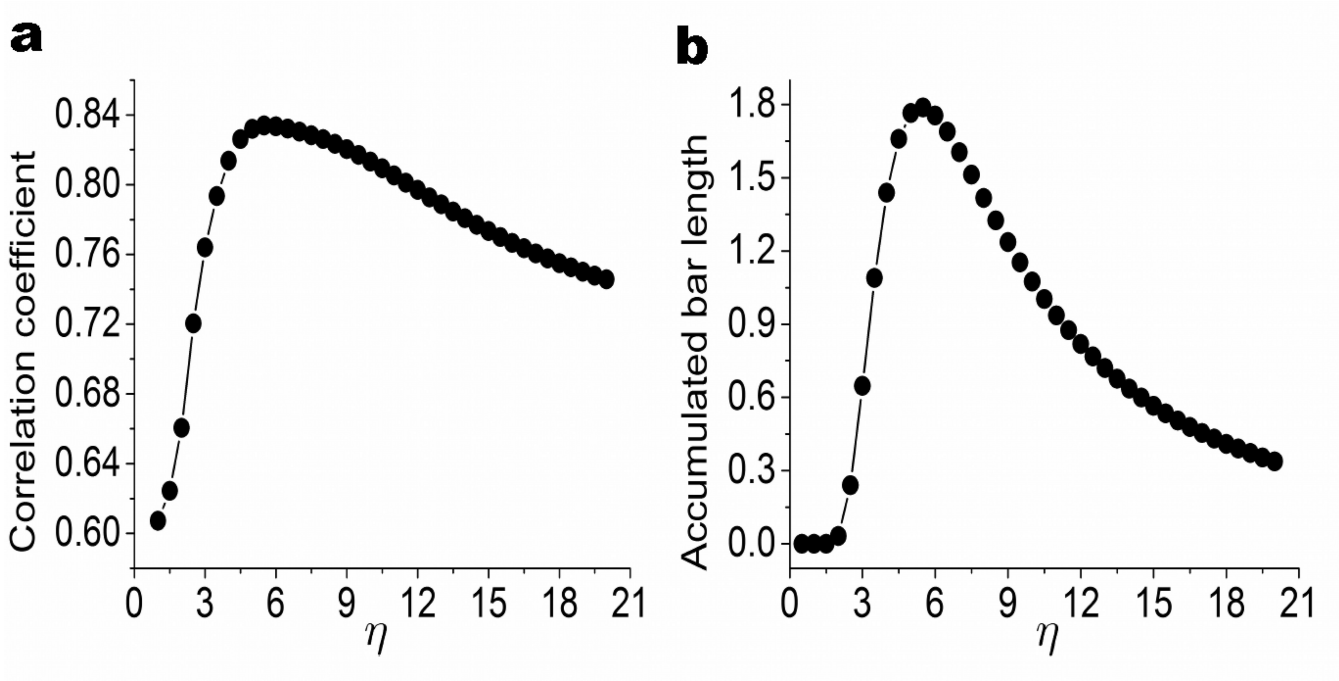

Figure 17.

The comparison between the correlation coefficient from the B-factor prediction by FRI (left chart) and accumulated bar length from persistent homology modeling (right chart) under various η values in Å for protein 1TZM. It is seen that correlation coefficient and accumulated bar length share a similar shape that their values increase dramatically at first and then gradually decrease. The common maximum near η = 6.0Å indicates the ability of persistent homology for the prediction of optimal characteristic distance.

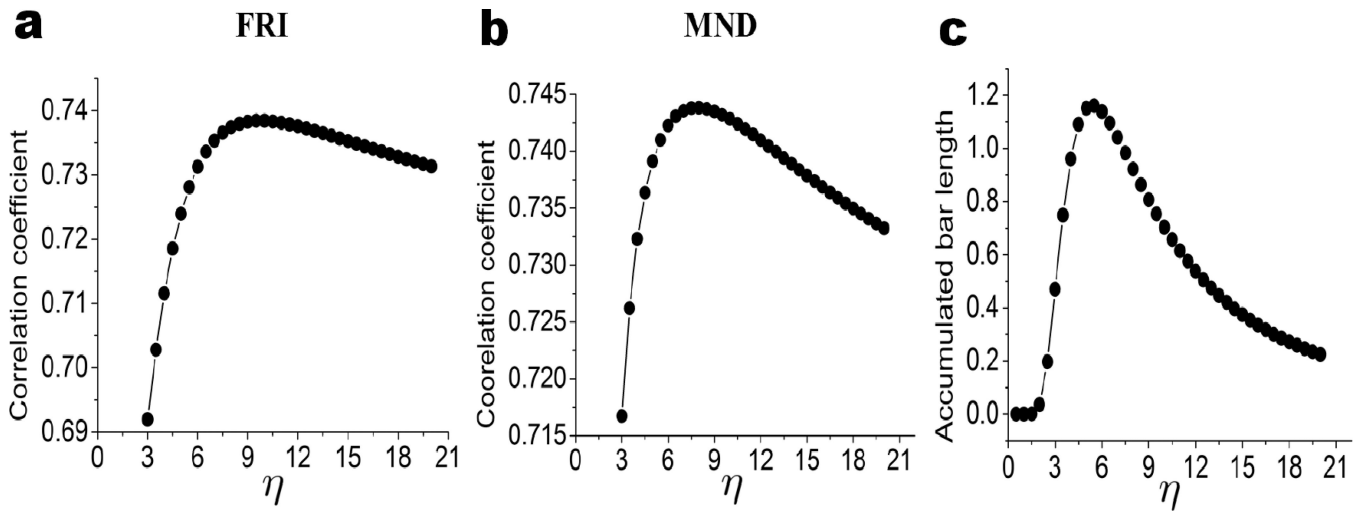

Figure 18.

Comparison between the patterns of average correlation coefficients for B-factor predictions and the shape of average accumulated bar length for a set of 30 proteins listed in Table 2. The average correlation coefficients obtained from FRI and MND are plotted over a range of characteristic distances η (in Å) in **a** and **b**, respectively. The average accumulated bar length A_1 is shown in **c**. All patterns share a similar trend. The highest correlation coefficients are reached around η = 9Å and 8Å for MND and FRI, respectively. While the highest accumulated bar length is found around η =6Å.

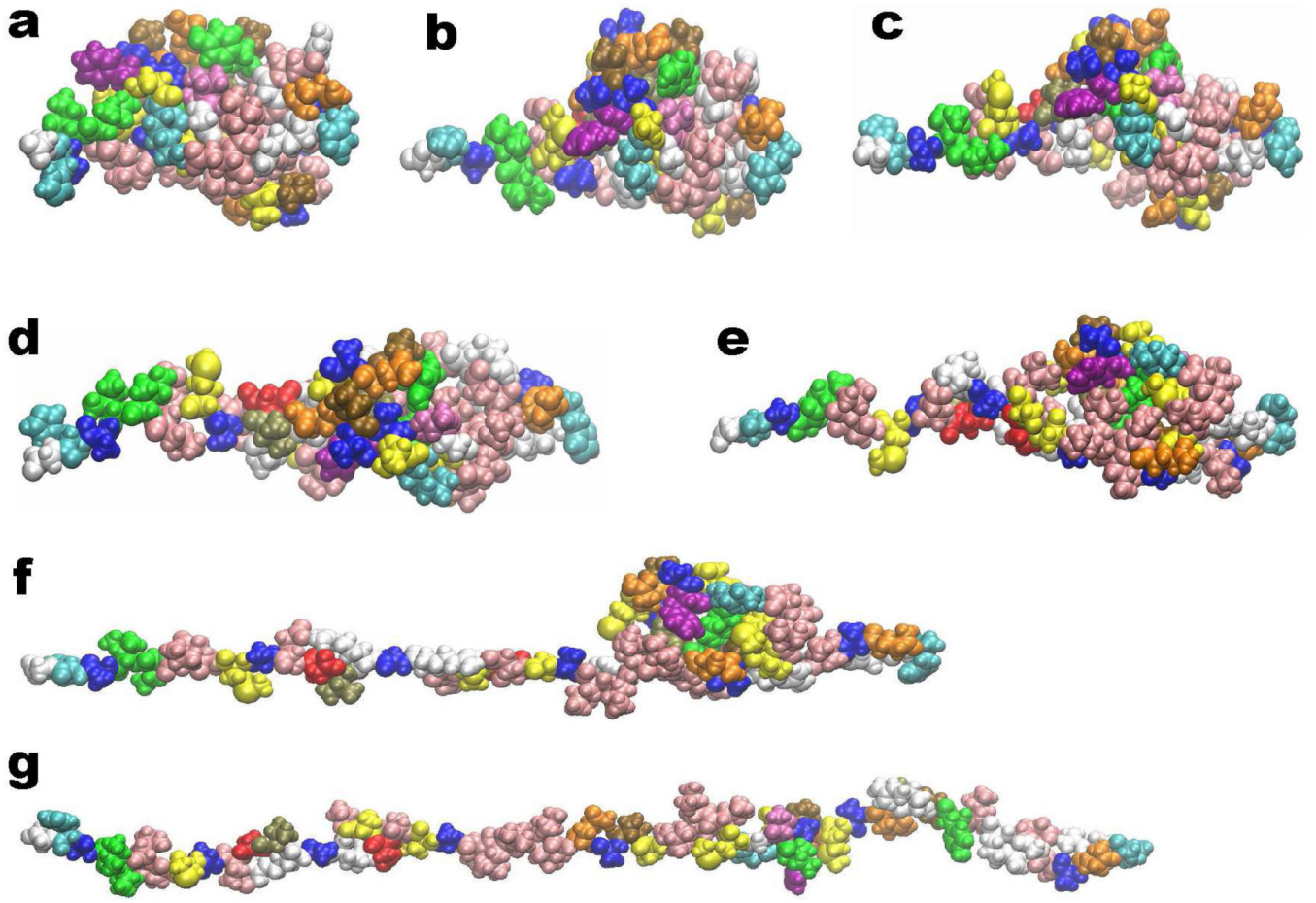

Figure 19.

The unfolding configurations of protein 1I2T obtained from the steered molecular dynamics with the constant velocity pulling algorithm. Charts **a, b, c, d, e, f** and **g** are the corresponding configuration frames 1, 3, 5, 7, 10, 20, and 30. Amino acid residues are labeled with different colors. From **a** to **g**, protein topological connectivity decreases, while protein total energy increases.

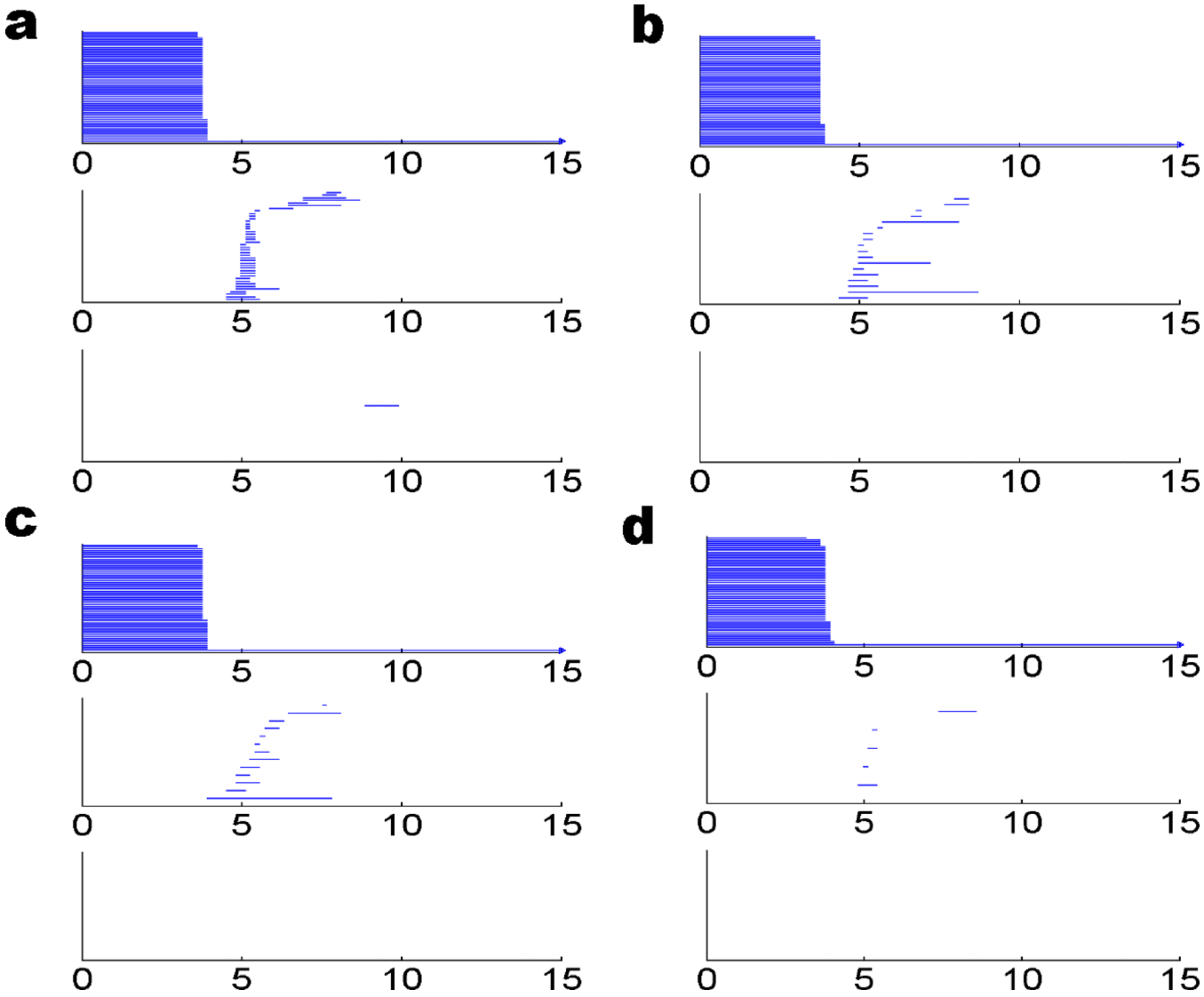

Figure 20.

Topological fingerprints of four configurations of protein 1I2T generated by using the distance based filtration. Charts **a, b, c**, and **d** are for frame 1, 10, 20 and 30, respectively (see Figs. 19a, 19e, 19f and 19g for their geometric shapes.). It can seen that as the protein unfolds, the (β_0 bars are continuously decreasing, which corresponds to the reduction of topological connectivity among protein atoms.

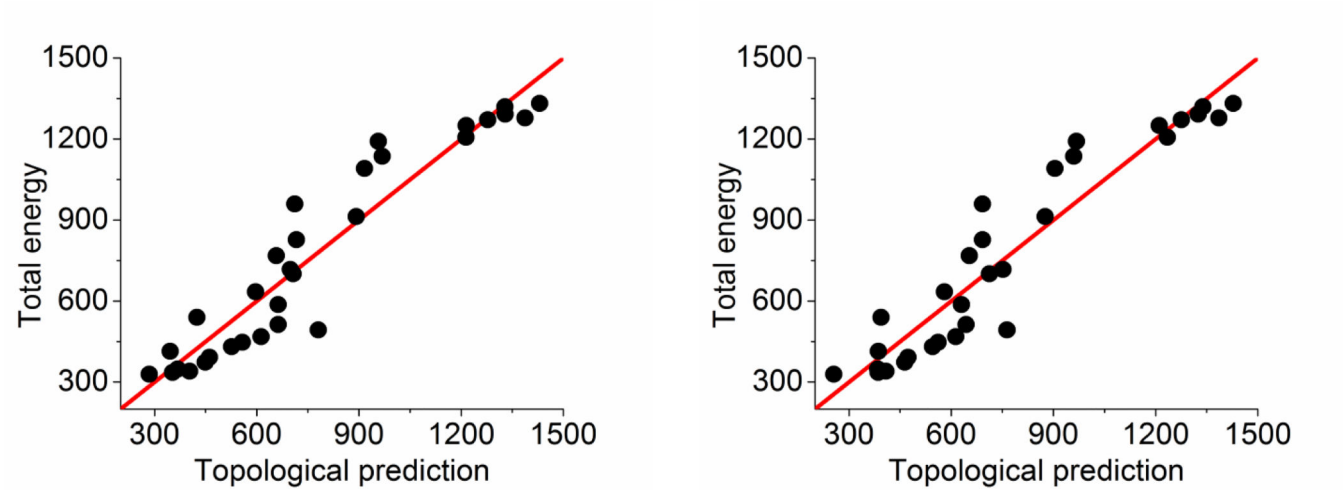

Figure 21.

Comparison between the total energies and the persistent homology prediction for 31 configurations of protein 1I2T. The unfolding configurations are generated by using the SMD. The negative accumulation bar length of $\beta_1(A_1^-)$ is used in the persistent homology prediction with both distance based filtration (left chart) and correlation matrix based filtration (right chart). Their correlation coefficients are 0.947 and 0.944, respectively. Clearly, there is a linear correlation between the negative accumulation bar length of $\beta_1(A_1^-)$ and total energy.

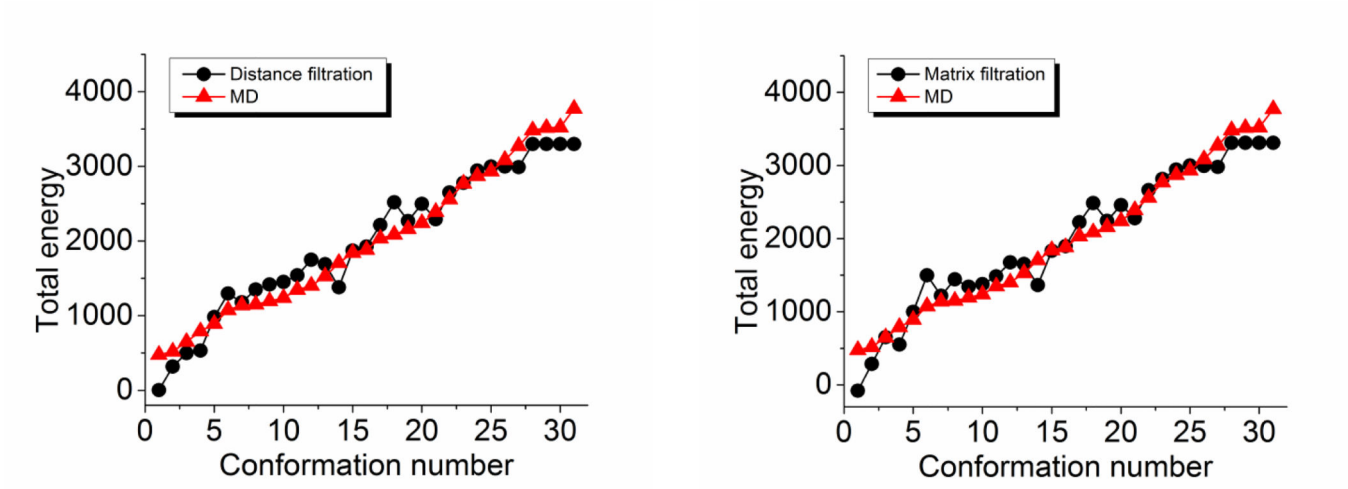

Figure 22.

Comparison between the total energy and the persistent honiology prediction for 31 configurations of protein 2GI9. The negative accumulation bar length of $\beta_1(A_1^-)$ is used in the persistent honiology prediction with both distance based filtration (left chart) and correlation matrix based filtration (right chart). Their correlation coefficients are 0.972 and 0.971, respectively. The linear correlation between the negative accumulation bar length and total energy is confirmed.

Table 1

A summary of Betti number and Euler characteristic in Fig. 3. Symbols V, E, F, and C stand for the number of vertices, edges, faces, and cells, respectively. Here χ is the Euler characteristic.

Simplex	***V***	***E***	***F***	***C***	β_0	β_1	$\beta 2$	χ
Figure 3$\mathbf{a}_1$	4	6	0	0	1	3	0	−2
Figure 3$\mathbf{a}_2$	4	6	4	0	1	0	1	2
Figure 3$\mathbf{a}_3$	4	6	4	1	1	0	0	1
Figure 3$\mathbf{b}_1$	8	12	0	0	1	5	0	−4
Figure 3$\mathbf{b}_2$	8	12	6	0	1	0	1	2
Figure 3$\mathbf{b}_3$	8	12	6	1	1	0	0	1

Table 2

The 30 proteins used in our B-factor prediction and persistent homology analysis of optimal characteristic distance.

PDB ID	PDB ID	PDB ID	PDB ID	PDB ID	PDB ID	PDB ID	PDB ID	PDB ID	PDB ID
1BX7	1DF4	1ETL	1FF4	1GK7	1GVD	1HJE	1KYC	1NKD	1NOT
1O06	1OB4	1OB7	1P9I	1PEF	1Q9B	1UOY	1VRZ	1XY1	1XY2
1YZM	2BF9	2JKU	2OL9	2OLX	3E7R	3MD4	3PZZ	3Q2X	4AXY