Bull Math Biol (2018) 80:1–31
https://doi.org/10.1007/s11538-017-0362-6

Society for Mathematical Biology

RESEARCH METHODS ARTICLE

Multiscale Persistent Functions for Biomolecular Structure Characterization

Kelin Xia[1,2] · **Zhiming Li**[3] · **Lin Mu**[4]

Received: 15 November 2016 / Accepted: 19 October 2017 / Published online: 2 November 2017

Abstract In this paper, we introduce multiscale persistent functions for biomolecular structure characterization. The essential idea is to combine our multiscale rigidity functions (MRFs) with persistent homology analysis, so as to construct a series of multiscale persistent functions, particularly multiscale persistent entropies, for structure characterization. To clarify the fundamental idea of our method, the multiscale persistent entropy (MPE) model is discussed in great detail. Mathematically, unlike the previous persistent entropy (Chintakunta et al. in Pattern Recognit 48(2):391–401, 2015; Merelli et al. in Entropy 17(10):6872–6892, 2015; Rucco et al. in: Proceedings of ECCS 2014, Springer, pp 117–128, 2016), a special resolution parameter is incorporated into our model. Various scales can be achieved by tuning its value. Physically, our MPE can be used in conformational entropy evaluation. More specifically, it is found that our method incorporates in it a natural classification scheme. This is achieved through a density filtration of an MRF built from angular distributions. To further validate our model, a systematical comparison with the traditional entropy evaluation model is done. It is found that our model is able to preserve the intrinsic topological features of biomolecular data much better than traditional approaches, particularly for resolutions in the intermediate range. Moreover, by comparing with

✉ Kelin Xia
xiakelin@ntu.edu.sg

1 Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore

2 School of Biological Sciences, Nanyang Technological University, Singapore 637371, Singapore

3 Key Laboratory of Quark and Lepton Physics (MOE) and Institute of Particle Physics, Central China Normal University, Wuhan 430079, China

4 Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

traditional entropies from various grid sizes, bond angle-based methods and a persistent homology-based support vector machine method (Cang et al. in Mol Based Math Biol 3:140–162, 2015), we find that our MPE method gives the best results in terms of average true positive rate in a classic protein structure classification test. More interestingly, all-alpha and all-beta protein classes can be clearly separated from each other with zero error only in our model. Finally, a special protein structure index (PSI) is proposed, for the first time, to describe the "regularity" of protein structures. Basically, a protein structure is deemed as regular if it has a consistent and orderly configuration. Our PSI model is tested on a database of 110 proteins; we find that structures with larger portions of loops and intrinsically disorder regions are always associated with larger PSI, meaning an irregular configuration, while proteins with larger portions of secondary structures, i.e., alpha-helix or beta-sheet, have smaller PSI. Essentially, PSI can be used to describe the "regularity" information in any systems.

Keywords Conformational entropy (CE) · Persistent entropy · Multiscale rigidity function (MRF) · Multiscale persistent function (MPF) · Multiscale persistent entropy (MPE) · Protein structure · Persistent homology

1 Introduction

In the past decade, persistent homology has been developed as a new multiscale representation of topological features (Edelsbrunner et al. 2002; Zomorodian and Carlsson 2005, 2008). As an essential part of topological data analysis, persistent homology models have undergone particular growth in the area of data analysis and opened up new opportunities for researchers from mathematics, computer sciences, computational biology, biomathematics, engineering, etc. Frosini and Landi (1999) have introduced the basic concepts in terms of size theory in shape recognition. A general form is further developed by Robins (1999), Edelsbrunner et al. (2002) and Zomorodian and Carlsson (2005), independently. To further implement and advance the theory, many efficient softwares, including JavaPlex (Tausz et al. 2011), Perseus (http://www.sas.upenn.edu/~vnanda/perseus), Dipha (Bauer et al. 2014), Dionysus (http://www.mrzv.org/software/dionysus), jHoles (Binchi et al. 2014), have been proposed (Bubenik and Kim 2007; Edelsbrunner and Harer 2010; Dey et al. 2008; Dey and Wang 2013; Mischaikow and Nanda 2013). Visualization methods, including persistent diagram (Mischaikow and Nanda 2013), persistent barcode (Ghrist 2008), persistent landscape (Bubenik 2015), have also been proposed. Persistent homology is deeply rooted in algebraic topology but finds great potential in the simplification of complex data (Frosini and Landi 1999; Robins 1999; Edelsbrunner et al. 2002; Zomorodian and Carlsson 2005). Unlike the traditional topological method, persistent homology provides a multiscale topological representation. It is able to measure the persistence of topological invariants and provide a bridge between geometry and topology. The essence of persistent homology is its filtration process. By systematically varying a filtration parameter, a series of topological spaces on various scales are generated. During a filtration, homology generators will be produced and further survive or persist for some time. Their lifespans or persistent times give a relative

geometric measurement of the associated topological properties. Persistent homology finds great success in topological characterization, identification and analysis. It has been used in a variety of fields, including shape recognition (Di Fabio and Landi 2011), network structure (Silva and Ghrist 2005; Lee et al. Dec 2012; Horak et al. 2009), image analysis (Carlsson et al. 2008; Pachauri et al. Oct 2011; Singh et al. 2008; Bendich et al. 2010; Frosini and Landi 2013), data analysis (Carlsson 2009; Niyogi et al. 2011; Wang et al. 2011; Rieck et al. 2012; Liu et al. 2012), chaotic dynamics verification (Mischaikow et al. 1999; Kaczynski et al. 2004), computer vision (Singh et al. 2008) and computational biology (Kasson et al. 2007; Yao et al. 2009; Gameiro et al. 2013). Recently, we have introduced persistent homology for structure characterization and mathematical modeling of fullerene molecules, proteins and other biomolecules (Xia and Wei 2014; Xia et al. 2015; Wang and Wei 2016). Consistent barcode patterns are extracted and defined as molecular topological fingerprint, which is used in the analysis of protein flexibility and protein folding (Xia and Wei 2014). We have also developed multiresolution and multidimensional persistence (Xia and Wei 2015a, b).

Although persistent homology has great potential for big data analysis, computationally it becomes unaffordable when data size gets very large. Various different definitions of complexes, including Alpha complex (Edelsbrunner and Mucke 1994) and witness complex (Carlsson 2014), are proposed to solve the problem. More recently, we have introduced a multiresolution/multiscale persistent homology (MPH) model by matching the resolution with the scale of interest so as to represent large-scale datasets with appropriate resolution. In this model, multiscale rigidity functions (MRFs) are employed to transform a point cloud data into various density maps in different resolutions. The MRF is derived from our flexibility and rigidity index (FRI), which is used to model the biomolecular flexibility (Xia et al. 2013; Opron et al. 2014). Our MRF has incorporated in it a special resolution/scale parameter; by appropriately tuning its value, we are able to focus the topological lens on the scale of interest. Our MPH model is validated by different types of point cloud data. It has been successfully used in the study of topological properties of a virus capsid with about 30,000 atoms, which would otherwise be inaccessible to common persistent homology models. More interestingly, by using density filtration, our MPH incorporates in it a clustering scheme. To be more specific, when we transform a point cloud data into a density map, regions with more points will have higher densities and will persist longer during the density filtration. In this way, the number together with the lengths of the β_0 barcodes provides a classification of the data. This lays the foundation for a more quantitative analysis by persistent functions.

In this paper, we propose a series of multiscale persistent functions for the quantitative analysis of MPH. We illustrate our idea with a multiscale persistent entropy (MPE) model. Persistent entropy (Chintakunta et al. 2015; Merelli et al. 2015; Rucco et al. 2016, 2017) is a new concept proposed recently. It has been used to find the "minimal" barcodes (Chintakunta et al. 2015), characterize complex systems (Merelli et al. 2015), study idiotypic immune networks (Rucco et al. 2016) and compare noisy signals (Rucco et al. 2017). The definition of persistent entropy is very natural. Simply speaking, each bar in the barcodes can be viewed as a certain "state," and its barlength represents the relative probability of this "state." In this way, persistent entropy is defined by using the Shannon entropy formula. Even though it is a very simple for-

mula, persistent entropy, based on the essential topological information, has a great promise to quantitatively characterize disorder of system or data. It is worth mentioning that persistent entropy is different from topological entropy (Bowen 1973), which is a topological invariant for dynamical systems. Unlike the previous persistent entropy model, our MPE incorporates in it a resolution/scale parameter. By turning this special parameter, we are able to evaluate the persistent entropy on any scale we interested. The essential idea of our MPE is to define the entropy from our MPH model. Since MPH is essentially a multiscale representation, which is based on MRF, our MPE model can characterize the disorder of the biomolecular data from various scales. To demonstrate the potential of our MPE, we consider the problem of evaluation of biomolecular conformational energy (CE).

The computational estimation of conformational energy is a long-standing problem and challenge in computational chemistry (Baron et al. 2009). Characterization of biomolecular conformations by some structural parameters has been proposed to estimate CE (Trbovic et al. 2008). Particularly, backbone dihedral angles and side chain rotamers are structural measurements that are widely used for protein CE evaluation. By the assumption that amino acid distributions in native states of proteins are comparable to that found in denatured states, Stites and Pranata (1995) propose a way to evaluate the relative CEs for different amino acids. They analyze the preferred distribution of amino acid residues by systematically studying about 12,000 residues from 61 non-homologous and high-resolution protein crystal structures. Ramachandran plots for various amino acid residues are obtained, and CEs are evaluated through the discretization of angle distributions with an uniform grid. The dihedral angle- and side chain rotamer-based parameterization has been widely used in biomolecular CE estimation (Doig and Sternberg 1995; Brady and Sharp 1997; Zhong et al. 2006; Zhang et al. 2008; Baruah et al. 2015).

The dihedral angle-based entropy evaluation method usually involves a discretization of angle distributions. It is found that the entropy calculated in this way is sensitive to the grid size (Stites and Pranata 1995; Trbovic et al. 2008). It is true that when the grid is very fine, angles with very similar values can still be classified into different categories. In contrast, when the grid is very coarse, even angles with huge different values tend to be classified into a same category. Therefore, dramatically different entropy values can be obtained from the same data under different discretizations. However, it has also been pointed out that correlation coefficients between entropies computed from different meshes are very high, suggesting that mesh-related bias does not systematically alter relative entropy values (Stites and Pranata 1995; Trbovic et al. 2008). But this consistence is highly related to the studied systems. For example, if the same type of amino acid is considered, their angles are highly concentrated in a particular area thus a more consistent entropy values. While if various types and numbers of amino acids are considered simultaneously, their angles will scatter around and inconsistent entropy values will be found. Researchers have realized that a lack of a robust classification poses challenge to a rigorous estimation of the entropies. Recently, a K-mean clustering method is proposed to deliver an optimized discrete k-state classification model (Zhang et al. 2008). In this method, the distribution of torsional angles is naturally classified into k clusters with irregular boundaries. In this way, it achieves an optimized classification and a high Silhouette value, indicating

very good quality of clustering, is obtained. Motivated by its success, we proposed a new way of CE evaluation by our MPE model.

To validate our MPE model, we have systematically compared it with the traditional model in various grid sizes. It has been found that similar results can be obtained in two extreme situations, i.e., when both grid spacing and resolution value are very small or very large. This corresponds with its physical implication very well. When a minimal grid spacing or resolution value is used, both entropies show log linear dependence on the total number of points in the angular plot. When a maximal grid spacing or resolution value is used, both entropies reduce to zero as all angular points are classified into one cluster. Further, in the middle range, where traditional entropy is usually evaluated, different results, however, are obtained. This means our MPE model differs from the traditional ones. To further explore the advantage and limitation of our model, we study a classic protein structure classification problem. In general, proteins can be classified into three categories, i.e., all-alpha (AA) proteins, all-beta (AB) proteins and mixed-alpha-and-beta (MAB) proteins. By comparing with traditional entropies from various grid sizes, bond angle-based methods and a persistent homology-based support vector machine method (Cang et al. 2015), we find that our MPE method gives the best results in terms of average true positive rate (TPR). More interestingly, with suitable threshold values, AA and AB protein classes have dramatically different MPE values and can be clearly separated from each other only by our model. Further, based on our MPE, a protein structure index (PSI) is proposed to describe the "regularity" of protein structures. The essential idea of PSI is to evaluate disorder in the angle distributions. Simply speaking, for a highly "regular" structure element like alpha-helix, its dihedral and bond angles are very consistent and thus contribute very little to the total entropy. Loops and intrinsically disorder regions are extremely "irregular" in terms of dihedral and bond angles and tend to contribute a large weight in the total entropy. In this way, PSI is able to provide a new way of structure regularity characterization. Essentially, PSI can be used to describe the regularity information in any system.

The paper is organized as follows: Section 2 is devoted to the introduction of basic method. The persistent homology analysis is reviewed in Sect. 2.1, which includes the basic theory of simplicial complex, filtration, complex construction and persistence. Section 2.2 is devoted for the multiscale/multiresolution persistent homology. In this section, we discuss the multiscale rigidity function, which is essentially a continuous version of our rigidity index in FRI model. Through a density filtration of MRF, a MPH model is constructed. By turning the resolution parameter, MPH can be used to study topological properties on various scales. Further, multiscale persistent functions are introduced in Sect. 2.3. Particularly, we propose the multiscale persistent entropy. To demonstrate the potential of our multiscale persistent functions, we use MPE as an example and discuss its application in conformational entropy calculation. Section 3 is devoted to the validation of our MPE. A brief introduction of CE, including conformation representation, traditional entropy models and our MPE for CE evaluation, is given in Sect. 3.1. Sections 3.2 and 3.3 are devoted for a classic protein structure classification test and the protein structure index. The paper ends with a conclusion.

2 Method

Algebraic topology is a very important mathematical tool for the study of global connectivity and topological invariants in the structure (Hatcher 2001; Munkres 1984; Edelsbrunner et al. 2002). It explores topological properties with algebraic tools, like Abelian group, quotient group, homomorphism, isomorphism, homology. Persistent homology is a newly invented algebraic topology method for structure characterization. It is able to bridge the gap between topology and geometry. In this section, we give a brief introduction of basic concepts in persistent homology. After that, we present our multiscale persistent homology and multiscale persistent functions. MPH is originally proposed to characterize topological information on various scales. Based on MPH, we introduce several multiscale persistent functions, particularly the multiscale persistent entropy. A more detailed description is given below.

2.1 Persistent Homology

As an important component of topological data analysis, persistent homology has attracted attention from researchers in various fields. Unlike the previous topological tools, it enables a geometric measurement of homology generators and demonstrates great power in not only qualitative but also quantitative characterization of structures. Two essential components of persistent homology are simplicial homology and filtration process.

2.1.1 Simplicial Homology

Simplicial complex is a finite set of simplices, which can be simply understood as vertices, edges, triangles and their high-dimensional counterparts. Simplicial complex, including geometric simplicial complex and abstract simplicial complex, is not a topological space, but it can be topologized as a subspace of $\mathbb{R}^n$ called polyhedron. Groups and group operations can be defined on simplices. In this way, algebraic tools, particularly homology analysis, can be used to analyze the topological properties.

Simplicial complex Simplices are building blocks for simplicial complex. A k-simplex is the convex hull of $k+1$ affinely independent points in $\mathbb{R}^n$ $(n > k)$. For a set of $k+1$ affinely independent points $v_0, v_1, v_2, \cdots, v_k$, a correlated k-simplex $\sigma^k = \{v_0, v_1, v_2, \cdots, v_k\}$ can be expressed as

$$\sigma^k = \left\{ \lambda_0 v_0 + \lambda_1 v_1 + \cdots + \lambda_k v_k \mid \sum_{i=0}^{k} \lambda_i = 1; 0 \leq \lambda_i \leq 1, i = 0, 1, \cdots, k \right\}. \quad (1)$$

A geometric k-simplex σ^k is a closed convex subspace of $\mathbb{R}^n$. Its i-dimensional face is the convex hull formed by $i+1$ 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.

A simplicial complex K is a finite set of simplices that satisfy two basic conditions, i.e., (1) any face of a simplex from K is also in K; (2) the intersection of any two simplices in K is either empty or shared faces. The dimension of a simplicial complex is the maximal dimension of its simplices. The polygon $|K|$ is a topological space formed by the union of all simplices of K. In order to study the topological space with algebraic tools, we need to introduce the concept of chain.

Homology A k-chain c is a linear combination of k-simplices, i.e., $c = \sum_i \alpha_i \sigma_i^k$. Coefficients α_i can be selected from various groups, including rational field $\mathbb{Q}$, integer field $\mathbb{Z}$ and prime integer field $\mathbb{Z}_p$ with prime number p. In computational topology, coefficients α_i is chosen from group $\mathbb{Z}_2$ for simplicity. An Abelian group $C_k(K, \mathbb{Z}_2)$ is formed by the set of all k-chains from the simplicial complex K together with addition operation (modulo-2).

The boundary operation is essential to the definition of homology. A boundary operator ∂_k is defined as $\partial_k : C_k \to C_{k-1}$. The boundary of an oriented k-simplex $[\sigma^k] = [v_0, v_1, v_2, \cdots, v_k]$ can be denoted as,

$$\partial_k[\sigma^k] = \sum_{i=0}^{k} (-1)^i [v_0, v_1, v_2, \cdots, \hat{v_i}, \cdots, v_k]. \tag{2}$$

Here $[v_0, v_1, v_2, \cdots, \hat{v_i}, \cdots, v_k]$ means a $(k-1)$ oriented simplex, which is generated by the elimination of vertex v_i. An oriented simplex is a simplex together with an orientation, i.e., ordering of its vertex set. Also we have $\partial_0 = 0$. From its definition, it can be found that if applying the boundary operation twice, any k-chain will be mapped to a zero element as $\partial_{k-1}\partial_k = 0$. Further, 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,

$$Z_k = \text{Ker } \partial_k = \{c \in C_k \mid \partial_k c = 0\}, \tag{3}$$

$$B_k = \text{Im } \partial_{k+1} = \{c \in C_k \mid \exists d \in C_{k+1} : c = \partial_{k+1} d\}. \tag{4}$$

Their elements are called the k-th cycle and the k-th boundary, respectively. It can be noticed that $B_k \subseteq Z_k$, as the boundary of a boundary is always zero $\partial_{k-1}\partial_k = 0$. 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: $H_k = Z_k / B_k$. The rank of k-th homology group is called k-th Betti number and it can be calculated by

$$\beta_k = \text{rank } H_k = \text{rank } Z_k - \text{rank } B_k. \tag{5}$$

If we consider a chain group $C_k(K, Z)$ in the field Z, based on the fundamental theorem of finitely generated Abelian group, homology group H_k can be further expressed as a direct sum,

$$H_k = Z \oplus \cdots \oplus Z \oplus Z_{p_1} \oplus \cdots \oplus Z_{p_n} = Z^{\beta_k} \oplus Z_{p_1} \oplus \cdots \oplus Z_{p_n}, \tag{6}$$

where the rank of free subgroups is the k-th Betti number β_k. Here Z_{p_i} are torsion groups with torsion coefficients $\{p_i | i = 1, 2, ..., n\}$.

Simply speaking, geometric meanings of Betti numbers in $\mathbb{R}^3$ are as follows: β_0 represents the number of isolated components, β_1 is the number of one-dimensional loops, circles or tunnels, and β_2 describes the number of two-dimensional voids or cavities. Together, the Betti number sequence $\{\beta_0, \beta_1, \beta_2, \cdots\}$ describes intrinsic topological properties of a system. Betti numbers are important topological invariants.

2.1.2 Persistent Homology Analysis

In computational geometry and topology, a big problem is to recover the original topological space from a sampled point cloud data. The simplest and also the most widely used method is to employ an open covering with a consistent radius parameter ϵ. However, how to find the best suitable ϵ for the underling space has puzzled researchers for a long time. It is true that when ϵ is too small, originally connected regions may not be fully recovered. But when ϵ is too large, originally non-connected regions may be mistaken as connected. To solve this problem, a brilliant idea has been proposed and it is known as filtration. In the filtration process, through a systematical investigation of a wide range of ϵ values, a series of topological spaces from various scales have been generated. It is found that some topological invariants last for a large range of ϵ values, but some invariants disappear very quickly when the scale changes. State different, the calculated topological invariant has a certain "lifespan" or persistence. This means that a special geometric measurement (range of ϵ) can be assigned to each topological invariant. This method is known as the persistent homology. It differs greatly from the traditional geometric and topological methods by the incorporation of geometric information into topological invariants. It can work as a bridge between geometry and topology (Bubenik and Kim 2007; Edelsbrunner and Harer 2010; Dey et al. 2008; Dey and Wang 2013; Mischaikow and Nanda 2013).

General filtration processes Filtration is of great importance to persistent homology. A suitable filtration is key to the persistent homology analysis. In practice, two filtration algorithms, Euclidean distance-based and density-based ones, are commonly used. These filtration processes can be modified in many different ways to address physical needs as shown in our previous papers (Xia and Wei 2014; Xia et al. 2015; Wang and Wei 2016).

The Euclidean distance-based filtration is straightforward. We assign each point in the data a sphere with an ever-increasing radius. When these spheres gradually overlap with each other, complexes can be constructed by using different definitions, such as Čech complex, Rips complex and Alpha complex (Edelsbrunner and Mucke 1994). More importantly, during a filtration, previously formed simplicial complexes are included into latter ones. We demonstrate the filtration process in Fig. 1. We consider a fullerene C_{60} molecule, which is constructed by 60 carbon atoms. We associate each carbon atom with a sphere. During the filtration process, the sphere is systematically increased.

Another important filtration process is the density-based filtration process. In this process, the filtration goes along the increase or decrease in the density value. In

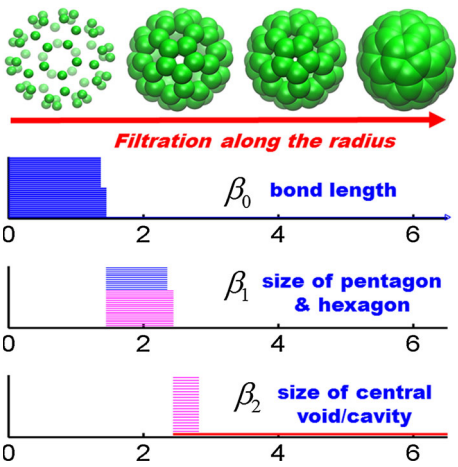

Fig. 1 Filtration process and barcode representation of a fullerene C_{60} molecule. During the filtration process, each carbon atom is associated with a sphere, whose radius increases systematically to generate topologies on various scales. In the molecular barcode representation, each bar represents a homology generator and has unique chemical or physical properties. For β_0 bars, they are related to atomic bonds. The pentagon and hexagon ring structures are represented by β_1 bars. The void or cavity structures are captured in β_2 bars (Color figure online)

this way, a series of isosurfaces are generated. Morse complex (Mischaikow and Nanda 2013) is used for the characterization of their topological invariants. Persistence information can be derived from these complexes. A more rigorous mathematical formulation is given in the following.

Persistent homology The filtration can be described as a nested sequence of its subcomplexes,

$$\varnothing = K^0 \subseteq K^1 \subseteq \cdots \subseteq K^m = K. \tag{7}$$

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 p-persistent k-th homology group at filtration time i can be represented as

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

Through the study of the persistent pattern of these topological features, the persistent homology is capable of capturing the intrinsic properties of the underlying space.

Barcode representation To visualize the persistent homology results, many elegant representation methods have been proposed, including persistent diagram (Mischaikow and Nanda 2013), persistent barcode (Ghrist 2008), persistent landscape (Bubenik 2015).

In this paper, we use barcode representation. Basically, barcodes are clusters of bars. Each of these bars represents a homology generator with "birth" and "death" time as its starting and ending points. In this way, the length of bar tells how long the homology generator "lives" or "persists." Figure 1 demonstrates the barcodes of fullerene C_{60}.

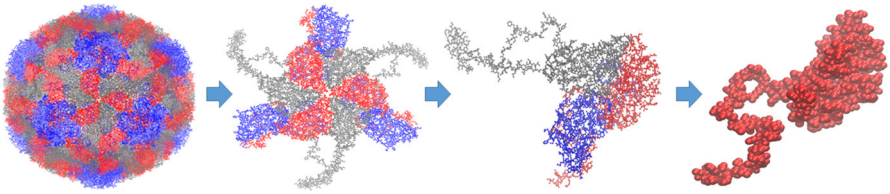

Fig. 2 Illustration of multiscale features in picornavirus capsid (ID: 5APM). The virus capsid is a highly symmetric structure made of protein complexes. Each protein complex is composed of several proteins. And a protein usually has several chains. If zoom in further, we can also arrive at residual level and atomic level (Color figure online)

It is worth mentioning that for molecular or biomolecular data, all bars, no matter long or short, are important. This is because that bars are tightly associated with structural, physical or chemical properties. For example, β_0 bars in Fig. 1 represent C–C bond lengths of fullerene C_{60} molecule. Actually, it can be seen that there are two types of C-C bonds in C_{60} molecule and their lengths are around 1.37 Å and 1.45 Å. This corresponds very well with the chemical properties of the C_{60} fullerene (Xia et al. 2015). Further, we can find that there are two types of β_1 bars. Structurally, they represent the pentagon and hexagon rings formed from the adjacent five or six carbon atoms. And their barlengths are measurement of the ring sizes. Finally, the longest bar in β_2 represents the void or cavity inside C_{60} molecule and its barlength gives the size.

2.2 Multiscale Persistent Homology

With the great advancement of experimental tools and the accumulation of gigantic amount of data, we are able to observe and study the world from various scales. Traditionally, we study biology at the level of population, tissue and cell. Nowadays, we can obtain biomolecular information from subcellular, molecular and atomic scale. Especially, in molecular biology and structural biology, enormous efforts have been devoted to acquire biomolecular information in atomic detail. This is because that most of important biological processes, including transcription, translation, enzyme interaction, protein folding, protein–protein interaction, ion transportation, happen at atomic level. However, the obtained biomolecular data often involves excessively high degrees of freedom and high dimensionality, thus, it is computationally prohibitively expensive. For example, if we want to explore the molecular mechanics of a human immunodeficiency virus (HIV) capsid, we are facing with 4.2 million atoms. With each atom moves in $\mathbb{R}^3$ space, this gives rise to a problem of 12.6 million degrees of freedom. On the other hand, large biomolecules are essential multiscale, ranging from atom, residue, domain, protein monomer, protein polymer to the whole biomolecular assembly. Figure 2 illustrates the multiscale property in the picornavirus capsid.

To address the challenge of the complexity and multiscale nature of the biomolecular data, multiscale modelings are widely used in biophysics, biochemistry and computational biology. More recently, persistent homology has been advocated as a new strategy for the topological simplification of complex data. Since topological representations can dramatically reduce geometric details leaving only the essential global

connectivity information, it provides a great promise for complexity and dimensionality reduction. However, the direct application of persistent homology analysis to large biomolecules, particularly macroproteins and protein assemblies is currently unfeasible. One of the major reasons is that a uniform resolution is used in the filtration. Therefore, cross-scale filtration at a high resolution is prohibitively expensive computationally. With this consideration, we proposed a multiresolution/multiscale persistent homology model. The essential idea is to match the scale of interest with appropriate resolution in the topological analysis. Simply speaking, this works like using a microscope. By tuning a specially designed resolution parameter to the proper value, we can focus our topological representation on the right scale. Since different resolutions give rise to topological information on different scales, our multiresolution persistent homology is essentially multiscale. So it is also called multiscale persistent homology and shares the same abbreviation MPH.

Multiscale rigidity function, which is derived from our flexibility and rigidity index (FRI) method (Xia et al. 2013; Opron et al. 2014), is essential to our MPH. With this function, a discrete point cloud data is converted into a continuous density function. The conversion is realized by using a kernel function with a resolution parameter. And this special parameter enables us to facilitate a multiscale analysis of complex data. More details will be discussed below.

Multiscale rigidity function Multiscale rigidity function is derived from our flexibility and rigidity index model (Xia et al. 2013; Opron et al. 2014), in which flexibility index and rigidity index are two essential components. In our previous works, we find that these two indexes can be used to study biomolecular flexibility properties, particularly in experimental B-factor prediction (Opron et al. 2015, 2016; Xia et al. 2015; Nguyen et al. 2016). The continuous version of rigidity index is rigidity function, which is originally devised for the representation of biomolecular densities (Xia and Wei 2015a; Xia et al. 2015). We find that a rigidity function with a special resolution parameter can deliver a multiscale representation of biomolecular structures (Xia and Wei 2015a; Xia et al. 2015). We call it multiscale rigidity function.

For a data set with a total N entries, which can be physical elements like atoms, residues and domains or data components like points, pixels and voxels, if we assume their generalized coordinates are $\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_N$, a multiscale rigidity function can be derived from the data as,

$$\mu(\mathbf{r}; \eta) = \sum_j^N w_j \Phi(\|\mathbf{r} - \mathbf{r}_j\|; \eta), \tag{9}$$

where w_j and $\Phi(\|\mathbf{r} - \mathbf{r}_j\|; \eta)$ are weight and kernel function for the j-th atom. The kernel function satisfies the following admissibility conditions,

$$\lim_{\|\mathbf{r}-\mathbf{r}_j\|\to 0} \Phi(\|\mathbf{r} - \mathbf{r}_j\|; \eta) = 1; \tag{10}$$

$$\lim_{\|\mathbf{r}-\mathbf{r}_j\|\to \infty} \Phi(\|\mathbf{r} - \mathbf{r}_j\|; \eta) = 0. \tag{11}$$

Here $\eta > 0$ is a resolution parameter that can be adjusted to achieve the desirable resolution for a given scale. Commonly used correlation functions are generalized Gaussian functions,

$$\Phi(\|\mathbf{r}-\mathbf{r}_j\|;\eta,\kappa) = e^{-(\|\mathbf{r}-\mathbf{r}_j\|/\eta)^{\kappa}}, \quad \kappa > 0, \tag{12}$$

or generalized Lorentz functions,

$$\Phi(\|\mathbf{r}-\mathbf{r}_j\|;\eta,\upsilon) = \frac{1}{1+\left(\|\mathbf{r}-\mathbf{r}_j\|/\eta\right)^{\upsilon}}, \quad \upsilon > 0. \tag{13}$$

It can be noticed that the larger the η value, the lower the resolution is.

The MRF of the data can be expressed as,

$$\mu(\mathbf{r};\eta) = \sum_{j}^{N} w_j \Phi(\| \mathbf{r}-\mathbf{r}_j \|;\eta). \tag{14}$$

Generally, parameter w_j is chosen as the atomic number. For example, its value is 6 for carbon atom and 8 for oxygen atom.

A multiscale geometric model can be naturally derived from our MRFs. To illustrate this idea, an example of protein complex (ID: 3JBL) is demonstrated in Fig. 3a–c. We use the coarse-grained representation with totally 9999 C_α atoms. Generalized Gaussian kernel $\Phi(\|\mathbf{r}-\mathbf{r}_j\|;\eta) = e^{-(\frac{\|\mathbf{r}-\mathbf{r}_j\|}{\eta})^2}$ with η =1Å , 4Å and 20Å is employed in the multiscale rigidity function. The demonstrated surfaces are extracted by using isovalue 1, 10 and 600, respectively. It can be seen that these surfaces represent protein geometries from different scales and they have dramatically different physical implications. Roughly speaking, Fig. 3a–c are on the atomic scale, subdomain/domain scale and protein scale, respectively.

Multiscale persistent homology To enable a more quantitatively comparison between various scales, we propose MPH. In our model, we linearly rescale all the rigidity function values to the region [0, 1] using formula

$$\mu^s(\mathbf{r};\eta) = 1.0 - \frac{\mu(\mathbf{r};\eta)}{\mu_{\max}}. \tag{15}$$

Here $\mu(\mathbf{r})$ and $\mu^s(\mathbf{r})$ are the original and normalized rigidity function, and $\mu_{\max}$ is the maximum value of the original rigidity function. The filtration is done along the normalized rigidity function value from small to large. Figure 3d–f demonstrates the barcode results from the previously mentioned protein example. When η = 1Å, we can observe the atomic information in β_0 bars. There are around 10000 β_0 bars, corresponding to virtual bonds between adjacent C_α atoms. When the value η increases to 4 Å , both β_0 and β_1 bars decrease dramatically. Actually, barcodes capture the topological information from the scale of protein domain and protein subdomain. Further, when we enlarge η to 20 Å , the most global scale begins to emerge. We can

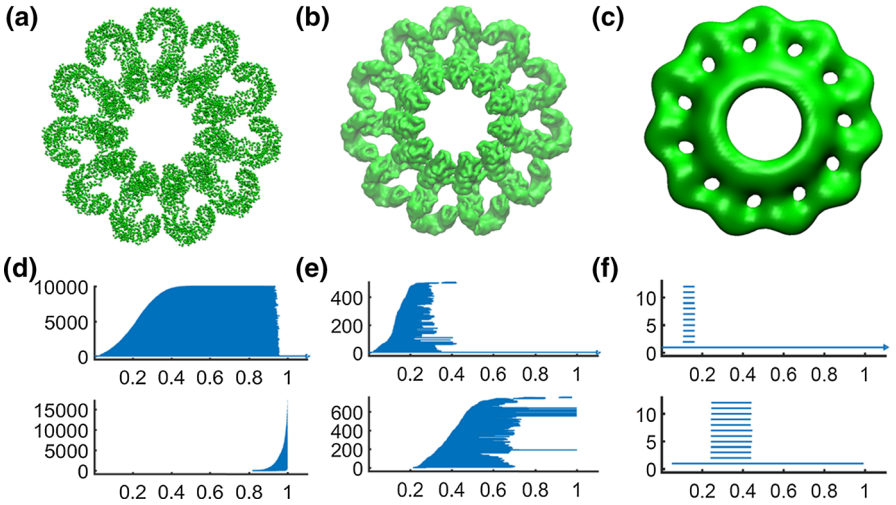

Fig. 3 Multiresolution persistent homology analysis of activated NAIP2-NLRC4 inflammasome structure (ID: 3JBL). **a**–**c** Isosurfaces of multiscale rigidity functions profiles generated at three resolutions $\eta = 1.0, 4.0$ and 20.0Å. The isosurface values are 1.0, 10.0 and 600.0, respectively. **d**–**f** The corresponding persistent barcode representations from multiscale persistent homology. Top and bottom panels are for β_0 and β_1 barcodes, respectively. The horizontal axes denote the normalized rigidity density value. It can be seen that with the decrease in resolution, the density profiles of protein isosurfaces shift from local details to global patterns

clearly identify 11 relatively small circles and 1 large circle in the structure. And the topological information is well captured in our persistent barcode. To sum up, based on MRF, our MPH is able to capture and characterize the topological information of the structure from various scales.

It should be noticed that our MPH is different from multidimensional persistent homology. Multidimensional persistent homology comes from higher-dimensional filtration process (Carlsson et al. 2009; Cohen-Steiner et al. 2006; Cerri and Landi 2013; Frosini and Landi 1999; Biasotti et al. 2008; Cerri et al. 2013). Its construction is nontrivial because theoretically, a complete discrete representation for multidimensional persistent module analogous to one-dimensional situation dose not exist (Carlsson and Zomorodian 2009). However, there are some computable incomplete invariants that can be used. For example, persistent Betti numbers (PBNs) (Edelsbrunner et al. 2002) have been proved to be stable in the constraint of certain marching distance (Cerri and Landi 2013). More recently, we have proposed two ways of constructing multidimensional persistent homology (Xia and Wei 2015a, b). The first one is constructed from some dynamic processes. In it, one more filtration parameter, which is along the evolution time, is considered. The second type is based on multiresolution process, and the resolution/scale parameter is considered as the new filtration parameter. Our MPH belongs to the second type. It should be noticed that, for a dynamic process, even though we can have multidimensional persistence, the whole system is characterized with a single scale and the associated persistent homology is not multiscale. It is also worth mentioning that our MPH is relatively stable under a small variation of resolu-

tion value. This is true because our MRF is continuous and a relatively small variation of resolution will only induce a small fluctuation in rigidity values. The stability of our model can be rigorously studied by using the "stability theorem" (Chazal et al. 2014).

With this brief introduction of MRF and MPH, now we are ready to discuss our MPFs.

2.3 Multiscale Persistent Functions

As stated above, a series of barcodes from various scales are generated in our MPH. We can represent them as follows:

$$\{L_{k,j}(\eta) = [a_{k,j}(\eta), b_{k,j}(\eta)]|k = 0, 1, 2, ...; j = 1, 2, 3, ..., N_k(\eta)\}, \quad (16)$$

where parameter k is the dimension of Betti number β_k, parameter j indicates the j-th bar, and N_k is the number of β_k barcodes. Further, we define the set of barcodes in the k-th dimension as,

$$L_k(\eta) = \{L_{k,j}(\eta), j = 1, 2, 3, ..., N_k(\eta)\}, \quad k \geq 0.$$

Based on these barcodes, various persistent functions can be defined. In our previous works (Xia and Wei 2014; Xia et al. 2015), we have defined the accumulated barlength (on a fixed resolution η_0) as,

$$A_b(L_k(\eta_0)) = \sum_j \left(b_{k,j}(\eta_0) - a_{k,j}(\eta_0)\right), \quad (17)$$

and used it to study the bond energy and the folding energy in protein unfolding simulation. In fullerene total curvature energy evaluation, the intrinsic barcode of β_2 (on a fixed resolution η_0) is,

$$I_b(L_2(\eta_0)) = \max\{b_{2,j}(\eta_0) - a_{2,j}(\eta_0)|j = 1, 2, ..., N_2(\eta_0)\}, \quad (18)$$

which is found to have important physical meanings (Xia et al. 2015). However, all these functions are defined on a single resolution, which has limited their power in characterizing topological properties of a multiscale system.

In this section, based on our MPF, we propose a series of MPFs, particularly MPE. These functions can be used for either visualization or quantitative analysis of multiscale topological properties of structures we interested.

2.3.1 Multiscale Persistent Entropy

Entropy is proposed for characterizing disorder of a system (Karplus and Kushick 1981; Brady and Sharp 1997). It measures the degrees of freedom for a system to evolve into various potential configurations. Entropy is a key property to understand a

wide variety of physical, chemical and biochemical phenomena. It plays very important roles in characterizing various biomolecular functions and interactions, including protein folding, protein–protein interaction, protein–ligand binding, chromosome configuration, DNA translation and transcription. For instance, the folding of a single peptide chain into a well-defined native structure is greatly facilitated by the reduction in its CE. Therefore, the study of entropy is very important in computational chemistry, biophysics and biochemistry.

Recently, persistent entropy has been proposed and is used to study complex systems (Merelli et al. 2015), networks (Rucco et al. 2016), noise signals (Rucco et al. 2017), etc (Chintakunta et al. 2015). Based on our MRF, we propose a MPE model as follows:

$$S_k(\eta) = \sum_{j}^{N_k(\eta)} -p_{k,j}(\eta) log(p_{k,j}(\eta)), \tag{19}$$

with the probability function

$$p_{k,j}(\eta) = \frac{b_{k,j}(\eta) - a_{k,j}(\eta)}{\sum_j \left(b_{k,j}(\eta) - a_{k,j}(\eta)\right)}. \tag{20}$$

The MPE can also be simplified as follows:

$$\begin{aligned} S_k(\eta) = log\left(\sum_{j}^{N_k(\eta)} \left(b_{k,j}(\eta) - a_{k,j}(\eta)\right)\right) \\ - \frac{\sum_j^{N_k(\eta)} \left((b_{k,j}(\eta) - a_{k,j}(\eta)) log(b_{k,j}(\eta) - a_{k,j}(\eta))\right)}{\sum_j^{N_k(\eta)} \left(b_{k,j}(\eta) - a_{k,j}(\eta)\right)}. \end{aligned} \tag{21}$$

Persistent entropy can be used to characterize the disorder of a system. Essentially, each bar in the barcodes can be viewed as an independent "state," and its length is the relative "probability" of this state. In this way, Shannon entropy concept can be naturally used to define persistent entropy. Generally speaking, our model is very similar to the previous persistent entropy, except that we have incorporated a resolution/scale parameter in our representation, which gives us more flexibility in structural analysis.

Other multiscale persistent functions Several other types of functions are also found to be interesting and useful. The first type is

$$f_1(x; L_k(\eta)) = \sum_j w_{k,j}(\eta) e^{-\left(\frac{x - \frac{b_{k,j}(\eta) + a_{k,j}(\eta)}{2}}{\sigma(\eta)(b_{k,j}(\eta) - a_{k,j}(\eta))}\right)^{\kappa}}, \quad \kappa > 0 \tag{22}$$

where $w_{k,j}(\eta)$ is the weight function for the j-th bar of β_k. Parameter $\sigma(\eta)$ is the characterization of "significance" of barcode in different lengths.

It should be noticed that there is no meaningful sequence arrangement for barcodes. However, for biomolecules, their barcodes are highly organized. Each bar or each type of bar has its unique structural, physical or chemical implication. With this consideration, we can assign a weight value $w_{k,j}(\eta)$ to each bar or each type of bars.

The second important type of persistent energy functions is

$$f_2(x; L_k(\eta)) = \sum_j w_{k,j}(\eta) \frac{1}{1 + \left(\frac{x - \frac{b_{k,j}(\eta) + a_{k,j}(\eta)}{2}}{\sigma(\eta)(b_{k,j}(\eta) - a_{k,j}(\eta))} \right)^{\upsilon}}, \quad \upsilon > 0. \tag{23}$$

It can be noticed that these two multiscale functions can be used to visualize and quantitatively characterize the barcodes information.

More interestingly, we can construct a persistent homology-based potential function (or energy function) E_{Top} as follows:

$$E_{\text{Top}} = \sum_{i=0}^{2} E(L_k(\eta)). \tag{24}$$

Here $E(L_k(\eta))$ is the energy contribution from β_k bars. Different physical models can be used. For example, we can choose a harmonic potential function and construct a persistent homology potential $E(L_k(\eta))$ as follows:

$$E(L_k(\eta)) = \frac{\gamma}{2} \sum_j |\Delta L_{k,j}(\eta)|^2, \tag{25}$$

where $\Delta L_{k,j}(\eta)$ is barlength variation for j-th bar of β_k, between the equilibrium and non-equilibrium structures. And γ is the spring constant. Ideologically, our persistent homology potential is similar to the statistical potential in protein structure prediction (Shen and Sali 2006). Instead of using the traditional angle and bond representation, statistical potential uses knowledge-based scoring function of residue contacts, while our persistent homology potential uses special topological invariants. The full potential of our persistent homology potential will require further investigation.

To demonstrate some potential applications of our MPFs, we use MPE as an example. In the following section, we will discuss the application of MPE in CE evaluation.

3 Application

In this section, we will give a brief review of conformational entropy. A comparison between the traditional entropy and our MPE will be discussed. To validate our model, we employ a classic test example, i.e., classification of all-alpha (AA), all-beta (AB) and mixed-alpha-and-beta (MAB) proteins. We find that in suitable resolutions, our MPE can be used to discriminate different protein configurations very efficiently. We further propose a protein structure index based on our entropy model.

3.1 Conformational Entropy

Biomolecular CE is of great importance (Frederick et al. 2007; Marlow et al. 2010) for the study of interaction between systems, like protein–protein, receptor–ligand, antigen–antibody, DNA–protein, RNA–ribosome (Gellman 1997; Brooijmans and Kuntz 2003; Janin et al. 2013). The understanding of interactions requires the characterization of binding process, in which CE plays an essential role (Frederick et al. 2007; Marlow et al. 2010). It is found that changes in protein CE can contribute significantly to the free energy of protein–ligand association (Frederick et al. 2007). Internal dynamics of the protein calmodulin varies significantly when binding to a target. The apparent change in the corresponding CE is linearly related to the change in the overall binding entropy. Also, CE of protein side chain is a major effect in the energetics of folding. The entropy of heterogeneous random coil or denatured proteins is significantly higher than that of the folded native state tertiary structure. In particular, CE of the amino acid side chains in a protein is thought to be a major contributor to the energetic stabilization of the denatured state and thus a barrier to protein folding. The reduction in the number of accessible main chain and side chain conformation when a protein folds into a compact globule, yields an unfavorable entropic effect. This reduction in CE counters the hydrophobic effect favoring the folded state and in part explains the marginal stability of most globular proteins.

Even though biomolecular conformational energy is a key property to understand a wide variety of physical, chemical and biochemical phenomena, its evaluation or calculation is very challenging both experimentally and computationally. Only recently, nuclear magnetic resonance relaxation methods for characterizing thermal motions on the picosecond–nanosecond timescale are developed and the resulting order parameters are used as a proxy for CE evaluation (Frederick et al. 2007; Trbovic et al. 2008; Sapienza and Lee 2010). Atomic force microscopy for unfolding has great potential in measuring the backbone CE for protein folding (Thompson et al. 2002). Neutron spectroscopy is also used to elucidate the role of CE upon thermal unfolding by observing the picosecond motions, which are dominated by side chain reorientation and segmental movements of flexible polypeptide backbone regions (Fitter 2003).

Computationally, entropy evaluation necessitates a characterization of biomolecular configuration spaces. Due to the limitation of Cartesian grid-based representation, structural parameters, particular dihedral angles and bond angles, are usually employed for structure description. Their distributions can be derived from various computational methods, including molecular dynamics, Monte Carlo simulation, normal mode analysis. Various microstates can be obtained by the discretization of the angle distribution, usually through an equal-spacing grid. And conformational entropy can be evaluated by using the classic Shannon entropy form. Even though it is suggested that the relative entropy is consistent in this procedural, a lack of a robust classification still poses a challenge to a rigorous estimation of entropy. A detailed discussion will be given below.

Conformation representation A quantitative evaluation of CE requires the characterization of biomolecular conformation spaces. Generally, dihedral angle models are employed. In these models, biomolecular structures are parameterized by their back-

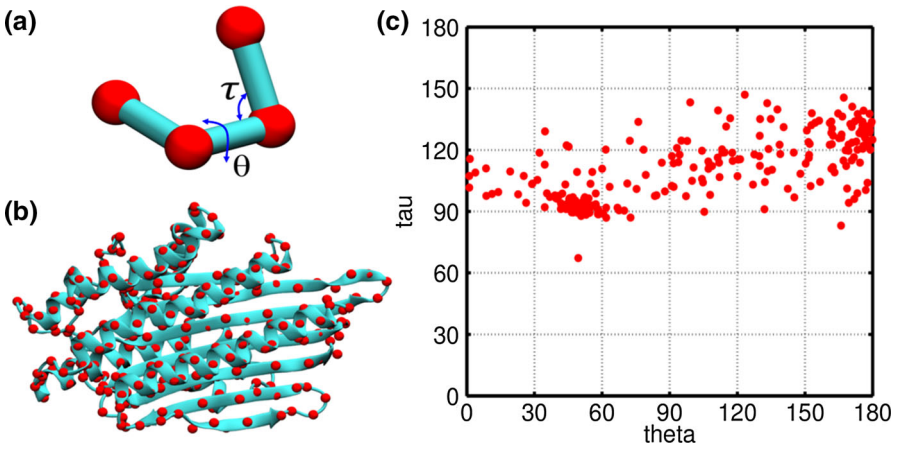

Fig. 4 Illustration of a (θ, τ) angle representation of protein 1VJU (chain A). The coarse-grained model with each residue represented by its C_α atom (red ball) is used. The virtual dihedral angle θ is rescaled to $[0^o, 180^\circ]$ by treating the parallel and antiparallel situations as the same. **a** Angle τ is the virtual bond angle between adjacent three C_α atoms, and angle τ is the dihedral angle formed by four C_α atoms. For four atoms, there are two τ angles. We only depict one. **b** The coarse-grained model of protein 1VJU (chain A). **c** The (θ, τ) angle distribution of protein 1VJU (chain A). It should be noticed that since for each four C_α atoms, there is only one θ angle but two τ angles. To make their number consistent, in (θ, τ) angle distribution, our bond angle τ is defined as the average of every two virtual bond angles (Color figure online)

bone dihedral angle ϕ and ψ, and side chain rotametric angles χ. And the probability distribution function can be expressed as $P(\phi, \psi, \chi)$ in this representation. More specifically, three unique dihedral angles can be found on the protein backbone (or protein peptide chain) and all of them are formed between the adjacent four backbone atoms. Dihedral angle ϕ is formed between atoms $\{C, N, C_\alpha, C\}$, angle ψ is between atoms $\{N, C_\alpha, C, N\}$, and angle ω is from atoms $\{C_\alpha, C, N, C_\alpha\}$. Due to the partial double-bond character, dihedral ω is within a peptide planar and its value is normally 180°. In this way, each residue can be associated with a pair of ϕ and ψ angles, and protein backbone configuration can be characterized by a two-dimensional vector composed of (ϕ, ψ) angles. And Ramachandran plot, a (ϕ, ψ) angle distribution graph, is commonly used to visualize energetically favorable regions.

However, for macromolecules with a large number of amino acids, the (ϕ, ψ) representation can be computationally inefficient. To reduce the complexity, many coarse-grained models are used. The most common one is the C_α model, in which a whole amino acid residue is represented by its C_α atom. Correspondingly, the backbone configuration can be characterized by virtual dihedral angle θ and virtual bond angle τ (Levitt and Warshel 1975; Korkut and Hendrickson 2013). To be more specific, virtual dihedral angle θ is evaluated from four consecutive C_α atoms. These four atoms can form two virtual bond angles, and our bond angle τ is defined as their average. In this way, for protein with N_{res} residues, we have $N_t = N_{res} - 3$ number of (θ, τ) points. Figure 4a illustrates the geometric meaning of angle θ and τ. In this representation, a C_α backbone of a protein is specified with (θ, τ) virtual angles. In analogy to the Ramachandran plot, we can use the distribution of (θ, τ) angles to explore structure

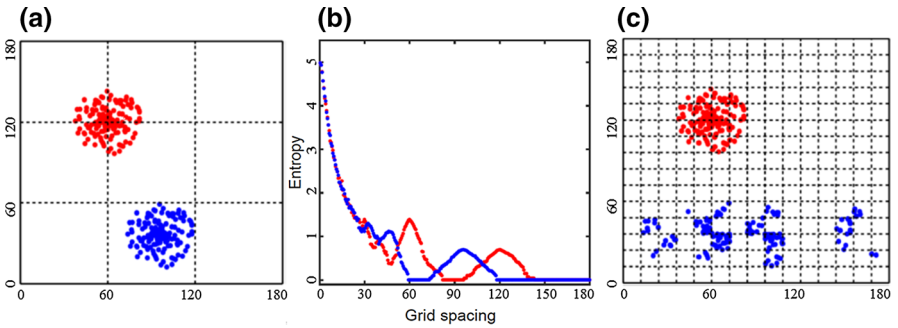

Fig. 5 Illustration of limitation of traditional entropy calculation. **a** Two identical data colored by red and blue have dramatically different entropy values in coarse grid. **b** The calculated entropy values for two identical data in various mesh spacing. The entropy curves are colored in the same way as the corresponding data sets. **c** The same entropy may represent dramatically different data distributions in a refined mesh (Color figure online)

properties. Figure 4b and c illustrates a (θ, τ) angle distributions for protein 1VJU (chain A). The angle distribution representation is essential to the CE evaluation.

Conformational entropy evaluation In CE evaluation, angular domain will be discretized into equal subdomains. Each subdomain is regarded as a configuration state, and its probability can be evaluated by counting the number of angle points within the region and dividing it by total number of points. In this way, a discrete CE formula can be employed, which is the widely used Shannon entropy:

$$S = -\sum p_i \log(p_i) \tag{26}$$

where p_i is the probability of the system being in state i and the sum is taken over all possible states of the system. It should be noticed that the Boltzmann constant is no longer exist in this definition, as Shannon entropy is entropy of information.

However, the partition of angular domain into different states is not unique and highly depends on the way of discretization. Figure 5 demonstrates the great importance of discretization. To evaluate the entropy, the whole region is discretized. Each grid box is regarded as an independent state. The probability p_i can be evaluated by counting the number of angular points in the i-th grid box and dividing it by the total number of points. The discretization procedure has no common standard and is a very subtle issue. For instance, we have two identical point sets colored by red and blue, but located in different regions of the angular domain. First, a coarse 3×3 mesh is used in Fig. 5a. With this discretization, two data have dramatically different entropies, i.e., one is 0.0 and the other is about 1.39. Then, we employ a mesh refinement and subdivide each grid box equally into four boxes. This time the entropy values are very close, i.e., one is 1.11 and the other is 1.39. With further mesh refinement, these two entropy values converge to the same value. To get a whole picture of the relationship between entropies and mesh sizes, various grid spacings are used and Fig. 5b shows the corresponding entropy values. To avoid confusion, the angular points are colored in the same way as the corresponding data. It can be seen that as the grid spacing

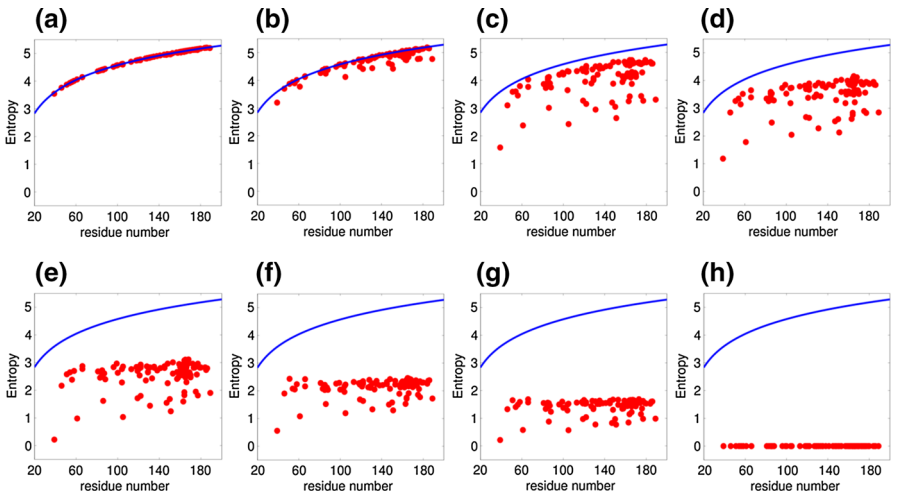

Fig. 6 Illustration of entropy values for the protein set in different grid spacing. The grid spacings **d** from **a** to **h** are 0.2°, 1.0°, 5.0°, 10.0°, 20.0°, 30.0°, 60.0° and 180.0°, respectively. The blue line represents the function $log(N)$ with N the residue number (Color figure online)

decreases (or grid size increases), two data begin to share the same entropy, whose value is not a constant but keeps increasing. Since finer meshes deliver same entropies for these two data, it seems that finer meshes are always preferred. This, however, is not true. To see it clearly, an example is given in Fig. 5c. We still use two sets of data colored in blue and read. The red-colored data is the same as in Fig. 5a. The blue-colored data is generated by redistributing the grid boxes contained with red points. To be more specific, the process is done as follows. First we "stabilize" the red data points in their grid boxes. Then we move these grid boxes together with the red points to new blank boxes. Finally, we change the red color into blue. We repeat this process for all boxes contained with red data points. In this way, no matter how refiner the meshes are at the beginning, CEs for these two sets will always be the same, as point distributions for correlated grid boxes are the same.

To further illustrate the great impact of grid spacing in biomolecular CE evaluation, we carefully choose a data set with 110 proteins. The protein IDs are listed in Table 2. All chosen proteins have resolutions smaller than 1.5 Å . Most structures have only one chain in it except eight structures marked by asterisk. For these data, we remove all the other chains from it leaving only the first chain (chain A) in the structure. In this way, there are no unphysical dihedral and bond angles in our (θ, τ) plot due to the dislocation between the ends of two chains. Figure 6 demonstrates the relation between the calculated CEs and grid spacings. To avoid confusion, the blue curve in the each subfigure of Fig. 6 represents function $\log(N_t)$. Again N_t is the total number of angle points and equals to $N_{\text{res}} - 3$ with N_{res} the total number of residues. It can be seen that when the grid spacing is small enough, CE simply converges to $\log(N_t)$. When grid spacing is large enough to incorporate all the points in a single grid, the entropy goes to zero. Essentially, grid spacing works as a resolution parameter and we can rewrite the entropy formula in Eq. (26) as

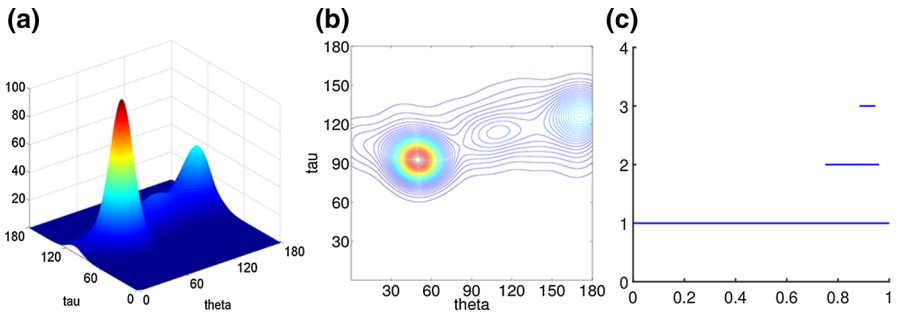

Fig. 7 Illustration of multiscale rigidity function and barcode results protein 1VJU (chain A). **a** and **b** The illustrated multiscale rigidity function is constructed by using generalized Gaussian kernel with a scale parameter $\eta = 8°$. **c** The barcodes representation of density filtration for protein 1VJU (chain A). The x-axis is the normalized density value. Three bars corresponded to the three peaks in the density map, and contour plot can be observed. And in this way, the points in Fig. 4c are naturally subdivided into three regions represented by three individual bars in barcodes. And the probability for each subdomain equals to the ratio between its barlength and the total barlength (Color figure online)

$$S(d) = -\sum p_i(d)\log(p_i(d)) \tag{27}$$

where parameter d is the grid spacing. We also have,

$$\lim_{d\to 0^\circ} S(d) = log(N_t); \quad \lim_{d\to 180^\circ} S(d) = 0. \tag{28}$$

This means in CE evaluation, we cannot use extremely fine or extremely coarse mesh. Instead, a "suitable" intermediate resolution is preferred.

Traditional discretization methods always use a regular mesh without any consideration of the angular distribution properties. More recently, K-mean clustering algorithm has been used to study CE by Zhang et al. (2008). In this method, angular points are classified into several clusters so that the summation of distances between points to their cluster centers is minimized. Inspired by this method, we propose a MPE-based CE evaluation method. Similar to Zhang's method, we do not use a regular mesh to discretize a configuration space; instead, we define configurational states by their angular distribution properties. In our method, angular distribution is transformed into an angular rigidity function. And configurational states are represented by topological invariants, particularly β_0 bars. With this representation, persistent entropy can be used as a proxy for CE. Since a resolution parameter is naturally incorporated into our persistent entropy, we will be able to deliver a multiscale entropy model and find the most "suitable" entropy for protein structure description. A detailed discussion is given below.

Multiscale persistent entropy The angular distribution data are point cloud data. Its classification problem can be transformed into a topological problem by using our MRF. To be more specific, we consider protein 1VJU (chain A) as demonstrated in Fig. 4b and c. Based on its (θ, τ) angular distribution, a rigidity function, as illustrated in Fig. 7a, is generated by using the generalized Gaussian kernel in Eq. (1) with $\kappa = 2$

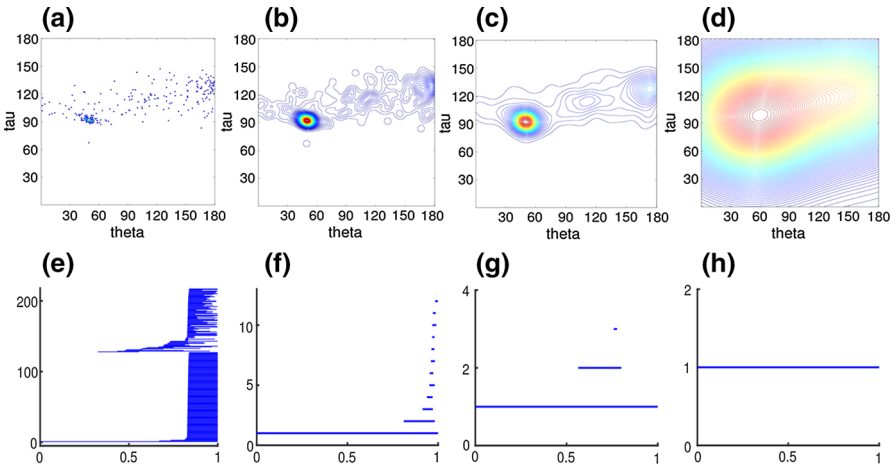

Fig. 8 Multiscale density function distributions and their corresponding barcodes for protein 1VJU (chain A). The (θ, τ) angular data are from Fig. 4c. The generalized Gaussian kernel in Eq. (1) is used with $\kappa = 2$. The values of resolution parameter η in **a**–**f** are 0.5°, 5.0°, 10° and 60.0°, respectively. The barcodes are for β_0. The number of bars represents how many clusters in the system. And the bar length represents the size of each cluster (Color figure online)

and $\eta = 8°$. After the comparison of this rigidity function and its angular distribution in Fig. 4c, it can be seen that topology of the rigidity function reveals the clustering information within the data. Simply speaking, if more angular points concentrate in a certain region, a higher "peak" will emerge in our rigidity function in the same region. If there is less or no points, a "valley" or "plane" will appear in the rigidity function. In Fig. 7a and b, we can observe three "peaks" in the rigidity function, meaning that the original angular points are concentrated in three clusters. Further, larger clusters result in higher "peaks", that is to say the size of clusters can be measured by relative heights of "peaks." All these angular distribution properties are naturally incorporated into our persistent barcodes illustrated in Fig. 7c. Basically, three bars represent three clusters and their barlengths measure the density or point numbers in the cluster. Simply speaking, each β_0 bar represents a cluster of the data and its barlength represents the relative size of this cluster. In this way, persistent entropy is a natural measurement of the disorder. To avoid confusion, all the persistent homology calculation in this paper is done with software Dipha (Bauer et al. 2014).

Further, with various resolution values, the above persistent homology analysis is able to give a full "spectrum" information of the data. Figure 8 illustrates MRF distributions and barcode results for protein 1VJU (chain A). The rigidity functions are generated with resolution parameter $\eta = 0.5°$, 5.0°, 10° and 60.0°. From the comparison of density function contours and barcode results, it can be seen that our β_0 bars capture topological features of density maps very well. When resolution value is small, local details of the density maps are revealed. So more short bars emerge. Correspondingly, we classify data into more clusters. When resolution value is large, local details are smoothed away leaving only a few long persisting bars. In this situation, data are classified into a very few clusters. Depending on the scale we interested,

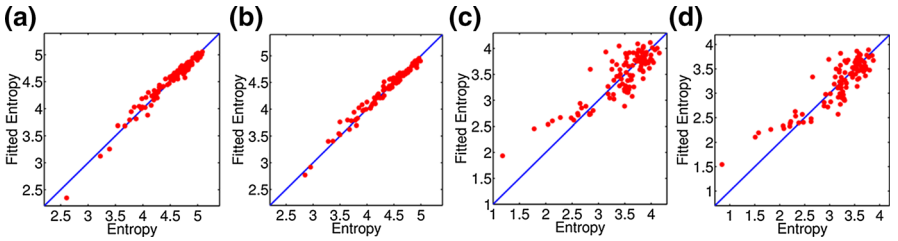

Fig. 9 Comparison of persistent entropies with traditional entropies at different resolutions. In **a** and **b**, we fix resolution value η as 1.0° and use two grid spacings 2.0° and 3.0°. The Pearson correlation coefficients between our persistent entropy and two traditional ones are 0.984 and 0.988 for **a** and **b**, respectively. In **c** and **d**, we fix the value of resolution parameter η as 5.0° and use two grid spacings 10.0° and 15.0°. The corresponding Pearson correlation coefficients are 0.848 and 0.863 for **c** and **d**, respectively. Since we use the Gaussian kernel, we chose grid spacing values based on 3η rule (Color figure online)

the resolution can be systematically adjusted. This gives us great flexibility in structure characterization. Moreover, unlike the traditional way of discretization, which employs a regular mesh, our discretization and classification is highly dependent on data structure, thus preserving more topological features in the data. To have a clear picture of the difference between the two approaches, we compare the our MPE with the general CE evaluation method. Again we use the same data set with 110 proteins as in Sect. 3.1, and protein IDs are listed in Table 2.

Generally speaking, our scale parameter η can be viewed as a counterpart of grid spacing in a traditional discretization scheme. Since dominant values in a Gaussian kernel have a range about 3η, we choose the size of the grad spacing to be about 2 to 3 times of η and make a comparison. It is seen that when grid spacing is very small, all entropy values approach to $log(N_t)$ as indicated in Fig. 6. This property is well-preserved in our scheme. It can be seen that there is a nice linear correlation for small η values as demonstrated in Fig. 9a and b. Further, when the grid spacings are extremely large, entropies in the traditional method approach to zero as demonstrated in Fig. 6h. This is also true in our scheme. When η value is very large, density map can have only one peak as indicated in Fig. 8d and the corresponding barcodes have only one long persisting bar as illustrated in Fig. 8f. In this way, our persistent entropy equals exactly to zero just as the same as in the traditional method. And a nice linear correlation can also been achieved.

More interesting is the situation when resolutions and grid spacing values are in the middle range. In this situation, two types of methods differ greatly. This is largely due to the reason that different ways of classification are used and angular data are discretized into dramatically different clusters. Figure 9c, d demonstrates this difference. The resolution parameter η is 5.0°. The grid spacings are 10.0° and 15.0° in Fig. 9c and d. The Pearson correlation coefficients decrease greatly compared with the ones in Fig. 9a and b. It should be noticed that in the traditional CE calculation, grid spacing value is always chosen in a range around 10° to 60°. And this is exactly the range in which our classification results are dramatically different from traditional ones. To further demonstrate the great power of our MPE, we use a classic protein structure classification test.

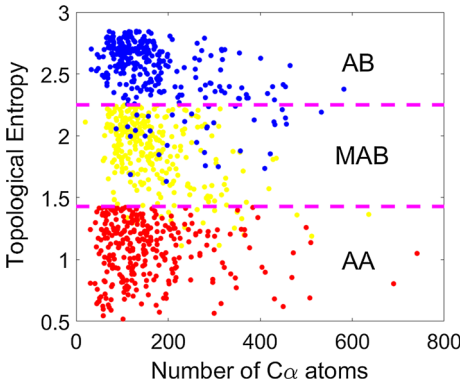

Fig. 10 Persistent entropy-based protein structure classification. The three protein classes are all-alpha (AA) protein, all-beta (AB) protein and mixed-alpha-and-beta (MAB) protein. Three hundred structures are chosen from each type of proteins, and they are colored in red (AA), blue (AB) and yellow (MAB), respectively. Two optimized thresholds are 1.45 and 2.25. It can be seen that AA type and AB type are clearly separated from each other in our method (Color figure online)

3.2 Protein Structure Classification Test

Based on secondary structures, proteins can be classified into three general categories, i.e., AA, AB and MAB proteins. To validated our MPE in protein structure classification, we employ a classic test dataset (Cang et al. 2015), which is downloaded from Structural Classification of Proteins—extended (SCOPe) database. In this test case, 300 structures are chosen from each type of proteins.

In our approach, for each protein, we calculate (θ, τ) angle distribution first, then evaluate the rigidity function and perform the persistent homology analysis on the function. For all protein structures, a generalized Gaussian kernel with $\kappa = 2$ and $\eta = 5^\circ$ is used. As stated in the previous section, similar to the traditional entropy method, we cannot use an extremely small or an extremely large resolution value. In this section, we choose $\eta = 5^\circ$ for protein structure classification. Other η values may also be suitable. Figure 10 shows our persistent entropy results. The red, blue and yellow points represent AA, AB and MAB samples, respectively. In general, The persistent entropies for AA are lower than those for AB. And there is a clear separation between two sets of values. This means persistent entropy can be used as a proper measurement to classify AA and AB proteins. In the mean time, persistent entropies for MAB proteins are distributed in the middle. To have a more quantitative comparison, we can choose two thresholds to classify the persistent entropies into three categories, corresponding to three types of proteins. For the test case, we find that the two best values are 1.45 and 2.25. We define a true positive rate (TPR) as the proportion of positives that are correctly identified as such. For instance, among the 300 MAB proteins, only 275 are predicted as MAB (with entropy value between the two thresholds). Therefore, the TPR for MAB is 91.7%. To avoid confusion, our thresholds are chosen to maximize the TPRs.

We have compared our persistent entropy results with the ones from traditional entropies (Stites and Pranata 1995; Baruah et al. 2015), persistent homology-based

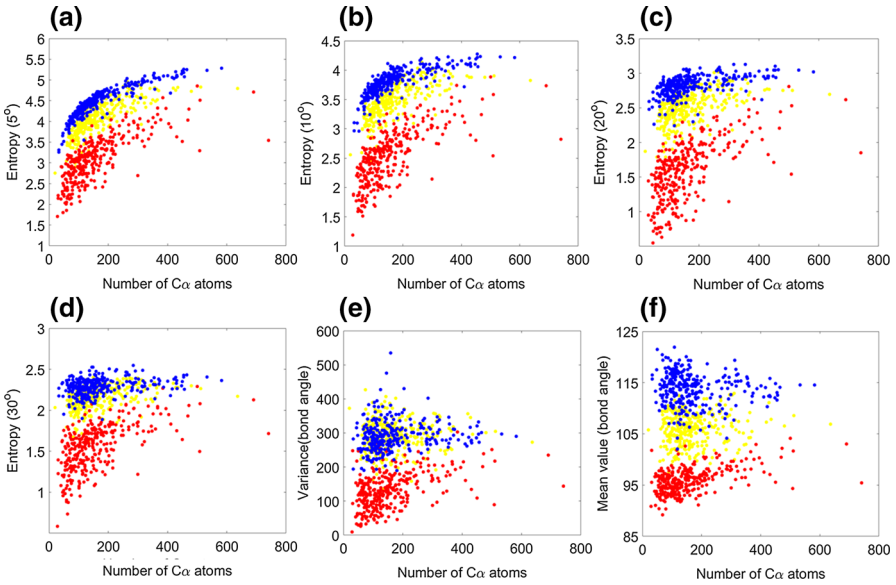

Fig. 11 Comparison of the performance of traditional entropies and virtual bond angle mean and variance method in protein structure classification. Again, red, blue and yellow points represent all-alpha (AA), all-beta (AB) and mixed-alpha-and-beta (MAB) proteins, respectively. Three hundred structures are chosen from each type. **a–d** Traditional entropies with grid sizes as 5°, 10°, 20° and 30°, respectively. **e** and **f** Variance and mean values of virtual bond angles, respectively (Color figure online)

Table 1 Comparison of the performance of our persistent entropy with traditional entropies, persistent homology-based support vector machine method and virtual bond angle mean and variance method, in protein structure classification

TPR	S(5°)	S(10°)	S(20°)	S(30°)	Mean	Variance	PH-SVM	PE
AA	82.3	85.7	89.0	91.0	93.3	95.3	90.7	100.0
AB	86.7	83.7	83.3	68.0	89.0	93.0	78.8	88.7
MAB	46.0	61.7	79.7	67.0	89.7	4.33	83.3	91.7
Average	71.7	77.0	84.0	75.3	90.7	64.2	84.93	93.4

Here TPR is true positive rate. AA, AB and MAB stand for all-alpha, all-beta, or mixed-alpha-and-beta proteins, respectively. $S(d)$ with grid spacing $d = 5°, 10°, 20°$ and $30°$ represents traditional entropy result. PH-SVM is a persistent homology-based support vector machine method (Cang et al. 2015) PE is our persistent entropy

support vector machine method (Cang et al. 2015) and virtual bond angle mean and variance method. The results are demonstrated in Fig. 11 and Table 1. It can be seen that our persistent entropy-based method gives the best result. It is worth mentioning that the classification results of traditional entropies can be further improved if a certain nonlinear threshold boundary is used. However, our persistent entropy will still prove to be better, as none of them are able to discriminate AA and AB with zero error.

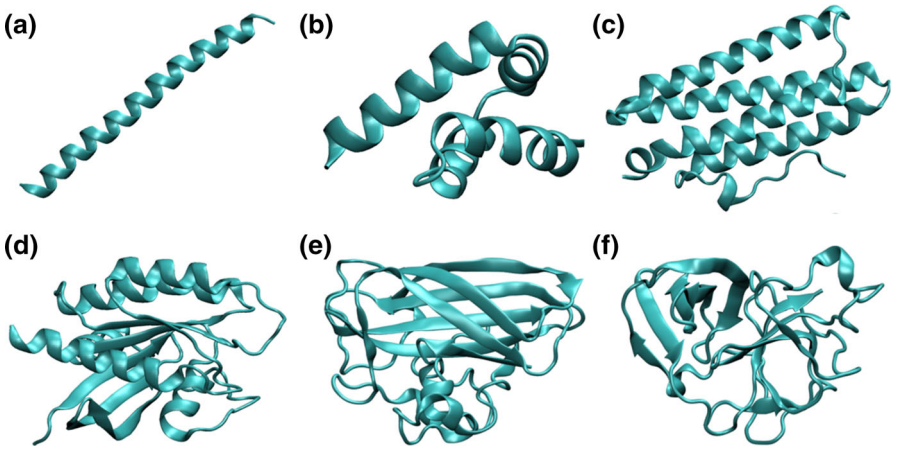

Fig. 12 Protein structure index derived from persistent entropy. The proteins and PSIs are 1GK7 (0.00), 1I2T (0.60), 4XQI (0.84), 4EPV (1.52), 4ALT (2.11) and 3ZFP (2.50). A low PSI means that the corresponding angular distribution is highly concentrated in a certain region and usually represents α-helix structure. As PSI value increases, the related angular distribution will be more diverse or scattered, meaning the structure has more twist β-sheets and loops

3.3 Protein Structure Index

Motivated by our success in protein structure classification, we propose a topological index based on our MPE to quantitatively characterize the protein structure "regularity." We call it protein structure index.

The essential idea of PSI is to evaluate the disorder of protein (θ, τ) angles. We find that when a protein structure contains only α-helix, its (θ, τ) plot is very regular with data points highly concentrated around $(50°, 90°)$. The corresponding persistent entropy is very low. The reason is that α-helix has a highly "regular" spiral conformation. On the other hand, β-sheet is also a very regular structure with (θ, τ) angles concentrated around $(195°, 117°)$. In contrast, β-sheets have various twist configurations so that (θ, τ) angles tend to be more scattered than the ones in α-helix. Persistent entropy for β-sheet is generally larger than that of β-sheet. Further, if protein structures have a large portion of loops or intrinsically disordered regions, (θ, τ) angles will be more diverse and the corresponding persistent entropy will be even higher. Physically, loops and intrinsically disordered regions are usually very unstable compared with α-helix and β-sheet. Intrinsically disorder regions even lack a fixed or ordered three-dimensional structure. In general, there is a strong correlation between angular distribution and structure disorder. And (θ, τ) angle-based persistent entropy provides a way to quantitative characterization of protein structure regularity.

To evaluate our PSI model, we use the 110 protein data set in Table 2. We systematically calculate their (θ, τ) angles, transform point cloud data into density representations and employ the persistent homology analysis. We use the generalized Gaussian kernel with $\kappa = 2$ and $\eta = 5°$ as previously. The results are demonstrated in Fig. 12 and Table 2. It can be seen that the protein structure index provides a comparably nice description of protein regularity. The smallest index indicates the most

Table 2 Protein structure index (PSI) for 110 proteins; the persistent entropy is evaluated by using Gaussian kernel with $\kappa = 2$ and scale parameter $\eta = 5°$. Most structures in the table have only one chain except eight structures marked by asterisk. For these data, we remove all the other chains from them, leaving only the first chain (chain A)

Protein ID	PSI	Protein ID	PSI	Protein ID	PSI	Protein ID	PSI	Protein ID	PSI
1GK7	0.000	1I2T	0.599	1TQG	0.694	3BT5	0.768	1JL6	0.786
$2CB8$*	0.833	4XQ1	0.842	2IMT	0.895	2F1S	0.907	2BBR	0.917
4W59	0.932	3SP7	0.934	2F2Q	0.997	3S0D	1.053	4DUI	1.092
4WOH	1.093	1LWB	1.154	4ETN	1.252	1IO0	1.275	2CLC	1.312
$4J20$*	1.353	4LD1	1.365	3DZE	1.395	2GKP	1.401	4NGJ	1.429
4PZZ	1.451	4DLR	1.466	1XEO	1.473	4EPV	1.518	3EYE	1.520
3X1X	1.520	2VIM	1.530	3L42	1.543	4LDJ	1.545	4B0D	1.587
1R2Q	1.660	3LF5	1.669	4J4Z	1.675	1RCF	1.681	4TKJ	1.696
1KMV	1.703	1SMU	1.754	3T3L	1.756	2NN5	1.772	4ICI	1.778
3CCD	1.804	4FXL	1.840	4BPY	1.841	4D0Q	1.858	3NXO	1.880
2I24	1.902	$4HIL$*	1.903	$4J46$*	1.911	3U7T	1.911	4BJ0	1.934
1TT8	1.971	2V1M	1.978	1TG0	1.997	2D4J	2.005	$2GBJ$*	2.015
4QT2	2.029	2C01	2.043	1IKJ	2.050	1ZK5	2.053	4DT4	2.057
3R54	2.060	1OD3	2.061	4WDC	2.062	4IL7	2.062	4WEE	2.066
2QDW	2.085	4ACJ	2.085	2GKT	2.087	3EY6	2.102	4ALT	2.110
2PND	2.114	4NI6	2.120	2BSC	2.132	4CST	2.133	4JGL	2.139
2ALL	2.142	4ALR	2.145	3KZD	2.166	2XOM	2.186	4P5R	2.213
$4TN9$*	2.224	2Y6H	2.228	1KT7	2.230	1T2I	2.263	3M0U	2.264
1XT5	2.267	4N1S	2.272	1HK0	2.274	4B9P	2.275	1ZIR	2.275
2FT7	2.278	$2Y72$*	2.280	2EU7	2.287	4JFM	2.292	2ON8	2.295
4R4T	2.340	4ZSD	2.343	4N1M	2.379	4DPC	2.383	1MFM	2.388
1OH4	2.395	2J1A	2.421	$3NEQ$*	2.428	3LL1	2.476	3ZFP	2.504

Protein with a smaller index value is more regular. Protein with larger index value usually has more loops or disordered regions

regular structure, i.e., a regular β-sheet. As the index increases, the protein structure gets more and more "irregular." Again, by "irregular," we mean large portion of loops and intrinsic disorder regions. It is worth mentioning that our PSI is different from Debye–Waller factor, also known as B-factor. Essentially, PSI is the general structure property with each protein one PSI value. B-factor is defined on each atom and it is used to describe the "compactness" (Halle 2002). Simply speaking, atoms with more neighbors will have a smaller B-factor than the ones with less neighbors.

Our structure index can also be used to describe the regularity in any system. Essentially, it provides a unique way of characterization disorder. It should noticed that, our protein structure index is only based on the β_0 persistent entropy. More interesting information can be derived from β_1 and β_2 topological entropies. This, however, will not be discussed in the current paper.

4 Conclusion Remarks

In this paper, we discuss our multiscale persistent homology analysis, particular multiscale persistent function, for biomolecular structure characterization. The multiscale persistent homology analysis is based on two methods, i.e., multiscale rigidity function and persistent homology. The multiscale rigidity function is essentially a continuous version of the rigidity index in our flexibility rigidity index model. It incorporates the multiscale information by using a specially designed resolution parameter. Further, multiscale barcode representation of the structure data can be achieved by a density filtration over multiscale rigidity functions. Multiscale persistent functions are defined on these barcode spaces. We discuss in great detail a particular function–multiscale persistent entropy and illustrate its applications in protein structure classification and characterization.

There are several significant characteristics of our multiscale persistent entropy. Firstly, we naturally divide the data into several clusters. This classification is based on the general topological features of the rigidity function derived from (θ, τ) angular points. Secondly, the classification information is embedded into our β_0 barcodes. Essentially, each β_0 bar represents a cluster and its length represents the relative size of the cluster. Thirdly, even through β_0-based clustering is very natural and reveals interesting structure information, it still does not differ greatly from the available methods, including hierarchical clustering, K-mean methods (Zhang et al. 2008), spectral graph theory (Chung 1997). And for persistent entropy, similar results can be obtained by the other clustering methods with some special kernels on the right scale. However, clustering and topological entropies defined from higher-dimensional homologies, i.e., β_1, β_2, will be dramatically different from all the previous clustering methods. Because these topological invariants are able to describe higher-dimensional global structure information. In this way, topological entropies defined by higher-dimensional barcodes will be able to provide more interesting intrinsic topological information of the structure. This interesting topic requires further investigation.

Acknowledgements This work was supported in part by Nanyang Technological University Startup Grant M4081842.110 and Singapore Ministry of Education Academic Research fund Tier 1 M401110000. Zhiming Li thanks the Chinese Scholarship Council for the financial support No. 201506775038. Lin Mu's

research is based upon work supported in part by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program under award number ERKJE45; and by the Laboratory Directed Research and Development program at the Oak Ridge National Laboratory, which is operated by UT-Battelle, LLC., for the U.S. Department of Energy under Contract DE-AC05-00OR22725.

References

Baron R, Hunenberger PH, McCammon JA (2009) Absolute single-molecule entropies from quasi-harmonic analysis of microsecond molecular dynamics: correction terms and convergence properties. J Chem Theory Comput 5(12):3150–3160

Baruah A, Rani P, Biswas P (2015) Conformational entropy of intrinsically disordered proteins from amino acid triads. Sci Rep 5:11740

Bauer U, Kerber M, Reininghaus J (2014) Distributed computation of persistent homology. In: Proceedings of the sixteenth workshop on algorithm engineering and experiments (ALENEX), 2014

Bendich P, Edelsbrunner H, Kerber M (2010) Computing robustness and persistence for images. IEEE Trans Vis Comput Gr 16:1251–1260

Biasotti S, De Floriani L, Falcidieno B, Frosini P, Giorgi D, Landi C, Papaleo L, Spagnuolo M (2008) Describing shapes by geometrical-topological properties of real functions. ACM Comput Surv 40(4):12

Binchi J, Merelli E, Rucco M, Petri G, Vaccarino F (2014) jHoles: a tool for understanding biological complex networks via clique weight rank persistent homology. Electron Notes Theor Comput Sci 306:5–18

Bowen R (1973) Topological entropy for noncompact sets. Trans Am Math Soc 184:125–136

Brady GP, Sharp KA (1997) Entropy in protein folding and in protein protein interactions. Curr Opinn Struct Biol 7(2):215–221

Brooijmans N, Kuntz ID (2003) Molecular recognition and docking algorithms. Ann Rev Biophys Biomol Struct 32(1):335–373

Bubenik P (2015) Statistical topological data analysis using persistence landscapes. J Mach Learn Res 16(1):77–102

Bubenik P, Kim PT (2007) A statistical approach to persistent homology. Homol Homot Appl 19:337–362

Cang ZX, Mu L, Wu KD, Opron K, Xia KL, Wei GW (2015) A topological approach to protein classification. Mol Based Math Biol 3:140–162

Carlsson G (2009) Topology and data. Am Math Soc 46(2):255–308

Carlsson G (2014) Topological pattern recognition for point cloud data. Acta Numerica 23:289

Carlsson G, Ishkhanov T, Silva V, Zomorodian A (2008) On the local behavior of spaces of natural images. Int J Comput Vis 76(1):1–12

Carlsson G, Singh G, Zomorodian A (2009) Computing multidimensional persistence. Algorithms and computation. Springer, Berlin, pp 730–739

Carlsson G, Zomorodian A (2009) The theory of multidimensional persistence. Discrete Comput Geom 42(1):71–93

Cerri A, Fabio B, Ferri M, Frosini P, Landi C (2013) Betti numbers in multidimensional persistent homology are stable functions. Math Methods Appl Sci 36(12):1543–1557

Cerri A, Landi C (2013) The persistence space in multidimensional persistent homology. Discrete geometry for computer imagery. Springer, Berlin, pp 180–191

Chazal F, De Silva V, Oudot S (2014) Persistence stability for geometric complexes. Geometriae Dedicata 173(1):193–214

Chintakunta H, Gentimis T, Gonzalez-Diaz R, Jimenez MJ, Krim H (2015) An entropy-based persistence barcode. Pattern Recognit 48(2):391–401

Chung F (1997) Spectral graph theory. American Mathematical Society, Providence

Cohen-Steiner D, Edelsbrunner H, Morozov D (2006) Vines and vineyards by updating persistence in linear time. In: Proceedings of the twenty-second annual symposium on Computational geometry, ACM. pp 119–126

Dey TK, Li KY, Sun J, David CS (2008) Computing geometry aware handle and tunnel loops in 3d models. ACM Trans Gr 27:45

Dey TK, Wang YS (2013) Reeb graphs: approximation and persistence. Discrete Comput Geom 49(1):46–73

Di Fabio B, Landi C (2011) A Mayer–Vietoris formula for persistent homology with an application to shape recognition in the presence of occlusions. Found Comput Math 11:499–527
Dionysus: the persistent homology software. Software available at http://www.mrzv.org/software/dionysus
Doig AJ, Sternberg MJE (1995) Side-chain conformational entropy in protein folding. Prot Sci 4(11):2247–2251
Edelsbrunner H (2010) Computational topology: an introduction. American Mathematical Society, Providence
Edelsbrunner H, Letscher D, Zomorodian A (2002) Topological persistence and simplification. Discrete Comput Geom 28:511–533
Edelsbrunner H, Mucke EP (1994) Three-dimensional alpha shapes. Phys Rev Lett 13:43–72
Fitter J (2003) A measure of conformational entropy change during thermal protein unfolding using neutron spectroscopy. Biophys J 84(6):3924–3930
Frederick KK, Marlow MS, Valentine KG, Wand AJ (2007) Conformational entropy in molecular recognition by proteins. Nature 448(7151):325–329
Frosini P, Landi C (2013) Persistent Betti numbers for a noise tolerant shape-based approach to image retrieval. Pattern Recognit Lett 34(8):863–872
Frosini Patrizio, Landi Claudia (1999) Size theory as a topological tool for computer vision. Pattern Recognit Image Anal 9(4):596–603
Gameiro M, Hiraoka Y, Izumi S, Kramar M, Mischaikow K, Nanda V (2015) A topological measurement of protein compressibility. Jpn J Ind Appl Math 32(1):1–17
Gellman SH (1997) Introduction: molecular recognition. Chem Rev 97(5):1231–1232
Ghrist R (2008) Barcodes: the persistent topology of data. Bull Am Math Soc 45(1):61–75
Halle B (2002) Flexibility and packing in proteins. PNAS 99:1274–1279
Hatcher A (2001) Algebraic topology. Cambridge University Press, Cambridge
Horak D, Maletic S, Rajkovic M (2009) Persistent homology of complex networks. J Stat Mech Theory Exp 2009(03):P03034
Janin J, Sternberg MJ (2013) Protein flexibility, not disorder, is intrinsic to molecular recognition. F1000 Biol Rep 5(2):1–7
Kaczynski T, Mischaikow K, Mrozek M (2004) Computational homology. Springer, Springer
Karplus M, Kushick JN (1981) Method for estimating the configurational entropy of macromolecules. Macromolecules 14(2):325–332
Kasson PM, Zomorodian A, Park S, Singhal N, Guibas LJ, Pande VS (2007) Persistent voids a new structural metric for membrane fusion. Bioinformatics 23:1753–1759
Korkut A, Hendrickson WA (2013) Stereochemistry of polypeptide conformation in Coarse Grained analysis. In: Biomolecular forms and functions: a celebration of 50 years of the Ramachandran Map, World Scientific Publishing. pp 136–147
Lee H, Kang H, Chung MK, Kim B, Lee DS (2012) Persistent brain network homology from the perspective of dendrogram. IEEE Trans Med Imaging 31(12):2267–2277
Levitt M, Warshel A (1975) Computer simulation of protein folding. Nature 253(5494):694–698
Liu X, Xie Z, Yi DY (2012) A fast algorithm for constructing topological structure in large data. Homol Homot Appl 14:221–238
Marlow MS, Dogan J, Frederick KK, Valentine KG, Wand AJ (2010) The role of conformational entropy in molecular recognition by calmodulin. Nat Chem Biol 6(5):352–358
Merelli E, Rucco M, Sloot P, Tesei L (2015) Topological characterization of complex systems: using persistent entropy. Entropy 17(10):6872–6892
Mischaikow K, Mrozek M, Reiss J, Szymczak A (1999) Construction of symbolic dynamics from experimental time series. Phys Rev Lett 82:1144–1147
Mischaikow K, Nanda V (2013) Morse theory for filtrations and efficient computation of persistent homology. Discrete Comput Geom 50(2):330–353
Munkres JR (1984) Elements of algebraic topology, vol 2. Addison-Wesley, Menlo Park
Nanda V Perseus: the persistent homology software. Software available at http://www.sas.upenn.edu/~vnanda/perseus
Nguyen D, Xia KL, Wei GW (2016) Generalized flexibility–rigidity index. J Chem Phys 144(23):234106
Niyogi P, Smale S, Weinberger S (2011) A topological view of unsupervised learning from noisy data. SIAM J Comput 40:646–663
Opron K, Xia KL, Burton ZF, Wei GW (2016) Flexibility rigidity index for protein nucleic acid flexibility and fluctuation analysis. J Comput Chem 37(14):1283–1295

Opron K, Xia KL, Wei GW (2014) Fast and anisotropic flexibility–rigidity index for protein flexibility and fluctuation analysis. J Chem Phys 140:234105
Opron K, Xia KL, Wei GW (2015) Communication: capturing protein multiscale thermal fluctuations. J Chem Phys 142(21):211101
Pachauri D, Hinrichs C, Chung MK, Johnson SC, Singh V (2011) Topology-based kernels with application to inference problems in alzheimer's disease. IEEE Trans Med Imaging 30(10):1760–1770
Rieck B, Mara H, Leitte H (2012) Multivariate data analysis using persistence-based filtering and topological signatures. IEEE Trans Vis Comput Gr 18:2382–2391
Robins Vanessa (1999) Towards computing homology from finite approximations. Topol Proc 24:503–532
Rucco M, Castiglione F, Merelli E, Pettini M (2016) Characterisation of the idiotypic immune network through persistent entropy. In: Proceedings of ECCS 2014, Springer. pp 117–128
Rucco M, Gonzalez-Diaz R, Jimenez MJ, Atienza N, Cristalli C, Concettoni E, Ferrante A, Merelli E (2017) A new topological entropy-based approach for measuring similarities among piecewise linear functions. Signal Process 134:130–138
Sapienza PJ, Lee AL (2010) Using NMR to study fast dynamics in proteins: methods and applications. Curr Opin Pharmacol 10(6):723–730
Shen MY, Sali A (2006) Statistical potential for assessment and prediction of protein structures. Prot Sci 15(11):2507–2524
Silva VD, Ghrist R (2005) Blind swarms for coverage in 2-d. In: Proceedings of robotics: science and systems, pp 01
Singh G, Memoli F, Ishkhanov T, Sapiro G, Carlsson G, Ringach DL (2008) Topological analysis of population activity in visual cortex. J Vis 8(8):11.1–18
Stites WE, Pranata J (1995) Empirical evaluation of the influence of side chains on the conformational entropy of the polypeptide backbone. Prot Struct Funct Bioinf 22(2):132–140
Tausz A, Vejdemo-Johansson M, Adams H (2011) Javaplex: a research software package for persistent (co)homology. Software available at http://code.google.com/p/javaplex
Thompson JB, Hansma HG, Hansma PK, Plaxco KW (2002) The backbone conformational entropy of protein folding: experimental measures from atomic force microscopy. J Mol Biol 322(3):645–652
Trbovic N, Cho JH, Abel R, Friesner RA, Rance M, Palmer AG III (2008) Protein side-chain dynamics and residual conformational entropy. J Am Chem Soc 131(2):615–622
Wang B, Summa B, Pascucci V, Vejdemo-Johansson M (2011) Branching and circular features in high dimensional data. IEEE Trans Vis Comput Gr 17:1902–1911
Wang B, Wei GW (2016) Object-oriented persistent homology. J Comput Phys 305:276–299
Xia KL, Feng X, Tong YY, Wei GW (2015) Persistent homology for the quantitative prediction of fullerene stability. J Comput Chem 36:408–422
Xia KL, Opron K, Wei GW (2013) Multiscale multiphysics and multidomain models—flexibility and rigidity. J Chem Phys 139:194109
Xia KL, Opron K, Wei GW (2015) Multiscale Gaussian network model (mGNM) and multiscale anisotropic network model (manm). J Chem Phys 143(20):204106
Xia KL, Wei GW (2014) Persistent homology analysis of protein structure, flexibility and folding. Int J Numer Methods Biomed Eng 30:814–844
Xia KL, Wei GW (2015) Multidimensional persistence in biomolecular data. J Comput Chem 36:1502–1520
Xia KL, Wei GW (2015) Persistent topology for cryo-EM data analysis. Int J Numer Methods Biomed Eng 31:e02719
Xia KL, Zhao ZX, Wei GW (2015) Multiresolution topological simplification. J Comput Biol 22:1–5
Yao Y, Sun J, Huang XH, Bowman GR, Singh G, Lesnick M, Guibas LJ, Pande VS, Carlsson G (2009) Topological methods for exploring low-density states in biomolecular folding pathways. J Chem Phys 130:144115
Zhang J, Lin M, Chen R, Wang W, Liang J (2008) Discrete state model and accurate estimation of loop entropy of rna secondary structures. J Chem Phys 128(12):125107
Zhong S, Moix JM, Quirk S, Hernandez R (2006) Dihedral-angle information entropy as a gauge of secondary structure propensity. Biophys J 91(11):4014–4023
Zomorodian A, Carlsson G (2005) Computing persistent homology. Discrete Comput Geom 33:249–274
Zomorodian Afra, Carlsson Gunnar (2008) Localized homology. Comput Geom Theory Appl 41(3):126–148