Research Article **Open Access**

Zixuan Cang, Lin Mu, Kedi Wu, Kristopher Opron, Kelin Xia, and Guo-Wei Wei*

A topological approach for protein classification

DOI 10.1515/mlbmb-2015-0009
Received October 16, 2015; accepted November 4, 2015

Abstract: Protein function and dynamics are closely related to its sequence and structure. However, prediction of protein function and dynamics from its sequence and structure is still a fundamental challenge in molecular biology. Protein classification, which is typically done through measuring the similarity between proteins based on protein sequence or physical information, serves as a crucial step toward the understanding of protein function and dynamics. Persistent homology is a new branch of algebraic topology that has found its success in the topological data analysis in a variety of disciplines, including molecular biology. The present work explores the potential of using persistent homology as an independent tool for protein classification. To this end, we propose a molecular topological fingerprint based support vector machine (MTF-SVM) classifier. Specifically, we construct machine learning feature vectors solely from protein topological fingerprints, which are topological invariants generated during the filtration process. To validate the present MTF-SVM approach, we consider four types of problems. First, we study protein-drug binding by using the M2 channel protein of influenza A virus. We achieve 96% accuracy in discriminating drug bound and unbound M2 channels. Secondly, we examine the use of MTF-SVM for the classification of hemoglobin molecules in their relaxed and taut forms and obtain about 80% accuracy. Thirdly, the identification of all alpha, all beta, and alpha-beta protein domains is carried out using 900 proteins. We have found a 85% success in this identification. Finally, we apply the present technique to 55 classification tasks of protein superfamilies over 1357 samples and 246 tasks over 11944 samples. Average accuracies of 82% and 73% are attained. The present study establishes computational topology as an independent and effective alternative for protein classification.

Keywords: persistent homology; machine learning; protein classification; topological fingerprint

1 Introduction

Proteins are essential building blocks of living organisms. They function as catalyst, structural elements, chemical signals, receptors, etc. The molecular mechanism of protein functions are closely related to their structures. The study of structure-function relationship is the holy grail of biophysics and has attracted enormous effort in the past few decades. The understanding of such a relationship enables us to predict protein functions from structure, amino acid sequence, or both, which remains to be a major challenge in molecular biology. Intensive experimental investigation has been carried out to explore the interactions among proteins or proteins with other biomolecules, e.g., DNAs and/or RNAs. In particular, the understanding of protein-drug interactions is of premier importance to human health.

***Corresponding Author: Guo-Wei Wei:** Mathematical Biosciences Institute, The Ohio State University, Columbus, Ohio 43210, USA
On leave from the Department of Mathematics, Michigan State University
E-mail: wei@math.msu.edu
Zixuan Cang, Kedi Wu, Kelin Xia: Department of Mathematics, Michigan State University, East Lansing, MI 48824, USA
Lin Mu: Oak Ridge National Laboratory, One Bethel Valley Road, P.O. Box 2008, MS 6211, Oak Ridge, TN 37831, USA
Kristopher Opron: Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824, USA

A wide variety of theoretical and computational approaches has been proposed to understand the protein structure-function relationship. One class of these approaches is biophysics. From the point of view of biophysics, protein structure, function, dynamics, and transport are, in general, dictated by protein interactions. Quantum mechanics (QM) is based on the fundamental principle, and offers the most accurate description of interactions among electrons, photons, atoms, and even molecules. Although QM methods have unveiled many underlying mechanisms of reaction kinetics and enzymatic activities, they typically are computationally prohibitive for large biomolecules. Based on classic physical laws, molecular mechanics (MM) [74] in combination with fitted parameters can simulate the physical movement of atoms or molecules for relatively large biomolecular systems like proteins quite precisely. However, it can be computationally intractable for macromolecular systems involving realistic biological time scales. Many time-independent methods like normal mode analysis (NMA) [12, 57, 71, 96], elastic network model (ENM) [3, 49, 68, 99], graph theory [61], and flexibility-rigidity index (FRI) [81, 82, 107] are proposed to capture features of large biomolecules. Variational multiscale methods [26–30, 102–104] are another class of approaches that combine atomistic description with continuum approximations. There are well developed servers for predicting protein functions based on three-dimensional (3D) structures [67] or models from the homology modeling (here homology is in biological sense) of amino acid sequence if 3D structure is not yet available [91].

Another class of important approaches, bioinformatical methods, plays a unique role for the understanding of the structure-function relationship. These data-driven predictions are based on similarity analysis. The essential idea is that proteins with similar sequences or structures may share similar functions. Also, based on sequential or structural similarity, proteins can be classified into many different groups. Once the sequence or structure of a novel protein is identified, its function can be predicted by assigning it to the group of proteins that share similarities to a good extent. However, the degree of similarity depends on the criteria used to measure similarity or difference. Many measurements are used to describe similarity between two protein samples. Typical approaches use either sequence, physical information, or both. Among them, sequence alignment can describe how closely the two proteins are related. Protein blast [63], clustalW2 [72], and other software packages can perform global or local sequence alignments. Based on sequence alignments, various scoring methods can provide the description of protein similarity [2, 58]. Additionally, sequence features such as sequence length and occurrence percentage of a specific amino acid can also be employed to compare proteins. Many sequence based features can be derived from the position-specific scoring matrix (PSSM) [95]. Moreover, structural information provides an efficient description of protein similarity as well. Structure alignment methods include rigid, flexible, and other methods. The combination of different structure alignment methods and different measurements such as root-mean-square deviation (RMSD) and Z-score gives rise to various ways to quantify the similarity among proteins. According to structure information, different physical properties such as surface area, volume, free energy, flexible-rigidity index (FRI) [81, 82, 107], curvature [46, 47], electrostatics [115], etc. can be calculated. A continuum model, Poisson Boltzmann (PB) equation delivers quite accurate estimation for electrostatics of biomolecules. There are many efficient and accurate PB solvers including PBEQ [62], MIBPB [25, 115], etc. Together with physical properties, one can also extract geometrical properties from structure information. These properties include coordinates of atoms, connections between atoms such as covalent bonds and hydrogen bonds, molecular surfaces [4, 5, 114] , and curvatures [46, 47, 105]. These various approaches reveal information of different scales from local atom arrangement to global architecture. Physical and geometrical properties described above add a different perspective to analyzing protein similarities.

Due to the advance in bioscience and biotechnology, biomolecular structure data sets are growing at an unprecedented rate. For example, the Protein Data Bank (PDB) has accumulated more than a hundred thousand biomolecular structures. The prediction of the protein structure-function relationship from such huge amount of data can be extremely challenging. Additionally, an eve-growing number of physical or sequence features are evaluated for each data set or amino-acid residue, which adds to the complexity of the data-driven prediction. To automatically analyze excessively large data sets in molecular biology, many machine learning methods have been developed [31, 48, 70, 76]. These methods are mainly utilized for classification, regression, comparison, and clustering of biomolecular data. Clustering is an unsupervised learning method which divides a set of inputs into groups without knowing the groups beforehand. This method can unveil hidden

patterns in the data set. Classification is a supervised learning method, in which, a classifier is trained on a given training set and used to do prediction for new observations. It assigns an observation to one of several pre-determined categories based on knowledge from training data set in which the labels of observations are known. Popular methods for classification include support vector machine (SVM) [13], artificial neural network (ANN) [75], deep learning [59], etc. In classification, each observation in the training set has a feature vector that describes the observation from various perspectives and a label that indicates to which group the observation belongs. A model trained on the training set indicates to which group a new observation belongs with feature vector and unknown label. To improve the speed of classification and reduce effect from irrelevant features, many feature selection procedures have been proposed [38] . Machine learning approach has been used successfully for protein hot spot prediction [37].

The data-driven analysis of the protein structure-function relationship is compromised by the fact that same protein may have different conformations which possess different properties or deliver different functions. For instance, hemoglobins have taut form with low affinity to oxygen and relaxed form with high affinity to oxygen; and ion channels often have open and close states. Different conformations of a given protein may only have minor differences in their local geometric configurations. These conformations share the same sequence and may have very similar physical properties. However, their minor structural differences might lead to dramatically different functions. Therefore, apart from the conventional physical and sequence information, geometric and topological information can also play an important role in understanding the protein structure-function relationship. Indeed, geometric information has been extensively used in the protein exploration. In contrast, topological information has been hardly employed in studying the protein structure-function relationship.

In general, geometric approaches are frequently inundated with too much geometric detail and are often prohibitively expensive for most realistic biomolecular systems, while traditional topological methods often incur too much reduction of the original geometric and physical information. Persistent homology, a new branch of applied topology, is able to bridge traditional geometry and topology. It creates a variety of topologies of a given object by varying a filtration parameter, such as a radius or a level set function. In the past decade, persistent homology has been developed as a new multiscale representation of topological features. The 0-th dimensional version was originally introduced for computer vision applications under the name “size function" [51, 52] and the idea was also studied by Robins [90]. The Persistent homology theory was formulated, together with an algorithm given, by Edelsbrunner et al. [43], and a more general theory was developed by Zomorodian and Carlsson [116]. There has since been significant theoretical development [8, 15, 16, 19, 23, 24, 32–34, 39], as well as various computational algorithms [6, 40, 78, 79, 83, 97]. Often, persistent homology can be visualized through barcodes [20, 56], in which various horizontal line segments or bars are the homology generators which survive over filtration scales. Persistence diagrams are another equivalent representation [42]. Computational homology and persistent homology have been applied to a variety of domains, including image analysis [7, 17, 53, 84, 93], chaotic dynamics verification [64, 77], sensor network [92], complex network [60, 69], data analysis [14, 73, 80, 89, 100], shape recognition [1, 41] , and computational biology [36, 55, 65, 85, 86]. For example, alpha shape has been utilized in protein function and binding site prediction [117, 118]. Compared with traditional computational topology [22, 66, 113] and/or computational homology, persistent homology inherently has an additional dimension, the filtration parameter, which can be utilized to embed some crucial geometric or quantitative information into the topological invariants. The importance of retaining geometric information in topological analysis has been recognized [10], and topology has been advocated as a new approach for tackling big data sets [9, 11, 14, 54, 56].

Recently, we have introduced persistent homology for mathematical modeling and prediction of nano particles, proteins , and other biomolecules [106, 108]. We have proposed molecular topological fingerprint (MTF) to reveal topology-function relationships in protein folding and protein flexibility [108]. We have employed persistent homology to predict the curvature energies of fullerene isomers [101, 106], and to analyze the stability of protein folding [108]. More recently, we have introduced resolution based persistent homology [111, 112]. Most recently, we have developed new multidimensional persistence, a topic that has attracted much attention in the past few years [18, 19], to better bridge geometry and traditional topology and achieve

better characterization of biomolecular data [109]. We have also introduced the use of topological fingerprint for resolving ill-posed inverse problems in cryo-EM structure determination [110].

The objective of the present work is to explore the utility of MTFs for protein classification and analysis. We construct feature vectors based on MTFs to describe unique topological properties of protein in different scales, states, and/or conformations. These topological feature vectors are further used in conjugation with the SVM algorithm for the classification of proteins. We validate the proposed MTF-SVM strategy by distinguishing different protein conformations, proteins with different local secondary structures, and proteins from different superfamilies or families. The performance of proposed topological method is demonstrated by a number of realistic applications, including protein binding analysis, ion channel study, and more.

The rest of the paper is organized as following. Section 2 is devoted to the mathematical foundations for persistent homology and machine learning methods. We present a brief description of simplex and simplicial complex followed by basic concept of homology, filtration, and persistence in Section 2.1. Three different methods to get simplicial complex, Vietoris-Rips complex, alpha complex, and Čech complex are discussed. We use a sequence of graphs of channel proteins to illustrate the growth of a Vietoris-Rips complex and corresponding barcode representation of topological persistence. In Section 2.2, fundamental concept of support vector machine is discussed. An introduction of transformation of the original optimization problem is given. A measurement for the performance of classification model known as receiver operating characteristic is described. Section 2.3 is devoted to the description of features used in the classification and pre-processing of topological feature vectors. In section 3, four test cases are shown. Case 1 and Case 2 examine the performance of the topological fingerprint based classification methods in distinguishing different conformations of same proteins. In Case 1, we use the structure of the M2 channel of influenza A virus with and without an inhibitor. In Case 2, we employ the structure of hemoglobin in taut form and relaxed form. Case 3 validates the proposed topological methods in capturing the difference in local secondary structure. In this study, proteins are divided into three groups, all alpha protein, all beta protein, and alpha+beta protein. In Case 4, the ability of the present method for distinguishing different protein superfamilies is examined. This paper ends with some concluding remarks.

2 Materials and Methods

This section presents a brief review of persistent homology theory and illustrates its use in proteins. A brief description of machine learning methods is also given. The topological feature selection and construction from biomolecular data are described in details.

2.1 Persistent homology

Points, edges, triangles and their higher dimensional counterparts are defined as simplices. A simplicial space is a topological space constructed from finitely many simplices.

Simplex A k-simplex denoted by σ^k is a convex hull of $k+1$ vertices which is represented by a set of points

$$\sigma^k = \{\lambda_0 u_0 + \lambda_1 u_1 + \ldots + \lambda_k u_k | \sum \lambda_i = 1, \lambda_i \geq 0, i = 0, 1, \ldots, k\}, \quad (1)$$

where $\{u_0, u_1, \ldots, u_k\} \subset \mathbb{R}^n$ is a set of affinely independent points. Geometrically, a 1-*simplex* is a line segment, a 2-simplex is a triangle, a 3-simplex is a tetrahedron, and a 4-simplex is a 5-cell (a four dimensional object bounded by five tetrahedrons). An m–face of the k-simplex is defined as a convex hull formed from a subset consisting of m vertices.

Simplicial complex A simplicial complex $\mathcal{K}$ is a finite collection of simplices satisfying two conditions. First, faces of a simplex in $\mathcal{K}$ are also in $\mathcal{K}$; Secondly, intersection of any two simplices in $\mathcal{K}$ is a face of both the simplices. The highest dimension of simplices in $\mathcal{K}$ determines dimension of $\mathcal{K}$.

Homology For a simplicial complex $\mathcal{K}$, a k-chain is a formal sum of the form $\sum_{i=1}^{N} c_i[\sigma_i^k]$, where $[\sigma_i^k]$ is oriented k-simplex from $\mathcal{K}$. For simplicity, we choose $c_i \in \mathbb{Z}_2$. All these k-chains on $\mathcal{K}$ form an Abelian group

which is called chain group and is denoted as $C_k(\mathcal{K})$. A boundary operator ∂_k over a k-simplex σ^k is defined as,

$$\partial_k \sigma^k = \sum_{i=0}^{k} (-1)^i [u_0, u_1, ..., \widehat{u_i}, ..., u_k], \tag{2}$$

where $[u_0, u_1, ..., \widehat{u_i}, ..., u_k]$ denotes the face obtained by deleting the ith vertex in the simplex. The boundary operator induces a boundary homomorphism $\partial_k : C_k(\mathcal{K}) \to C_{k-1}(\mathcal{K})$. A very important property of the boundary operator is that the composition operator $\partial_{k-1} \circ \partial_k$ is a zero map,

$$\begin{aligned} \partial_{k-1}\partial_k(\sigma^k) &= \sum_{j<i} (-1)^i (-1)^j [u_0, ..., \widehat{u_i}, ...\widehat{u_j}, ...u_k] + \sum_{j>i} (-1)^i (-1)^{j-1} [u_0, ..., \widehat{u_j}, ...\widehat{u_i}, ...u_k] \\ &= 0 \end{aligned} \tag{3}$$

A sequence of chain groups connected by boundary operation form a chain complex,

$$\cdots \longrightarrow C_n(\mathcal{K}) \xrightarrow{\partial_n} C_{n-1}(\mathcal{K}) \xrightarrow{\partial_{n-1}} \cdots \xrightarrow{\partial_1} C_0(\mathcal{K}) \xrightarrow{\partial_0} 0. \tag{4}$$

The equation $\partial_k \circ \partial_{k+1} = 0$ is equivalent to the inclusion $\mathrm{Im}\partial_{k+1} \subset \mathrm{Ker}\ \partial_k$, where Im and Ker denote image and kernel. Elements of $\mathrm{Ker}\partial_k$ are called kth cycle group, and denoted as Z_k=$\mathrm{Ker}\partial_k$. Elements of $\mathrm{Im}\partial_{k+1}$ are called kth boundary group, and denoted as B_k=$\mathrm{Im}\partial_{k+1}$. A kth homology group is defined as the quotient group of Z_k and B_k.

$$H_k = Z_k / B_k. \tag{5}$$

The kth Betti number of simplicial complex $\mathcal{K}$ is the rank of H_k,

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

Betti number β_k is finite number, since $\mathrm{rank}(B_p) \leq \mathrm{rank}(Z_p) < \infty$. Betti numbers computed from a homology group are used to describe the corresponding space. Generally speaking, the Betti numbers β_0, β_1 and β_2 are numbers of connected components, tunnels, and cavities, respectively.

Filtration and persistence A filtration of a simplicial complex $\mathcal{K}$ is a nested sequence of subcomplexes of $\mathcal{K}$.

$$\varnothing = \mathcal{K}_0 \subseteq \mathcal{K}_1 \subseteq \ldots \subseteq \mathcal{K}_m = \mathcal{K}. \tag{7}$$

With a filtration of simplicial complex $\mathcal{K}$, topological attributes can be generated for each member in the sequence by deriving the homology group of each simplicial complex. The topological features that are long lasting through the filtration sequence are more likely to capture significant property of the object. Intuitively, non-boundary cycles that are not mapped into boundaries too fast along the filtration are considered to be possibly involved in major features or persistence. Equipped with a proper derivation of filtration and a wise choice of threshold to define persistence, it is practicable to filter out topological noises and acquire attributes of interest. The p-persistent kth homology group of $\mathcal{K}_i$ is defined as

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

where $Z_k^i = Z_k(\mathcal{K}_i)$ and $B_k^i = B_k(\mathcal{K}_i)$. The consequent p-persistent kth Betti number is $\beta_k^{i,p} = \mathrm{rank}(H_k^{i,p})$. A well chosen p promises reasonable elimination of topological noise.

Vietoris-Rips complex Based on a metric space M and a given cutoff distance d, an abstract simplicial complex can be built. If two points in M have a distance shorter than the given distance d, an edge is formed between these two points. Consequently, simplices of different dimensions are formed and a simplicial complex is built. For a point cloud data, natural metric space based on Euclidean distance or other metric spaces based on alternative definition of distance can be used to build a Vietoris-Rips complex. For example, any correlation matrix can be used directly to form a Vietoris-Rips complex. Figure 1 illustrates growth of Vietoris-Rips complex along with increment of d over the point set of C_α atoms from M2 chimera channel.

There are many ways of constructing complex other than Vietoris-Rips complex, including Alpha complex, Čech Complex, CW complex, etc. In the present work, we used Vietoris-Rips complex in part because

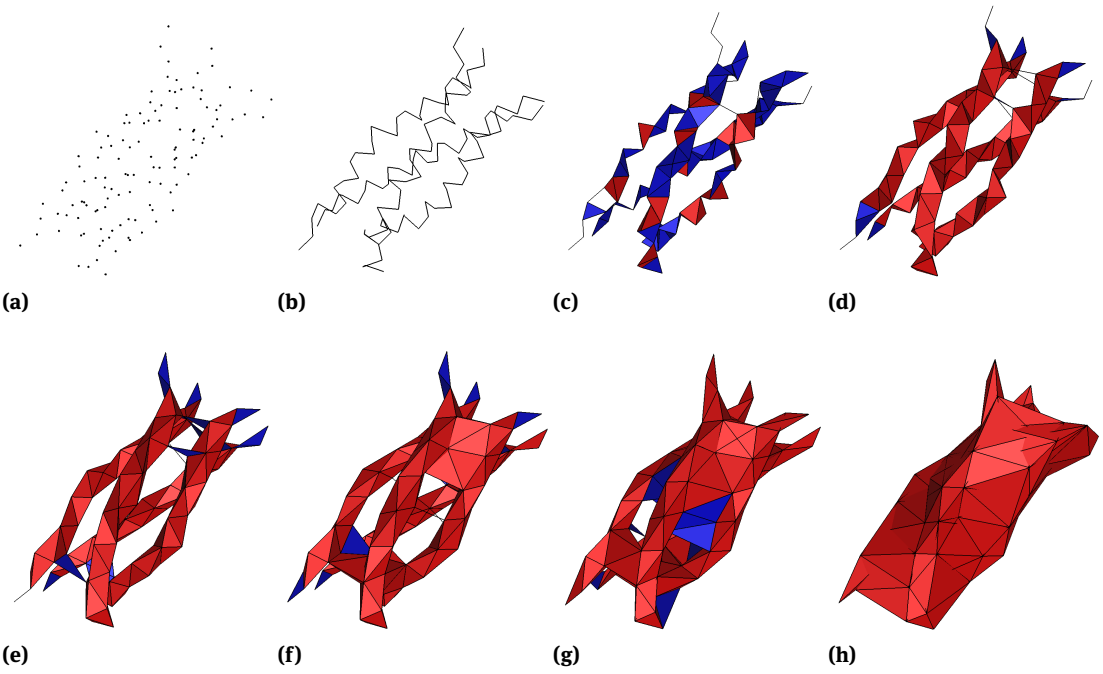

Figure 1: Filtration of Vietoris-Rips complex built on the α-carbon point cloud of the M2 chimera channel of influenza A virus (PDB ID: 2LJC [88]). The corresponding filtration values for each graph are (a) $d = 1.0$Å, (b) $d = 4.0$Å, (c) $d = 5.5$ Å, (d) $d = 6.0$Å, (e) $d = 7.0$Å, (f) $d = 8.0$Å, (g) $d = 9.0$Å, and (h) $d = 12.0$Å. Facets of 3-simplices are shown in red, 2-simplices are shown in blue, 1-simplices are shown as lines, and 0-simplices are shown as dots.

of its intuitive nature and in part because of the moderate size of the systems we studied. The computational topology package JavaPlex [98] was used for computation of persistent homology. The results were represented in the form of barcodes [56]. Figure 2 illustrates barcodes computed from point cloud data extracted from a C_α atom model and an all-atom model of the protein with PDB ID 2LJC.

2.2 Support vector machine

2.2.0.1 Basic theory

SVM is a machine learning method that can be applied to classification and regression problems. It computes a hyperplane which maximizes margin between positive and negative training sets. In this work, Classification SVM Type 1, also known as C-support vector classification (C-SVC) [35] is used. For the problem of classification, a classifier is built on a training data set with the description of samples and their pre-determined classes. The classifier can then predict the class of new observations based on their descriptions. The input for SVM is a set of samples. Each sample has a feature vector that describes the properties of the sample and a label that implies to which class the sample belongs. Given the input which is the training set, SVM will generate a hyperplane in the feature space or higher dimensional spaces depending on which kernel it uses that separates the classes. For two-class SVM, it looks for a hyperplane $\mathbf{w}^T\mathbf{x}+b = 0$ that separates the classes. The determination of the coefficients $\mathbf{w}$ and b breaks down to a constrained optimization problem as

$$\min_{\mathbf{w},b} \frac{1}{2}|\mathbf{w}|^2 + C\sum \xi_i, \tag{9}$$

subject to

$$\begin{cases} y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, i = 1, \ldots, n, \\ \xi_i \geq 0, i = 1, \ldots, n, \end{cases} \tag{10}$$

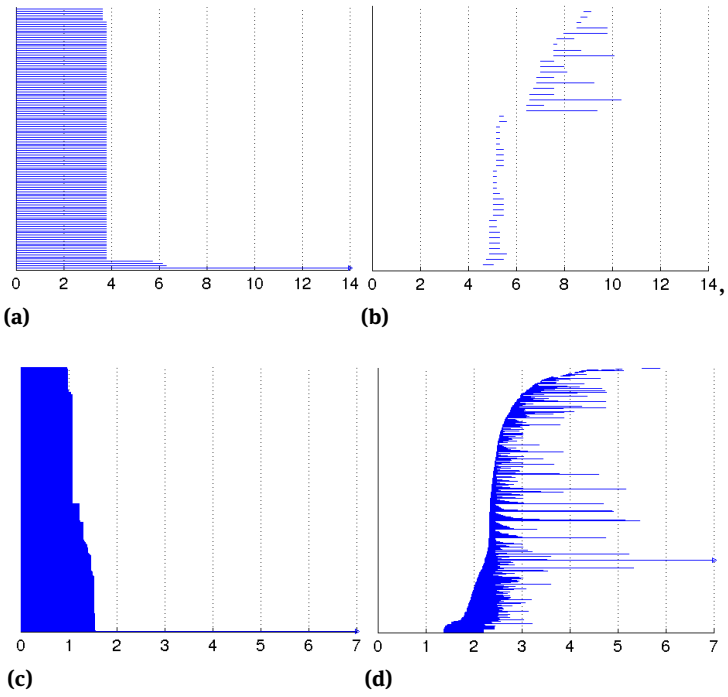

Figure 2: Bar code plots of persistent homology calculated for α-carbon and all atom point cloud of M2 chimera channel of influenza A virus based in Vietoris-Rips complex. (a) and (b) are respectively Betti 0 and Betti 1 bar code plots for point cloud of α-carbon. (c) and (d) are respectively Betti 0 and Betti 1 bar code plots for all atom point cloud.

where $\mathbf{x}_i$ denotes the feature vector of the ith sample, y_i is the label of the ith sample which takes value of either 1 or −1, and C is a penalty coefficient for misclassified points. To handle linearly inseparable data, one maps the data into a higher dimensional space as $\phi : \mathbb{R}^N \to \mathbb{R}^M$ with $N < M$. Since in the optimization problem and in scoring function of the classifier, the operator used is dot product, ϕ does not need to be explicitly found. A decaying kernel $K(\mathbf{x}_i, \mathbf{x}_j)$ function is used to represent $\phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j)$. Commonly used kernel functions include linear function: $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_j$, polynomial: $K(\mathbf{x}_i, \mathbf{x}_j) = (a\mathbf{x}_i^T\mathbf{x}_j+b)^d$, radial basis functions (RBFs) such as Gaussian $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma|\mathbf{x}_i-\mathbf{x}_j|^2)$, $\gamma > 0$. In fact, the admissible kernels of flexibility-rigidity index (FRI) [81, 82, 107] work too. In this work, The Gaussian kernel is used and a 5-fold cross validation was applied to search for optimized training parameters for problems with large amount of samples. To solve the optimization problem, the original problem is transformed into the corresponding Lagrange dual problem. For a constrained optimization problem

$$\begin{aligned} &\min_{\mathbf{x}} f(\mathbf{x}), \\ g_i(\mathbf{x}) \le 0, &\ i = 1, 2, \ldots, k_1 \\ h_i(\mathbf{x}) = 0, &\ i = 1, 2, \ldots, k_2 \end{aligned} \tag{11}$$

the Lagrange function of this problem is defined as

$$L(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_{i=1}^{k_1} \alpha_i g_i(\mathbf{x}) + \sum_{i=1}^{k_2} \lambda_i h_i(\mathbf{x}) \tag{12}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$ are Lagrange multipliers. The Lagrange dual problem is defined as

$$\begin{aligned} &\max_{\boldsymbol{\alpha},\boldsymbol{\lambda}} \theta(\boldsymbol{\alpha}, \boldsymbol{\lambda}) \\ &\alpha_i \ge 0 \end{aligned} \tag{13}$$

where $\theta(\boldsymbol{\alpha}, \boldsymbol{\lambda}) = \inf_{\mathbf{x}\in\Omega} L(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\lambda})$. The Lagrange function of the original optimization problem (9) is formulated as

$$\begin{aligned} L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) &= \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{i=1}^{n}(C - \alpha_i - \lambda_i)\xi_i - b\sum_{i=1}^{n}\alpha_i y_i + \sum_{i=1}^{n}\alpha_i y_i \mathbf{w}^T\phi(\mathbf{x}_i) \\ &= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j). \end{aligned} \tag{14}$$

Tthe corresponding dual problem with Karush-Kuhn-Tucker conditions is defined as

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) \\ &\alpha_i(y_i(\sum_{j=1}^{n}\alpha_j\phi(\mathbf{x}_j)^T\phi(\mathbf{x}_i) + b) - 1 + \xi_i)) = 0 \\ &\xi_i(\alpha_i - C) = 0 \\ &\sum_{i=1}^{n}\alpha_i y_i = 0 \\ &C \geq \alpha_i \geq 0 \end{aligned} \tag{15}$$

The dual problem can be solved with sequential minimal optimization (SMO) method [44].

2.2.0.2 Receiver operating characteristic (ROC)

ROC is a plot that visualizes the performance of a binary classifier [45]. A binary classifier uses a threshold value to decide the prediction label of an entry. In testing process, we define true positive rate (TPR) and false positive rate (FPR) for the testing set.

$$\begin{aligned} &\text{TPR} = (\text{number of positive samples predicted as positive})/(\text{number of positive samples}) \\ &\text{FPR} = (\text{number of negative samples predicted as positive})/(\text{number of negative samples}) \end{aligned} \tag{16}$$

An ROC space is a two dimensional space defined by points with x coordinate representing FPR and y coordinate representing TPR. In the prediction process of a binary classifier, a score is assigned to a sample by the classifier. A test sample may be labeled as positive or negative with different threshold values used by the classifier. Corresponding to a certain threshold value, there is a pair of FPR and TPR values which is a point in the ROC space. All such points will fall in the box $[0, 1] \times [0, 1]$. Points above the diagonal line $y = x$ are considered as good predictors and those below the line are considered as poor predictors. If a point is below the diagonal line, the predictor can be inverted to be a good predictor. For points that are close to the diagonal line, they are considered to act similarly to random guess which implies a relatively useless predictor. ROC curve is obtained by plotting FPR and TPR as continuous functions of threshold value. The area between ROC curve and x axis represents probability that the classifier assigns higher score to a randomly chosen positive sample than to a randomly chosen negative sample if positive is set to have higher score than negative. The area under the curve (AUC) of ROC is a measurement of classifier quality. Intuitively, a higher AUC implies a better classifier.

2.3 Topological feature selection and construction

In this work, algebraic topology is employed to distinguish proteins. Specifically, we compute MTFs through the filtration process of protein structural data. MTFs bear the persistence of topological invariants during the filtration and are ideally suited for protein classification. To implement our topological approach in the SVM algorithm, we construct protein feature vectors using MTFs. We select distinguishing protein features from MTFs. These features can be both long lasting and short lasting Betti 0, Betti 1, and Betti 2 intervals.

Table 1: A list of features used in support vectors

Feature #	Betti #	Description
1	0	The length of the second longest Betti 0 bar.
2	0	The length of the third longest Betti 0 bar.
3	0	The summation of lengths of all Betti 0 bars except for those exceed the max filtration value.
4	0	The average length of Betti 0 bars except for those exceed the max filtration value.
5	1	The onset value of the longest Betti 1 bar.
6	1	The length of the longest Betti 1 bar.
7	1	The smallest onset value of the Betti 1 bar that is longer than 1.5Å.
8	1	The average of the middle point values of all the Betti 1 bars that are longer than 1.5Å.
9	1	The number of Betti 1 bars that locate at $[4.5, 5.5]$Å, divided by the number of atoms.
10	1	The number of Betti 1 bars that locate at $[3.5, 4.5)$Å and $(5.5, 6.5]$Å, divided by the number of atoms.
11	1	The summation of lengths of all the Betti 1 bars except for those exceed the max filtration value.
12	1	The average length of Betti 1 bars except for those exceed the max filtration value.
13	2	The onset value of the first Betti 2 bar that ends after a given number.

Table 1 lists topological features used for classification. Detailed explanation of these features is discussed. The length and location value of bars are in the unit of angstrom (Å) for protein data.

- Feature 1: The length of the second longest Betti 0 bar indicates the onset in filtration that the simplices in the corresponding complex form one connected component.
- Feature 2: Similar to Feature 1, this value indicates the onset in filtration that the simplices form two connected components. For the more complicated point cloud, the more features of this kind may be utilized.
- Feature 3: Geometrically, the total length of Betti 0 bars describes how compactly the points are located.
- Feature 4: This averaged Betti 0 bar length shows similar property as that in Feature 3 with no correlation to atom number.
- Feature 5: This value shows the filtration value at which, the largest persistent loop is formed.
- Feature 6: The persistence of the longest Betti 1 bar reflects the size of the geometrically dominating loop.
- Feature 7: A Betti 1 bar with length larger than the threshold is considered to be important and this feature records the onset filtration value of such a long bar. In this work, a threshold of 1.5 is used for α-carbon point cloud data of proteins.
- Feature 8: This feature records the average location of midpoints of Betti 1 bars which are longer than the threshold value discussed in Feature 5. This value shows at which filtration value the loops are centered.
- Feature 9: This feature indicates the portion of alpha helices in a protein. For each four α-carbons on a alpha helix, they are likely to form a short Betti 1 bar around filtration value 5Å. A bar is considered to be short if it has length shorter than 0.5Å and to be around 5Å if the distance from its midpoint to 5Å is less than 0.6Å.
- Feature 10: Similar to Feature 7, this feature can be used to identify portion of beta sheets. Detailed discussion of Features 7 and 8 can be found in Ref. [108].
- Feature 11: A strong correlation between accumulation bar length of Betti 1 and total energy has been reported [108].
- Feature 12: The average value of Betti 1 bars correlates to the average loop size.
- Feature 13: The smallest onset value of the Betti 2 bar that ends after a given value. This feature gives information about birth and death of cavities in the complex through filtration.

Each feature is scaled to the interval $[0, 1]$ with linear mapping. With the same scale, all the features are equally considered by the classifier. For simplicity and due to the moderate size of the system, automated feature selection was not performed.

3 Results

In this section, we validate the proposed idea, examine the accuracy, and explore the utility of the proposed topology based classification of protein molecules. We consider four different types of problems. In our first case, we study a protein-drug binding problem, namely, the drug inhibition of Influenza A virus M2 channels. In our second case, we use MTFs to classify two type of conformations of hemoglobin proteins. Default parameters are used and brute force cross validation is performed for these first two cases due to their small size of samples. We further consider the classification of three types of protein domains, i.e., all alpha domains, all beta domains , and mixed alpha and beta domains. Finally, our method is tested on two problem sets, PCB00019 and PCB00020 from Protein Classification Benchmark Collection [94]. In the last two cases, a grid search with cross validation on training sets is performed to optimize SVM parameters. For the last case, different penalty parameters for positive and negative sets are applied to overcome the unbalanced data and an ROC analysis is used to evaluate the results.

Data for the M2 channels are all obtained from NMR experiments [88]. Data for hemoglobin structures are all collected from X-ray crystallography. Structure data for the last two test cases are mostly attained from X-ray crystallography. However, a few structures are determined by NMR techniques and thus have many alternatives. In this situation, we select the second structure for each sample in the data base.

In this work, we utilize JavaPlex [98] to compute MTFs. For implementation of support vector machine, LIBSVM is employed [21].

3.1 Protein-drug binding analysis

Proteins are vital to many processes in cells. In many biological processes, proteins bind to other molecules. Protein-protein interaction and protein-ligand interaction are of crucial importance to their functions and/or malfunction. These interactions have been intensively exploited in drug design. Specifically, many drugs bind to target proteins to modify their functions and activities. After binding to other molecule, a protein usually experiences a structure change at the binding site. In many cases, it may also undergo allosteric process with a global structural change upon the binding. We test our method for distinguishing proteins with drug bound from proteins without drugs.

We use M2 channel, which is a transmembrane protein found in influenza A virus [87], as an example. M2 channel equilibrates pH across the viral membrane during cell entry and plays a vital role in viral replication. Therefore, it is used as a target for the anti-influenza drugs, i.e., amantadine and rimantadine, which bind to the M2 channel pore and block the proton permeation. The drug binding creates a topological change to the M2 channel in the conventional sense. However, in the present work, it is not the topological change itself, rather that the binding induced geometric variation of the M2 channel that is converted into the change in the topological invariants. Such a change is recorded in our MTFs and utilized for protein classification.

The structures of chimera channels with and without rimantadine are used for classification. PDB IDs of the two structures are 2LJC for channel with the inhibitor and 2LJB for channel without the inhibitor [88]. The structures are shown in Figure 3a–(b). Note that inhibitor itself is not included in our filtration. A total of 15 snapshots from NMR for each structure (15 positive samples and 15 negative samples) are used to perform classification. Due to small size of instances, default parameters in C-SVC with penalty $C = 2$ and $\gamma = 1/(\text{number of features})$ are used. Each time, 10 instances from each class are set as the training set and the rest are set as the testing set. A brute-force cross validation is performed. The average accuracy for unbound form is 93.91% and accuracy for bound form is 98.31%. Due to small size of testing set, AUC value is not calculated in this example.

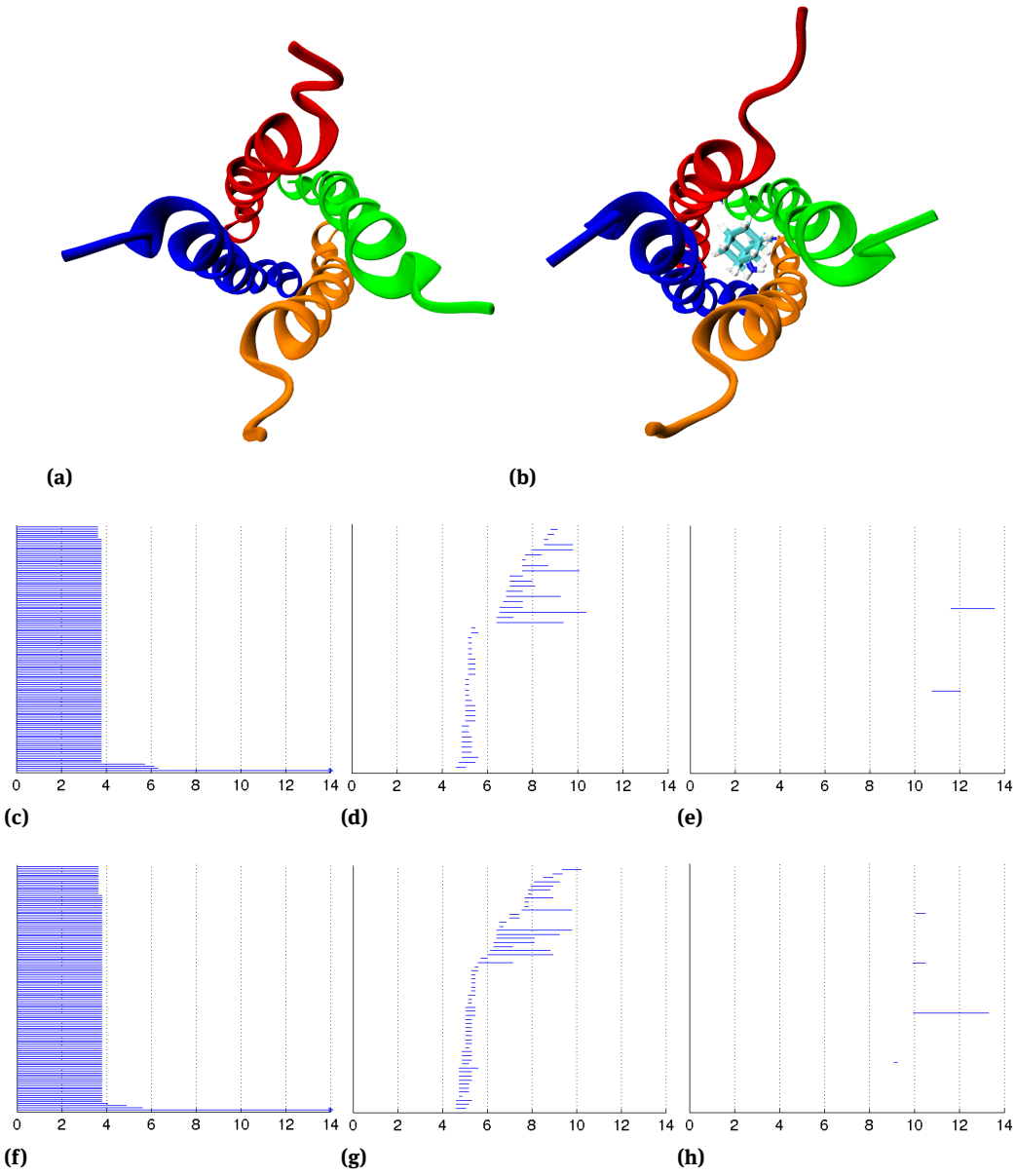

Figure 3: Protein structures used in M2 channel classification. (a) (PDB ID: 2LJB [88]) M2 channel of influenza A without inhibitor. (b) (PDB ID: 2LJC[88]) M2 channel of influenza A with inhibitor. The small molecule in the graph is for illustration and is not used in classification. (c), (d) and (e) are respectively Betti 0, Betti 1, and Betti 2 barcodes for (a). (f), (g) and (h) are respectively Betti 0, Betti 1, and Betti 2 barcodes for (b).

3.2 Discrimination of hemoglobin molecules in relaxed and taut forms

Hemoglobin is oxygen transport metalloprotein in red blood cells of most vertebrates. It carries oxygen from lungs or gills to other organs or parts in the body. Oxygen is released to tissues and is used for metabolism. Hemoglobin is also known to carry carbon dioxide in some cases. It exists in two forms, known as taut (T) form and relaxed (R) form. Examples of these two forms are shown in Figure 4a–(b). Relaxed form has a high oxygen binding affinity with which hemoglobin can better bind to oxygen in lungs or gills. Taut form has a low oxygen binding affinity which helps release the oxygen in the rest of the body. Many factors affect the conformation form of hemoglobin, such as pH value, concentration of carbon dioxide, and partial pressure in the system. Structurally, the two forms are slightly different. In this test case, we pick up 9 structures of

hemoglobin in R form and 10 structures of hemoglobin in T form from protein data bank. Table 2 lists PDB IDs of the proteins used.

Table 2: Protein molecules used for the Hemoglobin classification.

R-form	1HHO, 3A0G, 1LFQ, 1HBR, 1RVW, 2D5X, 1IBE, 1AJ9, 2W6V
T-form	2HHB, 2DHB, 1LFL, 2D5Z, 1GZX, 2HBS, 4ROL, 1O1J, 2DXM, 1KD2

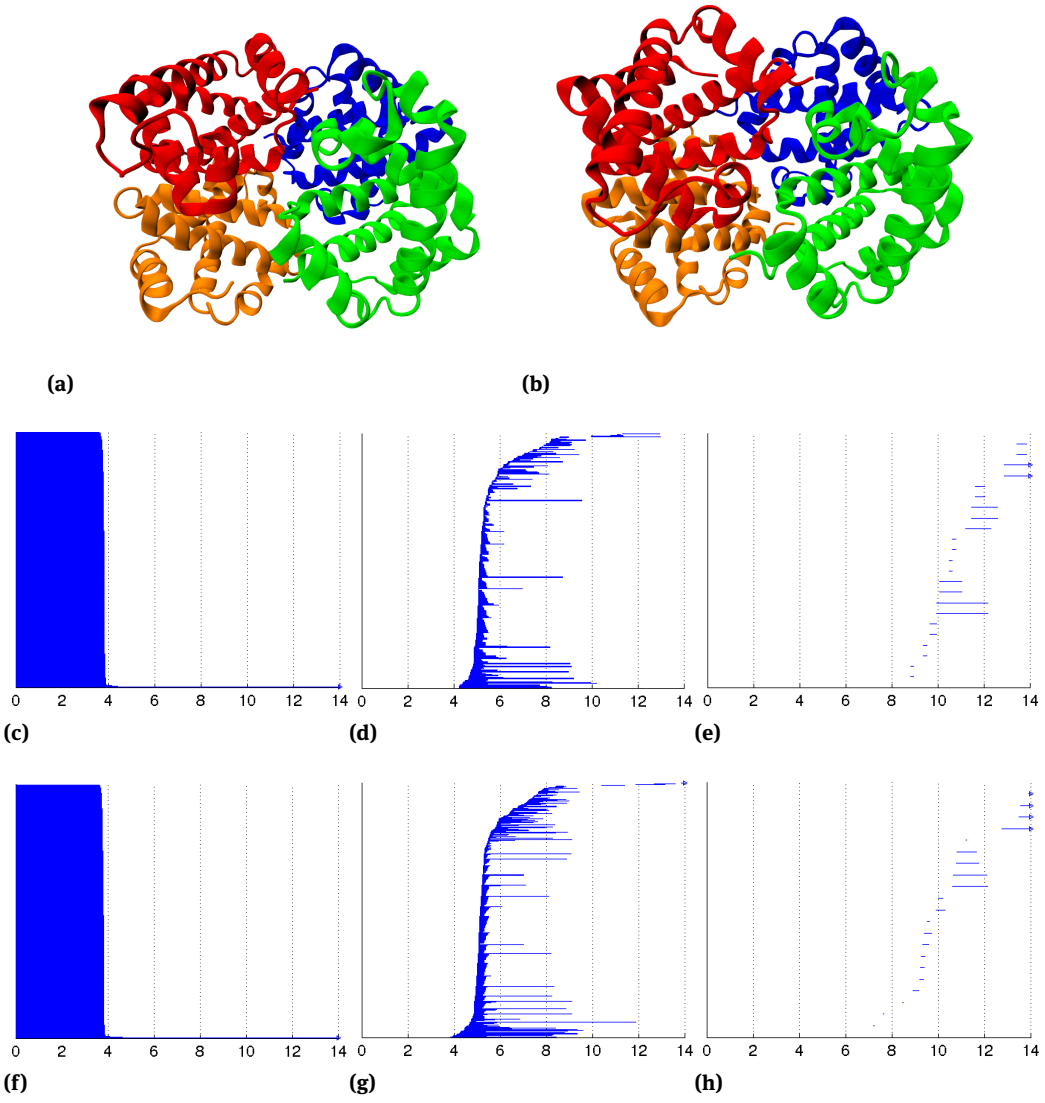

Figure 4: Protein structures used the Hemoglobin classification. (a) (PDB ID: 3A0G) Relaxed (R) form of hemoglobin which express high affinity to oxygen. (b) (PDB ID: 2HHB citeFermi:1984) Taut (T) form of hemoglobin which express low affinity to oxygen. (c), (d) and (e) are respectively Betti 0, Betti 1, and Betti 2 barcodes for (a). (f), (g) and (h) are respectively Betti 0, Betti 1, and Betti 2 barcodes for (b).

In this test case, as the number of instances is relatively small, a brute-force cross validation is performed with the same default parameters as in last case. Each time one instance from each class is picked as the

testing set leaving the rest instances as the training set. The average accuracy of the prediction for testing set is 84.50%. The average accuracy of R form is 77.16% and average accuracy of T form is 91.11%. Since testing set is small, ROC analysis is not applied in this case.

3.3 The classification of all alpha, all beta, and mixed alpha and beta protein domains

Protein secondary structures are three dimensional patterns of protein local segments. Common secondary structures include alpha helices and beta sheets. These local structures are formed by hydrogen bonds between amine hydrogen and carbonyl oxygen atoms in the backbone of a protein. Typically, secondary structures can be identified from amino acid sequence data. In this test example, we use only geometric data without sequence information to generate MTFs and then classify alpha helices and beta sheets. Instances of this example are taken from SCOPe (Structural Classification of Proteins-extended) database [50]. The SCOPe ID (SID) of samples used in this test case are listed in Tables 3,4, and 5.

In this test case, protein domains are separated into three classes, namely, all alpha helix domains, all beta sheet domains, and mixed alpha helix and beta sheet domains. Examples for each of three classes are shown in Figures 5a–(c) and their barcode plots are shown in Figures 5d–(i). For each class in SCOPe, 300 structures from different superfamilies are used for classification. Among the 900 instances, 60 from each class are used as testing set and the rest are used as training set. A 5-fold cross-validation is performed to test accuracy. In each training process, a 5-fold cross validation on the training set only is carried out to optimize training parameters. The overall accuracy is 84.93%. Specifically, the accuracy for all alpha helix domains is 90.67%, the accuracy for mixed alpha helix and beta sheet domains is 78.77%, and accuracy for all beta sheet domains is 83.31%.

3.4 Classification of protein superfamilies

A protein superfamily is the largest collection of proteins for which a common ancestor can be traced. Within a superfamily, similarity between amino acid sequences may not be easily observed. Therefore, a superfamily can be further divided into several families within which, similarity among amino acid sequences usually can be identified. Members in a protein superfamily share similar structure with the common ancestor though they may not have similar sequences. In this case, based on structure information, we test our method in classification of protein superfamilies. The samples in this test case are taken from Protein Classification Benchmark Collection [94]. The problems used in our test have the accession numbers of PCB00019 and PCB00020. The goal of these data sets is to classify protein domain sequences and structures into protein superfamilies, based on protein families. The problem set PCB00019 contains 1357 protein samples and 55 classification tasks. The problem set PCB00020 contains 11944 protein samples and 246 classification tasks. Detailed description and classification results using different scoring method and various classification methods can be found in Protein Classification Benchmark Collection website. In this test, we utilize only the structure information of α-carbon in protein backbones. For each task, we perform 5 fold cross validation on the training set to search for reasonable parameters. In most tasks, the sizes of positive set and negative set are unbalanced. To prevent unbalanced training results, different values for penalty parameters are used. Specifically, the ratio between positive penalty parameter and negative penalty parameter is set to equal the ratio between number of negative instances and number of positive instances in the training set. The criteria for cross validation is chosen to be the recall value which is defined as (true positive)/(true positive + false negative) to overcome the extremely unbalanced nature of the data set. The average accuracy for the positive testing set and the negative testing set in PCB00019 are 82.29% and 80.94%, respectively. The average accuracies for the positive testing set and the negative testing set in PCB00020 are 72.32% and 73.18%, respectively. The average AUC value for the 55 tasks in PCB00019 is 0.8954 and the average AUC value for 246 tasks in PCB00020 is 0.7813. The thresholds used to determine ROC curve are set to be a list of decision values corresponding to

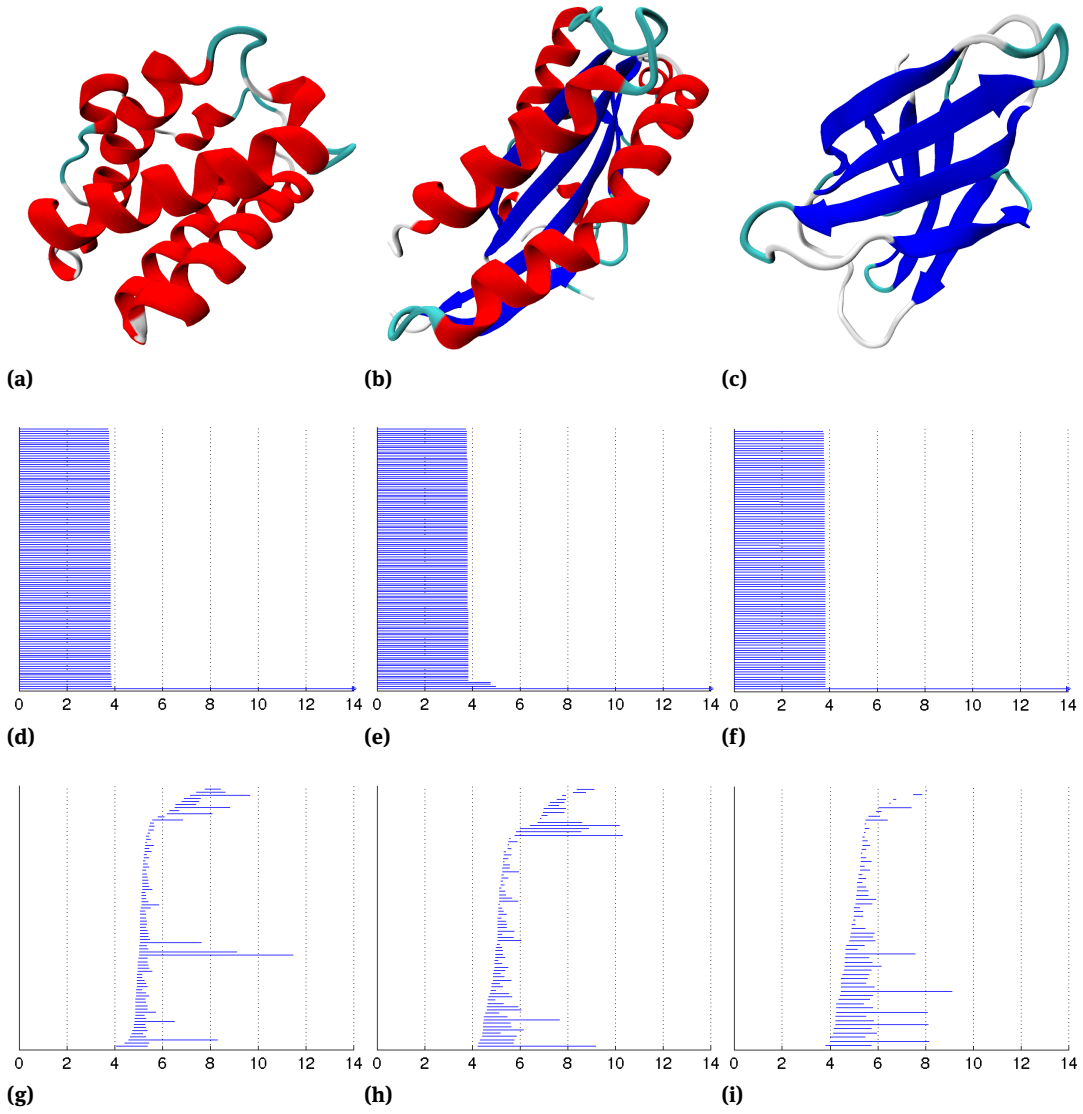

Figure 5: Example plots of different protein domains. (a) All alpha protein. (b) Alpha and beta protein. (c) All beta protein. (d) and (g) are respectively example Betti 0 and Betti 1 barcodes for all alpha protein. (e) and (h) are respectively example Betti 0 and Betti 1 barcodes for alpha+beta protein. (f) and (i) are respectively example Betti 0 and Betti 1 barcodes for all beta protein.

each instance in the testing set from the model. This makes sure that all the points are captured. The standard deviation of AUC for the 55 tasks in PCB00019 is 0.09 and standard deviation of AUC for the 246 tasks in PCB00020 is 0.2. The performance in PCB00020 is relatively poor compared to that in PCB00019. On the one hand, this is because that problem set PCB00020 is tougher. On the other hand, this may due to the fact that some information about differences between samples lies in sequence which is not contained in our model. Figure 6 shows plot of the ROC curves for the 55 tasks in PCB00019. The plot of ROC curves for the 246 tasks in PCB00020 is not shown due to unreadability of too many curves.

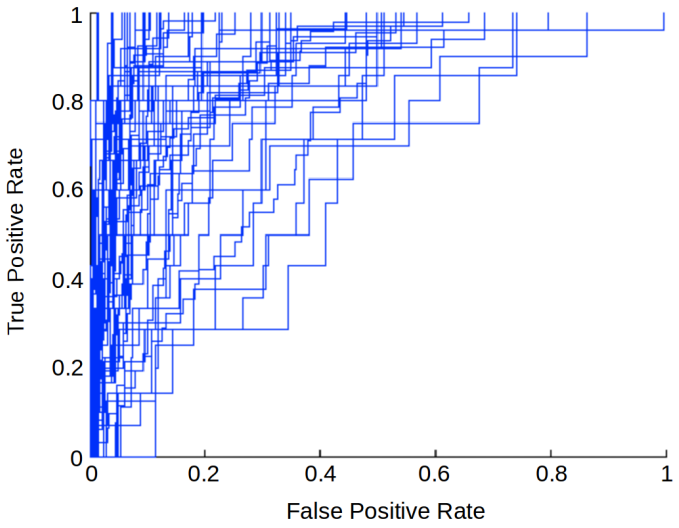

Figure 6: The ROC curves corresponding to the 55 tasks in problem set PCB00019 [94]. Plot was generated using LIBSVM tools [21].

4 Discussion and Conclusion

Persistent homology is a unique tool in computational topology and computational geometry. It explores the topology space by studying the evolution of simplicial complex over a filtration process of a given data set. A nested sequence of subsets are obtained by continuously increasing the filtration parameter. During the filtration process, birth and death of topological invariants are recorded. The lifespan of a topological invariant shows how significant it is geometrically. Persistent homology is capable of discovering the underlying topological feature of the space of interest and recognizing topologically small events. In other words, it gives not only information of global and significant topological features, but also perspective of local features of the underlying space. Persistent homology has been applied to computer graphics, geometric modeling, data analysis, and many other fields. A protein structure can be represented as point cloud in three dimension for atoms or graph with edges corresponding to different types of chemical bonds. This geometric nature of protein structures allows the application of persistent homology. In this work, we introduce the use of protein topological features captured by persistent homology for the protein classification. Our goal is to illustrate that molecular topological fingerprints (MTFs) can describe the structure of a protein from different perspectives and in different scales. This property of MTFs makes it possible to be used in protein classification from the topological point of view. We examine the performance of MTFs in several protein classification tasks with different emphasizes. We show that MTFs are a potential option for protein classification.

To introduce the topological features used in classification, we briefly review the definition of simplex and different types of simplicial complex. Basic concepts of filtration and persistence is recalled. We use α-carbon atoms in M2 proton channel of influenza A to illustrate the filtration of a simplicial complex. We also show the barcode plots for M2 channel in an all-atom model and an α-carbon model. Comparing these approach, it can be seen that all-atom model contains too many details which flood away useful information such as Betti 1 barcode representing alpha helices. Essentially, at the all-atom scale, different proteins have some common features due to the structures of amino acids. Using a coarse-grained model with α-carbon atoms reveals more information of the entire structure of the protein and dramatically reduces spatial complex and computational time. Therefore, we adapt coarse-grained model throughout this work. In some physical descriptions of proteins, an all-atom model may be preferred.

In persistent homology, a convention is to cherish long-persistent topological features which are presented as long-lived bars in a barcode representation. Whereas, short-lived barcodes are typically discharged as noise. In our case, the MTFs of proteins carry both global features and local traits. For protein analysis,

both global features and local traits are important. In other words, it takes both long-lived topological features and short-lived topological traits to effectively characterize different proteins. A fundamental reason is that biomolecular structure, function, dynamics and transport are governed by the interactions of wide range of scales, which lead to multiple characteristic length scales ranging from covalent bond, residue, secondary structure, and domain dimensions to protein sizes. Based on our understanding of protein characteristic length scales [109], we are able to identify the responding protein topological fingerprints and determine their relevance and importance in protein classification.

To use MTFs for the analysis of large scale biomoleculoar data, we have developed persistent homology based machine learning method. Essentially, we construct feature vectors by using MTFs. We utilize the support vector machine (SVM) algorithm, which is known for its robustness and high accuracy, in our study. The resulting MTF-SVM classifier is validated by four test cases. First, we explore the performance of the present MTF-SVM classifier for distinguishing drug bound M2 channels of influenza A virus from those of nature M2 channels. It is found that the proposed method does an excellent job in analyzing drug binding of M2 channels. A 96% prediction accuracy is recorded. In our second test, we consider the discrimination of hemoglobin molecules in their relaxed and taut forms. Again, the present approach works very well (80% accuracy) for this problem. We further employ our MTF-SVM classifier for the identification of all alpha, all beta, and alpha-beta protein domains. A total of 900 proteins is used in our study. Due to the relatively large sample size, a 5-fold cross-validation was carried out to optimize training parameters and validate the present method. In this study, the detailed local topological features facilitate the classification of proteins with different secondary structures. An average of 85% accuracy is found over three protein classes. Finally, we utilize the present method for the classification of protein superfamilies. We adapt two standard test sets, accession numbers PCB00019 and PCB00020, from Protein Classification Benchmark Collection [94]. PCB00019 involves 1357 samples and 55 classification tasks and PCB00020 involves 11944 samples and 246 classification tasks. A combination of both local and global topological features enables us to separate protein superfamilies. An average classification accuracy of 82% and an average AUC value of 0.89 are found on PCB00019 test set. An average classification accuracy of 73% and an average AUC value of 0.78 are found on PCB00020 test set.

The objective of the present work is to examine the utility, accuracy, and efficiency of computational topology for protein classification. As such, only topological information is employed. The extensive test study establishes topology as an independent and valuable option for large scale protein classification. Obviously, the present method can be improved in a variety of ways. Specifically, one can combine topological features with other more established features, namely, sequence features and physical features for protein analysis and classification. Indeed, MTFs computed from persistent homology differ sharply from sequence and physical based features. Therefore, a combination of topological features, sequence features, and physical features must be able to take advantages of these three classes of methods. This aspect is beyond the scope of the present work and will be explored in our future research.

In our earlier work, we have introduced computational topology for mathematical modeling and prediction, such as molecular stability prediction [108], protein folding analysis [112], and protein bond length prediction [109]. The present work indicates that the combination of machine learning and computational topology will create a new powerful strategy for topology based mathematical modeling and prediction.

Acknowledgement: This work was supported in part by NSF grants IIS-1302285, and DMS-1160352, and NIH Grant R01GM-090208. GWW thanks the Mathematical Biosciences Institute (MBI) for generous financial support.

Conflict of interest: Author state no conflict of interest.

References

[1] P. K. Agarwal, H. Edelsbrunner, J. Harer, and Y. Wang. Extreme elevation on a 2-manifold. *Discrete and Computational Geometry (DCG)*, 36(4):553–572, 2006.

[2] S. F. Altschul. A protein alignment scoring system sensitive at all evolutionary distances. *Journal of molecular evolution*, 36(3):290–300, 1993.

[3] I. Bahar, A. R. Atilgan, and B. Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 2:173 – 181, 1997.

[4] P. W. Bates, Z. Chen, Y. H. Sun, G. W. Wei, and S. Zhao. Geometric and potential driving formation and evolution of biomolecular surfaces. *J. Math. Biol.*, 59:193–231, 2009.

[5] P. W. Bates, G. W. Wei, and S. Zhao. Minimal molecular surfaces and their applications. *Journal of Computational Chemistry*, 29(3):380–91, 2008.

[6] U. Bauer, M. Kerber, and J. Reininghaus. Distributed computation of persistent homology. *Proceedings of the Sixteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, 2014.

[7] P. Bendich, H. Edelsbrunner, and M. Kerber. Computing robustness and persistence for images. *IEEE Transactions on Visualization and Computer Graphics*, 16:1251–1260, 2010.

[8] P. Bendich and J. Harer. Persistent intersection homology. *Foundations of Computational Mathematics (FOCM)*, 11(3):305–336, 2011.

[9] J. Bennett, F. Vivodtzev, and V. Pascucci, editors. *Topological and statistical methods for complex data: Tackling large-scale, high-dimensional and multivariate data spaces*. Mathematics and Visualization. Springer-Verlag Berlin Heidelberg, 2015.

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

[11] P. T. Bremer, V. P. I. Hotz, and R. Peikert, editors. *Topological methods in data analysis and visualization III: Theory, algorithms and applications*. Mathematics and Visualization. Springer International Publishing, 2014.

[12] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. States, S. Swaminathan, and M. Karplus. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4:187–217, 1983.

[13] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

[14] G. Carlsson. Topology and data. *Am. Math. Soc*, 46(2):255–308, 2009.

[15] G. Carlsson and V. De Silva. Zigzag persistence. *Foundations of computational mathematics*, 10(4):367–405, 2010.

[16] G. Carlsson, V. de Silva, and D. Morozov. Zigzag persistent homology and real-valued functions. In *Proc. 25th Annu. ACM Sympos. Comput. Geom.*, pages 247–256, 2009.

[17] G. Carlsson, T. Ishkhanov, V. Silva, and A. Zomorodian. On the local behavior of spaces of natural images. *International Journal of Computer Vision*, 76(1):1–12, 2008.

[18] G. Carlsson, G. Singh, and A. Zomorodian. Computing multidimensional persistence. In *Algorithms and computation*, pages 730–739. Springer, 2009.

[19] G. Carlsson and A. Zomorodian. The theory of multidimensional persistence. *Discrete Computational Geometry*, 42(1):71–93, 2009.

[20] G. Carlsson, A. Zomorodian, A. Collins, and L. J. Guibas. Persistence barcodes for shapes. *International Journal of Shape Modeling*, 11(2):149–187, 2005.

[21] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[22] H. W. Chang, S. Bacallado, V. S. Pande, and G. E. Carlsson. Persistent topology and metastable state in conformational dynamics. *PLos ONE*, 8(4):e58699, 2013.

[23] F. Chazal, D. Cohen-Steiner, M. Glisse, L. J. Guibas, and S. Oudot. Proximity of persistence modules and their diagrams. In *Proc. 25th ACM Sympos. on Comput. Geom.*, pages 237–246, 2009.

[24] F. Chazal, L. J. Guibas, S. Y. Oudot, and P. Skraba. Persistence-based clustering in riemannian manifolds. In *Proceedings of the 27th annual ACM symposium on Computational geometry*, SoCG '11, pages 97–106, 2011.

[25] D. Chen, Z. Chen, C. Chen, W. H. Geng, and G. W. Wei. MIBPB: A software package for electrostatic analysis. *J. Comput. Chem.*, 32:657 – 670, 2011.

[26] D. Chen, Z. Chen, and G. W. Wei. Quantum dynamics in continuum for proton transport II: Variational solvent-solute interface. *International Journal for Numerical Methods in Biomedical Engineering*, 28:25 – 51, 2012.

[27] D. Chen and G. W. Wei. Quantum dynamics in continuum for proton transport—Generalized correlation. *J Chem. Phys.*, 136:134109, 2012.

[28] Z. Chen, N. A. Baker, and G. W. Wei. Differential geometry based solvation models I: Eulerian formulation. *J. Comput. Phys.*, 229:8231–8258, 2010.

[29] Z. Chen, N. A. Baker, and G. W. Wei. Differential geometry based solvation models II: Lagrangian formulation. *J. Math. Biol.*, 63:1139– 1200, 2011.

[30] Z. Chen, S. Zhao, J. Chun, D. G. Thomas, N. A. Baker, P. B. Bates, and G. W. Wei. Variational approach for nonpolar solvation analysis. *Journal of Chemical Physics*, 137(084101), 2012.
[31] J. L. Cheng, M. J. Sweredoski, and P. Baldi. DOMpro: Protein domain prediction using profiles, secondary structure, relative solvent accessibility, and recursive neural networks. *Data Mining and Knowledge Discovery*, 13:1–10, 2006.
[32] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007.
[33] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Extending persistence using poincaré and lefschetz duality. *Foundations of Computational Mathematics*, 9(1):79–103, 2009.
[34] D. Cohen-Steiner, H. Edelsbrunner, J. Harer, and D. Morozov. Persistent homology for kernels, images, and cokernels. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA 09, pages 1011–1020, 2009.
[35] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
[36] Y. Dabaghian, F. Memoli, L. Frank, and G. Carlsson. A topological paradigm for hippocampal spatial map formation using persistent homology. *PLoS Comput Biol*, 8(8):e1002581, 08 2012.
[37] S. J. Darnell, L. LeGault, and J. C. Mitchell. Kfc server: interactive forecasting of protein interaction hot spots. *NUCLEIC ACIDS RESEARCH*, 36:W265–W269, 2008.
[38] M. Dash and H. Liu. Feature selection for classification. *Intelligent data analysis*, 1(1):131–156, 1997.
[39] V. de Silva, D. Morozov, and M. Vejdemo-Johansson. Persistent cohomology and circular coordinates. *Discrete and Comput. Geom.*, 45:737–759, 2011.
[40] T. K. Dey, F. Fan, and Y. Wang. Computing topological persistence for simplicial maps. In *Proc. 30th Annu. Sympos. Comput. Geom. (SoCG)*, pages 345–354, 2014.
[41] B. Di Fabio and C. Landi. A mayer-vietoris formula for persistent homology with an application to shape recognition in the presence of occlusions. *Foundations of Computational Mathematics*, 11:499–527, 2011.
[42] H. Edelsbrunner and J. Harer. *Computational topology: an introduction*. American Mathematical Soc., 2010.
[43] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete Comput. Geom.*, 28:511–533, 2002.
[44] R.-E. Fan, P.-H. Chen, and C.-J. Lin. Working set selection using second order information for training support vector machines. *The Journal of Machine Learning Research*, 6:1889–1918, 2005.
[45] T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.
[46] X. Feng, K. Xia, Y. Tong, and G.-W. Wei. Geometric modeling of subcellular structures, organelles and large multiprotein complexes. *International Journal for Numerical Methods in Biomedical Engineering*, 28:1198–1223, 2012.
[47] X. Feng, K. L. Xia, Y. Y. Tong, and G. W. Wei. Multiscale geometric modeling of macromolecules II: lagrangian representation. *Journal of Computational Chemistry*, 34:2100–2120, 2013.
[48] C. Fernandez-Lozano, E. Fernandez-Blanco, K. Dave, N. Pedreira, M. Gestal, J. Dorado, and C. R. Munteanu. Improving enzyme regulatory protein classification by means of svm-rfe feature selection. *Molecular Biosystems*, 10:1063–1071, 2014.
[49] P. J. Flory. Statistical thermodynamics of random networks. *Proc. Roy. Soc. Lond. A,*, 351:351 – 378, 1976.
[50] N. K. Fox, S. E. Brenner, and J.-M. Chandonia. Scope: Structural classification of proteins-extended, integrating scop and astral data and classification of new structures. *Nucleic acids research*, 42(D1):D304–D309, 2014.
[51] P. Frosini. A distance for similarity classes of submanifolds of a Euclidean space. *BUllentin of Australian Mathematical Society*, 42(3):407–416, 1990.
[52] P. Frosini and C. Landi. Size theory as a topological tool for computer vision. *Pattern Recognition and Image Analysis*, 9(4):596–603, 1999.
[53] P. Frosini and C. Landi. Persistent betti numbers for a noise tolerant shape-based approach to image retrieval. *Pattern Recognition Letters*, 34:863–872, 2013.
[54] I. Fujishiro, Y. Takeshima, T. Azuma, and S. Takahashi. Volume data mining using 3d field topology analysis. *IEEE Computer Graphics and Applications*, 20(5):46–51, 2000.
[55] M. Gameiro, Y. Hiraoka, S. Izumi, M. Kramar, K. Mischaikow, and V. Nanda. Topological measurement of protein compressibility via persistence diagrams. *Japan Journal of Industrial and Applied Mathematics*, 32:1–17, 2014.
[56] R. Ghrist. Barcodes: The persistent topology of data. *Bull. Amer. Math. Soc.*, 45:61–75, 2008.
[57] N. Go, T. Noguti, and T. Nishikawa. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci.*, 80:3696 – 3700, 1983.
[58] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.
[59] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.
[60] D. Horak, S. Maletic, and M. Rajkovic. Persistent homology of complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(03):P03034, 2009.
[61] D. J. Jacobs, A. J. Rader, L. A. Kuhn, and M. F. Thorpe. Protein flexibility predictions using graph theory. *Proteins-Structure, Function, and Genetics*, 44(2):150–165, AUG 1 2001.

[62] S. Jo, M. Vargyas, J. Vasko-Szedlar, B. Roux, and W. Im. Pbeq-solver for online visualization of electrostatic potential of biomolecules. *Nucleic Acids Research*, 36:W270 –W275, 2008.
[63] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, and T. L. Madden. Ncbi blast: a better web interface. *Nucleic acids research*, 36(suppl 2):W5–W9, 2008.
[64] T. Kaczynski, K. Mischaikow, and M. Mrozek. *Computational Homology*, volume 157 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.
[65] P. M. Kasson, A. Zomorodian, S. Park, N. Singhal, L. J. Guibas, and V. S. Pande. Persistent voids a new structural metric for membrane fusion. *Bioinformatics*, 23:1753–1759, 2007.
[66] B. Krishnamoorthy, S. Provan, and A. Tropsha. A topological characterization of protein structure. In *Data Mining in Biomedicine, Springer Optimization and Its Applications*, pages 431–455, 2007.
[67] R. A. Laskowski, J. D. Watson, and J. M. Thornton. Profunc: a server for predicting protein function from 3d structure. *Nucleic acids research*, 33(suppl 2):W89–W93, 2005.
[68] D. Lee, O. Redfern, and C. Orengo. Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology*, 8(12):995–1005, 2007.
[69] H. Lee, H. Kang, M. K. Chung, B. Kim, and D. S. Lee. Persistent brain network homology from the perspective of dendrogram. *Medical Imaging, IEEE Transactions on*, 31(12):2267–2277, Dec 2012.
[70] C. S. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble. Mismatch string kernels for discriminative protein classification. *Bioinformtics*, 20:467–476, 2004.
[71] M. Levitt, C. Sander, and P. S. Stern. Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. Mol. Biol.*, 181(3):423 – 447, 1985.
[72] W. Li, A. Cowley, M. Uludag, T. Gur, H. McWilliam, S. Squizzato, Y. M. Park, N. Buso, and R. Lopez. The embl-ebi bioinformatics web and programmatic tools framework. *Nucleic acids research*, page gkv279, 2015.
[73] X. Liu, Z. Xie, and D. Yi. A fast algorithm for constructing topological structure in large data. *Homology, Homotopy and Applications*, 14:221–238, 2012.
[74] J. A. McCammon, B. R. Gelin, and M. Karplus. Dynamics of folded proteins. *Nature*, 267:585–590, 1977.
[75] W. McCulloch and W. Pitts. A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943.
[76] P. Meinicke. UProC: tools for ultra-fast protein domain classification. *Bioinformtics*, 31:1382–1388, 2015.
[77] K. Mischaikow, M. Mrozek, J. Reiss, and A. Szymczak. Construction of symbolic dynamics from experimental time series. *Physical Review Letters*, 82:1144–1147, 1999.
[78] K. Mischaikow and V. Nanda. Morse theory for filtrations and efficient computation of persistent homology. *Discrete and Computational Geometry*, 50(2):330–353, 2013.
[79] V. Nanda. Perseus: the persistent homology software. Software available at http://www.sas.upenn.edu/~vnanda/perseus.
[80] P. Niyogi, S. Smale, and S. Weinberger. A topological view of unsupervised learning from noisy data. *SIAM Journal on Computing*, 40:646–663, 2011.
[81] K. Opron, K. L. Xia, and G. W. Wei. Fast and anisotropic flexibility-rigidity index for protein flexibility and fluctuation analysis. *Journal of Chemical Physics*, 140:234105, 2014.
[82] K. Opron, K. L. Xia, and G. W. Wei. Communication: Capturing protein multiscale thermal fluctuations. *Journal of Chemical Physics*, 142(211101), 2015.
[83] S. Y. Oudot and D. R. Sheehy. Zigzag Zoology: Rips Zigzags for Homology Inference. In *Proc. 29th Annual Symposium on Computational Geometry*, pages 387–396, June 2013.
[84] D. Pachauri, C. Hinrichs, M. Chung, S. Johnson, and V. Singh. Topology-based kernels with application to inference problems in alzheimer's disease. *Medical Imaging, IEEE Transactions on*, 30(10):1760–1770, Oct 2011.
[85] J. A. Perea, A. Deckard, S. B. Haase, and J. Harer. Sw1pers: Sliding windows and 1-persistence scoring; discovering periodicity in gene expression time series data. *BMC Bioinformatics*, 16:257, 2015.
[86] J. A. Perea and J. Harer. Sliding windows and persistence: An application of topological methods to signal analysis. *Foundations of Computational Mathematics*, 15:799–838, 2015.
[87] R. M. Pielak and J. J. Chou. Influenza m2 proton channels. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1808(2):522–529, 2011.
[88] R. M. Pielak, K. Oxenoid, and J. J. Chou. Structural investigation of rimantadine inhibition of the am2-bm2 chimera channel of influenza viruses. *Structure*, 19(11):1655–1663, 2011.
[89] B. Rieck, H. Mara, and H. Leitte. Multivariate data analysis using persistence-based filtering and topological signatures. *IEEE Transactions on Visualization and Computer Graphics*, 18:2382–2391, 2012.
[90] V. Robins. Towards computing homology from finite approximations. In *Topology Proceedings*, volume 24, pages 503–532, 1999.
[91] A. Roy, A. Kucukural, and Y. Zhang. I-tasser: a unified platform for automated protein structure and function prediction. *Nature protocols*, 5(4):725–738, 2010.
[92] V. D. Silva and R. Ghrist. Blind swarms for coverage in 2-d. In *In Proceedings of Robotics: Science and Systems*, page 01, 2005.

[93] G. Singh, F. Memoli, T. Ishkhanov, G. Sapiro, G. Carlsson, and D. L. Ringach. Topological analysis of population activity in visual cortex. *Journal of Vision*, 8(8), 2008.
[94] P. Sonego, M. Pacurar, S. Dhir, A. Kertész-Farkas, A. Kocsor, Z. Gáspári, J. A. Leunissen, and S. Pongor. A protein classification benchmark collection for machine learning. *Nucleic Acids Research*, 35(suppl 1):D232–D236, 2007.
[95] G. D. Stormo, T. D. Schneider, L. Gold, and A. Ehrenfeucht. Use of the 'perceptron' algorithm to distinguish translational initiation sites in e. coli. *Nucleic Acids Research*, 10:2997–3011, 1982.
[96] M. Tasumi, H. Takenchi, S. Ataka, A. M. Dwidedi, and S. Krimm. Normal vibrations of proteins: Glucagon. *Biopolymers*, 21:711 – 714, 1982.
[97] A. Tausz, M. Vejdemo-Johansson, and H. Adams. Javaplex: A research software package for persistent (co)homology. Software available at http://code.google.com/p/javaplex, 2011.
[98] A. Tausz, M. Vejdemo-Johansson, and H. Adams. JavaPlex: A research software package for persistent (co)homology. In H. Hong and C. Yap, editors, *Proceedings of ICMS 2014*, Lecture Notes in Computer Science 8592, pages 129–136, 2014.
[99] M. M. Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.*, 77:1905 – 1908, 1996.
[100] B. Wang, B. Summa, V. Pascucci, and M. Vejdemo-Johansson. Branching and circular features in high dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 17:1902–1911, 2011.
[101] B. Wang and G. W. Wei. Objective-oriented Persistent Homology. *ArXiv e-prints*, Dec. 2014.
[102] G. W. Wei. Differential geometry based multiscale models. *Bulletin of Mathematical Biology*, 72:1562 – 1622, 2010.
[103] G.-W. Wei. Multiscale, multiphysics and multidomain models I: Basic theory. *Journal of Theoretical and Computational Chemistry*, 12(8):1341006, 2013.
[104] G.-W. Wei, Q. Zheng, Z. Chen, and K. Xia. Variational multiscale models for charge transport. *SIAM Review*, 54(4):699 – 754, 2012.
[105] W. Wu, A. Srivastava, J. Laborde, and J. F. Zhang. An efficient multiple protein structure comparison method and its application to structure clustering and outlier detection. *IEEE, BIBM*, pages 69–73, 2013.
[106] K. L. Xia, X. Feng, Y. Y. Tong, and G. W. Wei. Persistent homology for the quantitative prediction of fullerene stability. *Journal of Computational Chemsitry*, 36:408–422, 2015.
[107] K. L. Xia, K. Opron, and G. W. Wei. Multiscale multiphysics and multidomain models — Flexibility and rigidity. *Journal of Chemical Physics*, 139:194109, 2013.
[108] K. L. Xia and G. W. Wei. Persistent homology analysis of protein structure, flexibility and folding. *International Journal for Numerical Methods in Biomedical Engineerings*, 30:814–844, 2014.
[109] K. L. Xia and G. W. Wei. Multidimensional persistence in biomolecular data. *Journal Computational Chemistry*, 36:1502–1520, 2015.
[110] K. L. Xia and G. W. Wei. Persistent topology for cryo-EM data analysis. *International Journal for Numerical Methods in Biomedical Engineering*, 31:e02719, 2015.
[111] K. L. Xia, Z. X. Zhao, and G. W. Wei. Multiresolution topological simplification. *Journal Computational Biology*, 22:1–5, 2015.
[112] K. L. Xia, Z. X. Zhao, and G. W. Wei. Multiresolution persistent homology for excessively large biomolecular datasets. *Journal of Chemical Physics*, in press, 2015.
[113] Y. Yao, J. Sun, X. H. Huang, G. R. Bowman, G. Singh, M. Lesnick, L. J. Guibas, V. S. Pande, and G. Carlsson. Topological methods for exploring low-density states in biomolecular folding pathways. *The Journal of Chemical Physics*, 130:144115, 2009.
[114] Q. Zheng, S. Y. Yang, and G. W. Wei. Molecular surface generation using PDE transform. *International Journal for Numerical Methods in Biomedical Engineering*, 28:291–316, 2012.
[115] Y. C. Zhou, M. Feig, and G. W. Wei. Highly accurate biomolecular electrostatics in continuum dielectric environments. *Journal of Computational Chemistry*, 29:87–97, 2008.
[116] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete Comput. Geom.*, 33:249–274, 2005.
[117] Jie Liang. Geometry of protein shape and its evolutionary pattern for function prediction and characterization. *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, 2324–2327,2009.
[118] Liang, Jie and Kachalo, Sema and Li, Xiang and Ouyang, Zheng and Tseng, Yan-Yuan and Zhang, Jinfeng. Geometric structures of proteins for understanding folding, discriminating natives and predicting biochemical functions. *The World is a Jigsaw*, 2009.

Appendix: Instances used in Section 3.3

In this appendix we list protein SCOPe IDs used in Section 3.3.

Table 3: All alpha proteins

d1ux8a_	d3p4pb2	d1grja1	d2zjrv1	d2qwob_	d1fxkc_	d1cxzb_	d1seta1	d1k4ta1	d1qoja_
d2jdih1	d1uera1	d1pv0a_	d1rfya_	d1tjla1	d1x4ta1	d2a26a1	d2f6ma_	d1z0pa1	d1z0jb1
d2g0ua1	d2hepa1	d1gu2a_	d9anta_	d1sfea1	d1c20a_	d1biaa1	d1opca_	d1hc8a_	d2qanr1
d1whua_	d1k6ya1	d1twfj_	d1jhga_	d1ku2a1	d1vz0a1	d1rq6a_	d1cuka1	d1veja1	d1r5la1
d1enwa_	d1eija_	d1sr9a1	d1t95a1	d1ufza_	d3ugja2	d1jjcb1	d1quua1	d2e2aa_	d1jnra1
d1g73a_	d1qsda_	d2qant1	d3ldqb_	d1j2jb_	d1rrza_	d1z8ua_	d1vcta1	d1vp7a_	d1wfda_
d1xdpa1	d2crba1	d2goma1	d1lp1a_	d1gvna_	d3bvua1	d1u00a1	d1oksa_	d1nu9c1	d1r8ia_
d2qkwa1	d2ahma1	d2gf4a1	d2oo2a1	d1ebdc_	d2erla_	d1x9ba_	d1hb6a_	d1gg3a1	d4i9oa_
d2fcwa1	d1ujsa_	d1tbaa_	d1aila_	d2hp8a_	d2enda_	d3lyna_	d2wkxa1	d2gzka1	d1kx5c_
d1fpoa2	d1eexg_	d1mtyg_	d1om2a_	d1jw2a_	d2gboa1	d2gsva1	d1nfoa_	d2asra_	d256ba_
d1i4ya_	d1cgme_	d3fapb_	d1ya7o1	d1h6ga1	d2a0ba_	d1he1a_	d1f1ma_	d1wgwa_	d1qvxa_
d3tk0a_	d1knya1	d1l3pa_	d1nzea_	d1orja_	d1v74b_	d1gm5a1	d1t6ua_	d1szia_	d1ug7a_
d1xzpa1	d2huja1	d2ap3a1	d2g3ka1	d2p61a1	d4lqha_	d1niga_	d1wa8a1	d2g38a1	d2gtsa1
d2nr5a1	d1cnt1_	d1f7ua1	d2af8a_	d1unka_	d2eiaa1	d4oufa_	d1u8va1	d1gkza1	d1rj1a_
d1r3ba_	d1tdpa_	d2etda1	d1v9va1	d2fefa1	d2fug11	d3dbya1	d2qzga1	d2hi7b1	d2rlda1
d2hgka1	d1nkda_	d1joya_	d1pd3a_	d1ufia_	d1skva_	d1zkea1	d2bzba1	d2az0a1	d3im3a_
d1nh2b_	d1ecia_	d1b0nb_	d1g2ya_	d1pzqa_	d1q2ha_	d1ic8a2	d1hq1a_	d1k1va_	d1nlwa_
d1pzra_	d1qx2a_	d2jpoa_	d2ciwa1	d1iioa_	d1sh5a1	d1lnsa1	d1wixa_	d1a26a1	d1v32a_
d1baza_	d1k0ma1	d3bula1	d1khda1	d2fzta1	d1uura1	d1fioa_	d1s2xa_	d3nyla_	d2fupa1
d1t98a2	d3buxb2	d3bi1a1	d1wjta_	d2b4jc1	d2okua1	d1f6va_	d2a73b3	d1v54h_	d1n89a_
d1aiea_	d1adua1	d1p71a_	d1rm6c1	d1dj8a_	d1af7a1	d2rmra_	d1sv0a_	d1cuka2	d1z3eb1
d3ldaa1	d1ci4a_	d4klia1	d1bgxt1	d1d8ba_	d1f44a1	d1zyma1	d1ryka_	d4klia2	d1m6ya1
d1q46a1	d2fj6a1	d3ci0k1	d2gola_	d1qgta_	d1aepa_	d1l9la_	d1o82a_	d1n00a_	d1tada1
d1ej5a_	d1skyb1	d1fkma1	d1q0qa1	d2o3la1	d1abva_	d2oeba1	d1g7da_	d2qtva1	d1dvka_
d1u7ka_	d3o0gd_	d1husa_	d2qq9a2	d3ezqb_	d1hw1a2	d1ey1a_	d1a5ta1	d1b79a_	d3ju5a1
d1tx9a1	d1llaa1	d1llaa2	d1by1a_	d1boua_	d1hbna1	d1bgfa_	d1dk8a_	d1a9xa1	d1apxa_
d1vq8p1	d1aa7a_	d2abka_	d1j09a1	d1rlra1	d1dnpa1	d2pgda1	d1zkra1	d1gaia_	d1dl2a_
d1qaza_	d1n1ba1	d1r76a_	d2g0da_	d1a59a_	d1io7a_	d1rqta_	d1iiea_	d1aora1	d1d2ta_
d1wb9a1	d1f5na1	d1bvp11	d1xa6a1	d1nvus_	d1jdha_	d1re0b_	d1lsha1	d1qsaa1	d2h6fa1
d2o8pa1	d1ihga1	d1inza_	d3ag3e_	d2grrb_	d1k8kg_	d4fhrb_	d1l5ja1	d1j1ja_	d1ldja2

Table 4: All beta proteins

d2giya1	d1x4za1	d1b4ra_	d1yq2a1	d1ex0a2	d1l3wa1	d1acxa_	d1ej8a_	d1pl3a_	d1kyfa1
d1p5va1	d1kbpa1	d1vzia1	d1f00i1	d1tyeb1	d1ifra_	d1l6pa_	d2cxka1	d1lmia_	d1o75a1
d1osya_	d1roca_	d1xq4a_	d1xaka_	d1xo8a_	d2nqda_	d2dpka1	d2itea1	d1edya_	d4qmea3
d3abza4	d1f0la1	d2xbda_	d1amxa_	d3d06a_	d1e2wa1	d1h6ea_	d2qtva2	d2dexx1	d2jqaa1
d1vema1	d1h8la1	d1lm8v_	d1tfpa_	d1d2oa1	d1dmha_	d1xpna_	d1c3ga1	d1ok0a_	d2ov0a_
d1kzqa1	d1rlwa_	d1p5va2	d1n10a1	d1dcea2	d2huha1	d3ivva_	d2hnua_	d2bb2a1	d1loxa2
d1kful2	d4elda_	d1slua_	d1wpxb1	d1gofa2	d1bvp12	d1aola_	d2j1kc1	d1c3ha_	d1sfpa_
d2f4la1	d2a73b6	d1hn0a3	d1pcva_	d1khua_	d1cq3a_	d1p35a_	d1viwb_	d1w6ga1	d1yq2a4
d1jmma_	d1h2ca_	d1s2ea_	d1g8kb_	d1biaa2	d1ycsb2	d1kk8a1	d1viea_	d1vq8q1	d3vuba_
d1ex4a1	d1hyoa1	d2qqra1	d2coza1	d1m9sa2	d3ptaa3	d1azpa_	d1r4ka_	d1sf9a_	d1y71a1
d1vbva1	d2eyqa1	d3dkma_	d2heqa1	d3u1ua_	d1pcqo_	d1l1ca_	d1t2ma1	d2g3pa1	d1i4k1_
d1ib8a1	d2d6fa1	d2qi2a1	d2f5tx1	d1whia_	d4dfaa_	d3chbd_	d3v96a_	d1eova1	d1e9ga_
d1guta_	d2ch4w1	d2p5zx1	d1sr3a_	d1nnxa_	d2f4ia1	d2exda1	d2k5wa1	d2ot2a1	d4n8x1_
d3rloa1	d2j8ch2	d1bara_	d2zqoa1	d1jlya1	d3bx1c_	d3llpa1	d1t9fa_	d3brda3	d1wd3a2
d1fuia1	d1n0ua1	d1ja1a1	d1n08a_	d1f60a2	d1yx2a2	d2fg9a1	d1ywua1	d2jn9a1	d2bw0a1
d1arba_	d1bcoa1	d1skyb2	d1bd0a1	d1yloa1	d1fmba_	d1ffya2	d1wc2a1	d1ppya_	d1dfup_
d1h9db_	d1w1ha_	d1ywya1	d1nh2c_	d1o6ea_	d1a3wa1	d1ik9a1	d1i4ua_	d1nqna_	d1smpi_
d1ei5a1	d1pbya5	d2djfa_	d1y0ga_	d2gtln1	d2f09a1	d2c3ba1	d1a1xa_	d2rl8a_	d1f3ub_
d1su3a2	d1tl2a_	d1gyha_	d1s1da_	d1v3ea_	d1crua_	d1h6la_	d2hqsa1	d1ijqa1	d3o4pa_
d1k32a2	d1ofza_	d1q7fa_	d1suua_	d1zgka1	d1gofa3	d2bbkh_	d1fwxa2	d1pgua1	d1a12a_
d3gc3b1	d2xdwa1	d3niga_	d1k32a3	d1jofa_	d1ri6a_	d1shyb1	d1sqja1	d1xksa_	d1flga_
d1qksa2	d1xfda1	d1m7xa2	d2ho2a1	d1aiwa_	d1rk8c_	d1dkga1	d1v9ea_	d3enia_	d1ospo_
d2f69a1	d1vmoa_	d1i5pa2	d1vboa_	d1dlpa1	d1pxza_	d1ezga_	d1hf2a1	d1ea0a1	d1k4za_
d1vh4a_	d1k7ia1	d2bm4a1	d2qiaa_	d1l0sa_	d1p9ha_	d2cu2a1	d1ep0a_	d1w9ya1	d1ft9a2
d2arca_	d1gtfa_	d1ig0a1	d3ar4a2	d1v1ha1	d1dcza_	d1w96a1	d1gpra_	d2zjrt1	d2zdra1
d1c5ea_	d4ubpb_	d1pkha_	d1tula_	d2ftsa1	d1ml9a_	d1at0a_	d1f39a_	d2fu5a1	d2rqaa_
d3ezma_	d1lkta_	d1r6na_	d4ubpc1	d2jdih2	d1f35a_	d2ag4a_	d3zdha_	d1iaza_	d4qmea1
d2oqza_	d1e44b_	d1fjra_	d2ftsa2	d1jhna3	d1xp4a1	d1wrua1	d1eara1	d1h6wa1	d1h09a1
d1rh1a1	d1p6va_	d1js8a2	d3gpua1	d1m1ha1	d4ce8a_	d1mkfa_	d1lnza1	d1o70a1	d1ko6.1
d1o75a3	d1pgsa1	d1hx6a1	d1nlqa_	d1qqp.1	d1m06f_	d1dzla_	d1stma_	d1iq8a3	d1nc7a_

Table 5: Alpha+beta proteins

d2c4bb1	d2baaa_	d1qdqa_	d2jb0b_	d1km8a_	d1xu0a_	d1y7ma2	d4ubpa_	d1el0a_	d2cs7a1
d1bb8a_	d1qmea1	d2zjrq1	d1fita_	d2r7ja1	d1n0ua3	d1v5oa_	d1ibxa_	d1v8ca1	d1czpa_
d2saka_	d2g9hd2	d1hz6a_	d1tifa_	d1htqa1	d1tkea1	d1mjda_	d3n20c_	d2fug13	d2gria1
d3coxa2	d1mola_	d1w6ga2	d1eeja2	d1idpa_	d1udii_	d1iq8a4	d1nnva_	d1pcfa_	d4f7ea1
d2f4za1	d4ijza1	d1xqma_	d1c8za_	d2pila_	d1p32a_	d1ix5a_	d3eipa_	d1w9pa2	d3bn0a1
d2qans1	d1vq8x1	d1b33n_	d1cr5a2	d1f9za_	d1qyna_	d1bm8a_	d1dk0a_	d1eyqa_	d2v8qe1
d1lo7a_	d3bria_	d1to2i_	d1rm6a1	d2i1oa1	d1brwa3	d1vq8h1	d1fm0e_	d1r29a_	d1xb2b2
d1wina_	d1uera2	d2zjq51	d1wiba_	d1xp8a2	d4ghla_	d1e8ob_	d1whqa_	d1pdaa2	d1dq3a2
d1j26a_	d2fgga1	d2z0sa2	d3proc1	d1gpma3	d3ieua2	d1ib8a2	d1mkya3	d1v9ja_	d1kkga_
d1veha_	d2qfia1	d2bh1x1	d2qanc2	d1kkoa2	d2zjrp1	d1kp8a3	d1ghha_	d1blua_	d4fyxb1
d1jqga2	d1mlia_	d2cz4a1	d3b6ba_	d2cq2a1	d1b3ta_	d1bxna2	d2acya_	d1n0ua4	d1gh8a_
d1jjcb4	d2j5aa1	d2qanj1	d2q66a3	d1cc8a_	d1sc6a3	d1z2la2	d1dqaa1	d1ekra_	d1f3va_
d1mlaa2	d1ffgb_	d1kp6a_	d1h72c2	d1regx_	d1fvga_	d1wc3a_	d3hd2a_	d1hbnc_	d1diqa1
d1m5ha1	d1qd1a1	d2akja1	d1eara2	d1gpja3	d1o8ba2	d1ivza_	d1m1ha2	d2vv5a2	d1oy8a1
d1pbua_	d1nxia_	d1yqha1	d1in0a1	d1j27a_	d1q8ka2	d1utaa_	d1wj9a1	d1rwua_	d2p8ia_
d2a1ba1	d2vjva1	d4tnoa_	d2av5a1	d3b8pa1	d3bpda1	d1zava1	d1vq8w1	d1r8ea2	d1jh6a_
d1f32a_	d1d8ia_	d2if1a_	d2p92a1	d1lbua2	d1c05a_	d1tkea2	d1f7ua3	d1ge9a_	d1uv7a_
d1tiga_	d1qmha2	d1pava_	d1rq8a_	d1nj8a2	d1nfja_	d1ug8a_	d2d9ia1	d1kpta_	d1ev0a_
d1dwka2	d1bxni_	d1ru0a_	d1b4ba_	d1bdfa1	d2d6fc2	d2phcb2	d1r0va3	d1wb9a4	d1syxb_
d2ckca1	d1iq4a_	d1dzfa2	d1nq3a_	d2r6r12	d1t0kb_	d1clia1	d3erna_	d1q8ra_	d1r1ma_
d2ohwa1	d3byqa1	d1bjpa_	d1nvmb2	d2iafa1	d2glza1	d2dm9a1	d1oaca4	d4ec2a_	d1sgoa_
d1v5ra1	d1ylxa1	d1ewfa1	d1usub_	d2sici_	d1frsa_	d1rf8a_	d1m6ia3	d1rm6b1	d1mnma_
d4nbpa_	d1vfra_	d1dt9a3	d1c7ka_	d1qbaa4	d2abla2	d1mo1a_	d1j5ya2	d3bp3a_	d2qojz1
d1a8ra_	d1vrma1	d1qb3a_	d1jtgb_	d3elga1	d1diva1	d2hbaa1	d2gpfa1	d2ba0d2	d1efnb_
d1cbya_	d1yfsa2	d1kyfa2	d1wfra_	d1jhsa_	d1m4ia_	d1d0na1	d2qtva4	d2gnxa2	d1acfa_
d1mc0a1	d1nwza_	d1ifqa_	d1l3la2	d1ojga_	d1z09a_	d2h28a1	d2a2la1	d2grga1	d1rc9a1
d1a6ja_	d4n1ta_	d1hp1a1	d1jcua_	d1wdva_	d1qzfa2	d2y28a_	d1cyoa_	d1vcca_	d2iwxa_
d1ixma_	d1iooa_	d1c4ka3	d1g62a_	d2v3za2	d1htqa2	d1ytba1	d1kfia4	d2q3qa_	d1f46a_
d1v8ca2	d1ul7a_	d1j3ma_	d1t6aa_	d1xsza2	d2fpna1	d2pwwa1	d1mxaa1	d1ok7a1	d1rm6a2
d3mm5a2	d1go4a_	d1byra_	d1hqia_	d1clia2	d1seia_	d1rl6a1	d1iowa2	d1a0ia2	d2cnqa_