Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2024, 2024:48
https://doi.org/10.28919/cmbn/8213
ISSN: 2052-2541

A TOPOLOGICAL APPROACH FOR ANALYZING THE PROTEIN STRUCTURE

ZAKARIA LAMINE[1,*], MOHAMMED WADIA MANSOURI[1], MY ISMAIL MAMOUNI[2]

[1]LAGA research laboratory, Department of Mathematics, Faculty of Sciences, Ibnou Tofail University, Kenitra, Morocco

[2]M@Da research team, Department of Mathematics, CRMEF Rabat, Morocco

Abstract. Persistent homology is a new tool from algebraic topology, showing until nowadays a lot of success when it comes to application in biology since this latest use metrics only for measuring similarities, Embedding the geometric details and focusing on the global shape is the key point making the success of persistent homology as an efficient topological data analysis tool. In this work we will be dealing with the following points to survey our hypothesis: the flexibility-rigidity index, a classic method used to simulate molecule movements and flexible behavior, when it comes to atomic rigidity functions. We will also analyze interesting patterns in the binding site of the beta sheet generated from the pdb file 2JOX. We will be detecting and giving a simple description of different patterns generated by using javaplex generating barcodes and linear statistical results as a summary statistics.

Keywords: flexibility-rigidity index; persistent homology; COILED SERINE; beta sheet; pdb file 2JOX.
2020 AMS Subject Classification: 55N31, 62R40.

1. INTRODUCTION

Until nowadays a protein is defined to be as the main building component of all cellular tissues in all living organisms, this definition holds thanks to Anfinsen's dogma [1] but facing a real challenge regarding the complexity of the folding path of a protein, Analysis of protein

*Corresponding author
E-mail addresses: zakarialamine2@gmail.com, zakaria.lamine1@uit.ac.ma
Received September 09, 2023

structure and development of summary statistics to find an accurate structure-function relationship have made an evolutionary steps during last decades thanks to the enormous available data generated from Xray crystallography, the availability of data gives birth to a new paradigm wich is ”the complexity of data” and computational topology seems to perfectly answer a lot of questions [3], We can be sure from the XYZ distribution since all the configurations follow physical laws, but we need a better way to link between atoms in a macroscopic level in order to catch up the other aspects of a protein —involving persistent homology in detection and analysis of protein folding path was investigated using topological feature vector [3] The choice of persistent homology comes from its capacity of neglecting metric details and capturing void, cavities and holes at different scales by using a filtration parameter [11] [12] which is the truly demanded function from the mathematical tools used in the analysis of protein structure and binding sites. The majority of the mathematical models used to study protein characteristics such as flexibility, folding and structure are geometrically based ones which level up the complexity of the algorithms, we mention several methods to compute those network metrics such as VisANT [10], CentiScaPe [2], CentiLib [4] and Visone [5], but all these models and tools can’t catch up the dynamical nature of the protein which is done perfectly when using the filtration parameter [5] [17] [18].

In this work we will analyze the topology structure of COILED SERINE, and giving a substitute of the optimal characteristic distance that can be used in the flexibility-rigidity index (a classic method used to simulate molecule movements and flexible behavior, when it comes to atomic rigidity functions). We will also analyze interesting patterns in the binding site of the beta sheet generated from the pdb file 2JOX and will be detecting and giving a simple description of different patterns generated by using javaplex generating barcodes and persistent diagrams as a summary statistics. We will witness through the results, the dynamical nature of this parameter, the protocol starts with a point cloud, topology gives us the ability to hide the algebraic invariant which comes out with a final shape, the elements we will be filtering are called homology groups, two shapes or in a better axiomatic way a main level of the previously defined (protein) called secondary structure is investigated, in a first sight ”the beta sheet” and

the "alpha helix" will be reconstructed, the observations we will be using statistics on to visualize a previously theoretically justified parameter (FRI) are the (x,y) couples indicating the life time of each homology group, we will reduce dimensions until getting our XY graph.

This paper is organized in four sections: firstly an introduction (see section 1). Secondly, in section 2 we summarize the mathematical material required, especially the persistent homology tools. In section 3, we present all details of our topological approach to analyze the COILED SERINE protein structure. Finally, in section 4 we make some conclusions and discuss some further possible research issues.

2. Mathematical Background

As mentioned here above, in this section we will summarize the tools that will be used in our topological view point to approach the structure of a COS-1 cell protein. We will give the keynotes of the notion of simplicial homology, and give more details about persistent homology. For more details about simplicial homology we refer the reader to [14]. The reference [11] and [9] are considered, by almost all topological data analysts, elementary and unavoidable to learn more about persistent homology.

2.1. Simplicial homology.

Homology is the branch of algebraic topology making the computing part of it a true realization, the main application is dimentionality reduction via interesting tools such as persistent homology.

Definition 1. *A $p-$dimensional simplex (or $p-$simplex $\sigma^p = [e_0, e_1, ..., e_p]$ is the smallest convex set in a Euclidean space $\mathbb{R}^m$ containing the $p+1$ points $e_0, ..., e_p$:*

$$\Delta^p = \{(t_0, ..., t_p) \in \mathbb{R}^{p+1} : \sum_{i=0}^{p} t_i = 1 \text{ and } t_i \geq 0 \text{ for all } i = 0, ..., p\}$$

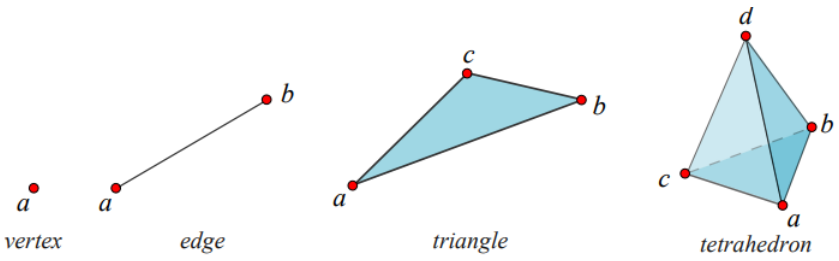

FIGURE 1. Illustration of p-simplices for p= 0, 1, 2, 3.

Definition 2. *Any simplex spanned by a subset of $e_0, \ldots, e_p$ is called face of the $p-$simplex*

from the previous figure, a face of a tetrahedron is a triangle, it can also be the union of triangles.

Definition 3. *A simplicial complex $\mathscr{K}$ is a finite set of simplices satisfying the following conditions:*

(1) For all simplices $A \in K$ with α a face of A, we have $\alpha \in K$.

(2) $A, B \in K \Rightarrow A, B$ are properly situated.

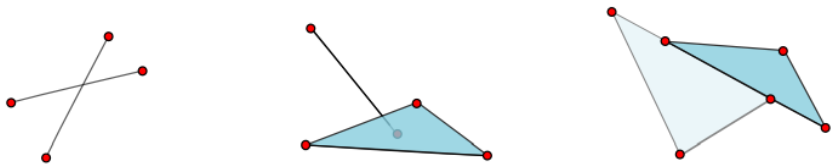

FIGURE 2. collection of simplices that do not form a simplicial complex

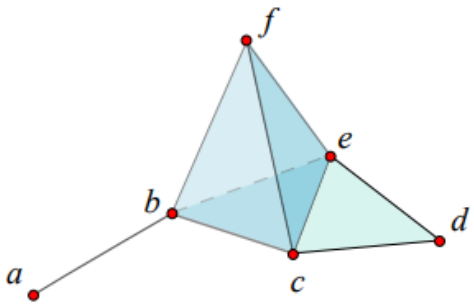

FIGURE 3. A well defined simplicial complex

Definition 4. *A p-chains is a formal sum*

$$c = \sum_{i=1}^{N_p} c_i \sigma_i^p$$

where σ_i^{pi} are p-simplicies in $\mathscr{K}$ and $c_i \in \mathbb{Z}$.

We define $(c_p + b_p)(\sigma^p) = c_p(\sigma^p) + b_p(\sigma^p)$, this induces over the set of $p-chains$ the structure of a free abelien groupe denoted $C_p(\mathscr{K})$

Definition 5. *The boundary operator is a homomorphism*

$$\partial_p : C_p(K) \to C_{p-1}(K)$$

well defined as level of generator as follows: For any p-simplex, $\sigma = [e_0, e_1, \cdots, e_p]$*, we associate the* $(p-1)$*-chain*

$$\partial\sigma = \sum_{i=0}^{p} (-1)^i [e_0, e_1, ..., \hat{e}_i, ..., e_p]$$

where $\hat{e}_i$ *is omitted.*

Thus, we obtain this chain complex

$$0 \xhookrightarrow{i} C_p(\mathscr{K}) \xrightarrow{\partial_p} C_{p-1}(\mathscr{K}) \xrightarrow{\partial_{p-1}} ... \xrightarrow{\partial_1} C_0(\mathscr{K}) \xrightarrow{\partial_0} 0$$

where $\hookrightarrow$ denotes the inclusion map. Elements of $Z_p(\mathscr{K}) = \ker \partial_p$ are called the p-*cycles*, while those of $B_p(\mathscr{K}) = \mathrm{Im}\partial_{p+1}$ are called a boundaries. The following fundamental result states the any boundary is a cycle. Indeed:

Theorem 1. *The boundary of a boundary vanishes, that is,*

$$\partial_p \circ \partial_{p+1} = 0$$

so we have $Im(\partial) \subset Ker(\partial)$

The p-th *simplicial homology group* is defined to be the quotient group

$$H_p(\mathscr{K}) = Z_p / B_p.$$

It measures the obstruction for a cycle to be a boundary. The $p-th$ Betti number is its rank:

$$\beta_p = rank(H_p).$$

For any topological space X, one way to define its homology is the following: Firstly one have to call a p-simplex of X, any continuous map

$$\sigma : \Delta^p \to X.$$

Then denote $\mathscr{K}_p(X)$ the $\mathbb{Z}$-module spanned by this p-simplicies. By this approach, one may associate to any topological space X, a simpicial complex $\mathscr{K}(X)$, unique up to homoemorphism. Secondly, one have to define the faces

$$\lambda_p : \Delta^p \to \Delta^{p-1},$$

by putting

$$\lambda_p[e_0, e_1, \ldots, e_i, \ldots, e_p] = [e_0, e_1, \ldots, \hat{e}_i, \ldots, e_p],$$

where $\hat{e}_i$ is omitted. And finally one have to define the boundaries on $\partial_p \mathscr{K}_p(X) \to \mathscr{K}_{p-1}(X)$, as follows:

$$\partial_p \sigma := \sigma \circ \lambda_p.$$

Hence the simplicial homology of X, none other than that of $\mathscr{K}_p(X)$. Mathematically speaking

$$H_p(X) := H_p(\mathscr{K}_p(X)).$$

The simplicial homology of topological space is known to be a homotopical invariant, In other word two homotopic topological spaces, have the same homology. The inverse is known to be in general false, however it can be used to prove that two topological space are not homotopic, whenever the have not the same homology. The key contribution of the simplicial homology is to compute the number of holes of a given dimension for a topological spaces. Connected components is the case of dimension 0. For example

- for a point: $H_0(pt) = \mathbb{Z}$, while $H_p(pt) = 0$ for $p > 0$;
- for a sphere: $H_0(S^n) = H_n(S^n) = \mathbb{Z}$, while $H_p(S^n) = 0$ for all other p;
- for a torus: $H_0(T) = \mathbb{Z}, H_1(T) = \mathbb{Z} \oplus \mathbb{Z}, H_2(T) = \mathbb{Z}$, while $H_p(S^n) = 0$ for all other p.

2.2. Persistent homology. Theoretically, the term persistence is for the first time introduced in [10]. It was describing an abstract definition as a natural extension of homology on filtered simplicial complexes. For applied purposes persistent homology is working as a statistical tool destined to rebuild the manifold supporting the point cloud already mentioned in the introduction, the manifold is the hidden space from which data has been extracted. the result making computing part a true realization is that persistent homology of filtered complex is nothing but the regular homology of a graded module over a polynomial ring [1]. Another interesting and explicit description of persistent homology via visualization of barcodes can be found in [9]. We suggest here a concise precise definition via classification theorem:

Remark 1 (Persistence modules)**.** *We apply the "homology functor" to the filtered chain complexes [11], so we get our "homology groups" category, which can be viewed as:*

$$0 \xhookrightarrow{i} H_p(\mathscr{K}) \xrightarrow{\partial_p} H_{p-1}(\mathscr{K}) \xrightarrow{\partial_{p-1}} \ldots \xrightarrow{\partial_1} H_0(\mathscr{K}) \xrightarrow{\partial_0} 0$$

where $\hookrightarrow$ *denotes the inclusion map.*

For a finite persistence module C with filed F coefficients

$$H_*(C;F) \cong \oplus_i x^{t_i}.F[x] \oplus (\oplus_j x^{r_j}.(F[x]/(x^{S_j}.F[x]))),$$

that are the quantification of the filtration parameter over a field. A clear description can be found in [13].

Definition 6. *The p-persistence k-th homology group*

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

well defined since B_k^{l+p} *and* Z_k^l *are subgroups of* C_k^{l+p}

To visualize efficiently the method one need to use metrics, for that aim let's define a metric on our topological space:

Definition 7. *The open vietoris-rips complex* $VR_r(X)$ *is the simplicial complex with vertices the points of* X *and* $p-$*simplicies the subsets of* X *with diameter less than* r

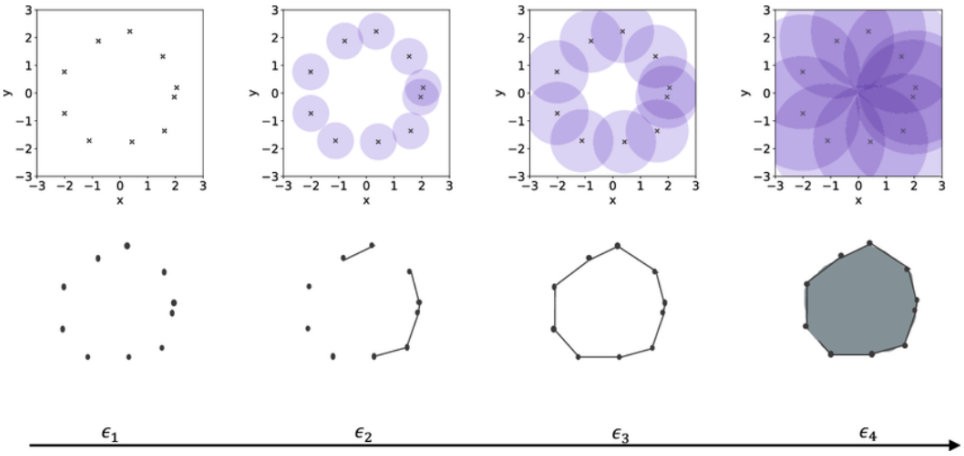

FIGURE 4. A vietoris Rips illustration

the lifetime of each homology group, which means the algebraic length of intervals (l, p) together with the values of k can be summarized and visualized using barcodes, since $\mathbb{R}$ is the perfect set to be describing an interval for analytical purposes, one needs to define homology on vector spaces to be able to use a field F, this may gives a clear definition ready to be exploited for applied purposes.

Definition 8. *A barcode is a multiset of intervals in $\mathbb{R}$, filling the previous description.*

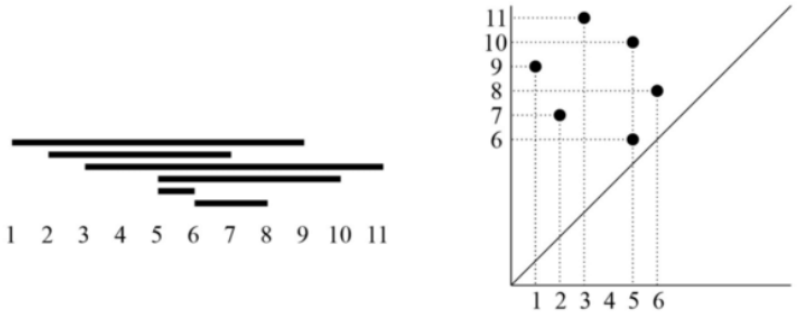

FIGURE 5. illustration of the birth and the death of a data through barcodes visualization

If our topological space X is a totaly bounded metric, one can write the barcode as: $barc_k^{VR}(X, F)$ to separate interleaving components one also needs to calculate distance between barcodes, this gives the following definition:

Definition 9. *Giving the decomposition:*

$$\oplus \mathbb{I}_x := (b(x), d(x))$$

of the persistent module, the set of $\mathbb{R}^2$ points $(b(x), d(x))$

is the persistent diagram of the barcode (l, p)

To be able to reattach intervals so the continuous property of the filtration can be filled, one needs to use a distance on the set of persistence diagrams, one way to do it is by using Wasserstein distance.

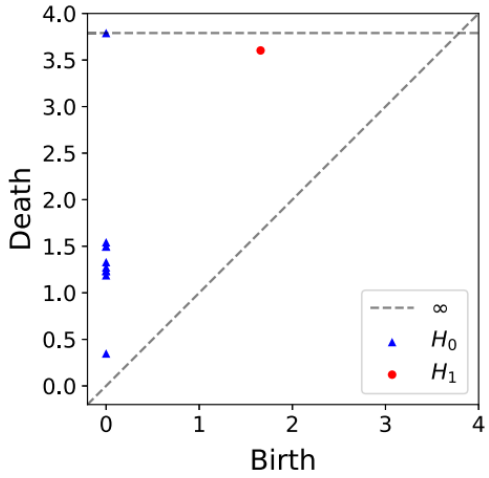

FIGURE 6. A persistent diagram for two first homologies.

As we can remark from the figure 5 each barcode can be represented by a persistent diagramme.

Definition 10. *Giving two diagrams $Dgm_k(F)$ and $Dgm_k(G)$ the (p,q)-Wasserstein distance is:*

$$W_(p,q)(Dgm_k(F),Dgm_k(G)) = inf_M(\sum_{x\in Dgm_k(K(F))}((|x-M(x)|))_{p,q})^{\frac{p}{q}}$$

where M is a bijection defined on the points of the diagonal.

The data often comes with noise since we sample from an unknown space (a probability distribution), for that reason an interesting proposition to survey and correct final results when comes the computing part is the stability theorem.

Theorem 2. *Let $f,g:\mathbb{K}\longrightarrow\mathbb{R}$ be monotone functions. Then*

$$W_p(Dgm_k(f),Dgm_k(g)) \leq |f-g|_p$$

for a homology dimension k we have:

$$W_p(Dgm_k(f),Dgm_k(g))^p \leq \sum_{dim(\sigma)\in k,k+1}|f(\sigma)-g(\sigma)|^p$$

One reason the previous theorem is called stability theorem is the contractibility of the wasserstein distance, this guarantee theoritically the mapping between data and associated persistent diagrams is a well defined homeomorphism.

3. Topological Data Analysis of the Protein

The most popular way TDA is exploited is for clustering purposes through persistent homology since this was the immediate extension of applied statistics in TDA, this comes from the intrinsic property of a point cloud, even said the axiomatic presentation seems to hide greater strategies [15], the field of application making until nowadays a great success is molecular biology since this latest doesn't fit into geometric representations when comes serious investigations or interesting behaviours such as flexibility and folding of proteins, plus the extremely expensive and complicated computation power needed, an interesting application is the protein binding analysis [16]. before we present parameters used to generate a suitable filtration one needs to comprehend in depth the notion of a protein, what is making it such an interesting concept and how modern models has been shaped through accumulation of interesting results and surveys, we need to mention that with the evolution in mathematical tools and computation power a lot of theoretical hypothesis made it to a well defined quantified results, the first step to protein structure definition and analysis start with a nobel prize in 1972 for his work on the connection between the amino acid sequence and the biologically active conformation, CHRISTIAN ANFINSEN gives to this conformation the first and last definition of a protein as a concept as well as a hypothesis to be investigated, which means all the researches made in proteins analysis are about questioning between the amino acid sequence and the active conformation, we must wait until 1994 when critical assessment of structure prediction becomes a true valued enterprise, the challenge starts when the relation structure,prediction takes place, the only way to do the calculations was through quantum mechanics which is not a sustainable way, for that reason after gathering an interesting amount of data the only way to complete databases was through dealing with the structure¼prediction question, this demanded a comprehension of the folding path, then naturally rises the works and results on protein flexibility and rigidity using mathematical statistical methods rather than experimental geometric ones, we cite [6] [7], for more enlightenment through an interesting detailed investigation of topologyfunction relationship paradigm of proteins.

3.0.1. *Topological fingerprints of alpha helix and beta sheet.* This section is devoted to an application part, the protocol is statistical inference for observations that are barcodes with the

aim to derive a comparison answering if the topological method gives the same result as the geometrical one, we will be using existing data from the freely protein data bank existing on the net, then we will smoothly be reading results as any statistical study, rearranging data will take place when barcodes seems noisy and difficult to compute, we will consider the Gaussian noise to set up our point cloud data set then an accumulated bar lengths to define the topological method in aim to make visual comparison.

To analyze an alpha helix structure, we download a protein of PDB ID: 1COS which can be viewed as an alpha helix chain with 30 residues. In the all-atom model, atoms are considered the same, each atom is associated with the same radius in the distance based filtration. The stream will be constructed for the point cloud data which is the *xyz* coordinates of the all atom representation. The size is not too large to choose a landmark selector, so we will simply build a Vietoris-Rips stream. We can choose a better filtration but for the limited computation power we stick with the value of 8. In this case a Vietoris-Rips simpicial complex is largely sufficient to decipher the topological fingerprints (a small data set) so their is no need to use a landmark selector, which can be seen in the code shown below.

```
>> size (ecos)

ans =

    696       3

>> max_dimension = 3;
>> max_filtration_value = 8;
>> num_divisions = 1000;
>> stream = api.Plex4.create
VietorisRipsStream(ecos, max_dimension,
... max_filtration_value,
num_divisions);
```

```
>> num_simplices = stream.getSize()

num_simplices =

      3259289

>> persistence = api.Plex4.get
ModularSimplicialAlgorithm(max_dimension,
2);
>> options.filename = 'cois';
>> options.max_filtration_value
= max_filtration_value;
>> options.max_dimension = max_dimension
- 1;
>> options.side_by_side = true;
>> plot_barcodes(intervals, options);
```

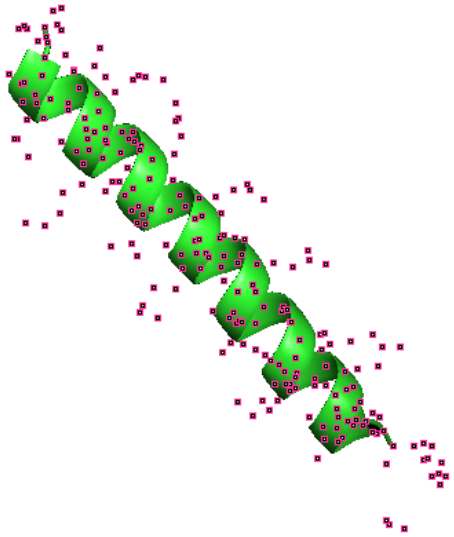

FIGURE 7. The all-atom representation of an alpha helix

We obtain the topological representation of our data in the form of a barcode, which can be called a topological fingerprints.

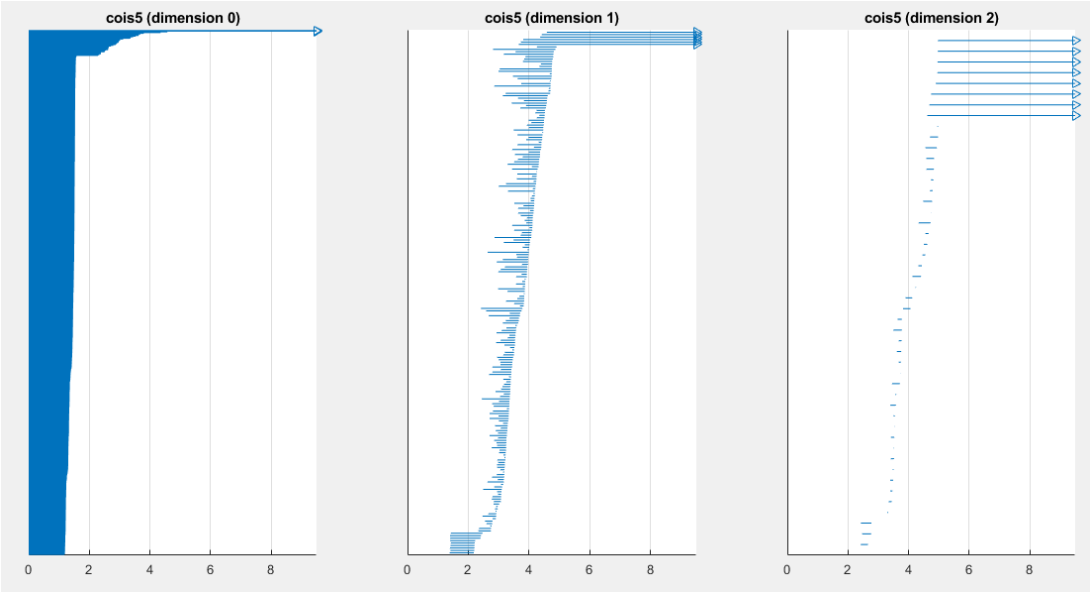

FIGURE 8. alpha helix related topological fingerprints

The β_0 bars reveal the bond length information, the filtration starts by identifying connected components, growing balls continue until they intersect leaving behind them the bond length information, starting then by identifying the 1-dimensional holes. physically, for protein molecule, the bond length is between 1 to 2 Å, in order to get an adequate filtration the bond length is reflected in the distance based filtration.

β_1 and β_2 are due to the loop, hole and void type of structures (it is difficult to directly decipher this high dimensional topological information).

To detect more topological details of the helix structure, we utilize the CG with each amino acid represented by its C_α atom. The simplices are constructed which is helpful for the detection of the helix structure So the corresponding barcode is simplified. As the last construction a Vietoris-Rips stream is largely sufficient to decipher the topological features of our data which is a 18 points in a 3-dimensional space. A part of the Matlab© code is shown below.

```
>> load ecos1
>> size (ecos)

ans =

    18     3

>> max_dimension = 2;
>> max_filtration_value = 23;
>> num_divisions = 1000;
>> stream = api.Plex4.createVietorisRipsStream
(ecos, max_dimension, ...
max_filtration_value, num_divisions);
>> options.filename = 'coiis2';
>> options.max_filtration_value =
max_filtration_value;
>> options.max_dimension = max_dimension - 1;
>> persistence = api.Plex4.get
ModularSimplicialAlgorithm(max_dimension, 2);
>> options.side_by_side = true;
>> intervals = persistence.computeIntervals
(stream);
>> plot_barcodes(intervals, options);
```

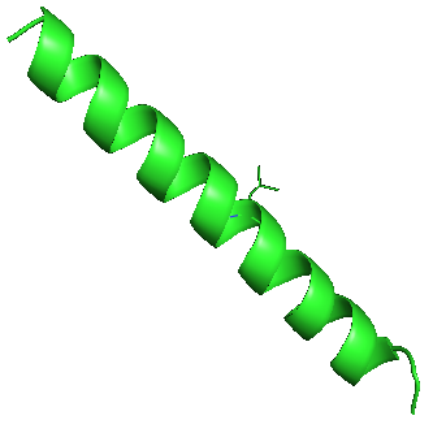

FIGURE 9. CG representation of an alpha helix generated from a protein of pdb ID 1COS

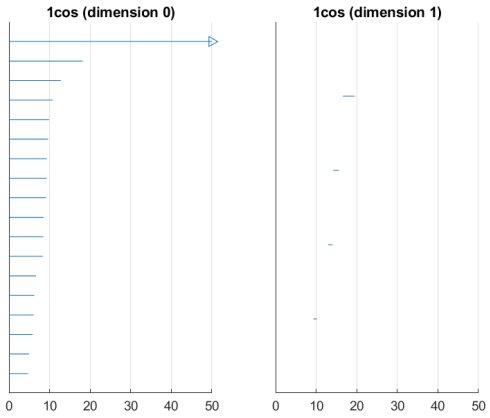

FIGURE 10. fingerprint of CG representation of an alpha helix generated from a protein of pdb ID 1COS

As we've already mentioned in the literature, all these barcodes are significant they can hide a tremendous information about our level of structure or the result of a particular configuration ..., it is up to us to catch the topological meaning of these bars in order to find accurate statistical tests, for this aim, we will calculate the helix homology, but this time we will be slicing a piece of 4 C_α atoms from the back bone and study its persistent homology behavior, then one more C_α atom is added at a time. We repeat the process as the figure shows:

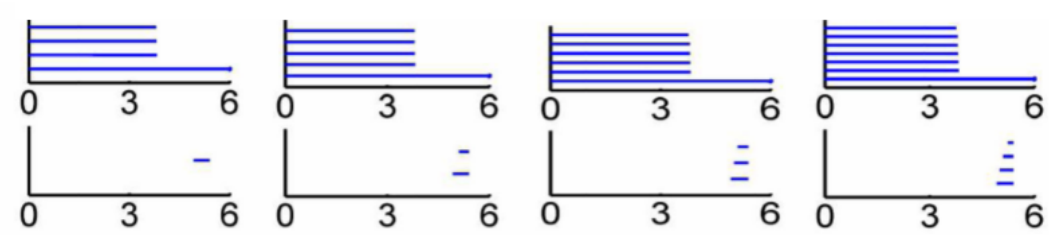

FIGURE 11. Each 4 carbon atoms represent one dimensional loop

This time it can be seen clearly that each four C_α atoms in the alpha helix form a one-dimensional loop, corresponding to a β_1 bar. By adding more C_α atoms, more loops are created and more β_1 bars are obtained. Finally, 16 residues in the alpha helix produce exactly 4 loops as seen in Fig. 4.4. In the case of beta sheets things still similar to the alpha helix, in the all-atom representation, the generated barcode has a complicated pattern due to excessively many residual atoms The barcode of the CG model, on the other hand, is much simpler with only 7 individual β_1 bars. A part of the matlab code is shown below.

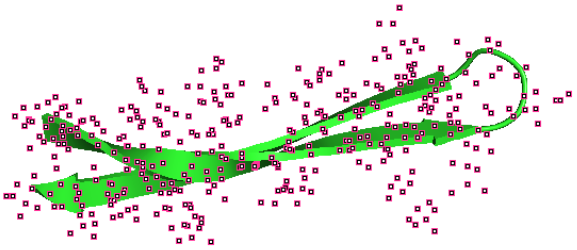

FIGURE 12. An all-atom representation of the beta sheet structure generated from PDB 2JOX

```
>> load finbeta
>> max_dimension = 2;
>> max_filtration_value = 5;
>> num_divisions=1000
num_divisions =
        1000
>> stream = api.Plex4.create
VietorisRipsStream(beti001,
max_dimension, ...
max_filtration_value, num_divisions);
>> num_simplices = stream.getSize()
num_simplices =
     149776
>> persistence = api.Plex4.get
ModularSimplicialAlgorithm(max_dimension,
2);
>> intervals = persistence.computeIntervals
(stream);
>> options.filename = '1bet';
>> options.max_filtration_value =
max_filtration_value;
>> options.max_dimension = max_dimension
- 1;
>> options.side_by_side = true;
>> plot_barcodes(intervals, options);
```

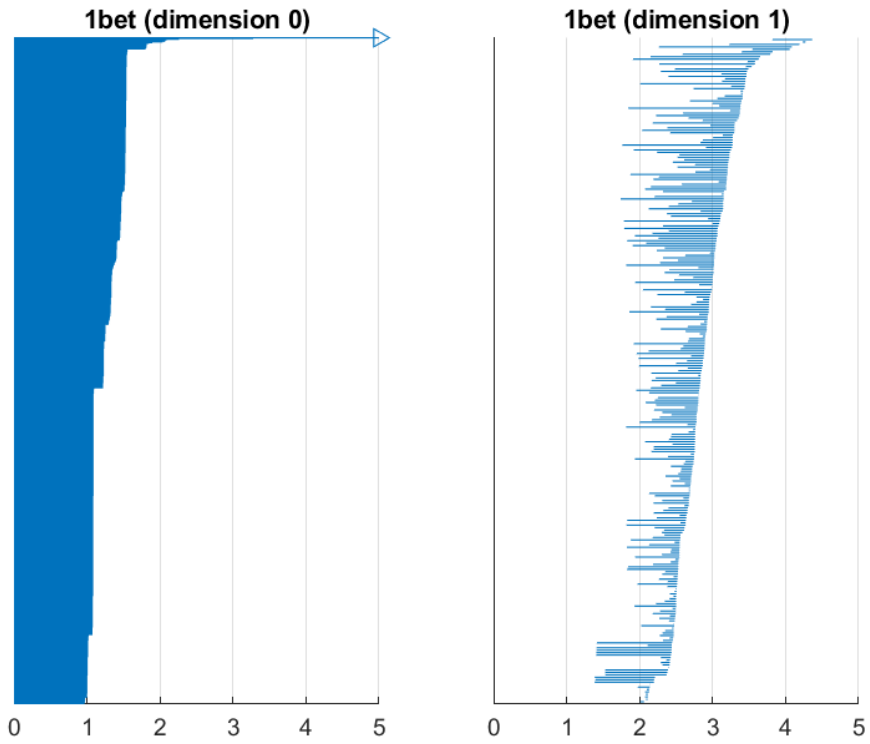

FIGURE 13. Topological fingerprint of the all-atom representation of the beta sheet structure generated from PDB 2JOX

For the Coarse grained model we consider only the alpha carbon atoms, to catch up the structure of the backbone we will be constructing a Vietoris-Rips stream.

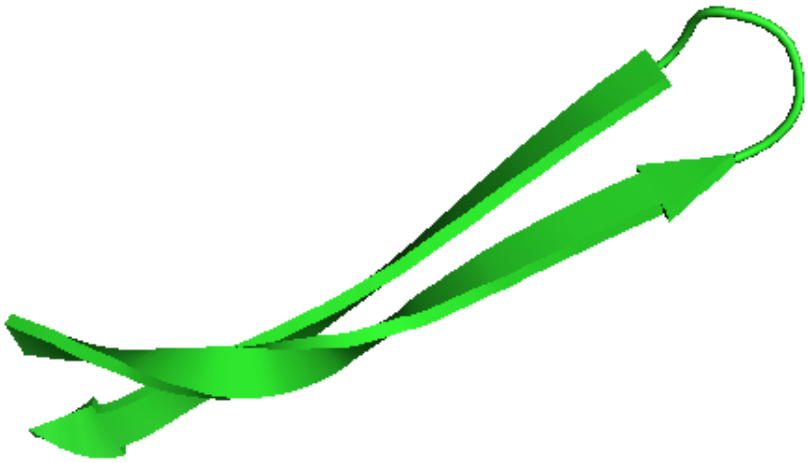

FIGURE 14. Backbone of the beta sheet structure generated from PDB 2JOX

```
>> load bety
>> size(beti01)
ans =

    24     3

>> max_dimension = 2;
>> max_filtration_value = 20;
>> num_divisions=1000
num_divisions =
        1000
>> stream = api.Plex4.
createVietorisRipsStream
(beti01, max_dimension, ...
max_filtration_value, num_divisions);
>> num_simplices = stream.getSize()
```

```
num_simplices =
          410
>> persistence = api.Plex4.get
ModularSimplicialAlgorithm(max_dimension, 2);
>> intervals = persistence.computeIntervals
(stream);
>> options.filename = 'bet';
>> options.max_filtration_value =
max_filtration_value;
>> options.max_dimension = max_dimension - 1;
>> options.side_by_side = true;
>> plot_barcodes(intervals, options);
```

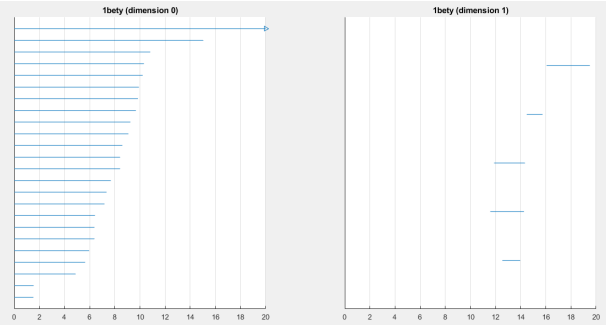

FIGURE 15. barcode for the coarse grained model for beta sheet generated from the protein of pdb ID 2JOX

First, as the filtration begins, adjoined C_α atoms in the same strand form 1-simplex. After that, adjacent C_α atoms in two different strands connect with each other as the filtration continues, which leads to one-dimensional circles and β_1 bars. The further filtration terminates all the β_1 bars. There is no β_1 bar in the CG representation of beta sheet structures.

As mentioned in the literature. We assume that the four levels are well defining the full structure of the main building component (protein) and we let our protocol catch up the results.

3.0.2. *Parameters used to generate a suitable filtration.* Proteins possess an intrinsic flexibility that allows them to function through molecular interactions within the cell, among cells and even between organisms. Many models have been proposed such as the molecular non-linear dynamic (MND) and flexibility-rigidity index (FRI) to analyze protein flexibility [8], the fundamental assumption of these methods is that. MND and FRI not only offer protein flexibility analysis, but also provide correlation matrix based filtration for the persistent homology analysis of proteins, An easy example defining the distance matrices for persistent homology uses can be found in [9]. One of the techniques that are utilized in the present flexibility analysis is Molecular non-linear dynamics: we denote the coordinates of atoms in the molecule studied as $r_1, r_2, \ldots, r_i, \ldots, r_N$, where $r_i \in R^3$ is the position vector of the j^{th} atom. The Euclidean distance between i^{th} and j^{th} atom r_{ij} can be calculated. We can easily construct our topological connectivity matrix serving as the input point cloud for our "barcode statistical inference" with monotonically decreasing radial basis functions. The general form is:

$$c_{ij} = \omega_{ij}\Phi(r_{ij}, \eta_{ij})$$

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

A generalized exponential kernel has the form $\Phi(r, \eta) = e^{-(r/\eta)^k}$, $k > 0$. And the Lorentz type of kernels is: $\Phi(r, \eta) = \frac{1}{1+(r/\eta)^\nu}$, $\nu > 0$.

The parameters k, ν, and η are adjustable. We usually search over a certain reasonable range of parameters to find the best fitting result by comparing with experimental B-factors [6]. It is assumed that each particle in a protein can be viewed as a non-linear oscillator and its dynamics can be represented by a non-linear equation. The interactions between particles are represented by the correlation matrix (c_{ij}). Therefore, for the whole protein of N particles, we set a non-linear dynamical system as:

$$\frac{du}{dt} = F(u) + Eu$$

Where $u = (u_1, u_2, ..., u_i, ..., u_N)^T$ is an array of state functions for N non-linear oscillators (T denotes the transpose),

$$u_j = (u_{j1}, u_{j2}, ..., u_{ji}, ..., u_{jN})$$

is an n-dimensional non-linear function for the j^{th} oscillator, $F(u) = (F(u_1), F(u_2), ..., F(u_N)^T$ is an array of non-linear functions of N oscillators, and

$$E = \varepsilon C \bigotimes \Gamma$$

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

The transverse stability of the MND system gradually increases during the protein folding from disorder conformations to their well-defined natural structure.

3.1. persistent homology analysis of the characteristic distance. We consider a folding protein that constitutes N particles and has the spatio-temporal complexity of $R^{3N} * R^{+}$. We Assume that our system can be described as a set of N nonlinear oscillators of dimension $R^{nN} * R^{+}$, where n is the dimensionality of a single nonlinear oscillator. One of the keys to MND model is the characteristic distance η to weight the distance effect in the geometry to topology mapping [6] [7], as shown in equations. $\Phi(r,\eta) = e^{-(r/\eta)^k}$ and $\Phi(r,\eta) = \frac{1}{1+(r/\eta)^v}$. Persistent homology can provide a quantitative prediction of optimal characteristic distances in MND and FRI. The optimal characteristic distance varies from protein to protein. An adequate filtration process is the essence of persistent homology analysis, for that a filtration matrix based on a modification of the correlation matrix of MND is proposed:

$$M_{ij} = \begin{cases} 1 - \Phi(r_{ij}, \eta_{ij}) & \text{if } i \neq j \\ 0 & \text{if } i = j. \end{cases}$$

Where $0 \leqslant \Phi(r_{ij}, \eta_{ij)} \leqslant 1$ is defined previously .with using the exponential kernel with parameter $K = 2$. When characteristic length varies, the formation of simplicial complex or topological connectivity changes too. To illustrate this point, a protein of pdb ID 1COS is used as an example. The related persistent connectivity patterns in term of β_1 are depicted in Fig. 4.9.

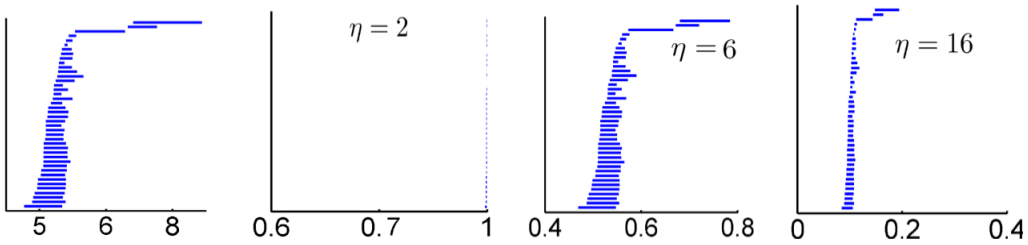

FIGURE 16. Comparison of β_1 behaviors in different filtration

Comparison of β_1 behaviors in different filtration settings for protein 1COS Cα point cloud data. Distance based filtration is shown in the first barcod representation in figure 4.12. The correlation matrix based filtration with exponential kernel ($\kappa = 2$) is used in the second, the third and the fourth. The η is chosen to be $\overset{A}{\text{deg}}$, 6Å and 16Å in the second, third and fourth barcode representations, respectively. The β_1 bar patterns are very similar but have different persistent values. The β_1 bar pattern in the seconde differs much from the rest due to a small characteristic distance $\eta = 2\mathring{A}$ global behavior is captured in all cases and the local connectivity is not overemphasized, this shows the efficacy of the correlation matrix based filtration, however some missing bars shows the underestimation of certain protein.

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

$$A_j = \sum_{i=1} P_G(L_i^{\beta_j})$$

where P_G is the Gaussian probability measure and $L_i^{\beta_j}$ is the length of the i_{th} bar of the j_{th} Betti number. We vary the value of η from 1Å to 21Å, for protein 1COS and compare the accumulated bar length A1 with the CC values obtained with FRI over the same range of η:

$$CC = \frac{\sum_{i=1}^{n}((B_i)^e - (\overline{B})^e)((B_i)^t - (\overline{B})^t)}{[\sum_{i=1}^{n}((B_i)^e - (\overline{B})^e)^2 \sum_{i=1}^{n}(((B_i)^t - (\overline{B})^t)^2)]^{\frac{1}{2}}}$$

Where $B_i^t, i = 1,2,...,N$ are a set of predicted B-factors by using the proposed method and $B_i^e, i = 1,2,\ldots,N$ are a set of experimental B-factors extracted from the PDB file.

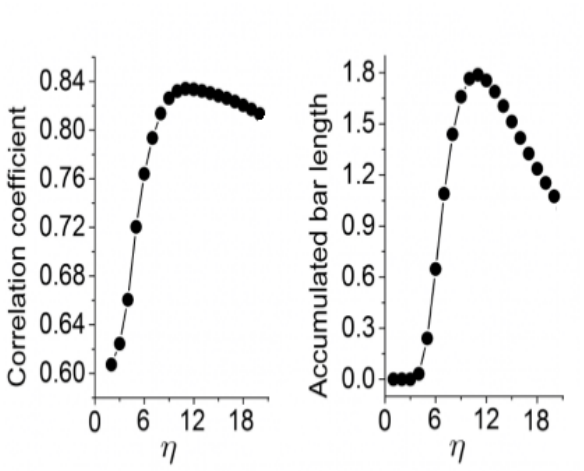

FIGURE 17. The comparison between the correlation coefficient from the B-factor prediction by FRI (left chart) and accumulated bar length from persistent homology modeling (right chart) under various η value

PDB ID	**PDB ID**	**PDB ID**	**PDB ID**	**PDB ID**	**PDB ID**
__1BX7	1DF4	2JOX	1COS	1DGV	7UG5
1GK7	1NKD	2OL9	3MD4	1DBP	7UI2
4AXY	1NOT	3Q2X	IHJE	1DK8	7UJ2

TABLE 1. Proteins used in persistent homology analysis of optimal characteristic distance

Protein data bank labels for 30 Complexes of coiled serine.

The two approaches share the same general trend in their behavior as η is increased. Both CC and A1 reach their maximum around $\eta = 12\mathring{A}$. The further increase of η leads to the decrease of both CC and A1. An optimal η in FRI model offers a best prediction of protein flexibility. In the correlation matrix filtration η impacts the birth and death of each given k-complex. For example, a pair of 2-complexes that do not coexist at a given cutoff distance in the distance based filtration might coexist at an appropriate η value in the correlation matrix based filtration. Once this achieved, the simplicies coexist, and the only component left is a 3-dimensional loop

describing the largest value of CC which means the end of the flexible behavior, the length of this loop is exactly the characteristic length of our molecule Since the same kernel and the same η are used in the FRI model and the persistent homology model (i.e., accumulation bar length).

4. Conclusion and Discussion

This work is showing an easy application of persistent homology, with the main focus of presenting a road map to get familiarized with the axiomatic idea, yet with a spectacular result, it was out of the scope of this proposition to theoretically justify the use of statistical tests on the set of barcodes, but the application shows clearly that the method can surpass a simple statistical approach, and instead of conducting a molecular dynamic simulation it is easier to use existing information from models to construct a quantified sequence of barcodes then to look for its convergence limit, we can find interesting productions in the literature but none exploited fully persistent homology far from being a statistical tool, an interesting attempt by using dynamical distances was made by Peter Bubenik and collaborators [16] , but couldnt theoretically justify barcodes as a statistical observation, instead it gives birth to a new functional tool wich is persistent landscapes,

Conflict of Interests

The authors declare that there is no conflict of interests.

References

[1] V. Prasolov, Elements of homology theory, American Mathematical Society, Providence, Rhode Island, 2007. https://doi.org/10.1090/gsm/081.

[2] H. Edelsbrunner, D. Morozov, Persistent homology: theory and practice, in: Proceedings of the European Congress of Mathematics, 31–50, 2012.

[3] Z. Hu, J.H. Hung, Y. Wang, et al. VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology, Nucleic Acids Res. 37 (2009), W115–W121. https://doi.org/10.1093/nar/gkp406.

[4] J. Gräßler, D. Koschützki, F. Schreiber, CentiLib: comprehensive analysis and exploration of network centralities, Bioinformatics. 28 (2012), 1178–1179. https://doi.org/10.1093/bioinformatics/bts106.

[5] D. Bramer, G.W. Wei, Atom-specific persistent homology and its application to protein flexibility analysis, Comput. Math. Biophys. 8 (2020), 1–35. https://doi.org/10.1515/cmb-2020-0001.

[6] K. Xia, K. Opron, G.W. Wei, Multiscale multiphysics and multidomain models-flexibility and rigidity, J. Chem. Phys. 139 (2013), 194109. https://doi.org/10.1063/1.4830404.

[7] K. Xia, G.W. Wei, Stochastic model for protein flexibility analysis, Phys. Rev. E. 88 (2013), 062709. https://doi.org/10.1103/physreve.88.062709.

[8] K. Opron, K. Xia, Z. Burton, et al. Flexibility–rigidity index for protein–nucleic acid flexibility and fluctuation analysis, J. Comput. Chem. 37 (2016), 1283–1295. https://doi.org/10.1002/jcc.24320.

[9] G. Carlsson, Topology and data, Bull. Amer. Math. Soc. 46 (2009), 255–308.

[10] T. Ichinomiya, I. Obayashi, Y. Hiraoka, Protein-folding analysis using features obtained by persistent homology, Biophys. J. 118 (2020), 2926–2937. https://doi.org/10.1016/j.bpj.2020.04.032.

[11] J. Liu, K.L. Xia, J. Wu, et al. Biomolecular topology: modelling and analysis, Acta. Math. Sin.-English Ser. 38 (2022), 1901–1938. https://doi.org/10.1007/s10114-022-2326-5.

[12] M. Buchet, F. Chazal, S.Y. Oudot, et al. Efficient and robust persistent homology for measures, Comput. Geom. 58 (2016), 70–96. https://doi.org/10.1016/j.comgeo.2016.07.001.

[13] A. Zomorodian, G. Carlsson, Computing persistent homology, in: Proceedings of the Twentieth Annual Symposium on Computational Geometry, (2004), 347–356. https://doi.org/10.1145/997817.997870.

[14] M.S. Lee, Q.C. Ji, eds., Protein analysis using mass spectrometry: accelerating protein biotherapeutics from lab to patient, Wiley, 2017. https://doi.org/10.1002/9781119371779.

[15] K. Xia, X. Feng, Y. Tong, G.W. Wei, Persistent homology for the quantitative prediction of fullerene stability, J. Comput. Chem. 36 (2014), 408–422. https://doi.org/10.1002/jcc.23816.

[16] V. Kovacev-Nikolic, P. Bubenik, D. Nikolić, et al. Using persistent homology and dynamical distances to analyze protein binding, Stat. Appl. Genetics Mol. Biol. 15 (2016), 19–38. https://doi.org/10.1515/sagmb-2015-0057.

[17] Z. Lamine, M.I. Mamouni, M.W. Mansouri, A topological data analysis of the protein structure, Int. J. Anal. Appl. 21 (2023), 136. https://doi.org/10.28924/2291-8639-21-2023-136.

[18] M.I. Mamouni, Z. Lamine, M.W. Mansouri, A topological approach for analyzing the protein structure, preprint, (2023). https://doi.org/10.21203/rs.3.rs-3269454/v1.