International Journal of Analysis and Applications

A Topological Data Analysis of the Protein Structure

Zakaria Lamine[1,2,*], **My Ismail Mamouni**[2], **Mohammed Wadia Mansouri**[3]

[1] *Department of Mathematics, Research Laboratory LAGA, Faculty of Sciences, Ibnou Tofail University, Kenitra, Morocco*

[2] *Department of Mathematics, Research Team: M@DA, CRMEF Rabat, Morocco*

[3] *Department of Mathematics, Research Laboratory LAGA, Faculty of Sciences, Ibnou Tofail University, Kenitra, Morocco*

[*] *Corresponding author:* zakarialamine2@gmail.com

Abstract. Persistent homology is a tool from a set of methods called Topological data analysis, showing until nowadays a lot of success when it comes to application in biology since this latest uses 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, this will be investigated in the paper since enormous work already done in the field and results seems to be endless, as an efficient topological data analysis tool. In this work we will be confirming the latest assumption (topology embeds geometry) by displaying the structure of COILED SERINE which is a protein estimated to constitute 3-5 percent of the encoded residues in most genomes, 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. 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.

1. Introduction

Giving interpretation of statistical summaries is being the main building block of any conclusion of an applied mathematical result, the paradigm of a well defined figure reflecting the starting hypothesis is the key point to any valuable scientific work, for that reason we will be investigating recent works

Received: Aug. 7, 2023.

2020 *Mathematics Subject Classification.* 55N31, 62R40.

Key words and phrases. persistent homology; protein structure topology; binding site; COILED SERINE.

https://doi.org/10.28924/2291-8639-21-2023-136

ISSN: 2291-8639

on topological data analysis since this latest is until nowadays being considered as the perfect tool on the two levels, algorithmic and axiomatic to be approaching a convincing answer. We will be dealing with the following points to survey our hypothesis:

- categories : parametric vs non parametric in TDA.
- parametric models: 1st methodology in TDA: Modeling and replicating statistical topology.
- Space of persistence diagrams, first definition, wasserstein distance measure.
- KERNEL DENSITY ESTIMATOR.
- point cloud, filtration parameter.
- statistical tests.
- Norm of a persistent diagram.

Different exploitations of TDA in molecular biology can be resumed in the following points:

- protein structure prediction.
- protein structure analysis.
- protein-ligand binding : integration of "element specific persistent homology" also called "multicomponent persistent homology" for protein-ligand binding affinity prediction.
- protein binding site analysis.

And different statistical methods can be summarized in :

- linear statistical approach using persistent homology.
- element specific persistent homology (ESPH).
- atom specific persistent homology (ASPH).
- Electrostatic persistence (EP).
- multi component persistence homology (MCPH).
- multi level persistent homology (MLPH).
- Electrostatic persistence.
- topological descriptors for an unsupervised learning approach.

A linear statistical approach for persistent homology is known to be the best way for none mathematical theoreticians to be deriving interesting results in the field of protein structure prediction and analysis, the first time TDA was used for such purpose is via persistent homology [1], with a comparison between topological and geometrical methods through a construction of different physical methods Xia, K and collaborators gives interesting results giving itself a confirmation to the well defined paradigm of topology-function rather than the geometry-function one, a detailed description of the method using persistent landscapes as the statistical summary was giving in [2] for the analysis of a "protein binding", the introduction of that tool comes from the restricted theoritical frame of barcodes to be defining a statistical observation since the statistical study involves calculating frechet means and a bijection between a frechet means and the set of barcodes seems to clearly be difficult to realize, barcodes are known to be the popularized statistical summary of persistent homology, using accumulated bar length is another interesting way where persistent homology can be exploited as shown in [3] to figure out exponential kernel of molecular dynamic simulations, highlighting topological signature of an atom in a macro molecule or " weighted persistent homology" [4] (ASPH) can even be more precise on the macromolecular level.

1.1. **Persistent homology, clustering an atoms point cloud.** 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 [5] but facing a real challenge regarding the complexity of the folding path of a protein, Analysis of protein 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 [6], 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 [7] 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 [8] [9] 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 [11], CentiLib [12] and Visone [2], 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 [13].

As already mentioned we will analyze the topology structure of COILED SERINE, and giving a clear response of how 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) can be replaced by a simple topological descriptors. 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 linear statistical 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 the previously defined molecule. 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 [15] and [16] 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 2.1. *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 $e_0, ..., e_p$:*

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

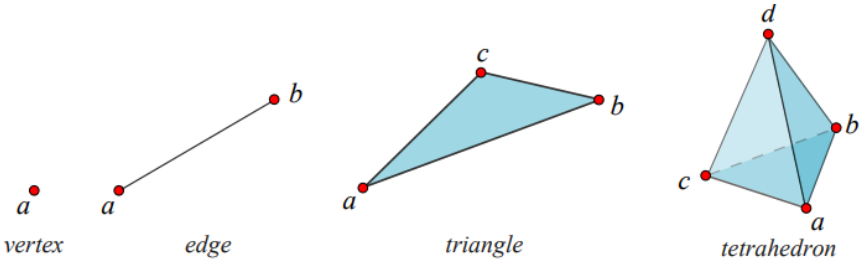

Figure 1. Illustration of p-simplices for p= 0, 1, 2, 3.

Definition 2.2. *Any simplex spanned by a subset of $e_0, ..., e_p$ 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 2.3. *A simplicial complex $\mathcal{K}$ 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.*

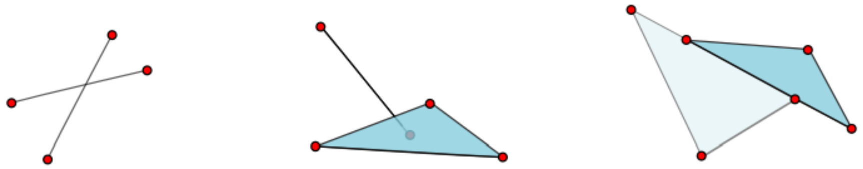

Figure 2. collection of simplices that do not form a simplicial complex

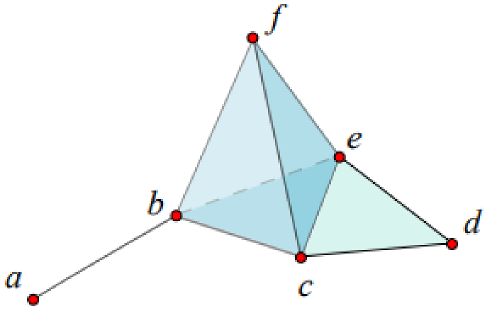

Figure 3. A well defined simplicial complex

Definition 2.4. *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 $\mathcal{K}$ 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 $C_p(\mathcal{K})$

Definition 2.5. *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

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

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

Theorem 2.1. *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

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

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 $\mathcal{K}_p(X)$ the $\mathbb{Z}$-module spanned by this p-simplicies. By this approach, one may associate to any topological space X, a simpicial complex $\mathcal{K}(X)$, 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 $\partial_p\mathcal{K}_p(X) \to \mathcal{K}_{p-1}(X)$, as follows:

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

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

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

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 : $H_0(pt) = \mathcal{Z}$, while $H_p(pt) = 0$ for $p > 0$;
- for a sphere : $H_0(S^n) = H_n(S^n) = \mathcal{Z}$, 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 [11]. 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 [17]. Another interesting and explicit description of persistent homology via visualization of barcodes can be found in [18]. We suggest here a concise precise definition via classification theorem :

Remark 2.1 (Persistence modules)**.** *We apply the "homology functor" to the filtered chain complexes [2], so we get our "homology groups" category, which can be viewed as :*

$$0 \overset{i}{\hookrightarrow} H_p(\mathcal{K}) \xrightarrow{\partial_p} H_{p-1}(\mathcal{K}) \xrightarrow{\partial_{p-1}} \dots \xrightarrow{\partial_1} H_0(\mathcal{K}) \xrightarrow{\partial_0} 0$$

where $\hookrightarrow$ denotes the inclusion map.

Theorem 2.2. *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 2.6. *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 2.7. *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 .

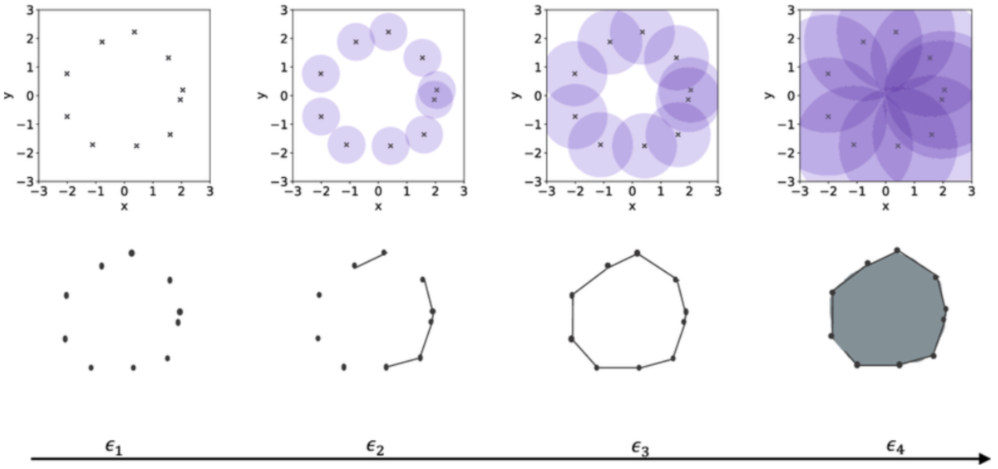

Figure 4. A vietoris Rips illustration

The lifetime of each homology group, which means the algebraic length of intervals (I, 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 2.8. *A barcode is a multiset of intervals in $\mathbb{R}$, filling the previous description.*

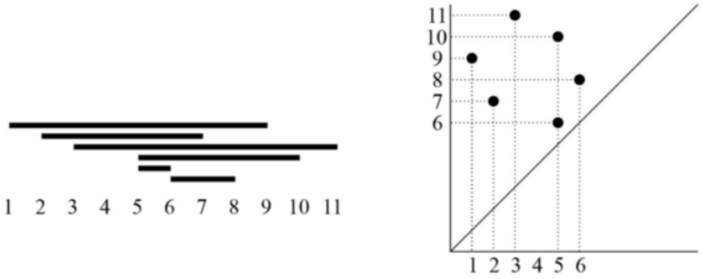

Figure 5. illustration of the birth and the death of a data through barcodes visualization

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 2.9. *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 (I, 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.

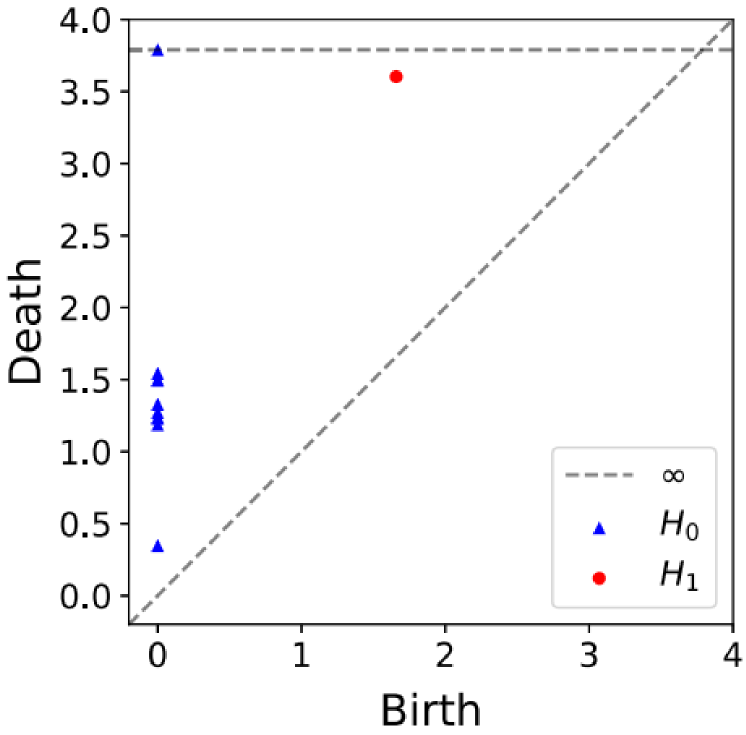

Figure 6. A persistent diagram as shared by software

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

Definition 2.10. *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 2.3. *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 theoretically the mapping between data and associated persistent diagrams is a well defined homeomorphism.

3. Topological data analysis of the protein

The most popular way Topological data analysis (TDA in short) 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 [19]. 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 behaviors such as flexibility and folding of proteins, plus the extremely expensive and complicated computation power needed, an interesting application is the protein binding analysis [20]. 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 have 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 structure prediction 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. For more enlightenment Let's consider the geometric representation shown in the figure below, This is the best of what how the visualization of a protein can be made, clearly a lot of complicated patterns even with the heavy costly computing demanded power, which also mean no real geometric representation for at least a well folded protein, we conclude since a protein shape dictates its function, one can tell from the figure why the paradigm of geometry.function is not a practical way to be adopted for a learning process purposes, being said rises the question of how the transition to the theoretically confirmed topology.function paradigm can take place so naturally we can be sure from the learning process.

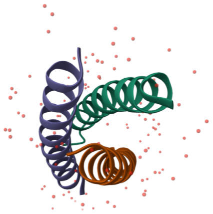

Figure 7

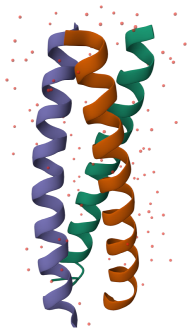

Figure 8

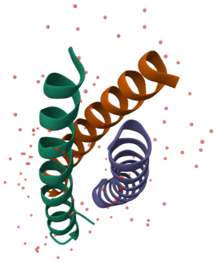

Figure 9

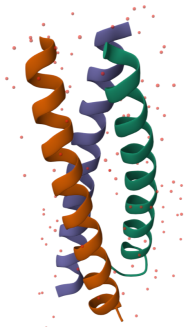

Figure 10

Figure 11. Representation of a folded protein with 1IJ3 ID

3.0.1. *Topological descriptors of an already defined beta sheet.* This section is devoted to an application part : the protocol is statistical inference for observations that are barcodes. 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 contained in $\mathbb{R}^{3.700}$

Let's illustrate the visualization part by a simple example, our point cloud is the set of atoms lying in a 3 dimensional space downloaded from a pdb file with 1COS ID, 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 descriptors (a small data set) so their is no need to use a landmark selector.

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

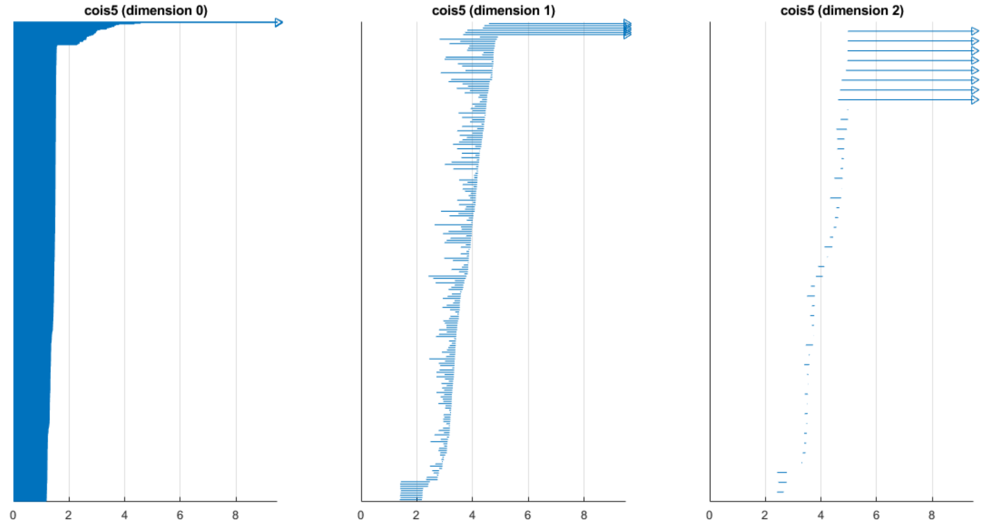

Figure 12. alpha helix related topological descriptors

As we can decipher from the resulted barcodes, 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. The two first betti numbers are describing the loop, hole and void type of structures, clearly with less clarity because of the high level of high level of the topological information. To detect more topological details of the alpha carbon point cloud used from the Alpha helix identity 1COS, we utilize the AC 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 an 18 points in a 3-dimensional space.

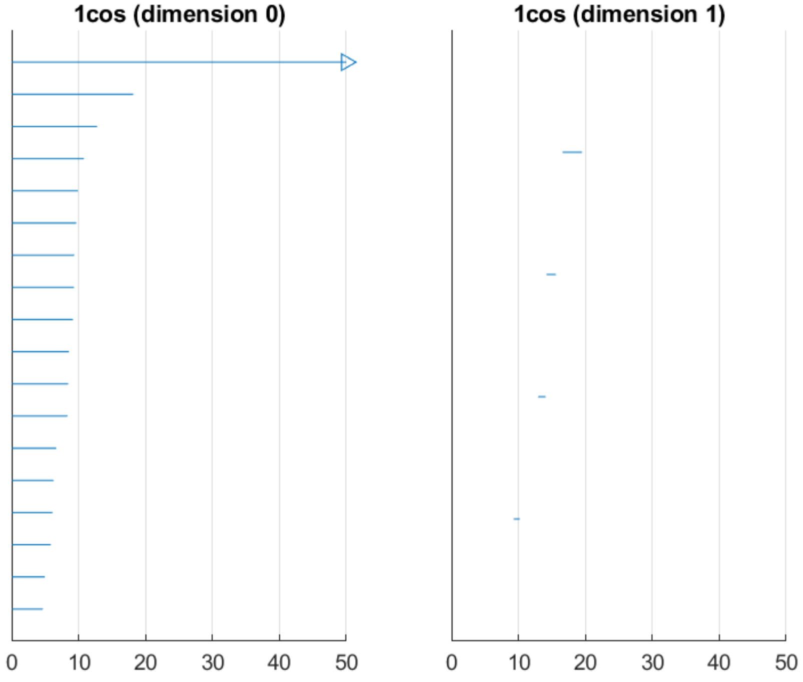

Figure 13. descriptors of alpha carbon distribution of an alpha helix generated from a protein of pdb ID 1COS

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

As it is naturally described, the filtration begins which makes around each C_α atom a topological 3 dimentional sphere, spheres start to overlap which form one dimensional loop, this is exactly our one dimensional simplex, then the first betti number is calculated and depicted in the barcode as a four little bars from 0.2 to 1.5 Angstrom.

As mentioned in the literature. We assume that the four levels are well defining the full structure of the main building component (protein).

3.0.2. *Parameters used to generate a suitable filtration.* different probabilistic methods and tools are used to simulate molecular behaviour emerging from atomistic level, we mention that no theoritical frame is giving, which also means only with a learning process we can find and interpret results, a clear description can be found in [21].

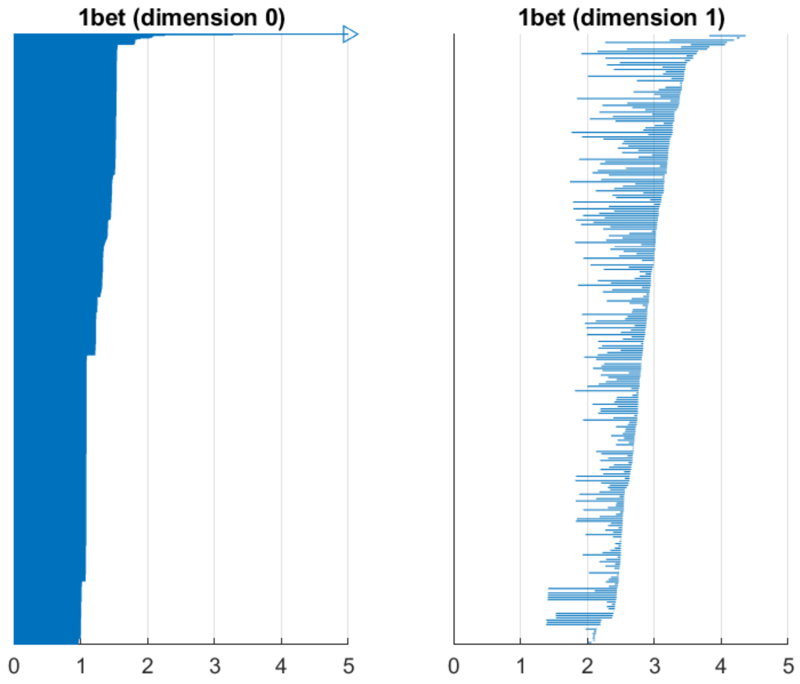

Figure 14. beta sheet all atom point cloud related topological descriptors

3.0.3. *Description of the folding process.* Proteins possess an intrinsic flexibility that allows them to function through molecular interactions within the cell, among cells and even between organisms. Many powerful models have been proposed such as the molecular non-linear dynamic (MND) and flexibility-rigidity index (FRI) to analyze protein main propreties such as folding and flexibility, the fundamental assumption of these methods is only through Physical laws and immediately following the mathematical description which is a truly demanding and complicated way one can achieve a model to be computed, being said, we have through the accumulated results tremendous data to be simplified into creative statistical tools, we will be using the main models MND and FRI to confirm the latest view point, which immediately means providing 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 [7]. One of the techniques that are utilized in the 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

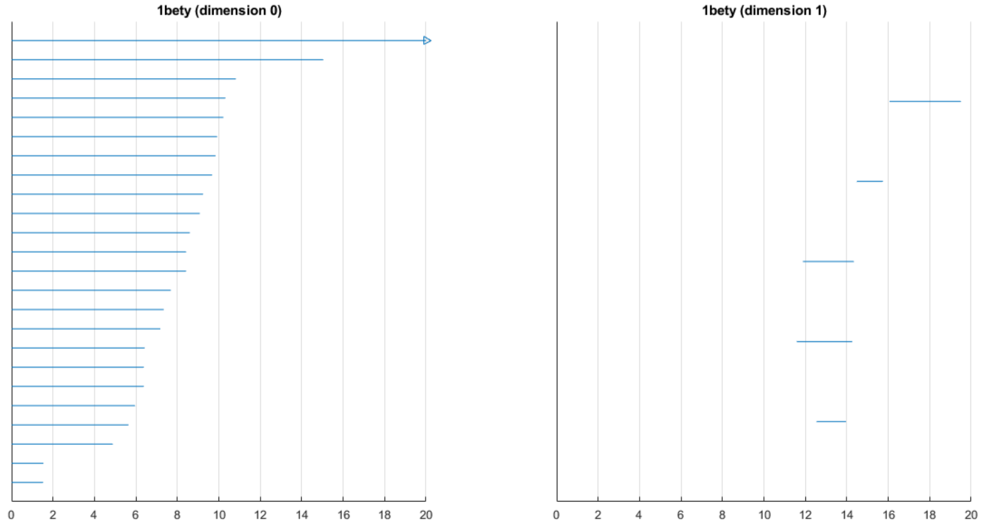

Figure 15. beta sheet alpha carbon point cloud related topological descriptors

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.

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

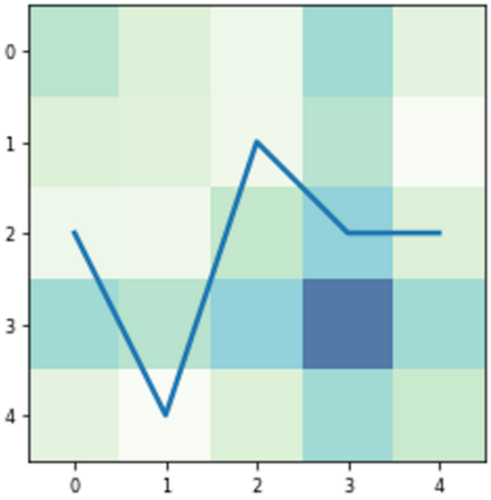

Figure 16. Behaviour of the folding process through filtration

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. As shown in equations. $\Phi(r, \eta) = e^{-(r/\eta)^k}$ and $\Phi(r, \eta) = \frac{1}{1+(r/\eta)^{\nu}}$. 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 $K = 2$. We slightly vary the filtration parameter of the AC point cloud for the Alpha helix from the 1COS identity, the formation of simplicial complex or topological connectivity changes too.

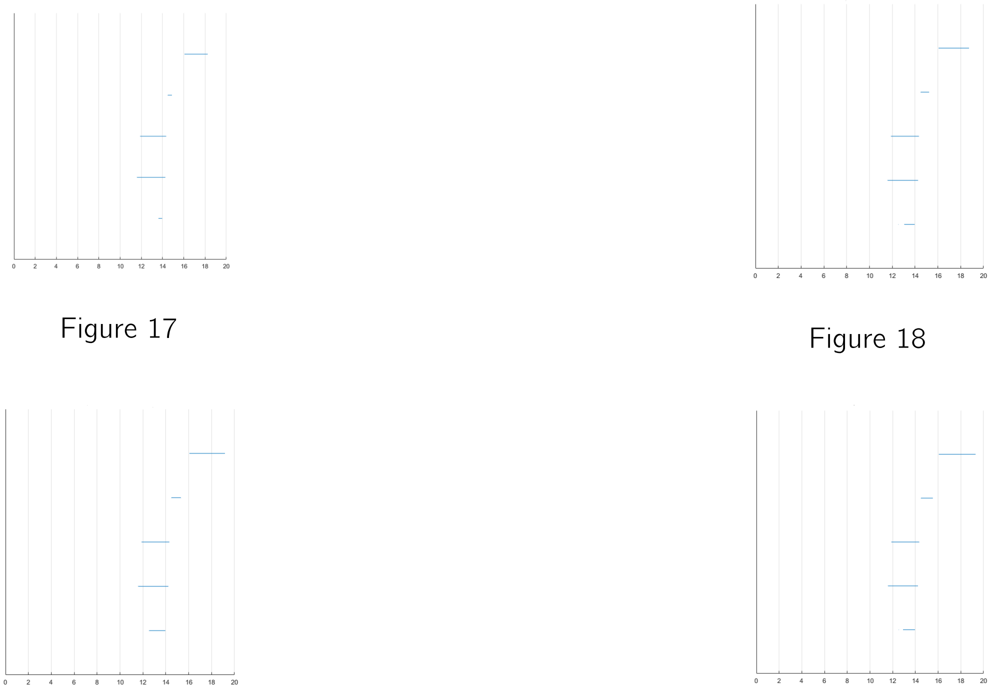

Figure 17

Figure 18

Figure 19

Figure 20

Figure 21. Connectivity patterns of the alpha carbon point cloud

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, but couldnt theoretically justify barcodes as a statistical observation, instead it gives birth to a new functional tool wich is persistent landscapes.

Conflicts of Interest: The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] 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.

[2] V. Kovacev-Nikolic, P. Bubenik, D. Nikolić, G. Heo, Using Persistent Homology and Dynamical Distances to Analyze Protein Binding, Stat. Appl. Genet. Mol. Biol. 15 (2016), 19–38. https://doi.org/10.1515/sagmb-2015-0057.

[3] 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.

[4] 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.

[5] Z. Cang, G. Wei, Analysis and Prediction of Protein Folding Energy Changes Upon Mutation by Element Specific Persistent Homology, Bioinformatics. 33 (2017), 3549–3557. https://doi.org/10.1093/bioinformatics/btx460.

[6] Z. Cang, L. Mu, G.W. Wei, Representability of Algebraic Topology for Biomolecules in Machine Learning Based Scoring and Virtual Screening, PLoS Comput. Biol. 14 (2018), e1005929. https://doi.org/10.1371/journal.pcbi.1005929.

[7] M. Buchet, F. Chazal, S.Y. Oudot, D.R. Sheehy, Efficient and Robust Persistent Homology for Measures, Comput. Geom. 58 (2016), 70–96. https://doi.org/10.1016/j.comgeo.2016.07.001.

[8] G. Carlsson, Topology and Data, Bull. Amer. Math. Soc. 46 (2009), 255–308. https://doi.org/10.1090/s0273-0979-09-01249-x.

[9] H. Edelsbrunner, D. Morozov, Persistent Homology: Theory and Practice, in: R. Latała, A. Ruciński, P. Strzelecki, J. Świątkowski, D. Wrzosek, P. Zakrzewski (Eds.), European Congress of Mathematics Kraków, 2 - 7 July, 2012, European Mathematical Society Publishing House, Zuerich, Switzerland, 2013: pp. 31-50. https://doi.org/10.4171/120-1/3.

[10] 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.

[11] Z. Hu, J.H. Hung, Y. Wang, Y.C. Chang, C.L. Huang, M. Huyck, C. DeLisi, 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.

[12] 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.

[13] M.S. Lee, Q.C. Ji, Protein Analysis Using Mass Spectrometry: Accelerating Protein Biotherapeutics From Lab to Patient, Wiley, Hoboken, 2017.

[14] J. Liu, K.L. Xia, J. Wu, S.S.T. Yau, G.W. Wei, Biomolecular Topology: Modelling and Analysis, Acta. Math. Sin.-English Ser. 38 (2022), 1901–1938. https://doi.org/10.1007/s10114-022-2326-5.

[15] K. Opron, K. Xia, Z. Burton, G. Wei, 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.

[16] V.V. Prasolov, Elements of Homology Theory, American Mathematical Society, Providence, R.I, 2007.

[17] 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.

[18] 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.

[19] 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.

[20] A. Zomorodian, G. Carlsson, Computing Persistent Homology, Discr. Comput. Geom. 33 (2004), 249–274. https://doi.org/10.1007/s00454-004-1146-y.

[21] R. Jing, Y. Wang, Y. Wu, Y. Hua, X. Dai, M. Li, A Research of Predicting the B-factor Base on the Protein Sequence, J. Theor. Comput. Sci. 1 (2014), 111. https://doi.org/10.4172/2376-130x.1000111.