Persistent homology analysis of protein structure

Adam Pratt

University of Illinois at Chicago

April 28, 2023

Introduction

- Proteins are one of the most fundamental types of macromolecules in biological systems
- Understanding the geometric structure of these macromolecules is an important problem in biophysics and molecular biology
- Mathematical models for protein structure frequently become very computationally intense
- Topological data analysis (TDA) gives us a convenient way to study protein structure from a computational perspective
- In [9], the authors consider individual atoms (or individual amino acids) as points in three-dimensional space, and attempt to use topological methods to understand the shape of proteins

Topology

- Topology is the mathematical study of certain “large-scale” geometric phenomena
- Classical geometric notions such as distance and angle are not considered in topology
- Instead, one considers geometric features that are invariant under continuous deformations of shapes, such as connectivity or the number of n-dimensional holes in a space

Topology, continued

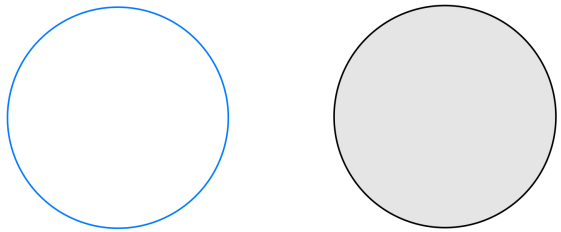


Figure. The circle on the left, denoted S^1, is topologically distinct from the disk on the right, denoted D^2, as the former has exactly one 1-dimensional hole, whereas the latter has no holes of any dimension.

- The main objects of study in topology are *topological spaces*: informally, sets with some abstract notion of "closeness" for points in the set
- Topological spaces are related to one another by *continuous maps*, an abstraction of the notion of a continuous map familiar from introductory calculus courses

Algebraic topology

- Computers are frequently unable to handle the abstract and "continuous" nature of topological problems
- *Algebraic* topology seeks to remedy this problem by associating linear algebraic objects, such as vector spaces, to topological spaces
- These algebraic objects allow us to retain some of the geometric information of the original space while still having access to the computationally-convenient tools of linear algebra (e.g. matrices)
- For example, the analysis on the previous slide can be made rigorous using an algebraic invariant of spaces known as *homology*

Simplicial complexes

One way in which we can simplify the notion of a topological space for the purposes of computation is to consider the less general notion of a *simplicial complex*:

Definition

A simplicial complex is a set of points in $\{v_0, \ldots, v_m\}$ in $\mathbb{R}^n$, called *vertices*, together with a collection of subsets of these points called the *k-simplices*, subject to some geometric constraints that allow us to think of these complexes as higher-dimensional generalizations of polyhedra

- 0-simplices are just vertices/points; 1-simplices are line segments that connect two vertices, 2-simplices are triangles defined by three vertices, 3-simplices are tetrahedra defined by four vertices, and so on

Simplicial complexes, continued

- The geometric constraints on these subsets guarantee that any two k-simplices meet in a k'-simplex, for $k' < k$
- For example, any two triangles in a simplicial complex must meet along a line segment in the complex or a single shared vertex; we can't have two triangles making contact along anything other than a shared edge/vertex
- Additionally, subsets of simplices must also be simplices, meaning all of the various faces, edges, vertices, etc. that bound a simplex must be included in the complex

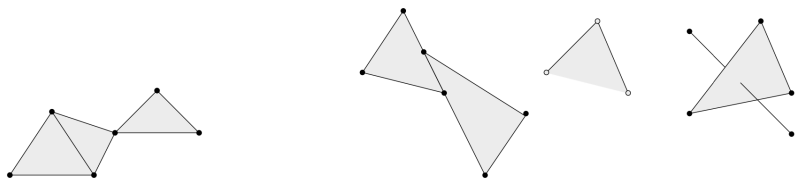


Figure. On the left, an example of a simplicial complex; The three shapes on the right are not examples of simplicial complexes, as the do not meet along simplices or do not include all of their subsimplices

Simplicial homology

- Every k-simplex can be sent to a linear combination of $(k-1)$-simplices that form its *boundary*; for example, a solid triangle would be sent to the sum of the three line segments that comprise its faces
- In topology, *homology* is the study of the n-dimensional holes of a space
- A 1-dimensional hole in topology is essentially what we think of when we think of a hole; for example, the hole in the middle of a circle
- 0-dimensional holes can be thought of as different connected pieces of a space; that is, there is a 0-dimensional hole between two components of a space if they are not connected by a path within the space
- 2-dimensional holes can be visualized as ”voids”, such as the empty space bound by the surface of a sphere

Simplicial homology, continued

- A *k-chain* in a simplicial complex is a linear combination of k-simplices; the set of k-chains in a complex is denoted C_k
- A *k-cycle* in a simplicial complex is a k-chain whose boundary is empty; the set of all k-cycles is denoted Z_k
- The subset of the k-cycles which are the boundary of a $(k+1)$-cycle are called the *k-boundaries*, denoted B_k
- A k-dimensional hole is then given by a k-cycle which is not a k-boundary

Simplicial homology, continued

- Algebraically, we can associate vector space structures to each of the sets described on the previous slide, and taking boundaries of k-chains gives a linear map $C_k \to C_{k-1}$
- The *k-dimensional homology* of a space X, denoted $H_k(X)$, is the vector space of k-cycles which are not k-boundaries
- For our purposes it will suffice to describe the dimension of this vector space, the *kth Betti number* of the space, denoted β_k
- By the Rank-Nullity Theorem, we have

$$\beta_k = \dim Z_k - \dim B_k$$

- The kth Betti number of the space corresponds exactly to the number of k-dimensional holes in the space

Persistent homology

- For the purposes of TDA, we need to associate simplicial complexes to a set of points in $\mathbb{R}^n$
- Need to consider how ”coarsely” the complex describes the topology of the data set
- For any real number r, we can define a simplicial complex X_r, where k points are connected by a k-simplex if each of the points are contained in a ball of radius r around any one of the points

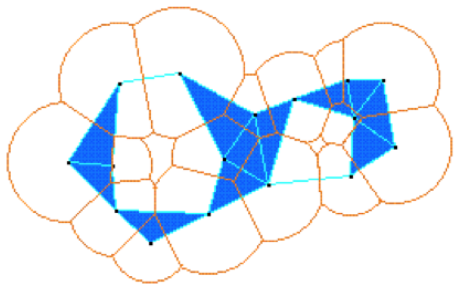


Figure. An example of a 2-dimensional simplicial complex built out of a point cloud in this manner.

Persistent homology, continued

- This gives us a family of simplicial complexes $\{X_r\}_{r\in\mathbb{R}}$ parameterized by the radius r
- As r grows, X_r gives a coarser description of the topology of the data set
- When $r \leq 0$, X_r is just the discrete set of points in the data set; as $r \to \infty$, X_r becomes increasingly connected, eventually becoming a single high-dimensional simplex
- We extract geometric information about the data set by studying the way the topology of X_r changes as the radius r grows

Persistent homology, continued

- Taking the homology of each complex in the family, we get a family of vector spaces $\{H_k(X_r)\}_{r\in\mathbb{R}}$, known as the *persistent homology* of the data set
- It is convenient to consider only the Betti numbers of each complex, giving us in each dimension k a family of natural numbers $\{\beta_{k,r}\}_{r\in\mathbb{R}}$
- The family of all of the Betti numbers of the data set for all k and r is known as the *persistence barcode* of the data set
- These persistence barcodes for various protein structures are the main object of study in [9]

The model

- The goal is to compute the persistence barcode of certain small amino acid chain shapes that occur frequently in proteins
- Data for these structures is taken from the protein databank (PDB)
- The authors consider two separate models: an all-atom model in which every atom in the molecule is represented as a single point in 3-space, and a “coarse-grained” (CG) model in which points correspond only to the C_α atom in each amino acid residue
- The CG model is significantly simpler, making it more amenable to in-depth topological analysis

Alpha helices

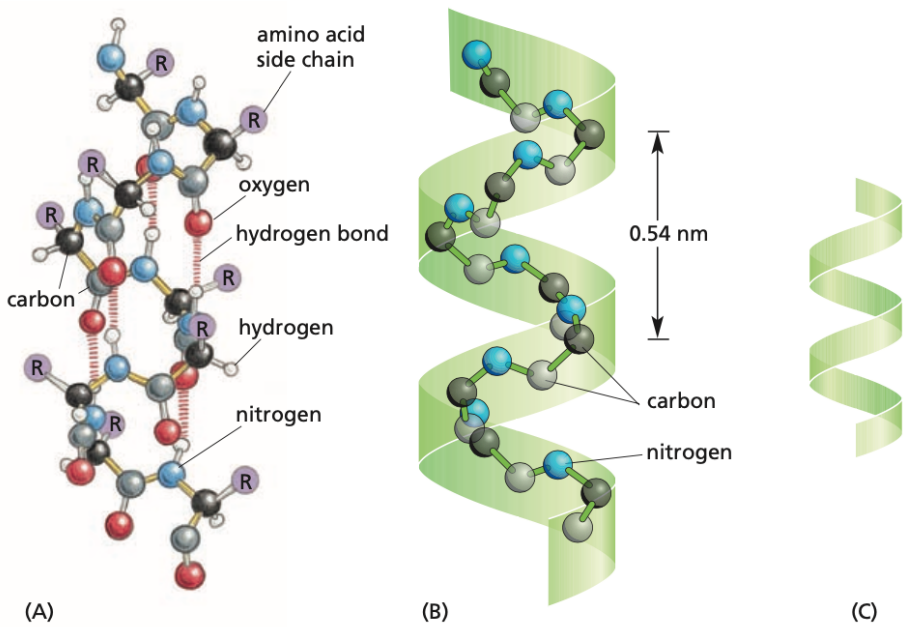


Figure. A ball and stick all-atom model, a hybrid model, and a ribbon diagram representation of an alpha helix.

- Alpha helices are one of the main substructures that occur throughout proteins
- Spirals (typically right-handed), stabilized by hydrogen bonds between N-H groups and the C=O group from the amino acid four residues earlier in the chain
- Approximately 3.6 amino acid residues per turn of the spiral

Topological fingerprint of alpha helices

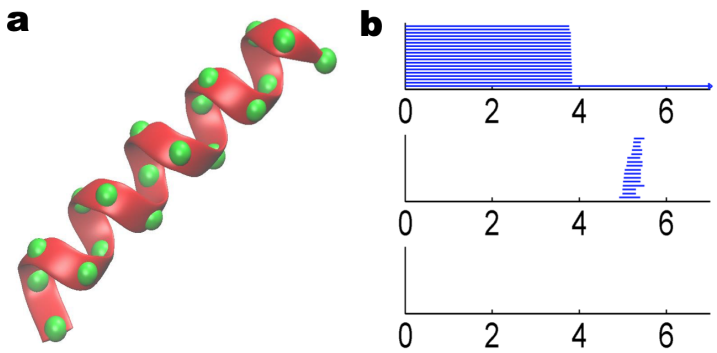


Figure. The CG model of the alpha helix in question on the left, along with its persistence barcode on the right.

- In the CG model of the alpha helix, there are 19 residues
- The β_0 corresponds to the distance between the residues, i.e. the bond length (around 3.8Å)
- There are 19 0-dimensional holes initially, representing the 19 different residues as distinct data points

Topological fingerprint of alpha helices, continued

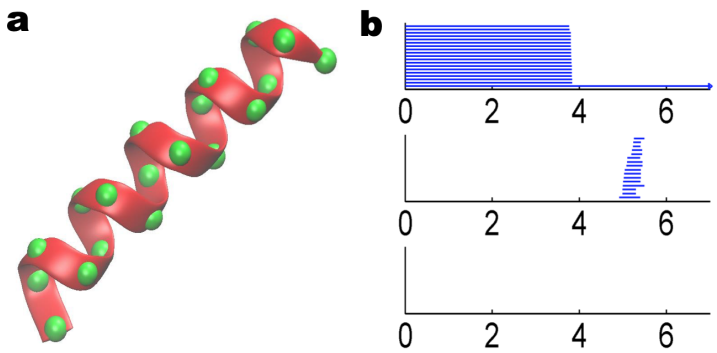


Figure. The CG model of the alpha helix in question on the left, along with its persistence barcode on the right.

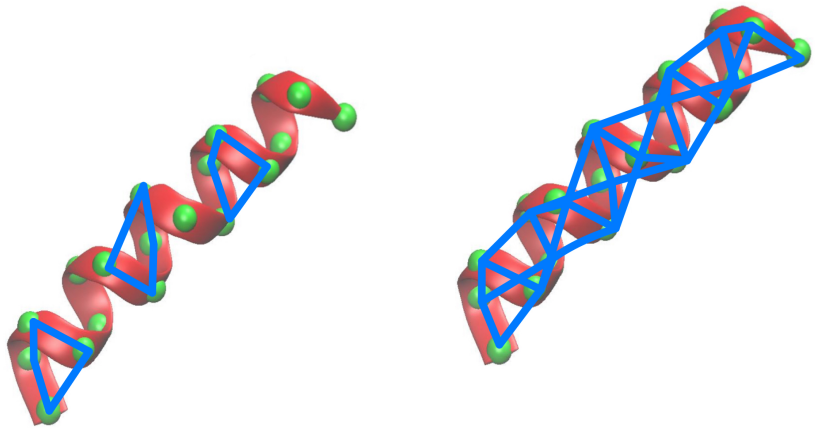


Figure. For a short range of radii, there are 16 1-dimensional holes, as when the radius is approximately 5Å, every chain of 4 consecutive residues forms a single loop (left). A few of the individual loops are shown for clarity on the right.

Topological fingerprint of alpha helices, continued

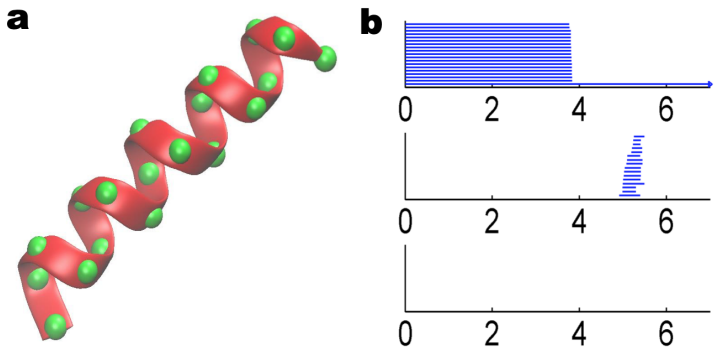


Figure. The CG model of the alpha helix in question on the left, along with its persistence barcode on the right.

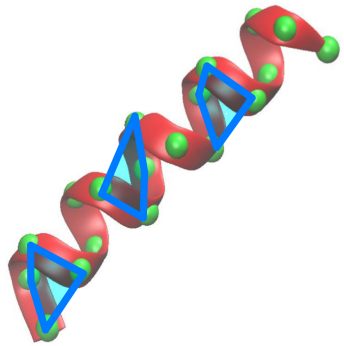


Figure. When the radius is over 5 Å, the β_1 bars vanish, as each of the 2-simplices shown in the figure on the right of the previous slide get filled in. β_2 bars never occur, as there is no radius at which a void forms.

Beta sheets

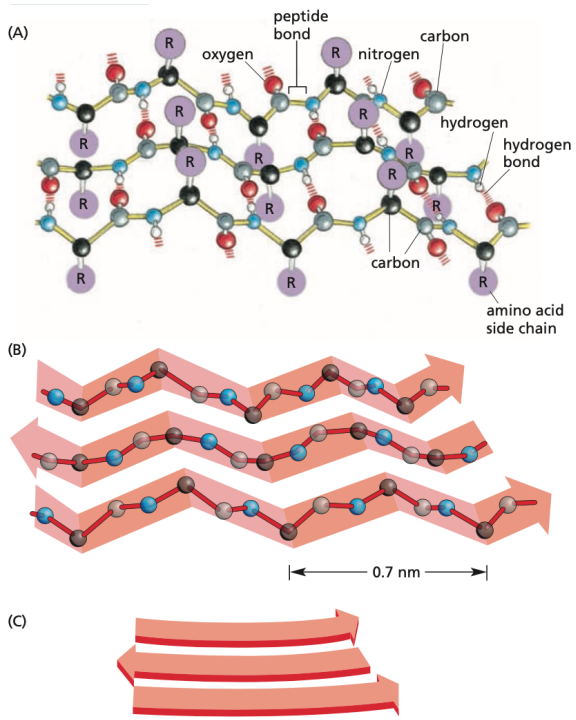


Figure. A ball and stick all-atom model, a hybrid model, and a ribbon diagram representation of an alpha helix.

- Beta sheets are another common substructure that occurs throughout proteins
- Amino acid chains in which parallel or antiparallel strands form hydrogen bonds between their backbones, forming a sheet
- Parallel and antiparallel sheets are slightly different in geometry due to the location of the hydrogen bonds

Topological fingerprint of beta sheets

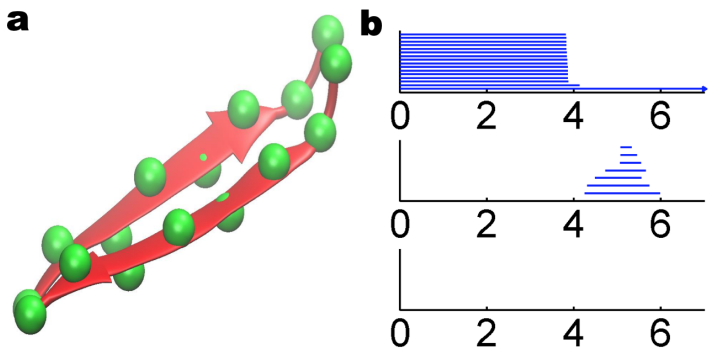


Figure. The CG model of the beta sheet in question on the left, along with its persistence barcode on the right.

- When the radius is zero, there is only a single 0-simplex for each amino acid residue
- In the beta sheet, there are 16 total residues, corresponding to the 16 β_0 bars
- One β_0 bar survives for all radii, corresponding to the eventual single connected component

Topological fingerprint of beta sheets, continued

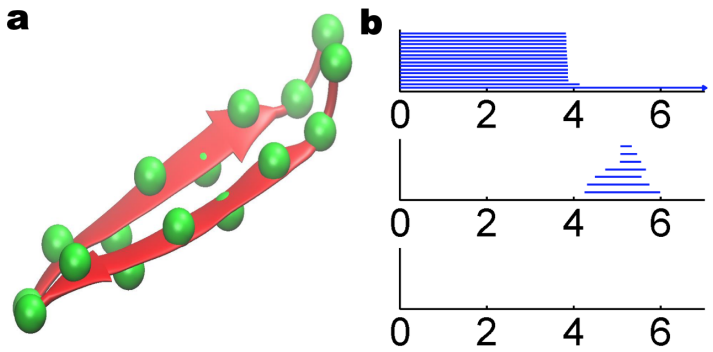


Figure. The CG model of the beta sheet in question on the left, along with its persistence barcode on the right.

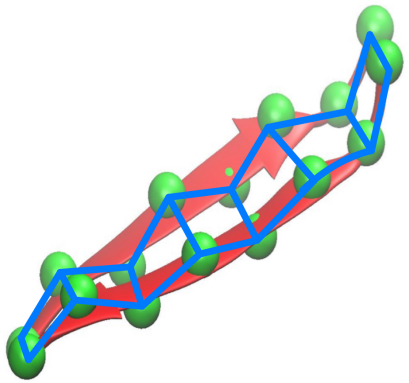


Figure. When the radius is between 4-6Å, the β_1 bars are the loops that are created when two parallel 1-simplices in the individual strands are connected, creating loops for each pair of pairs of adjacent residues on the respective strands, for a total of 7 loops.

Topological fingerprint of beta sheets, continued

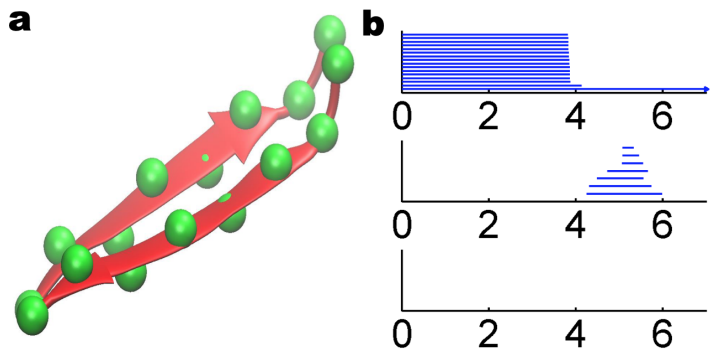


Figure. The CG model of the beta sheet in question on the left, along with its persistence barcode on the right.

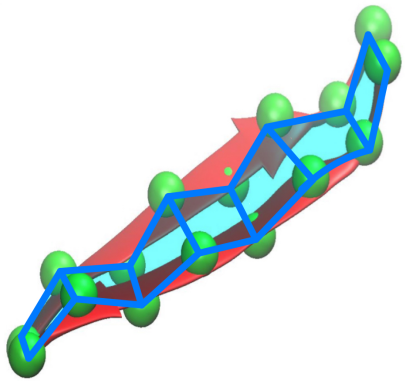
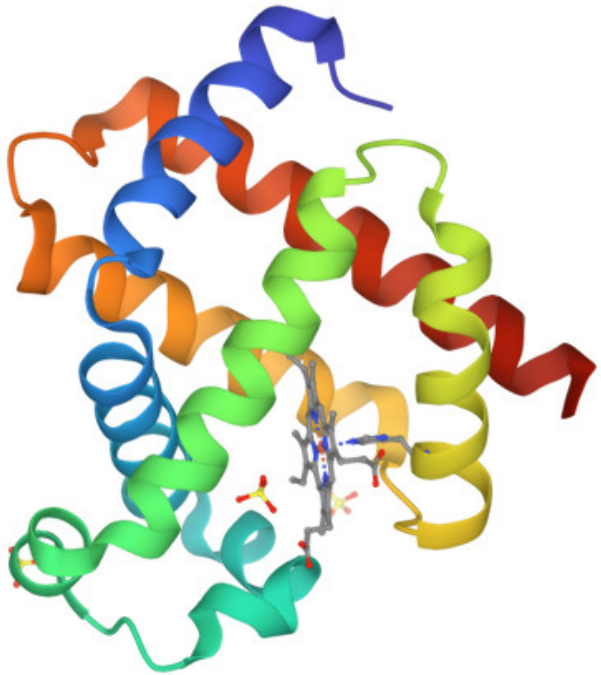


Figure. When the radius is greater than 6Å, the β_1 bars vanish, as each of the 2-simplices shown in the figure get filled in. β_2 bars never occur, as there is no radius at which a void forms.

Conclusion

- In [9], the authors use persistent homology to study the geometry of common shapes found within protein structures
- This work lays the foundation for further work studying protein structure computationally using persistent homology, considering simplified geometric models of complex proteins, allowing for faster, more accessible computation
- In the latter parts of the paper, the authors apply these models to study flexibility and folding of proteins
- The structures studied in the first part of the paper are quite simple, allowing for an in-depth analysis of the topology; however, this invariant can be computed for much larger protein structures as well
- The complexity of the data grows considerably with larger molecules as shown on the next slide

Conclusion, continued



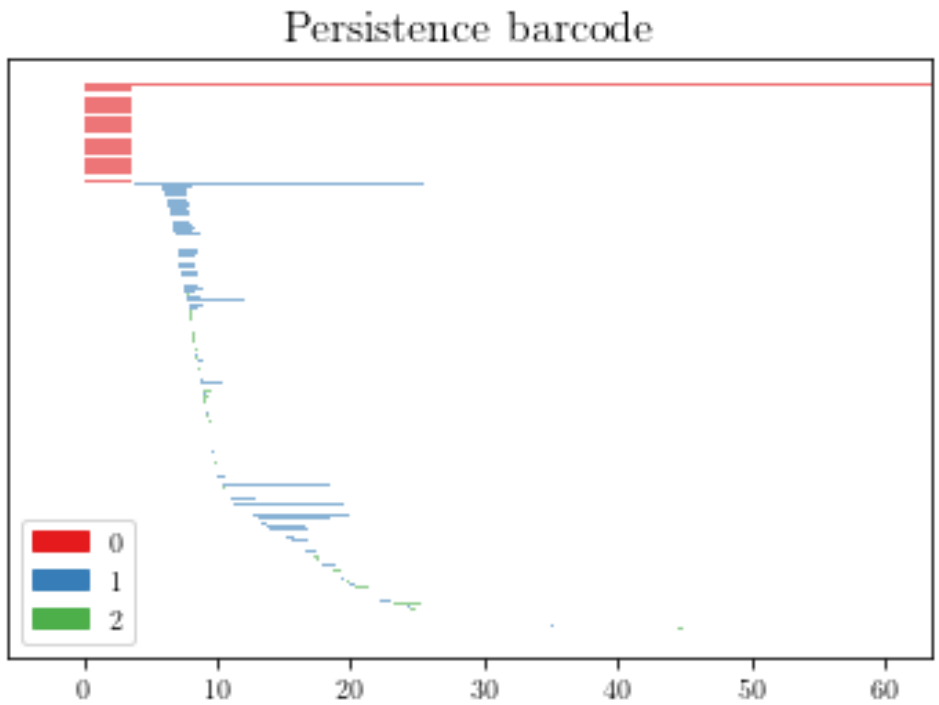


Figure. On the left, Crystal Structure of Human Myoglobin Mutant K45R (PDB ID: 3RGK); on the right, a persistence barcode of the CG model of the same protein (generated using Biopython, GUDHI, and Matplotlib).

Bibliography

[1] B. Alberts, K. Hopkin, A. Johnson, D. Morgan, M. Raff, K. Roberts, and P. Walter. Essential Cell Biology. W. W. Norton & Company, Fifth edition, 2019.

[2] P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11):1422–1423, 2009.

[3] J. Fraleigh. A First Course in Abstract Algebra. Pearson, 7th edition, 2002.

[4] A. Hatcher. Algebraic Topology. Cambridge University Press, 1st edition, 2001.

[5] S. R. Hubbard. Crystal Structure of Human Myoglobin Mutant K45R. https://doi.org/10.2210/pdb3RGK/pdb, 2011.

[6] S. R. Hubbard, S. G. Lambright, S. G. Boxer, and W. A. Hendrickson. X-ray crystal structure of a recombinant human myoglobin mutant at 2.8 A resolution. J Mol Biol, 20:215–218, 1990.

[7] J. D. Hunter. Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3):90–95, 2007.

[8] The GUDHI Project. GUDHI User and Reference Manual. GUDHI Editorial Board, 2015.

[9] K. Xia and G. Wei-Wei. Persistent homology analysis of protein structure, flexibility, and folding. Int J Numer Method Biomed Eng, 30(8):814–44, August 2014.

[10] A. J. Zomorodian. Topology for Computing. Number 16 in Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2009.