Algebraic Topological Applications in Data Modeling and Analysis

Authors: Rachel Laing, Jenny Lu, Elijah Stroud

Mentor: Dr. David Snyder

Summer 2017

Abstract

With increases in volume and complexity of individual data sets, it is becoming difficult to perform analysis using numerically-based methods. Algebraic topology provides the useful tool of persistent homology to analyze large, complex data sets. Persistent homology tracks the homology groups of simplicial complexes throughout a filtration process. Filtrations are a sequence of simplicial complexes such that each successive complex wholly contains each complex before it. By examining which homology groups in the complexes have the longest lifespans, underlying structures in the data can be exposed that traditional methods of data analysis would be blind to. Through the application of persistent homology to the structure of myoglobin 101mA, it is hoped that the efficacy of persistent homology in protein analysis can be demonstrated and new applications in biological and medical research are implemented. We are investigating topological differentiation within the myoglobin cluster's molecular surfaces and the extent to which it is significant.

1 Introduction

Big data is omnipresent in today's day and age. The quantity of available data (protein and DNA databases, financial information, geographic information systems, economic metrics, etc.) has proliferated due to improved means of data collection (sensors, text digitization, etc.). The ability to leverage these data collections using numerically-based methods has become strained by the size of these data collections, despite increases in processing and storage interaction speeds. The sheer scope and volume of data is, thus, a double-edged sword, as with greater size also comes greater complexity, and with greater complexity, far greater difficulty in the interpretation and determination of useful applications of the data. As a result, developing efficient methods of working with this data while still maintaining accuracy and reliability is critical.

A recent field of inquiry has arisen which integrates combinatorial topology with data analysis. Topology examines preserved properties of spatial objects under deformations, twistings, and stretchings. The method of topological data analysis abstractly models portions of data sets taken in their point cloud formation using geometric objects called *simplices*. Our research project topologically models a large data set (or an interesting subset of the data) as a collection of simplices (a *simplicial complex*) and then applies concepts and techniques from algebraic topology to analyze the resulting simplicial complex.

The field of algebraic topology defines a concept of simplicial homology, which algebraically computes what corresponds to "holes" in the structure (holes can have different dimensions). The algebra for computing the homology of a simplicial complex is based on algebraically combining subcomplexes of the complex so that any two such combined subsets bound a higher dimensional subcomplex will be considered equivalent. A collection of simplices is referred to as a simplicial complex (Definition 23 states formal definition). Creating simplicial complexes from a given set of data points allows for the homology of a data set to be computed, so one can identify and quantify connections, or lack thereof, within the data. Such features of the data can be identified by using a filtration, which is an increasing

sequence of topological spaces, each contained in the next. Building simplicial complex K from subcomplexes along filtration F cause various intermediate features to appear and disappear [1]. Persistent homology compares the homology groups of a simplicial complex at each level of a filtration. By tracking consistent subcomplexes throughout increasing dimensions, we can exhaust the data set for persistent homology. This allows for us to distinguish meaningful patterns in the data from temporary outlines.

We have selected a large data set for analysis and will use existing software, including Matlab and Mathematica, and existing software packages, to implement persistent homology techniques when analyzing our large data set.

2 Technical Background

In each of the following, n is an integer greater than or equal to 1.

Definition 1 A set of $n+1$ distinct points is called an *abstract n-simplex*.

Definition 2 The *convex hull* of a set of points S in n dimensions is the intersection of all convex sets containing S. For N points $p_1, p_2, ..., p_N$, the convex hull C is then given by the expression [2]

$$C = \left\{ \sum_{j=1}^{N} \lambda_j p_j : \lambda_j \geq 0 \text{ for all j and } \sum_{j=1}^{N} \lambda_j = 1 \right\}.$$

Definition 3 A set of $k+1$ points, $u_0, u_1, ..., u_k$, is *affinely independent* if the k vectors, $u_1 - u_0, u_2 - u_0, ..., u_k - u_0$, are linearly independent.

Intuitively, the set of k linearly independent vectors point to $k+1$ points, of which no three points lie on a common line, no four points lie in a common plane, *etc.*

Definition 4 *A k-simplex* is a generalization of a tetrahedral region of space to k dimensions, namely, a k-simplex is the convex hull of $k+1$ affinely independent points [3].

Definition 5 *A simplicial complex* K is a finite collection of simplices such that

1. for every simplex $\sigma \in K$, every face of σ is in K;
2. for every two simplices $\sigma,\tau \in K$, the intersection $\sigma \bigcap \tau$, is either empty or a face of both simplices.

Definition 6 The *underlying space* $|K|$ of the complex K is the union of simplices $|K| = \bigcup_{\sigma \epsilon K} \sigma$

Definition 7 A *subcomplex* is a subset of simplices that is itself a simplicial complex.

Definition 8 *Homology* is an extension of topological methodologies that uses linear algebra to to detect "holes" of all dimensions in a simplicial complex [1].

Definition 9 Two topological objects are said to be *homeomorphic* if they can be continuously deformed into each other.

Definition 10 A *homology group* is a collection of algebraic objects $H_d(K)$ that is associated to each simplicial complex K, where d ranges over the dimensions of simplices encountered in K [1].

Definition 11 The *n-th Betti number* is the rank of the n-th homology group of a topological space. In low dimensions, the 0, 1, 2-dimensional Betti numbers count the connected components, tunnels, and cavities of the underlying simplicial complex, respectively [1, 4].

Theorem 2.1 One can compute an unambiguous interval $[b_x, d_x)$ for each non-trivial homology class $[x]$ in $H_d(F_mK)$ using persistent homology techniques where the *birth* $b_x \leq m$ and *death* $d_x > m$. The following conditions hold true:

1. b_x is the smallest l with some homology class $[y]$ in $H_d(F_lK)$ with $[\phi_d^{l \to m}(y)] = [x]$;
2. d_x is the smallest n with $[\phi_d^{m \to n}(x)]$ the trivial homology class $[0]$ in $H_d(F_nK)$.

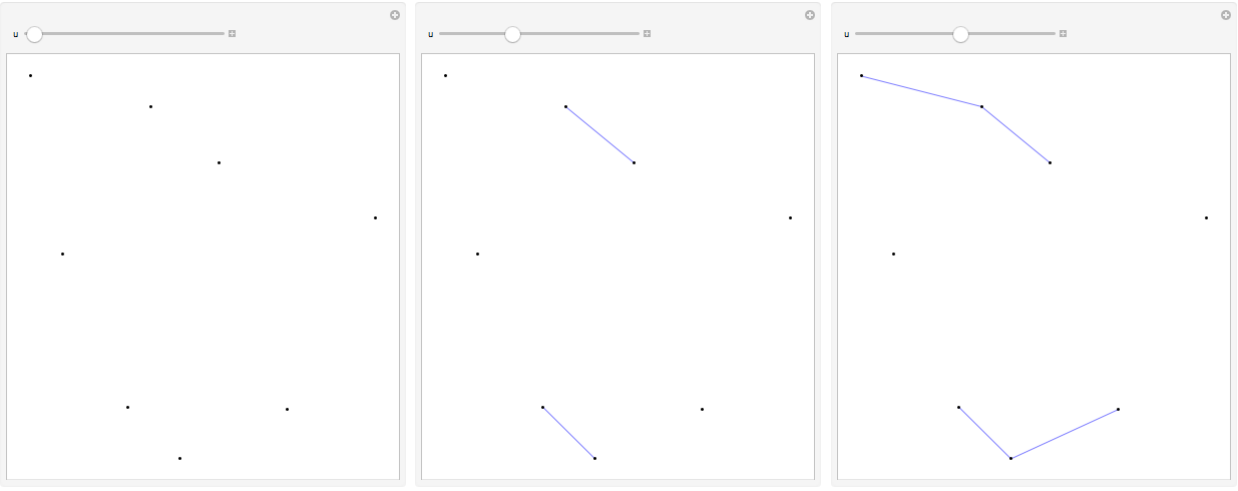
For each x, the length $(d_x - b_x)$ is the lifespan of the homology class $[x]$ across the filtration [1].

2.1 Filtrations

A filtration is a totally ordered set of the subcomplexes of a simplicial complex, indexed by non negative integers. Rather, a filtration is an increasing sequence of simplicial complexes in which simplexes are connected to each other if they are within a certain distance of one another. As dimension is increased, if one were to overlay all of the simplicial complexes of increasing dimension over each other, all the common components would no longer be visible. Thus, filtrations can be used to track the births and deaths of patterns, such as holes, in the data [5].

2.1.1 Vietoris Rips Filtration

A Vietoris Rips Filtration is a collection, or complex, of simplicial complexes that is defined as the vertex set of the data, and if two vertices a and b are within a specified maximum distance of each other, edge $[ab]$ is included in the complex. For higher dimensional simplexes, if all of its edges are already in the complex, then that higher dimensional simplex will be included as well [6].



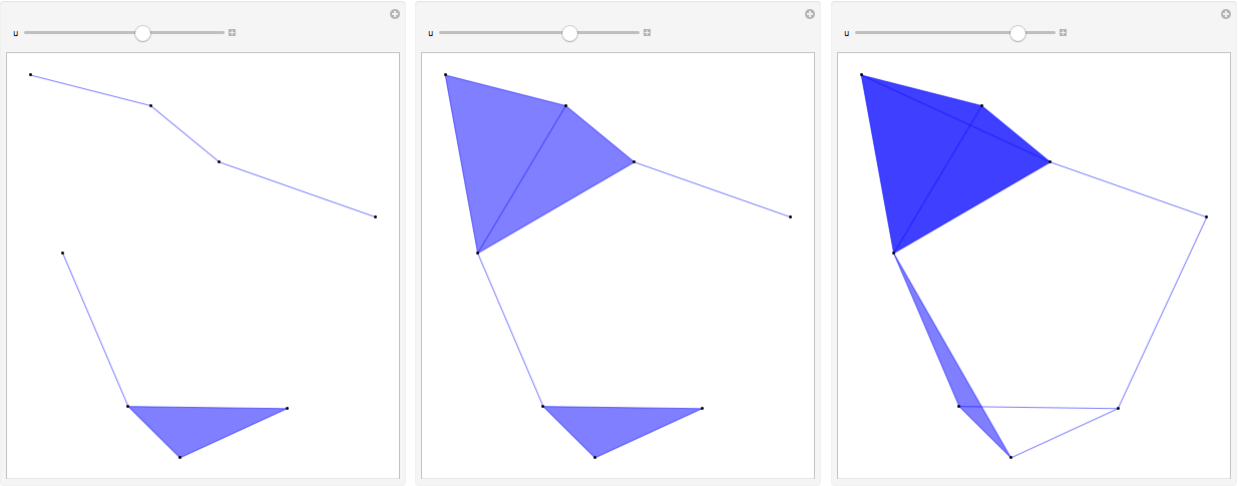


Figure 1: This series of images is a simple visual representation of a Vietoris Rips Filtration. The slider at in the top left corner of each image indicates the dimension of the filtration. As the slider approaches the right, vertices are connected by an edge when they are within a certain distance of each other. Eventually, holes are shaded in as the filtration value is increased.

2.1.2 Witness Filtration

If every point of a large dataset were used for a Vietoris Rips Filtration, the computation process would become incredibly inefficient. A Witness Filtration addresses the issue of computational power by using landmark points, selected from the dataset, as the vertices for the simplicial complexes. Landmark points can either be selected randomly, or selected using an inductive process called sequential maxmin. In the process of sequential maxmin, the first landmark point is chosen randomly, and all the landmarks that follow are chosen to be the farthest away from the rest of the points in order to create a subset that is much more spread out across and representative of the original data [7].

2.1.3 Weak Witness Filtration

A Weak Witness Filtration, or Lazy Witness Filtration, is very similar to a Witness Filtration, but it has an extra parameter on the witness points that are not selected to be in the landmark subset.

2.2 Prior Applications

Researchers in a wide range of fields have applied topological models to their data in order to attain a more comprehensive level of results. The applicability of such methods suggests the underlying mathematical foundations of the other sciences.

2.2.1 Computational Biology

Persistent homology was an integral part of the discovery of a new class of breast cancer, c-MYB+. Using algebraic topology on breast cancer transcription data, Nicolau, Levine, and Carlsson discovered a new subtype of breast cancer that both expresses high levels of c-MYB and is estrogen receptor positive (ER+). This newly discovered subtype of breast cancer was shown to have a 100 survival rate [8].

2.2.2 Politics

Persistent homology has been used to analyze Brexit and its potential impact on the EU and Europe as a whole. The results suggested that Britain choosing to leave the EU would have strong consequences for Ireland and its involvement in the EU. A non-trivial network loop was discovered surrounding Switzerland. The paper also utilized persistent homology to find significant disparities between geographically close locations in their voting habits. It was discovered that London, the Cotswolds, Reading, and the region around Oxfordshire voted differently than their neighboring areas [9].

2.3 Molecules

2.3.1 Proteins

From a biological standpoint, proteins really have four levels of structure. The first level consists of the sequence of amino acids, which are fundamentally the building blocks of proteins. Amino acids typically have two functional groups (amine and carboxyl) and a

side chain that is specific to each amino acid. These side chains are often referred to as "R groups", which will become important when examining the tertiary structure. The second level of protein structure derives from various foldings in the amino acid chains. The two variations of secondary structure are the α-helix and the β-pleated sheet; the former involves hydrogen bonding within a polypeptide while the latter involves hydrogen bonding between polypeptides.

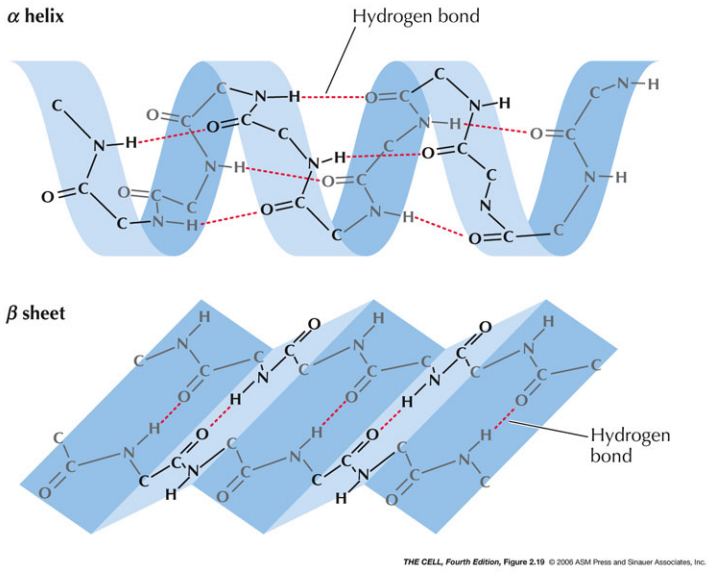


Figure 2: Hydrogen bonds contribute greatly to the secondary structure of proteins. [10]

Tertiary structure is developed from R groups interacting with each other, and quaternary structure occurs when multiple protein subunits come together. Understanding the different levels of protein structure is critical to contextualizing slight geometric variations that arise from flexibility. As proteins are dynamic structures, simplistic models for molecular binding are less accurate. [11]

2.3.2 Myoglobin

The protein analyzed here is sperm whale myoglobin from the cluster 101mA (5 clusters were randomly selected for analysis), which is classified as a low flexibility protein with a maximum Root Mean Square Deviation (RMSD) of 2.476, indicating that the various components of

Figure 3: 5 selected myoglobin clusters. From left to right: 1duoA.101mA, 4qauA.101mA, 105mA.101mA; 2mgmA.101mA, 1bzrA.101mA

the protein will not become extremely distant [12]. The function of the myoglobin protein is to store oxygen, specifically in the muscle cells. Aquatic mammals have evolved with more myoglobin as to maximize oxygen storage for extended periods of time underwater.

3 Methodology

3.1 Mathematica

The protein cluster files downloaded directly from the PDBFlex website were not compatible with the computations we planned on executing in Matlab, so we first imported the Protein Database files into Jmol. Then, we exported the STL 3D Model data for each of the selected clusters from Jmol. With that exported 3D Model data, we used Mathematica to extract the vertex data for the 3D models of the clusters.

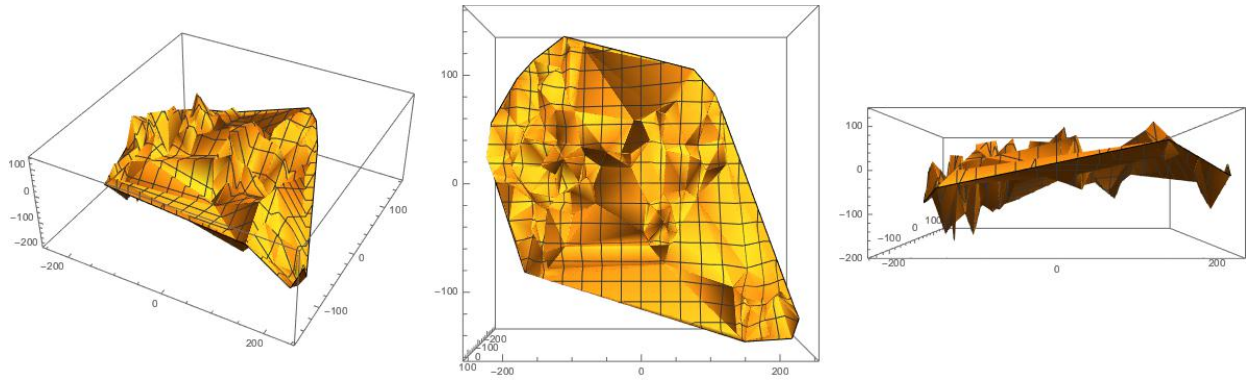


Figure 4: Angled, top, and front views of the 3D plot of the extracted vertex data for myoglobin cluster 5mbnA .101mA

3.2 JavaPlex

We are using the `JavaPlex` library [13] for Matlab in order to compute persistent homology of the data sets. Additionally, the powerful computational advantages of Matlab allow for the calculations of large distance matrices and construction of barcode images. The `JavaPlex` tutorial [6] outlines commands and processes that are necessary for topological data analysis.

3.2.1 Commands

The class `ExplicitSimplexStream` generates simplicial complexes from manual input through the following commands:

`stream = api.Plex4.createExplicitSimplexStream()` creates an empty stream with filtration values of 0.

`stream.addVertex` and `stream.addElement` construct the components in the complex.

`stream.getSize()` calculates the number of simplices in the entire complex.

`persistence.computeIntervals(stream)` computes the Betti number intervals (N.B. the `persistence` algorithm must be defined first).

Data from point clouds forms a finite metric space because the concept of distance between points holds. From here, we can increase the specificity of the metric space by constructing either a Euclidean metric space or an explicit metric space. The Euclidean metric space is based on numerical (x, y) coordinates while the explicit metric space is defined by pairwise distances between points.

For small data sets, coordinates can be entered manually to construct a Euclidean metric space `point_cloud = [`x_1, y_1`;` x_2, y_2`;` x_3, y_3`; ...` x_n, y_n`]` for some $n \in N$.

`metric.impl.EuclideanMetricSpace(point_cloud)` creates the metric space from the given coordinates, so we can then find distances between points and produce scatter plots.

Explicit metric spaces can be created from distance matrices and the `JavaPlex` command.

4 Results and Discussion

For each molecule, we constructed a witness stream or a lazy witness stream and computed the persistent homology of each such filtrated simplicial complex. The results are easiest to understand in the form of so-called barcodes. We explain this in more detail below.

4.1 Barcodes from Witness Streams

The x axis is the filtration value. The $Betti_k$ number of the stream at filtration value x is the number of intervals in the dimension k plot that intersect a vertical line through x [6].

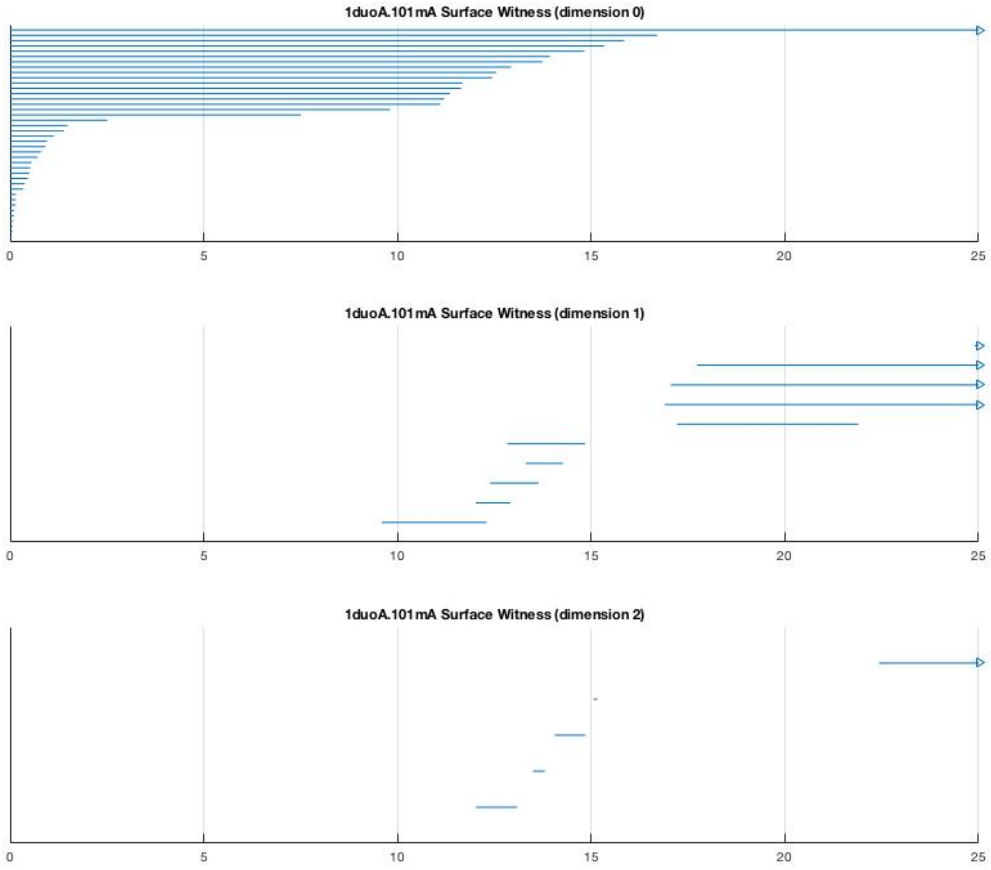


Figure 5: Witness stream barcodes for the 1duoA.101mA cluster.

In dimension 0, at filtration value 0, all of the lines in the barcode indicate the formation of connected components. By filtration value 17, all of the points in dimension 0 have been connected into 1 large clump, indicated by the single bar that persists. The dimension 1 barcode reveals that 5 1-dimensional holes are born and die between filtration values of 10 and 15. Another 1 dimensional hole appears at filtration value 18, and quickly dies by filtration value 22. 4 bars in the 1-dimensional graph continue past filtration value 25, suggesting that there are also larger 1-dimensional holes in the structure of the 1duoA.101mA cluster. In dimension 2, there are 4 short bars between filtration values of 10 and 15 that are part of the bars in the dimension 1 graph, indicating that there are 2 dimensional holes within the 1 dimensional holes in the cluster. Then, there is one bar in dimension 2 that persists from filtration value 22, past 25. This is a cavity in the structure of the cluster that lives almost forever. The obtained Betti numbers are $Betti_0 = 1, Betti_1 = 4, Betti_2 = 1$.

Out of the other 4 randomly selected clusters of myoglobin, 4qauA.101mA is the only cluster that also only has one eternal bar in dimension 0. This means that the vertex points of the 1duoA and 4qauA datasets are more close to each other than those of the other clusters, since they are connected into one component around filtration value 18. The 6 enduring bars in the dimension 1 barcode for 4qauA reveals that it has more 1-dimensional topological components than 1duoA. This difference indicates more potential topological disparities in higher dimensions for the 2 clusters. This is apparent in the dimension 2 barcode for 4qauA, as there are 14 total bars showing the birth and death of components, but none of them persist past filtration value 25, indicating that 4qauA does not have any topologically significant 2-dimensional holes, setting it topologically apart from 1duoA. Thus, the obtained Betti numbers are $Betti_0 = 1, Betti_1 = 6, Betti_2 = 0$.

The barcodes for the 105mA.101mA cluster show that in dimension 0, before filtration value 15, it is very similar to 1duoA.101mA and 4qauA.101mA. The differences in the flexibility of the clusters are more apparent in dimensions 1 and 2, as there are more bars of

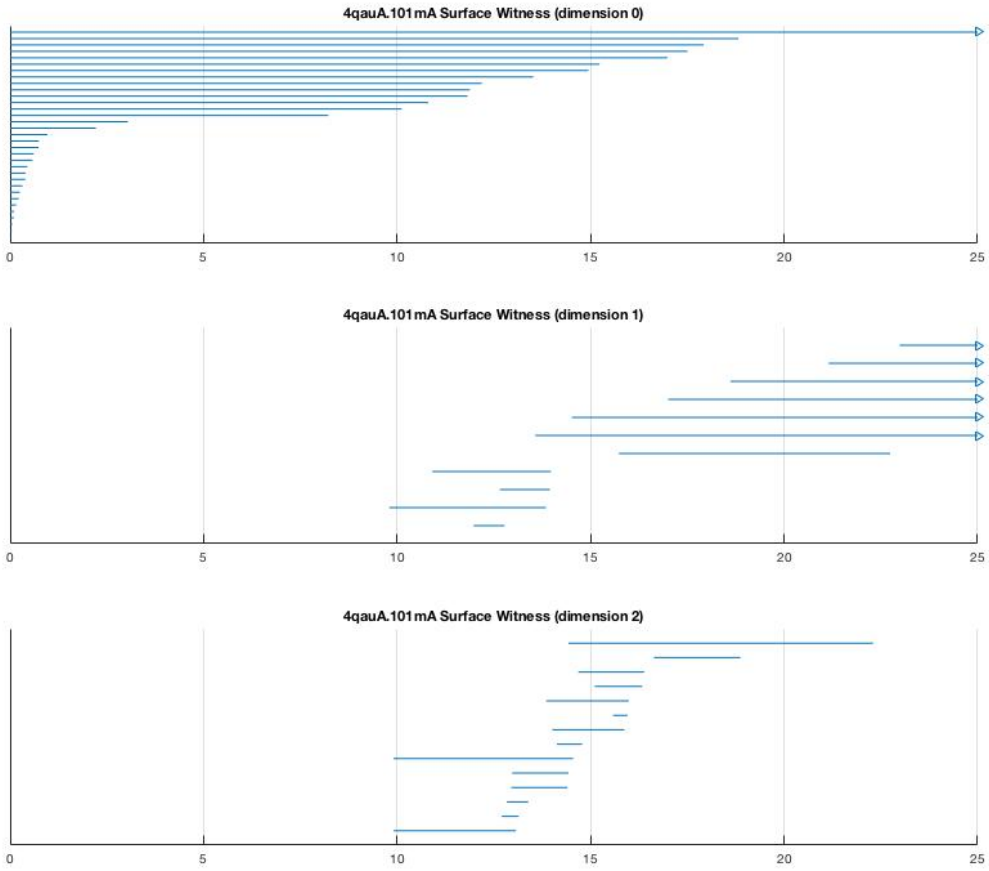


Figure 6: Witness stream barcodes for the 4qauA.101mA cluster.

varying length in the barcodes for 105mA than in those of 1duoA and 4qauA. In dimension 1 for 105mA, before filtration value 5, there is already the birth of a 1-dimensional hole, whereas for 1duoA and 4qauA, the first birth of a hole is not until right before filtration value 10. The 1-dimensional holes of 105mA persist for longer and are more abundant than those of 1duoA, but less abundant than those of 4qauA. 105mA also has 2 1-dimensional holes that are born and die within a filtration value of 1 between filtration values 20 and 25 that can be disregarded as noise since the features die so quickly. Recall that one of the ideas of persistent homology is that long intervals should correspond to real topological features, while short intervals are considered to be noise [6]. Thus, the most topologically

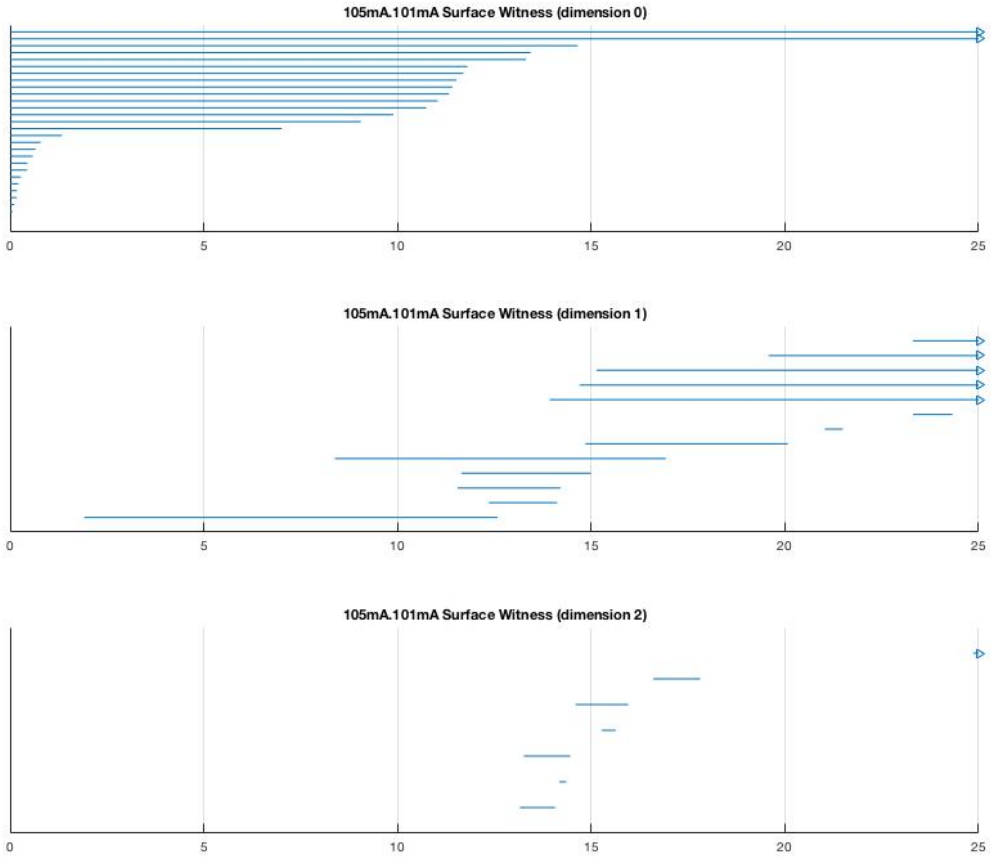


Figure 7: Witness stream barcodes for the 105mA.101mA cluster.

important aspect of the dimension 1 barcode for 105mA would be the 5 bars that persist past filtration value 25. In dimension 2, 1duoA and 105mA are topologically very similar since there is only one bar that persists past the set maximum filtration value of 25, and while the short bars are different for the 2 clusters, they all appear around filtration value 15, and the 2-dimensional holes indicated by the bars are all born and die so quickly that the short bars can be disregarded as noise. The only slight difference in dimension 2 for 1duoA and 105mA is that the shared 2-dimensional hole is born at a higher filtration value than it was for 1duoA. The obtained Betti numbers are $Betti_0 = 2, Betti_1 = 5, Betti_2 = 1$.

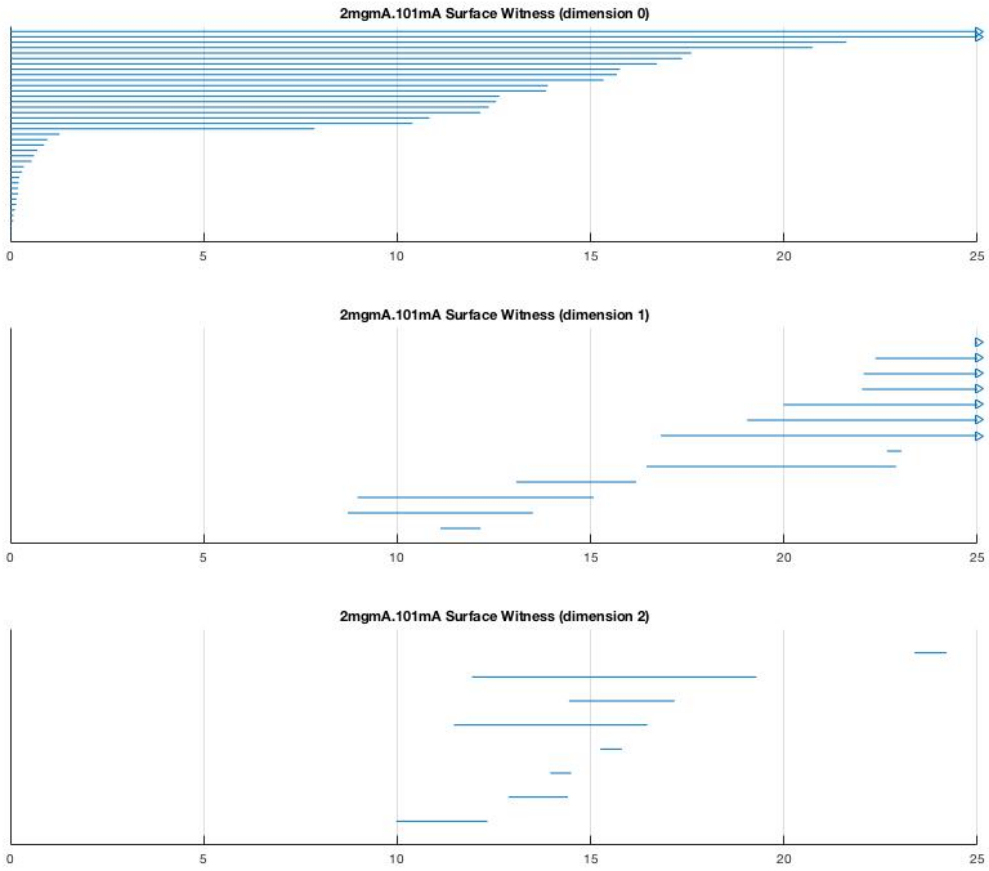


Figure 8: Witness stream barcodes for the 2mgmA.101mA cluster.

In dimension 0, the 2mgmA.101mA cluster is almost topologically equivalent to 105mA.101mA as it also has 2 bars that persist past filtration value 25. In other words, the points of the vertex data extracted using Mathematica are almost the same for the two clusters, which is to be expected since they are both clusters of Myoglobin. The only main notable difference is that among the 1duoA, 4qauA, 105mA, and 2mgmA barcodes, 2mgmA is the only cluster with more than 2 bars persisting past filtration value 20 in dimension 0. This minor difference in the vertices themselves within this cluster of myoglobin indicates more significant differences in the larger structure of the protein. This holds out to be true as it is apparent in the dimension 1 barcode that 2mgmA has 7 bars that persist past filtration value 25, which means

that 2mgmA has more topologically important 1-dimensional holes than 1duoA, 4qauA, and 105mA. In dimension 2, similar to 4qauA, the molecule 2mgmA does not have any bars that persist, indicating that this cluster does not have any topologically significant 2-dimensional cavities in its structure. While 2mgmA does have 8 bars of varying length occurring at varying filtration values in dimension 2, these are all cavities that are born and die, so they are not as significant in our understanding of the structure and flexibility of the 2mgmA cluster of myoglobin. The obtained Betti numbers are $Betti_0 = 2, Betti_1 = 7, Betti_2 = 0$.

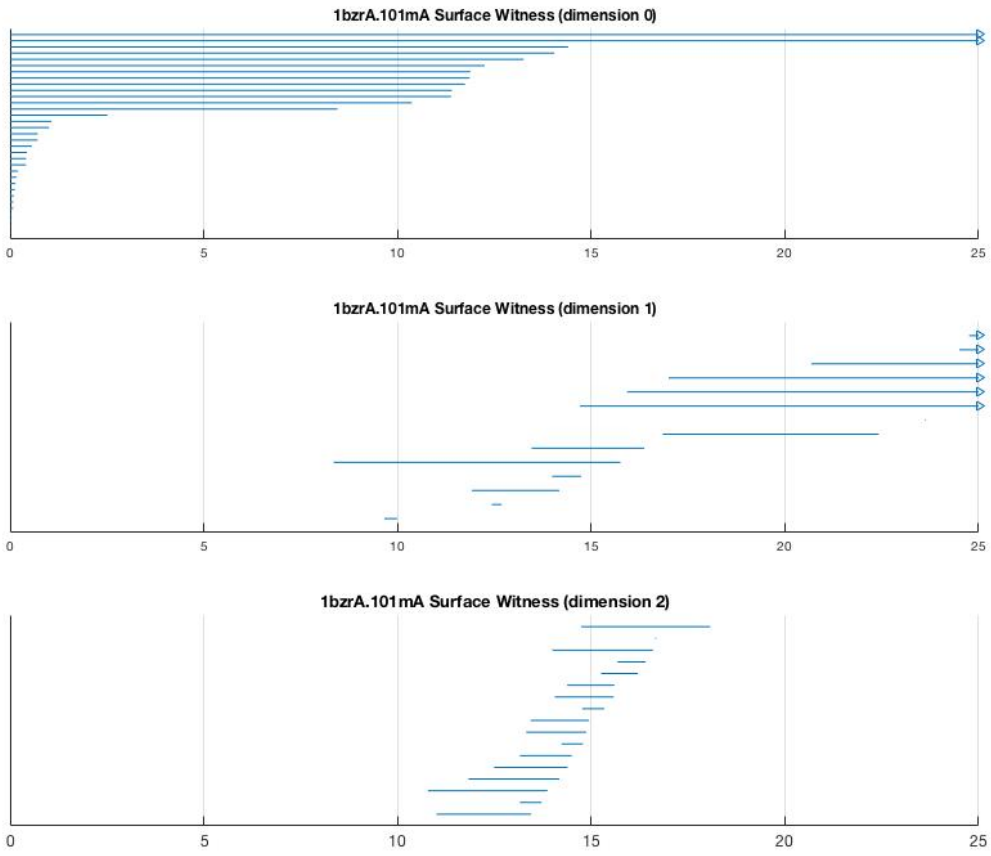


Figure 9: Witness stream barcodes for the 1bzrA.101mA cluster.

Again, in dimension 0, the 1bzrA.101mA cluster is almost topologically equivalent to the other clusters. In dimension 1, 1bzrA appears to be very similar to 2mgmA and 4qauA

as it has approximately the same number of birth and death intervals that appear between filtration values 10 and 15, and it has 6 bars that persist past filtration value 25, revealing the topologically significant 1-dimensional holes in the cluster. The dimension 2 barcode for 1bzrA is especially interesting as it is drastically different from those of 1duoA and 105mA. While similar to 4qauA and 2mgmA in that it does not have any 2-dimensional hollows that persist forever, the dimension 2 barcode of 1bzrA has 16 bars, which is twice as many bars as 2mgmA that all occur between filtration values 10 and 20, signifying the sudden birth and death of cavities in the cluster in that interval that may not be significant to the overall topology of the protein, but may be useful in helping us identify where the protein is more flexible and why, in the future. The obtained Betti numbers are $Betti_0 = 2, Betti_1 = 6, Betti_2 = 0$.

4.2 Barcodes from Lazy Witness Streams

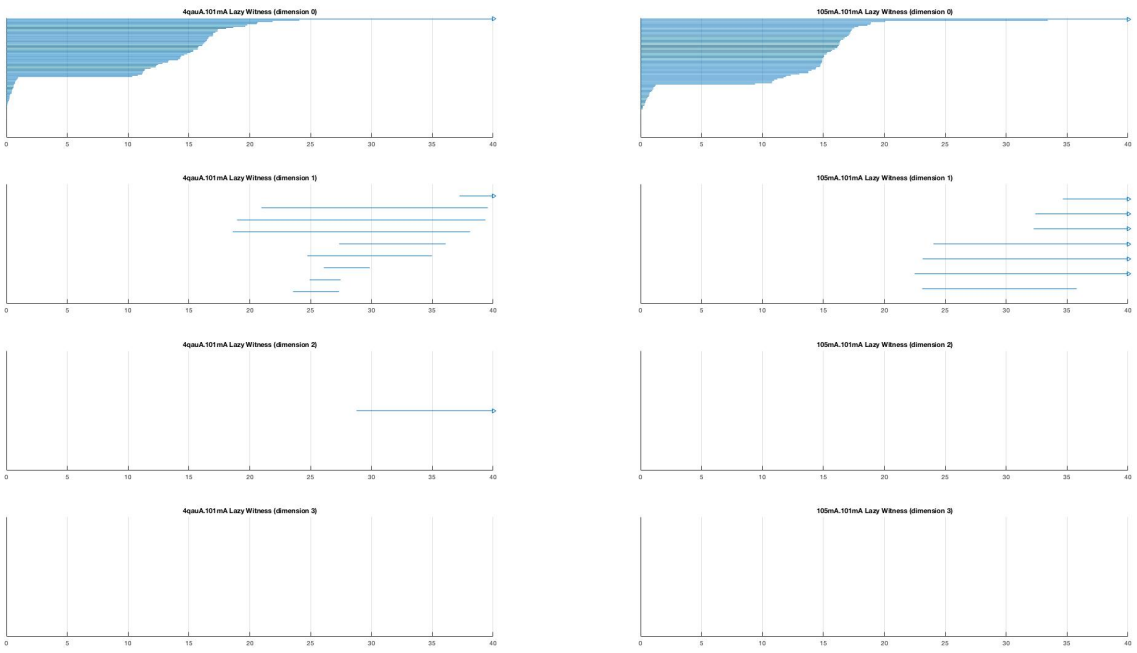


Figure 10: Lazy witness stream barcodes for the 4qauA.101mA(left) and 105mA.101mA(right) clusters.

Lazy witness streams give more computational freedom as they are more computationally efficient, so we generated barcodes using larger maximum filtration values for the 2 most interesting of our 5 randomly selected clusters: 4qauA and 105mA. The plots reveal that at a longer range, the obtained Betti numbers for the 4qauA cluster are actually $Betti_0 = 1, Betti_1 = 1, Betti_2 = 1, Betti_3 = 0$, which matches the homology groups of a pinched torus. Thus, the 4qauA cluster is well-approximated by a well-known topological object that is the pinched torus. And the Betti numbers obtained for the 105mA cluster are $Betti_0 = 1, Betti_1 = 6, Betti_2 = 0, Betti_3 = 0$, so the 105mA cluster would be well-approximated by a planar region with 6 holes.

4.3 Discussion

The 4qauA and 105mA clusters are not topologically equivalent. This shows that there are significant differences in the structures of the clusters, which suggests towards the flexibility of myoglobin. PDBFlex measures and evaluates the flexibility of proteins based on RMSD of atomic positions [14]. Approaching and analyzing the data through persistent homology provides a new perspective on the flexibility of proteins. This allows us to categorize them based on their homology, and analyze flexibility accordingly.

5 Conclusion

Using persistent homology, we analyzed 5 unique data sets of more than 174,000 values each of the myoglobin protein with the assistance of barcodes that we created in Matlab. We were able to show that the 4qauA cluster is homeomorphic to the common torus,and that the 105mA cluster can be well-approximated by a space with 6 holes. These results not only have potential in shedding new light on the research in the flexibility of proteins, but also have potential in both simplifying and complicating the way that we analyze the various interactions of proteins.

6 Future Work

In the future, we hope to analyze more protein data sets in higher dimensions. New and relevant cycles may exist in higher dimensions, giving deeper insight into the data. We would also like to develop a new filtration algorithm specific to analyzing protein data that has the potential of allowing for a different perspective on the flexibility of proteins. When combining these methods, it is possible to gain a more holistic understanding of the protein structure that can help empower research into treatment of protein related diseases and disorders by comparing the barcodes and homology groups of healthy protein data versus diseased protein data. For instance, it was recently discovered that in malignant pancreatic cancer cells, PAK4 interacts with p85 alpha to further disease progression [15]. Using persistent homology to analyze these two proteins and their interaction could enable the development of a drug that would inhibit this interaction, greatly accelerating the timeline for development of a treatment for pancreatic ductal adenocarcinoma.

References

[1] Vidit Nanda and Radmila Sazdanović. Simplicial models and topological inference in biological systems. In *Discrete and topological models in molecular biology*, pages 109–141. Springer, 2014.

[2] Eric W. Weisstein. Convex hull. Visited on 11/07/17.

[3] Eric W. Weisstein. Simplex. Visited on 11/07/17.

[4] Margherita Barile and Eric W. Weisstein. Betti number. Visited on 11/07/17.

[5] Herbert Edelsbrunner. *A Short Course in Computational Geometry and Topology.* Springer, 2014.

[6] Henry Adams and Andrew Tausz. Javaplex tutorial. Available at https://github.com/appliedtopology/javaplex.

[7] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, April 2009.

[8] Gunnar Carlsson Monica Nicolau, Arnold J. Levine.

[9] Bernadette J. Stolz, Heather A. Harrington, and Mason A. Porter. The topological "shape" of brexit. *CoRR*, abs/1610.00752, 2016.

[10] Bioinformatics: Protein structure.

[11] Daniel Alvarez-Garcia and Xavier Barril. Relationship between protein flexibility and binding: Lessons for structure-based drug design. *Journal of Chemical Theory and Computation*, 10(6):2608–2614, 2014. PMID: 26580781.

[12] Cluster 101ma. Visited on 11/09/17.

[13] Andrew Tausz, Mikael Vejdemo-Johansson, and Henry Adams. JavaPlex: A research software package for persistent (co)homology. In Han Hong and Chee Yap, editors, *Proceedings of ICMS 2014*, Lecture Notes in Computer Science 8592, pages 129–136, 2014. Software available at http://appliedtopology.github.io/javaplex/.

[14] Thomas Hrabe, Zhanwen Li, Mayya Sedova, Piotr Rotkiewicz, Lukasz Jaroszewski, and Adam Godzik. Pdbflex: exploring flexibility in protein structures. *Nucleic Acids Research*, 44(D1):D423–D428, 2016.

[15] Andrew Whale Prabhu Arumugam Hesham Eldaly Hemant M. Kocher Claire M. Wells Helen King, Kiruthikah Thillai.