Algebraic Topological Methods for the Analysis and Modeling of Protein Data

David F. Snyder[1], Jason Yuan[2], and Crystal Wang[2]
[1]Department of Mathematics, Texas State University, San Marcos, Texas, USA
[2]Honors Summer Math Camp, Texas State University, San Marcos, Texas, USA

Abstract—*Persistent Homology is a method used to find the most significant topological features in a space. It is derived from algebraic topology and has many applications in the analysis of noisy or large data sets. In this study, we find a novel application of persistent homology in protein structure analysis. We analyzed Cyclin-Dependent Kinase 6 (CDK6), an enzyme of great medical significance due to its relation to cancer. The fact that a protein's structure heavily influences its function motivated our research. We modeled the molecular surface as simplicial complex and used an open-source MATLAB package, JavaPlex, to perform persistent homology and compute the topological features. Our results showed that the Betti numbers in dimensions zero, one, and two are one, four, and one, respectively. From this, we conclude the surface is a two-holed torus. Our also study verifies that persistent homology is an effective method for protein data analysis.*

Keywords: proteins, biomolecules, modeling, computational biology, molecular biology, topology

1. Introduction

With newly-discovered methods of collecting large, complex data sets, efficient and effective means of data analysis techniques are becoming more critical. One such technique is the use of algebraic topology to model and analyze data. Topology is the study of spatial properties preserved under continuous deformations. A primary (often distinguishing) characteristic of any topological space (such as the solvent-accessible surface of a biomolecule) is the number of cavities, holes or connected components it contains. Algebraic topology leverages abstract and linear algebra to obtain such topological invariants. Persistence homology, a relatively new development in computational topology, has the potential to perform powerful data analytics due to its ability to simplify vast quantities of data by extracting essential geometric information from the data. To be able to apply this method, the data is first must be represented as a point-cloud in a topological space (for example, 3-dimensional space). In particular, solvent-accessible surfaces of biomolecules are typically represented as point-clouds in 3-dimensional space.

The volume and prevalence of outlier points often decrease the efficiency of computing the persistence homology of a given point cloud but without necessarily giving any extra useful information. In order to address this, we use *dense core subsets* (defined below in Definition 13) to reduce the data set size without losing pertinent topological features of the resulting point cloud. A dense core subset is a smaller percentage of the point cloud and is constructed using a parameter k, where lesser values of k give a more local view of the topological features of the data, while greater values give a more global.

In the project discussed here, we analyzed the solvent-accessible surface of a protein (we chose Cyclin-Dependent Kinase 6 (CDK6)) using persistent homology. We first stored the solvent-accessible surface as a point cloud, then select several dense core subsets from the data with varying k. By applying persistent homology techniques to multiple core subsets, we performed fast and insightful computations that revealed information about the structure that other methods of data analysis might overlook. We used existing algorithms that implement persistent homology in a computationally efficient manner.

2. Background

2.1 Definitions

See Edelsbrunner's text [1] for further details of the following definitions.

Definition 1. A collection of $k+1$ vectors $u_0, u_1, u_2 \dots u_k \in \mathbb{R}^n$ is *affinely independent* if the vectors $u_1 - u_0, u_2 - u_0, \dots, u_k - u_0$ are linearly independent.

In short, no vector in the collection is a weighted sum of the remaining vectors.

Definition 2. A *k-simplex* is the set of points determined by $k+1$ affinely independent points $u_0, u_1, u_2, \dots u_k \in \mathbb{R}^n$ determined an inequality:

$$\{\lambda_0 u_0 + \lambda_1 u_1 + \dots + \lambda_k u_k \mid \sum_{i=0}^{k} \lambda_i = 1, \text{ and } \lambda_i \geq 0 \text{ for all } i\}$$

The plural of 'simplex' is 'simplices.' In application, a 0-simplex is a point, a 1-simplex is a line segment, a 2-simplex is a triangle, a 3-simplex is a tetrahedron, and so on. When

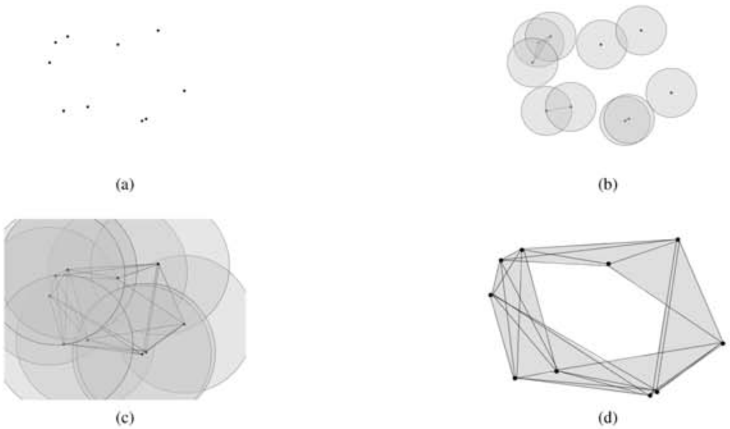


Fig. 1

THIS IS A VISUALIZATION OF A VIETORIS-RIPS FILTRATION ON A POINT CLOUD OF 10 VERTICES, WHERE t IS THE RADIUS OF THE CIRCLES IN EACH FIGURE. MORE SIMPLICES ARE FORMED AS t INCREASES, REVEALING THE HOLE IN THE CENTER OF THE POINTS.

the context is clear, we often refer to a k-simplex simply as a simplex.

Definition 3. A *face* of a k-simplex σ is a simplex determined by any subset of the vertices of σ besides the empty set. Thus, a k-simplex contains $2^{k+1} - 1$ faces.

Definition 4. A *complex* is a finite collection of simplices, K, such that: for each simplex $\sigma \in K$, every face of σ is in K; and for each pair of simplices $\sigma, \tau \in K$, the intersection $\sigma \cap \tau$ is either empty or a face of both σ and τ.

Definition 5. The *underlying space* $|K|$ of the complex K is the union of all simplices in K. Often the term complex is used to refer to what is actually the complex's underlying space, when the context is clear.

Definition 6. Two topological spaces $\mathbb{X}$ and $\mathbb{Y}$ are said to be *homeomorphic* if there is a bijective function f between the two spaces such that both f and f^{-1} are both continuous. More informally, $\mathbb{X}$ and $\mathbb{Y}$ are homeomorphic if one can be continuously deformed into the other.

Definition 7. A *homology group* is a collection of algebraic objects $H_d(K)$ that is associated to the complex K, where d ranges over the dimensions of simplices contained in K. A technical definition of these groups is beyond the scope of this paper, but may be found in [1].

Definition 8. The *Betti number in dimension* n, denoted β_n, is the rank of the n^{th} homology group of a complex.

We also call β_n the n^{th} *Betti number*. Intuitively, β_n counts the number of n-dimensional "gaps" in a space that are *not* bounded by an $(n+1)$-dimensional space. For example, β_0 measures the number of connected components ("pieces") of the underlying space of the complex, β_1 the number of tunnels through that space, and β_2 the number of cavities inside.

The next definition indicates the approach taken to estimate the Betti numbers of a point cloud: we build a sequence of complexes that, at each stage, reflect critical information within the data cloud.

Definition 9. A *filtration* of a simplicial complex K is a collection of subcomplexes $\{K(t) | t \in \mathbb{R}\}$ of K such that $K(t) \subset K(t')$ for all $t \leq t'$.

Now we define Betti intervals, which are used to understand how the homology of $K(t)$ changes with t.

Definition 10. A k-dimensional *Betti interval*, with endpoints $[t_{\text{start}}, t_{\text{end}}]$ (the birth and death time), corresponds to a k-dimensional hole that appears at filtration value t_{start}, remains open for $t_{\text{start}} \leq t < t_{\text{end}}$, and closes at value t_{end}].

Persistence homology is functorial: for $t \leq t'$, the inclusion map $i : K(t) \rightarrow K(t')$ of complexes induces a map $i_* : H_k(K(t)) \rightarrow H_k(K(t))$ of homology groups.

Definition 11. The *lifespan* of a Betti interval with birth t_{start} and death t_{end} is $t_{end} - t_{start}$. Betti intervals with longer lifespans often correspond to more significant topological features.

Definition 12. Given $k \in \mathbb{Z}^+$ and a vertex z in a simplicial complex, $\rho_{\mathbf{k}}(\mathbf{v})$ is defined as the distance from z to its k-th closest neighbor.

Definition 13. The *dense core subset* $X(k, p)$ of a data set is the subset of vertices that includes the $p\%$ densest vertices given density estimator $\frac{1}{\rho_k(z)}$ (see [2]).

2.2 Filtrations

Filtrations on a point cloud can be used to track the birth and death of Betti intervals, which are represented graphically in a *figure* (see Figures 3 and 2) . We now give a brief explanation; see the JavaPlex tutorial written by Adams and Tausz [2] for more details.

2.2.1 The Vietoris-Rips Complex

Given a set of vertices, the *Vietoris-Rips complex* is built by connecting any two vertices a and b in an edge if the two are within a specified maximum distance t from each other, as in Figure 1 . If all the edges of a higher-dimensional simplex exist in the stream, then the simplex does as well (see [2] for details). A *Vietoris-Rips stream* is the collection of Vietoris-Rips complexes that result as t is increased in value. The stream is a collection of topological images of the point cloud but across a spectrum of resolutions.

2.2.2 Landmarks

To reduce larger data sets, we often select a subset of all the points in the set to be landmarks. Streams are built on the set using only landmark points as vertices. Landmark points can be selected randomly from the total point cloud, or through a process called *sequential maxmin* (see [3]). In this case, the first landmark point is selected randomly, and each subsequent point selected is the farthest point in the cloud from any of the already-existing landmark points.

2.2.3 The Witness Filtration Complex

The following is based on the exposition found in [3]. Let $L = \{l_0, l_1, \ldots, l_i\}$ be the set of landmark points and let Z be the set of all data points. We define $m_k(z)$ to be the distance from a vertex $z \in Z$ to the $(k+1)$-th closest landmark point. Using a witness filtration, we build our simplicial complex as follows: for $k \in \mathbb{Z}^+$, the k-simplex $[l_0, l_1, \ldots, l_k]$ is included in the witness stream complex if there exists a witness point z such that

$$\max\{|l_0 - z|, |l_1 - z|, \ldots, |l_k - z|\} \leq t + m_k(z).$$

3. Prior Work

A wide variety of fields have applied the mathematical tools of algebraic topology with great success. For example, Nicolau, Levine, and Carlsson applied algebraic topology to analyze breast cancer transcriptional data in [4]. They were able to identify a unique subgroup of Estrogen Receptor-positive (ER+) in breast cancers that exhibited high levels of normal animal cellular myeloblastosis (c-MYB) and had a 100 percent survival rate.

Geometric concepts form the basis of much prior work in the computational modeling and analysis of protein structures. These methods typically model the atoms within the protein in 3-d coordinate space and calculate lengths, angles, and areas. One such example is the Method of Minimal Molecular Surfaces, which can be used to measure internal and open cavities [5]. This method is built upon the theory of differential geometry and uses iterative procedures and numerical algorithms in order to compute mean curvature of the protein surface. Geometric methods like these may entail inessential details leading to inefficient use of computing resources. Also, many outputs of geometric methods (such as the measuring of internal cavities) can be computed using more efficient topological methods. Computing these as topological (as opposed to geometric) features using persistence homology increases computational efficiency.

4. Proteins

We now briefly summarize relevant facts about proteins, much of which can be found in a standard text such as [6]. We also briefly discuss Cyclin-Dependent Kinase 6, and what is known about its relation to cancer.

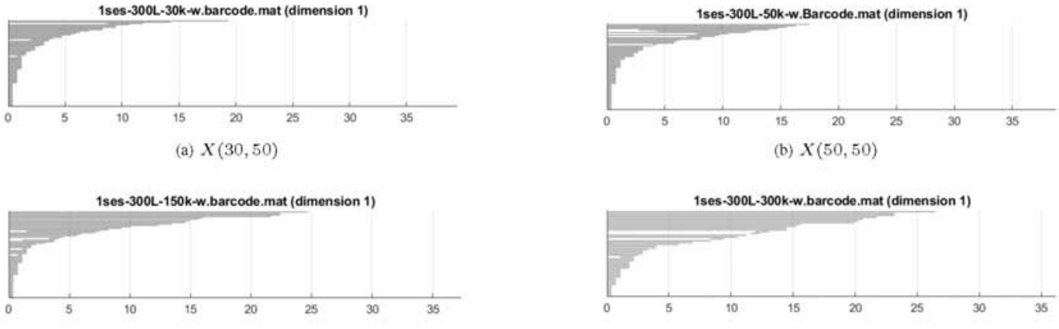
4.1 Function and Structure

Proteins have four levels of biological structure, but for this paper, only the first three are necessary for the reader's understanding. The primary structure is the protein's polypeptide chain, the sequence of amino acids that make up the protein. Each amino acid includes an amine group, a carboxyl group, and a side chain called an R-group which is unique to that amino acid. The protein's secondary structure is three-dimensional and focuses only on small sections of amino acids. The hydrogen bonds between the amine and carboxyl groups on separate amino acids cause them to twist into structures known as alpha helices and beta pleated sheets. Tertiary structures form by the interactions between the R-groups of each amino acid, which causes protein folding and determines the polypeptide chain's entire three-dimensional shape. Tertiary structure is unique from protein to protein.

Proteins carry out their functions by binding to specific substrate molecules on its surface in a kind of lock-and-key model. Therefore, changes in the tertiary structure correspond to differences in protein function.

4.2 Cyclin-Dependent Kinase 6

Cyclin-dependent kinase 6 (CDK6) is an enzyme of much medical significance due to its relation to cancer. CDK6 regulates cell cycle progression at the restriction, or R, phase in the cell's growth [7]. The presence of high concentrations of CDK6 is the final signal a cell needs to continue the cell cycle and enter the S phase of its life, where it replicates its DNA in preparation for cell division. Typically, cells that contain mutated DNA engage in apoptosis, or programmed



(a) $X(30, 50)$

(b) $X(50, 50)$

(c) $X(150, 50)$

(d) $X(300, 50)$

Fig. 2
THE BARCODES IN DIMENSION ONE, FOR VARIOUS VALUES OF k.

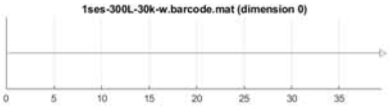


Fig. 3
BARCODE FOR DIMENSION ZERO, FOR $X(30, 50)$.

cell death, to prevent this DNA from being replicated and passed down to daughter cells. Cancer occurs when cells with mutated DNA divide and proliferate without regulation and form tumors. Because of the way CDK6 encourages cell division, unnaturally high amounts of the enzyme are often found in certain types of cancer cells [8]. Researchers are still evaluating CDK6's specific functions and importance to cancer treatment.

The model of CDK6 we chose to analyze shows the surface structure of the protein when binded to its inhibitor, the p16INK4a tumor suppressor. The inhibitor prevents CDK6 from carrying out its function by changing the structure of the cyclin-binding and ATP-binding sites of CDK6, and possibly changing the protein's topology as well.

5. Methods

We obtained 3-dimensional models for protein 1bi7A through the online Protein Data Bank pdbflex.org. Then, we read the protein file `.pdb` into Jmol, an open-source Java-based chemical structure viewer. This program allowed us to export the protein file into a 3-dimensional mesh in `.stl` format. Finally, since the `.stl` file contains extra information, we parsed that file through a custom MATLAB package in order to extract the vertex data.

5.1 Dense Core Subsets

For our dense core subsets $X(k, p)$, we used $k = 30, 50, 150, 300$ and $p = 50\%$. The naive method for computing dense core subsets is to calculate all distances between points and sort the distances to find the kth nearest neighbor. However, this is much too slow, with time complexity of $O(n^2 \log n)$. To compute the dense core subsets efficiently, we utilized a special data structure known as a k-d tree, which supports the following operations: building the k-d tree in $O(n \log^2 n)$ and querying the kth nearest neighbor from a point in $O(k \log n)$. This modified method of obtaining dense cores subsets allowed us to compute the dense core subsets far more efficiently.

5.2 JavaPlex

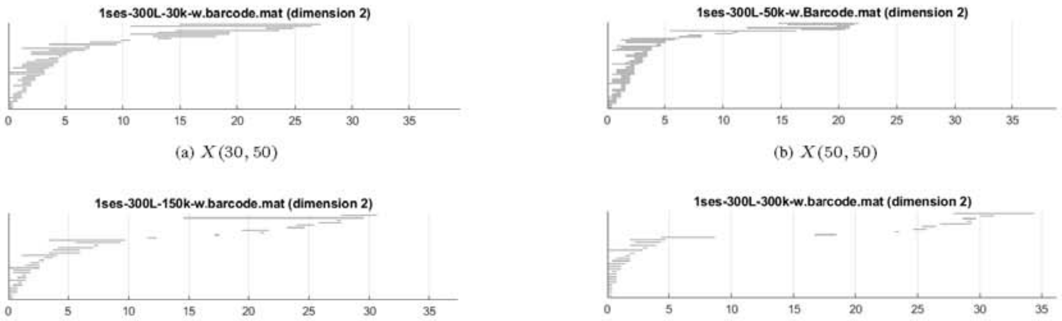
We used JavaPlex (see [2]), a powerful MATLAB package developed originally by Gunnar Carlsson and his students, to run witness streams for each dense core subset. Doing so generates barcodes which are used to determine the Betti numbers of the surfaces. The JavaPlex Package allows for several key functions such as a landmark selector, a witness stream constructor, and a persistence algorithm.

6. Results

The following barcode images are all results of a witness stream complex with 300 landmark points constructed on a variety of dimensions and dense core subsets of 1bi7A.

6.1 Dimension 0 Barcodes

Refer to Figure 3. All dense core subsets of protein 1bi7A had $\beta_0 = 1$ for values of t between 0 and 40. This means that 1bi7A contains one connected component throughout the filtration.



(a) $X(30, 50)$

(b) $X(50, 50)$

(c) $X(150, 50)$

(d) $X(300, 50)$

Fig. 4
THE BARCODES IN DIMENSION TWO, FOR VARIOUS VALUES OF k

6.2 Dimension 1 Barcodes

Although the four barcodes shown in Figure 2 vary, each one appears to have four bars that persist significantly longer than the others, indicating that $\beta_1 = 4$ is the best estimate for the 1-dimensional Betti number of the solvent accessible surface.

6.3 Dimension 2 Barcodes

In Figure 4a, though a few longer-lived cavities appear between $t = 15$ and $t = 20$, there are only two with a lifespan greater than ten and hence are much more significant. When we increase the value of k to 30 in Figure 4b, there is now only one bar with a life span longer than ten. Even up to $k = 150$ (see Figure 2c), we still only see one bar longer than ten across. This indicates that the two long-lived cavities indicated in Figure 4a was actually spurious, due to $k = 30$ providing a resolution level that is "too grainy" or "foggy". By the time k has reached the value of 300 (Figure 4d), there are no cavities with a life-span more than 7.5, indicating that this resolution level is being too indiscriminate.

6.4 Analysis

For all dense core subsets for protein 1bi7a, there was only one bar for $t \in [0, 40]$. Thus $\beta_0 = 1$, which represents the single connected component that forms the solvent accessible surface, confirming what we know about the structure.

In the barcodes for Betti numbers of dimension one, there are a significant number of features detected in the filtration across all dense core subsets. As the parameter of k for the dense core subsets increased, the number of 1-dimensional holes ("tunnels" through the surface) increased as well. This observation affirms our understanding of varying the value of k: increasing the value of k selects points that are farther from each other in the dense core subset which, in turn, gives a more global view of the surface. For both $k = 30$ and $k = 50$, the longest persisting feature was less than $t = 20$, but for $k = 150$ and 300, features were detected until around $t = 25$. From the four sets of barcodes, we determine that the solvent accessible surface has $\beta_1 = 4$. The barcodes for the dense core subset of $k = 50$ shows 4 distinct features that persisted from $t = 0$ to around $t = 15$, significantly longer than the next longest feature. The barcodes for the dense core subset of $k = 150$ also agrees upon this finding, with the 4 longest features lasting until $t = 22$ while the largest after these lasts barely over $t = 16$. Barcodes for dense core subsets of $k = 30$ and $k = 300$ do not agree as cleanly with the other two but still align with the trend.

In the dimension 2 barcodes, there is a significant amount of noise that is detected from $t = 0$ to $t = 10$. For all four subsets, there are many small cavities detected in this range, none of which persists at all. In the dense core subset of $k = 30$, there are two significant cavities detected from $t = 11$ to $t = 27$. When $k = 50$, there is one feature that is significantly longer than the others with a lifespan of greater than 10. For $k = 150$, there is a clear single cavity detected from $t = 15$ to $t = 29$, giving a distinct $\beta_2 = 1$. The barcodes for $k = 300$, are much less clear, there is a longer feature detected at $t = 27$, but none of the features detected persisted for a significant time. This lack of persistence can be the result of choosing a k value that is too large; the points selected for the dense core subset are too spread out to depict the surface accurately.

These findings align very closely with the barcodes from the dense core subsets of $k = 50$ and $k = 150$. The barcodes for dense core subsets made from $k = 30$ and $k = 300$ do not agree as determinedly, and this is the result of using more extreme values of k when selected the core subsets. From this collection of information, we estimate the Betti numbers of the solvent accessible surface to be $\beta_0 = 1, \beta_1 = 4$, and $\beta_2 = 1$, which coincide with the Betti numbers of a two-holed torus.

7. Conclusion

We used persistent homology, a powerful tool in algebraic topology, to analyze CDK6, a protein in cluster 1bi7A. We constructed dense core subsets and witness filtrations for our proteins using MATLAB. In doing so, we discovered the presence of one local cavity, and four tunnels on in our protein's solvent accessible surface. We also demonstrated the effects of changing specific parameters to give us a more local or global view. Finally, we demonstrated the potential of persistent homology as a fast and computable analytical method for successfully analyzing a salient feature of a protein.

8. Future Work

In the future, we hope to develop a novel filtration algorithm tailored to analyze proteins. We also hope to experiment with more approaches to selecting dense core subsets that are more representative of our data set. Finally, we hope to apply persistent homology to analyze protein folding.

Acknowledgment

The last two authors would like to thank the first author for his mentorship and guidance in the development of this project, undertaken at the 2018 Texas State University Honors Summer Math Camp. They also thank the referees for their valuable suggestions.

References

[1] H. Edelsbrunner, *A short course in computational geometry and topology*, ser. SpringerBriefs in Mathematical Methods. Springer, 2014.

[2] A. Tausz, M. Vejdemo-Johansson, and H. Adams, "JavaPlex: A research software package for persistent (co)homology," in *Proceedings of ICMS 2014*, ser. Lecture Notes in Computer Science 8592, H. Hong and C. Yap, Eds., 2014, pp. 129–136, software available at http://appliedtopology.github.io/javaplex/.

[3] V. De Silva and G. Carlsson, "Topological estimation using witness complexes," in *Proceedings of the First Eurographics Conference on Point-Based Graphics*, ser. SPBG'04. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2004, pp. 157–166. [Online]. Available: http://dx.doi.org/10.2312/SPBG/SPBG04/157-166

[4] M. Nicolau, A. J. Levine, and G. Carlsson, "Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival," *Proceedings of the National Academy of Sciences*, p. 201102826, 2011.

[5] P. Bates, G.-W. Wei, and S. Zhao, "Minimal molecular surfaces and their applications," *Journal of Computational Chemistry*, vol. 29, no. 3, pp. 380–391, 2008.

[6] J. B. Reece, L. A. Urry, M. L. Cain, S. A. Wasserman, P. V. Minorsky, R. B. Jackson, *et al.*, *Campbell biology*. Pearson Boston, 2014, no. s 1309.

[7] M. Aarts, S. Linardopoulos, and N. C. Turner, "Tumour selective targeting of cell cycle kinases for cancer treatment," *Current Opinion in Pharmacology*, vol. 13, no. 4, pp. 529–535, 2013.

[8] S. Tadesse, M. Yu, M. Kumarasiri, B. T. Le, and S. Wang, "Targeting cdk6 in cancer: state of the art and new insights," *Cell Cycle*, vol. 14, no. 20, pp. 3220–3230, 2015.