Available online at www.HEFJournal.org

ISSN: 2785-2997

Journal of
Human, Earth, and Future

Vol. 3, Special Issue, 2022

"Current Status and Future Trends in BioScience and BioTechnology"

Characterizing Protein Conformational Spaces using Efficient Data Reduction and Algebraic Topology

Arpita Joshi [1*], Nurit Haspel [1], Eduardo González [2]

[1] *Department of Computer Science, University of Massachusetts, Boston, MA, United States*

[2] *Department of Mathematics, University of Massachusetts , Boston, MA, United States*

Received 19 January 2022; Revised 28 March 2022; Accepted 12 May 2022; Published 31 May 2022

Abstract

Datasets representing the conformational landscapes of protein structures are high-dimensional and hence present computational challenges. Efficient and effective dimensionality reduction of these datasets is therefore paramount to our ability to analyze the conformational landscapes of proteins and extract important information regarding protein folding, conformational changes, and binding. Representing the structures with fewer attributes that capture the most variance in the data makes for a quicker and more precise analysis of these structures. In this study, we make use of dimensionality reduction methods for reducing the number of instances and for feature reduction. The reduced dataset that is obtained is then subjected to topological and quantitative analysis. In this step, we perform hierarchical clustering to obtain different sets of conformation clusters that may correspond to intermediate structures. The structures represented by these conformations are then analyzed by studying their high-dimensional topological properties to identify truly distinct conformations and holes in the conformational space that may represent high energy barriers. Our results show that the clusters closely follow known experimental results about intermediate structures as well as binding and folding events.

Keywords: Dimensionality Reduction; Hierarchical Clustering; Betti Numbers; Protein Folding; BioScience.

1. Introduction

Characterizing intermediate protein conformations is difficult to do experimentally due to the fleeting nature of these structures. Doing so is necessary for understanding and characterizing protein function and dynamics. This work uses dimensionality reduction and algebraic topology to extract meaningful information inherent in these structures. Protein structure and dynamics are essential to their function. Therefore, by understanding the connection between structure, dynamics, and function, we can better understand cellular processes involving proteins. The question of how the structure and dynamics of proteins relate to their function has challenged scientists for several decades but still remains open. Conformational search methods aim to characterize the conformational space of proteins in order to find low energy regions corresponding to highly populated or intermediate structures [1-3]. These intermediate states are transient and therefore hard to detect experimentally. However, they may be essential to understanding dynamic events such as folding, protein-protein interactions, and various cellular processes. The potential energy landscape of a protein is often rugged and has a large number of local minima [4]. This makes conformational exploration especially

* Corresponding author: arpita.joshi@isbscience.org

http://dx.doi.org/10. 28991/HEF-SP2022-01-01

challenging. The problem becomes even more challenging due to the high dimension of the problem. A typical protein can contain several hundred amino acids or several thousand atoms. Therefore, the search space made up of all possible conformations that a protein can assume is large and its enumeration is practically impossible. Existing physics-based computational methods that sample the conformational space of proteins include Molecular Dynamics (MD) [5], Monte Carlo (MC) [6] and their variants, as well as approximate methods based on geometric sampling [2, 7-10], Elastic Network Modeling [11-13], normal mode analysis [14, 15], morphing [16, 17], and several other methods. Even after the conformational space is sampled, it should be filtered and clustered to extract meaningful information. Several clustering methods have been designed for protein conformational space [7, 18, 19]. Most clustering methods for high-dimensional data incorporate metric functions that evaluate the distance between objects in the dataset, or a lower-dimensional representation of these objects, often trying to detect outliers [20].

1.1. Problem Statement

The protein folding problem, which aims at predicting the correct protein structure from its sequence, it is an important problem in Biology. The conformational search problem is a related problem that aims at characterizing the conformational space of proteins in order to find minimum energy regions corresponding to highly populated structures and characterize dynamic events such as folding or docking [21, 22]. The potential energy landscape of a protein is immense. A typical protein molecule has hundreds of amino acids and several thousand atoms. Consequently, the number of configurations that a molecule can attain is extremely large and practically impossible to enumerate computationally. This has given rise to a foray of methods that attempt to model the actual pathways that a protein undertakes while transitioning from one conformation to another that ultimately help in elucidating the highly populated conformations generated in the entire process. In this work we present a method to analyze the conformational space of proteins by first reducing the dimensionality of molecular conformations datasets that represent the molecule and then use a number of filtering techniques taken from algebraic topology to identify clusters of intermediate conformations. The molecules used range from small (Oxytocin and Vasopressin) to medium (Cdc42) to large (GroEL). Details about them can be found in the Results section. The main contributions of the paper can be summed up as under:

- Data representation – Using works described in sections 2.2.1 and 2.2.2, we create a space efficient way to represent molecular data. The datasets in use have just enough instances and attributes that capture the maximum amount of variance.
- Use versions of hierarchical clustering, depending on the molecule, to sample different conformations of a molecule. Details can be found in section 2.3.1.
- Use algebraic topology methods to analyze distinctiveness among various clusters of conformations and extract the topological properties of the clusters.

1.2. Feature Reduction

It is often helpful to obtain a lower-dimensional representation of the data that preserves as much of the variance in the original data as possible. It is especially useful in protein conformations, since the mutual constraints between atoms in the protein molecule makes the ”true” dimensionality of a protein structure much smaller than the number of parameters required to represent a 3 dimensional protein molecule. Existing algorithms for data instance reduction are broadly divided into, incremental, decremental, batch and mixed [23-27]. The incremental algorithms begin with a null set and data instances are added to it depending on the result of the algorithm. The decremental algorithms, on the contrary, begin with the entire set of instances and depending on the decision offered by the selection algorithms, instances are taken out from the set one had at the beginning. The batch algorithms function in a way that each instance is first analyzed and then a decision is made as to which ones to keep. Mixed algorithms begin with a pre-selected set of instances and the process then continues to figure whether instances should be deleted or added. An evaluation of the age-old techniques of instance reduction is explained by Wilson and Martinez [23].

Linear dimensionality reduction like PCA and its variants may not capture the complex, non-linear nature of protein conformational landscape. Dimensionality reduction techniques are broadly classified based on the solution space they generate, as convex and non-convex [28]. Techniques described by Maaten [29] give explicit details of the various well established non-convex methods. These methods are further sub-divided into Full Spectral Techniques, the ones that perform the eigenvector decomposition of a full matrix and Sparse Spectral Techniques, the ones that do the same for a sparse matrix. The latter ones have better time-complexity but these approaches are local. They attempt at retaining only the local structure that the sparse portions of the dataset present. On the other hand, Full Spectral Techniques, capture the covariance between all the data instances and form a more thorough representation of the structure as a whole. The *Isomap* algorithm [30] is a non-linear dimensionality reduction method that falls into the Convex Full Spectral category. It takes as input the distances between points in a high-dimensional observation space, and outputs their coordinates in a low-dimensional embedding that best preserves their intrinsic geodesic distances. The original dimensions of the matrix is $M \times N$, where N is the number of instances and M is the size of each

instance. The output matrix is of size $N \times m$, where $m << M$. It has been shown by Haspel [2] that for protein datasets, Isomap produces much better results, due to protein conformational changes being non-linear and complex. Despite its advantage in efficient representation of molecular data [31, 32], Isomap is computationally expensive, especially with very large, multi-dimensional datasets. To overcome this, we use a version of the Mode-III of Isomap [55]. Improvements over Isomap are presented in [33, 34]. A similar approach is adopted for this work but in a way that is more suited for protein data.

1.3. Algebraic Topology

The conformational space of proteins is highly complex and high dimensional. Obtaining a full, analytical description of it is essentially computationally impossible. Different methods for characterizing the conformational landscape try to obtain a coarser, more approximate description of the landscape while still capturing its global, essential characteristics while preserving important features. Below we give a brief survey of the tools used in this paper.

1.3.1. Persistent Homology

Persistent homology is an algebraic topological tool for computing features of a space at (essentially) different spatial resolutions [19]. To find the persistent homology of a space *X*, presented as a data set, we first assign a *simplicial complex*. Moreover, using a distance function on the underlying space, we can build a *filtration* of the simplicial complex, thatis, a nested sequence of complexes. We follow the notationsin [35]:

Simplicial Complexes: A simplicial complex *K* is given by the following datum:

- A set K_0 of vertices or 0-simplices. We will also use the notation $K_0 = Z$ for consistency with the literature.
- For each $i \geq 1$, a set K_i of *i*-simplices $\sigma = [z_0 z_1 ... z_i]$,where $z_j \in Z$.
- Each *i*-simplex has $i+1$ *faces* $\sigma_j, j = 0,..., i$ obtained by deleting the *j*-th element in the list. We require that the faces σ_j are in K_{i-1}.

We think of 0-simplices as vertices, 1-simplices as edges, 2-simplices as triangular faces, and 3-simplices as tetrahedra, etc.

Homology and Betti Numbers: For a simplicial complex *K* as above, we associate the group of *i*-chains *Ci* with coefficients on a ring *R* (we will take $R = \mathbf{Z}_2$ to avoid issues with orientations) as the group generated by the elements of *Ki*, that is the set of formal sums $\sum rs\sigma s$, $\sigma s \in Ki$, $rs \in R$. The boundary map ∂i: $Ci \rightarrow Ci - 1$ takes an *i*-simplex $\sigma = [z0, ..., zi] \in Ki$ to the formal alternating sum $\sum(-1)j[z0, \ldots, z\hat{}j, \ldots, zi]$, and it is extended by linearity over *Ci*. This map satisfies $\partial i \circ \partial i + 1 = 0$. That is, the *boundaries* given by the image $Bi = \text{im}(\partial i + 1) \subset Ci$ lies in the subgroup of *cycles*, given by the kernel $Zi = \ker(\partial i) \subset Ci$. This makes *C***.** a chain complex. The *i*-th homology $Hi(K)$ is defined as the quotient *Zi/Bi*, this is a module over *R*.

The *k-th Betti Number* $Betti_k(K)$ of the complex is the rank of the *k*-th homology of the complex $H_k(K)$ as an *R*-module. Roughly speaking, this gives a count of the number of *k*-dimensional holes in the complex. In particular, $Betti_0(K)$ is the number of connected components of *K*. For instance, a *k*-dimensional sphere, has all $Betti$ numbers equal to zero except for $Betti_0 = Betti_k = 1$. For a topological space *X*, approximated by a complex K_X, its homology $H_i(X) = H_i(KX)$ and associated Betti numbers are classical invariants. Figure 1 has some examples, for instance a torus has one connected component so $Betti_0$=1, two nonhomologous one-dimensional cycles – one equatorial and one meridional, so $Betti_1$=2 and a single 2-dimensional hole enclosed within the surface, $Betti_2$=1. Similarly, for a circle, $Betti_0 = Betti_1 = 1$, and for a point $Betti_0 = 1$.

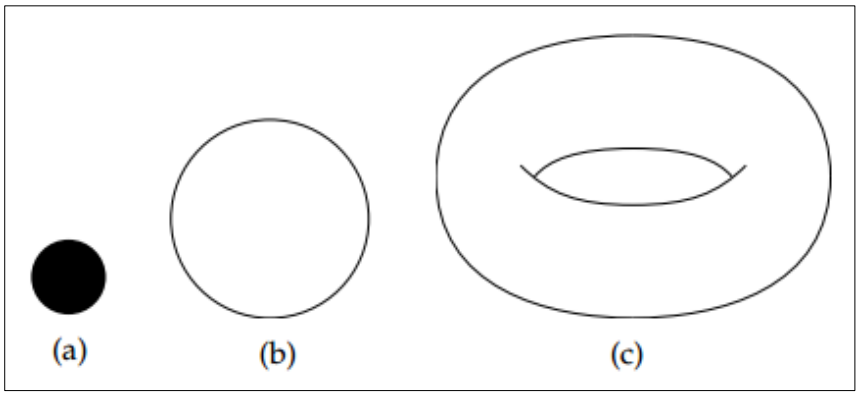


Figure 1. Representation of Betti Numbers as holes: (a) A dot – 1 connected component, (b) A circle – 1 connected component and 1 loop, (c)A torus – 1 connected component, 2 loops and a 2-dimensional face

Data Sets and Persistent Homology: An important example for this paper consists of the simplicial complex associated to a data set or point cloud. Let Z be a set of points in an Euclidean space $\mathbb{R}^n$. To this data, and to any radius $\epsilon > 0$ we can associate a simplicial complex K_ϵ as follows. The 0-simplices consist of the points $\{z\}$ in the data set themselves. A set of two points $\{z_0, z_1\}$ is declared to be a 1-simplex if the $\epsilon - balls$ centred at z_i intersect, that is if $B_\epsilon(z_1) \cap B_\epsilon(z_1) \neq \emptyset$. Analogously the k-simplices will be defined by those sets of $k + 1$ points $\{z_0, \ldots, z_k\}$ for which the intersection $B_\epsilon(z_1) \cap \ldots \cap B_\epsilon(z_k) \neq \emptyset$ is not empty. Certainly for small ϵ the complex only contains 0-complexes and for large ϵ the complex yields the *degenerate* complex for which all sets of $k + 1$ points define a k-simplex since all the points will be inside a single ball. Varying the ϵ yields a filtration of this complex. By a filtration of a simplicial complex K we mean a collection of sub-complexes $\{K(\epsilon)|\epsilon \in \mathbb{R}\}$ of K such that $K(t) \subset K(s)$ whenever $t \leq s$. A filtration yields *Persistent Homology*, an ϵ dependent homology theory H_ϵ and hence a family of Betti numbers which depend on ϵ. This is what its called *persistent bar codes*. For all purposes, we will interpret the radius ϵ as a time parameter and we will sometimes denote it as t.

We are interested in the study of how the topological invariants of a dataset varies with t and which *Betti* numbers and generators of homology persist with this variation. Several computational tools to compute Persistent Homology and associated bar codes, such as *JavaPlex* [36] and the Topological Data Analysis *TDAToolBox* [37] for R and [38] for Python were used. In general one has to apply dimensionality reduction methods to effectively compute any topological analysis.

In this work, on the embeddings produced by dimensionality reduction, we perform both hierarchical clustering and compute its persistent homology. The $Betti_0$ of the embeddings produces the same number as the number of relevant clusters in hierarchical clustering. The results are presented in the next section. Furthermore, these clusters are then subjected to further topological analysis. Each of these clusters (of a molecule) is a connected component of the embedding. Consequently, the $Betti_0$ of each of these is 1. We then compute $Betti_1$ and $Betti_2$ of each of these clusters to find out about the topological characteristic ofthe distinct clusters.

Algebraic topology in Bioinformatics: Algebraic topology approaches have recently been explored in thecontext of sampling biological data. The work cited in [39] describes the use of topological signatures, which the authors call *Evolutionary Homology* (EH) barcodes, reveal the topology-function relationship of the network and thus giverise to the quantitative analysis of nodal properties. The proposed EH is applied to protein residue networks for protein thermal fluctuation analysis, rendering accurate B-factor prediction for a number of proteins. A thorough review of the emerging applications of topological methods to genomics is presented in [40]. An insightful piece presenting concisely the ramifications of these approaches are elucidated in [41]. Algebraic topology and persistent homology were also used for protein-protein interaction [42] and protein conformational dynamics [19]. It is documented that intermediate structures can be detected in protein conformational spaces using algebraic topology andhierarchical clustering [3], but we take it further by exploring higher order $Betti$ numbers that may give us more information about the conformational space and analyzethe results by filtering the conformations with low potential energy. In this contribution we fill the gap by analysing the conformational spaces of multiple proteins of different sizesusing persistent homology, clustering and dimensionality reduction techniques. Our results match experimental data when it is available.

2. Methods

2.1. Generation of Data

The molecules chosen with in this work are diverse and very different in their function and dynamics. The conformational spaces for oxytocin, vasopressin, hGalanin, pGalanin and Cdc42 were sampled using Molecular Dynamics (MD) simulations. All systems were represented at full atomic resolution. All the above proteins except Cdc42 are shorter peptides with no crystal structures. The initial linear peptide was modelled using the Chimera software [43] as an idealized helix. Hydrogens and solvent were added using the AMBER software package [5]. Simulations were performed in constant volume (NVT) in an orthorhombic box usingthe TIP3 solvent model [44] to simulate infinite dilution. Periodic boundary conditions were applied using the nearest image convention. The overall charge of the system was kept neutral for the use of particle mesh Ewald summation to calculate electrostatic charges [45]. The simulations were carried out with the NAMD package [46] using the AMBERff03 force field [47]. To sample the conformational spaces of the peptides we used a simulated annealing based search [6]. For Cdc42, we used long MD simulations as described in [48]. The GroEL molecule is larger and cannot be easily sampled using MD simulations. The conformations were sampled using a Monte Carlo (MC) [6] based conformational pathway search. The protocol is detailed in Haspel [3]. The search begins with the PDB (Protein Data Bank) formatof one conformational extreme and expands following a biased Rapidly-expanding Random Tree (RRT) algorithm to simulate the pathway that can be undertaken to reach the goal conformation [49]. At every iteration a parent protein conformation is chosen from pool of new conformations,the one selected is the one which has its energy below a threshold. The new conformation is added to the pool[2] if its RMSD (Root Mean Square Deviation) is closer to the goal.

2.2. Dimensionality Reduction

Data Instance Reduction

The first step is to perform dimensionality reduction on the data. We use a data instance reduction method to decidethe information content offered by each data instance. We first apply spherical PCA as described in [50-52]. The input is a data matrix and the number of dimensions (principal components) desired of all the data instances.The output is a lower dimensional projection of the data. We used two or three dimension projections, in order to better visualize structurally significant data. It is observed that over eighty percent of the variance is explained by the first three principal components in all the types of data we used, although it is not always the case in other domains. This algorithm removes data instances that are deemed not to contribute to the variance in the data. More details about the algorithm itself and the results it produced can be found in [53, 54]. The result of this step produces a non-redundant representation of the dataset which now has fewer instances but its important properties that account for maximum variance present in it are preserved.

Low-Dimensional Embedding

The next step is to obtain an embedding of the conformations in the reduced dataset in a lower dimension. We used a parallelized version of the Isomap algorithm that produces the same results and works much faster [55]. The algorithm can be used to produce an embedding in as many dimensions as the attributes of the dataset, we pick the first three dimensions to work with because they capture over 80% of the variance inherent in the data.

1. It Is Predefined, Takes Care Of Whether A Conformation Would Be Feasible, Keeping In Mind Atom Collisions.
2. The Conformational Pathway.

2.3. Topological Analysis

After obtaining a reduced embedding that represents the dataset, the next step is to slot the data to sample different conformations inherent in these embeddings. To ascertain the number of conformations that can be extracted, we use two methods, namely, hierarchical clustering and persistent homology. Persistent Homology is the next step in the process but here it was used as a proof of validation,as both, hierarchical clustering and persistent homology revealed the same structural complexities in each dataset. We used the R package TDA's (Topological Data Analysis) function *calculate homology* [56] and the *TDAToolBox* [37] to generate the topological analysis.

Hierarchical Clustering

Hierarchical Clustering is a known method for identifying similar groups in a dataset. Built-in function in R, *hclust* is used for the purpose. We chose hierarchical clustering over k-means clustering to prevent having to pre-declarethe number of clusters sought, and present the results coherently in the form of a dendrogram. Depending on how similar or dissimilar the data instances are, hierarchicalclustering can be divided into two categories:

- Agglomerative Hierarchical Clustering: It works ina bottom-up fashion. Each data instance is considered a single element cluster to begin with. At each step of the algorithm, two most similar clusters are combined into one. The process continues until all instances have been combined into one big cluster containing all the data instances.
- Divisive Hierarchical Clustering: This method works complementary to the previous one. Here, all the data instances are considered a point of one big cluster at the beginning. At each iteration of the process, the most heterogeneous cluster is divided into two. The process continues until all the data instances are a single element cluster.

Agglomerative form of hierarchical clustering is more suited for more heterogeneous data that has a large number of small clusters to identify. On the other hand, divisive clustering as expected, is better at isolating big clusters. We use both these approaches depending on which molecule we are analyzing. The molecular datasets that are known to undergo large scale conformational changes like GroEl in this work was subjected to agglomerative hierarchical clustering and the others to the divisive one. To isolatethe clusters, we plot the results and the dendrogram tree suggests where it can be cut for major clusters. The function, *cutree* serves the purpose and the function *table* computes the size of each of these clusters. Then the clusters withat least 100 instances are picked. For example, in Figure 2 the height of the clusters is suggestive of their sizes, if the *cutree* function is called with an argument greater than 2, one of the two clusters splits in a way such that one child cluster has fewer than 100 points.

The entire workflow can be summed up as under:

1. Obtain a subset of the dataset generated by MD simulations using the PCA based data instance reduction.

2. Obtain the low dimensional embedding by performing Isomap on this subset.
3. Perform hierarchical clustering on this embedding and compare to the $Betti_0$ value for estimating the number of clusters.
4. Compute $Betti_1$ and $Betti_2$ of the clusters to establish the different topological properties of distinct clusters.

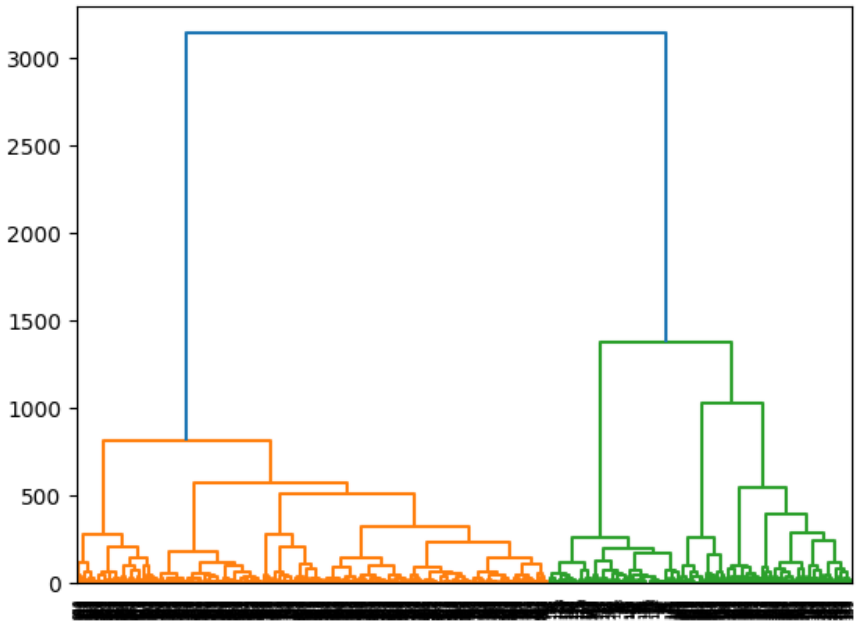


Figure 2. Clustering hierarchy for the inactive GDP-bound conforma- tion of Cdc42. The conformational space breaks into two clearly defined clusters

3. Results And Discussion

In this work we chose a number of medium to large proteins with different conformational dynamics. As mentioned earlier, to generate clusters from the reduced dimension embedding various methods were used depending on the molecular dataset. A thorough analysis for the various molecules is as under.

3.1. Cdc42

Cell Division Control-42 is a GTP (Guanosine Triphosphate) binding protein [57]. Human Cdc42 is a small molecule with 191 amino acids. It belongs to the *Rho* protein family, which regulates signaling pathways that control a wide range of cellular functions, for example, cell migration and cell cycle progression. It switches between cycles an active GTP-bound state and an inactive GDP (Guanosine Diphosphate)-bound state. Recently, Cdc42 has been shown to actively assist in cancer progression and metastasis. Several studies have established the basis for this and hypothesized about the underlying mechanisms [58, 59]. The data for all of the molecules was procured using Molecular Dynamics trajectories from a recent work by Haspel et al. [48].

Inactive form (GDP Bound): Both forms of hierarchical clustering produced the same number of clusters here, shown in Figure 2. There were two significant clusters. We call a cluster significant if it has at least 100 data instances. Figure 3 shows the embedding obtained by the feature reduction algorithm, where each cluster is shown in different color. To visualize the topology of the conformational landscape we used the kernel density estimation method and persistence diagram as described in the *TDAToolBox* package [60], Figure 4(a) shows the density estimates over the embedding, highlighting two peaks, Figure 4(b) shows the persistence diagram for the conformational space of this molecule. At the given ϵ (as defined and elaborated in section 1.3), here $Betti_0 = 2$, tantamount to two relevant clusters of this molecule. (Both methods of analysis: hierarchical clustering and algebraic topology produce corroborative results for this and all of the molecules in this study.).

One conformation representative of each of the significant clusters is shown in Figure-5 (a) and (b). Both these conformations are selected as the data point closest to the centroid of the two clusters. The PDB (Protein Data Bank) representations shown in Figure 3 were obtained using VMD (Visual Molecular Dynamics) [61]. Figure 5(c) shows a superimposition of the two clusters with the regions associated with binding and activation highlighted in color:The Switch-I region (residues 25-37) in blue, the Switch-II region (residues 57-75) in orange and the insert region (residues 122-135) in red. It can be shown that these three regions represent the highest variability, consistent

with thefact that they are involved with binding and activation [62,63]. These two clusters are then compared to each other to deduce conformational information.

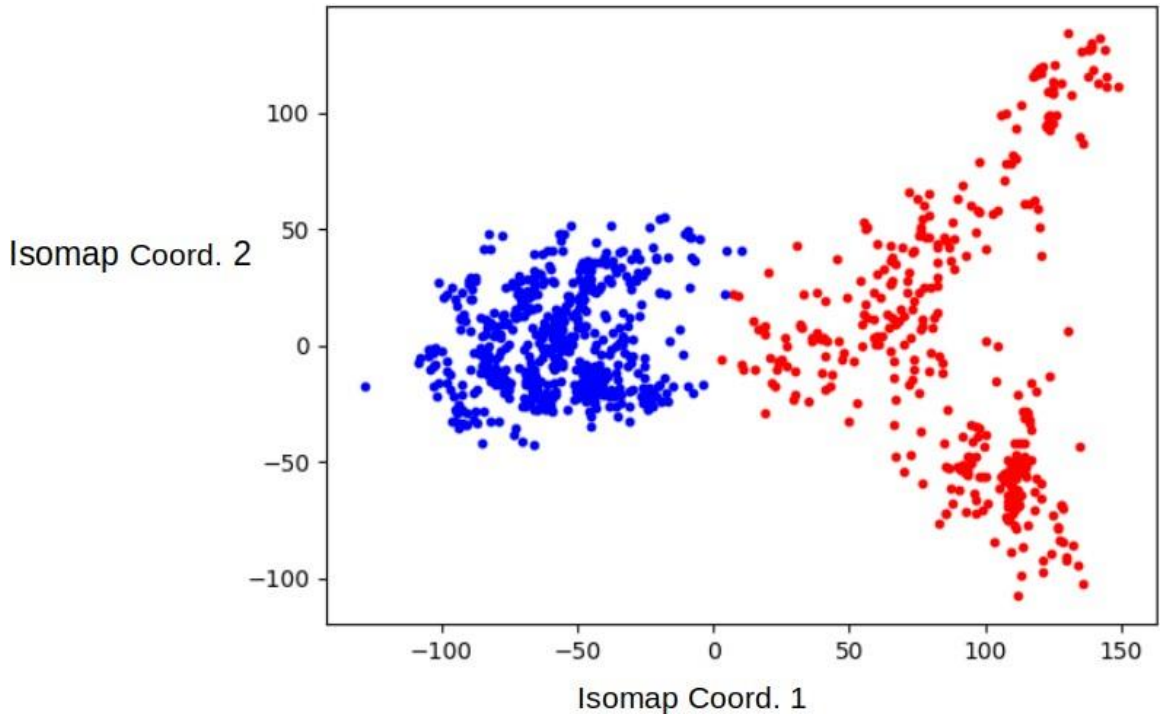
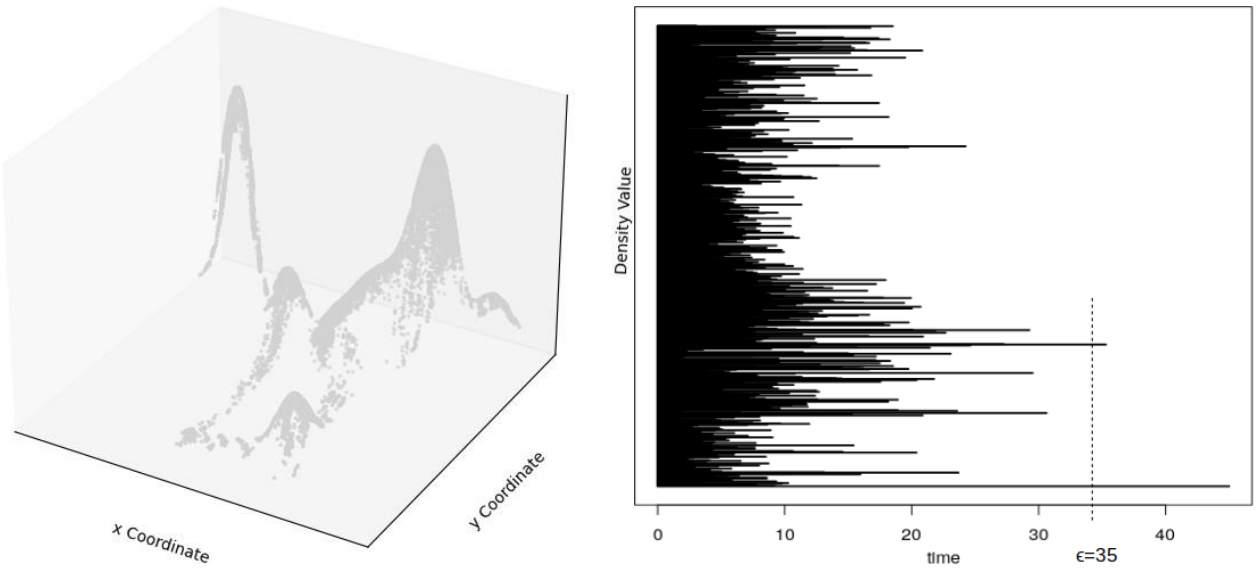


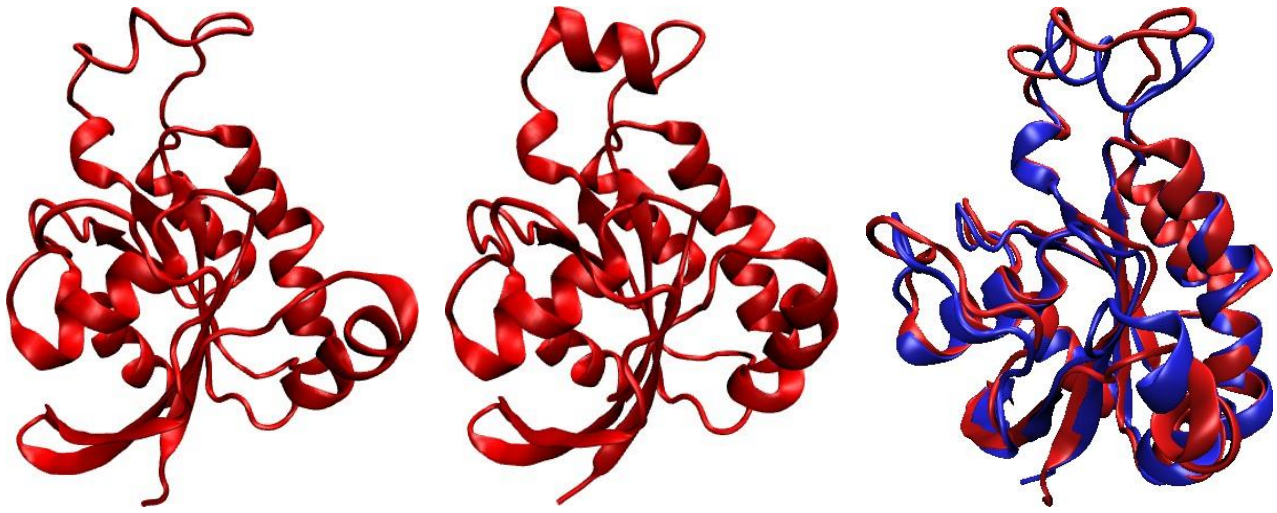
Figure 3. Embedding of the two clusters of the inactive Cdc42 (PDB:4did). Each cluster is highlighted in a different color



(a) **(b)**

Figure 4. Computations over Inactive Cdc42 (4did): (a) Density Esti- mate over 3-D space, (b) Persistence barcodes: two clusters persist at the given level of filtration for conformational space (ϵ = 35)

Next, we examined the topological features of each cluster separately. As mentioned earlier, all of the clusters (in all the molecules) are one connected component of a larger structure, so $Betti_0$ of all of them is 1. A quantitative representation of the persistence barcodes for the two clusters is shown in Figure-6. It is another way of visualizing persistent homology. Each feature is depicted by a single point with the horizontal axis representing feature birth andthe vertical axis representing feature death. The line $y = x$ isincluded as a reference. Since feature birth always precedes feature death ($x<y$), all points in a persistence diagramlie above the reference line. Just like topological barcodes, feature dimension is coded as the point's color, the red dots represent dimension 0 ($Betti_0$), the blue dots, dimension 1 ($Betti_1$) and the green dots represent dimension 2 ($Betti_2$).



(a) **(b)** **(c)**

Figure 5. (a) and (b) Two conformations of the inactive form of Cdc42 (PDB 4did)(c) A superimposition of the two conformations, the regions with maximum difference stand out at the top of the figure

In Figure 6, we see a number of 0-cycles (red dots); it is difficult to differentiate between most individual 0-cycles due to all of them being a part of the same connected component. However, there is a number of loops (blue dots)- in Figure 6 (b) three of which are significant andpersist in the entirety of the conformational space, there is also a green dot (a void) that persists. This represents the bigger, more branched and perforated cluster of Figure-3, the other cluster of this figure is smaller and structurally not as complex, as is also evident from its $Betti_1$ distribution in Figure-6(a), none of the loops in this conformation are significant enough to span the entire conformational space, and it has no voids (its $Betti_2 = 0$). The significant thing to note here is that both of these conformations have different distributions of the higher $Betti$ number at the same levelof filtration (ϵ) which means that conformations representedby these clusters are distinctively different.

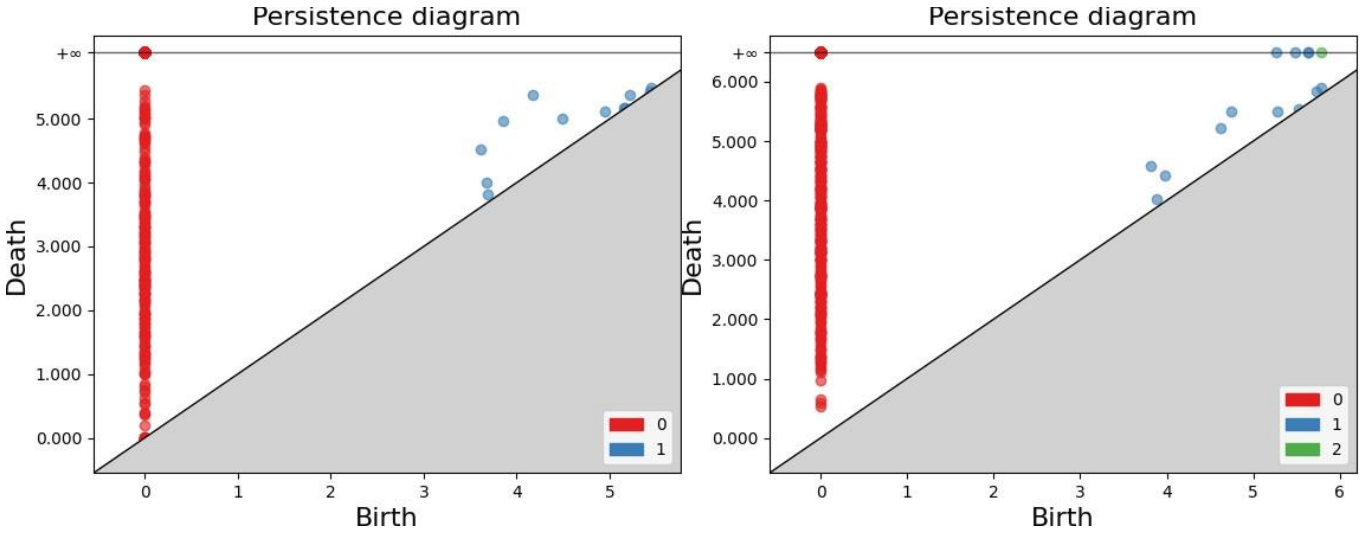


Figure 6. Quantitative representation of the persistence barcodes for the two clusters of the inactive (GDP bound) state of Cdc42: Red dots are $Betti_0$, the blue dots represent $Betti_1$, and green $Betti_2$

Active Form: This active (GTP bound) form of Cdc42 was subjected to divisive form of hierarchical clustering that indicated that the molecule has only one significant cluster. The clustering hierarchy, Isomap embedding, persistence diagram and the density estimates for this form of Cdc42 can be found respectively in Figure 7, all of which corroborate the existence of only one conformation. Figure 7 (a) clearly shows that there is only one significant cluster, the right partof the dashed line is a number of data points much less than100 and hence is not significant.

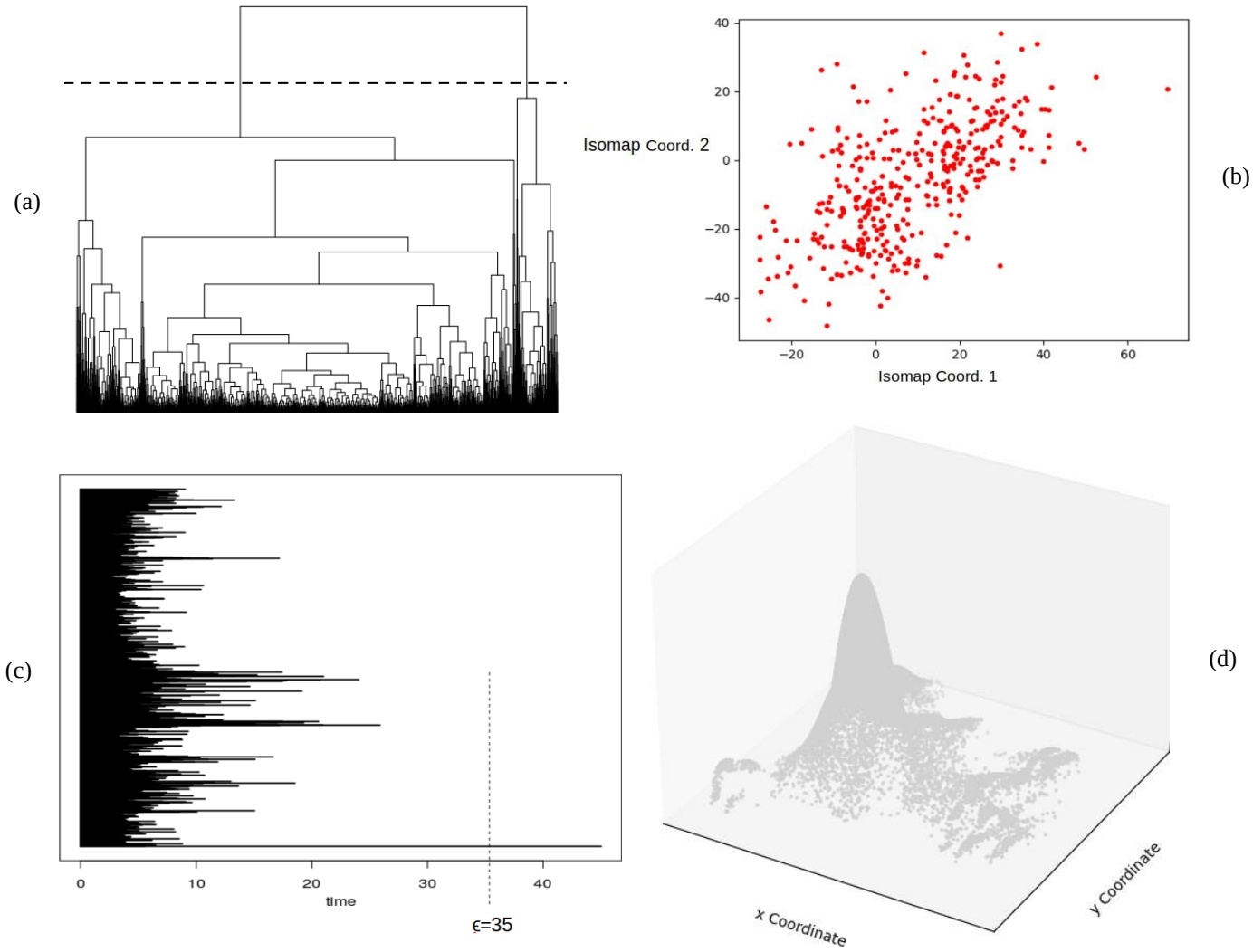


Figure 7. Active form of Cdc42-(PDB 4js0): (a)- Divisive Clustering hierarchy, (b)- Isomap Embedding, (c)- Persistence diagram with ϵ = 35, (d)- Density estimates

3.2. Oxytocin and Vasopressin

Oxytocin and Vasopressin are both small peptide hormones, both with 9 amino acids [64]. They differ only in two positions in their amino acid sequence, as shown under:

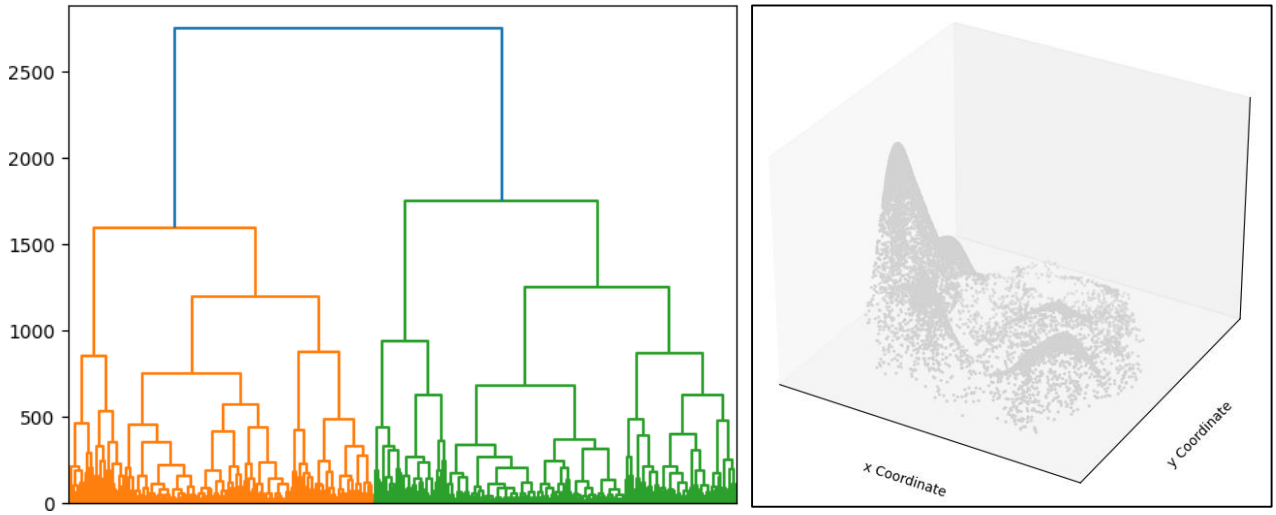
Oxytocin: Cys-Tyr-**Ile**-Gln-Asn-Cys-Pro-**Leu**-Gly

Vasopressin: Cys-Tyr-**Phe**-Gln-Asn-Cys-Pro-**Arg**-Gly

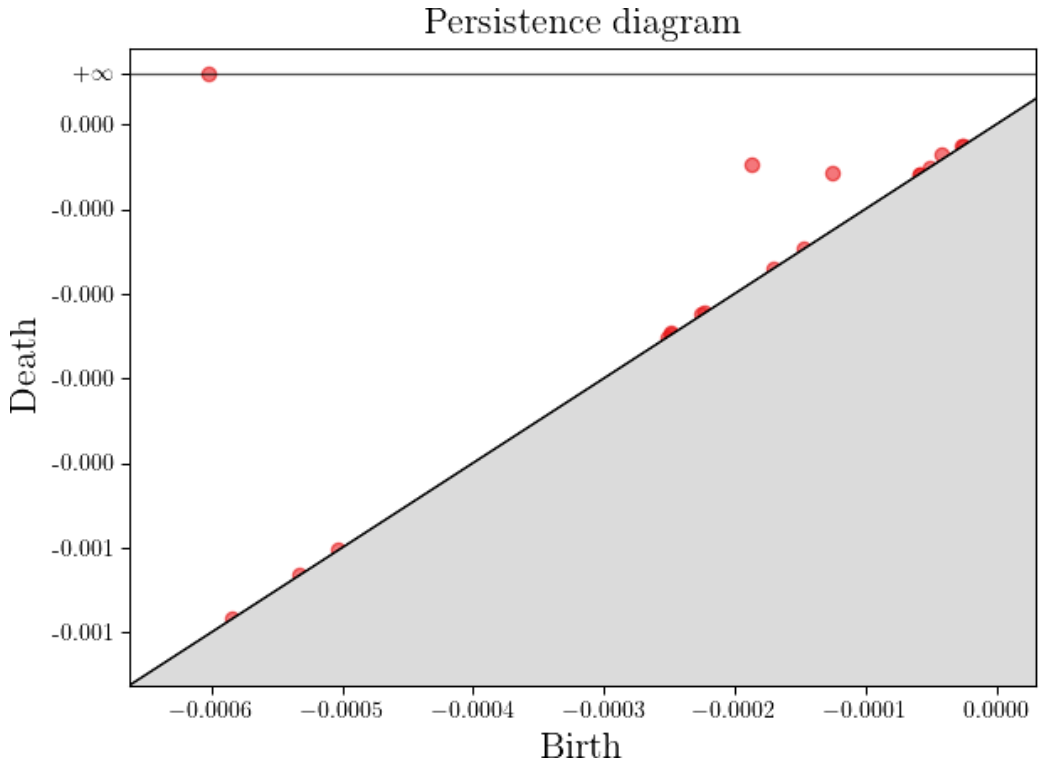
These are smaller peptides that do not have one stable structure. So, intuitively, divisive hierarchical clustering works better to sample conformations here. Nonetheless, for oxytocin we performed agglomerative clustering and obtained a number of clusters but they were very small and exhibited the same higher *Betti* numbers. The first three *Betti* numbers of these five clusters are in Table 1. These numbers indicate that there really are just two different clusters here: a result that was obtained when divisive form of hierarchical clustering was performed which resulted in just two significant clusters. This shows that the choice of the type of hierarchical clustering is critical to the study of conformational landscape. The hierarchy produced by divisive clustering is shown in Figure 8(a). As seen in Figure 8(b), the kernel density estimations for Oxytocin also resulted in two peaks (one major and another smaller that protrudes from it). Figure 8(c) the persistence diagram for the entire conformational space shows three red dots that ultimately represent two clusters - one big and one small, smaller of which is made of two even smaller clusters that merged during the process. The persistence diagrams for the two clusters of Oxytocin and the one in Vasopressin are shown in Figure 9. The first cluster of oxytocin had two loops and no voids. The second cluster on the other hand had three loops and a void. For Vasopressin, there was only one significant cluster produced by divisive hierarchical clustering. The only cluster in Vasopressin is close in its topological structure with the known Oxytocin conformation which was expected given the similarity in the sequences of the amino acids of the two molecules. The conformations representative of the two clusters of Oxytocin and the one of Vasopressin are shown in Figure 10.

Table 1. *Betti* Numbers for clusters of Oxytocin produced by AgglomerativeHierarchical Clustering

Cluster No.	$Betti_0$	$Betti_1$	$Betti_2$
1	1	1	0
2	1	1	0
3	1	1	0
4	1	4	0
5	1	0	0



(a) **(b)**



(c)

Fig. 8. Oxytocin: (a)- Divisive Clustering hierarchy, (b)- Density Estimate, (c)- the persistence diagram, showing one persistent cluster and two others (on the top right) that soon merge into the main one

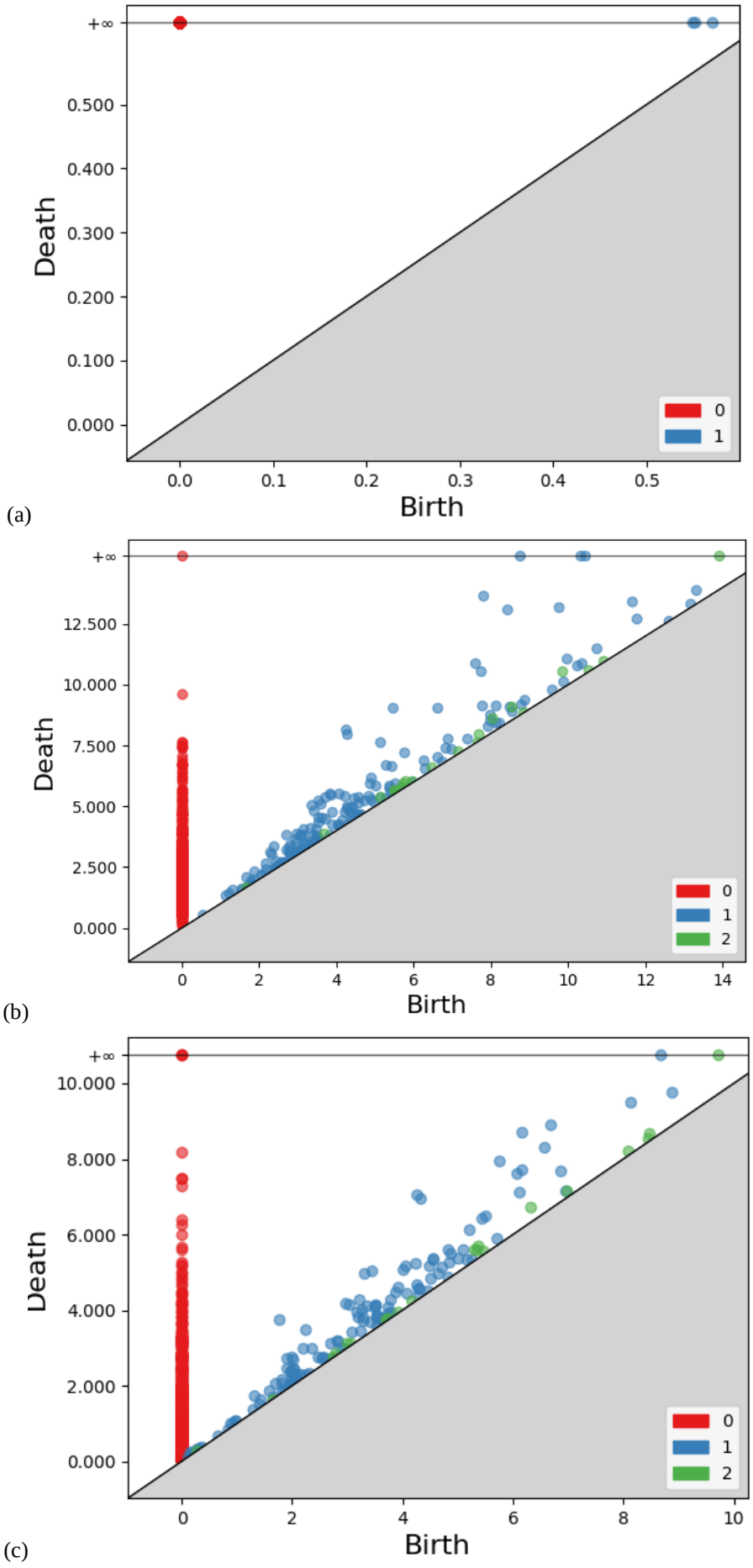
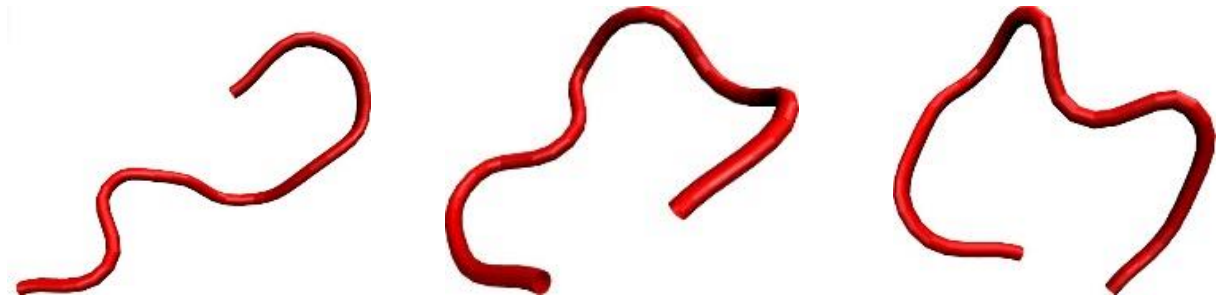


Figure 9. Persistence Diagrams of: (a)- Oxytocin, first cluster, (b)- Oxytocin, second cluster, (c) Vasopressin. (b) and (c) have similar high dimensional topological features



(a) (b) (c)

Figure 10. PDB Conformations of: (a)- Oxytocin, first cluster, (b)- Oxytocin, second cluster, (c) Vasopressin

3.3. Human and Porcine Galanin

Galanin is a neuropeptide which is encoded by the GAL gene. The gene is expressed in the parts of central nervous system, and gut of humans and some other mammals [65]. Porcine Galanin (pGalanin) and Human Galanin (hGalanin) are two variants of Galanin that differ by 5 outof their amino acids. pGalanin contains 29 amino acids and hGalanin has 30. Both of these molecules also do not have one stable 3D structure due to their small size. Divisive hierarchical clustering on hGalanin resulted in two major clusters (shown in Figure 11) and that of pGalanin producedonly one cluster.

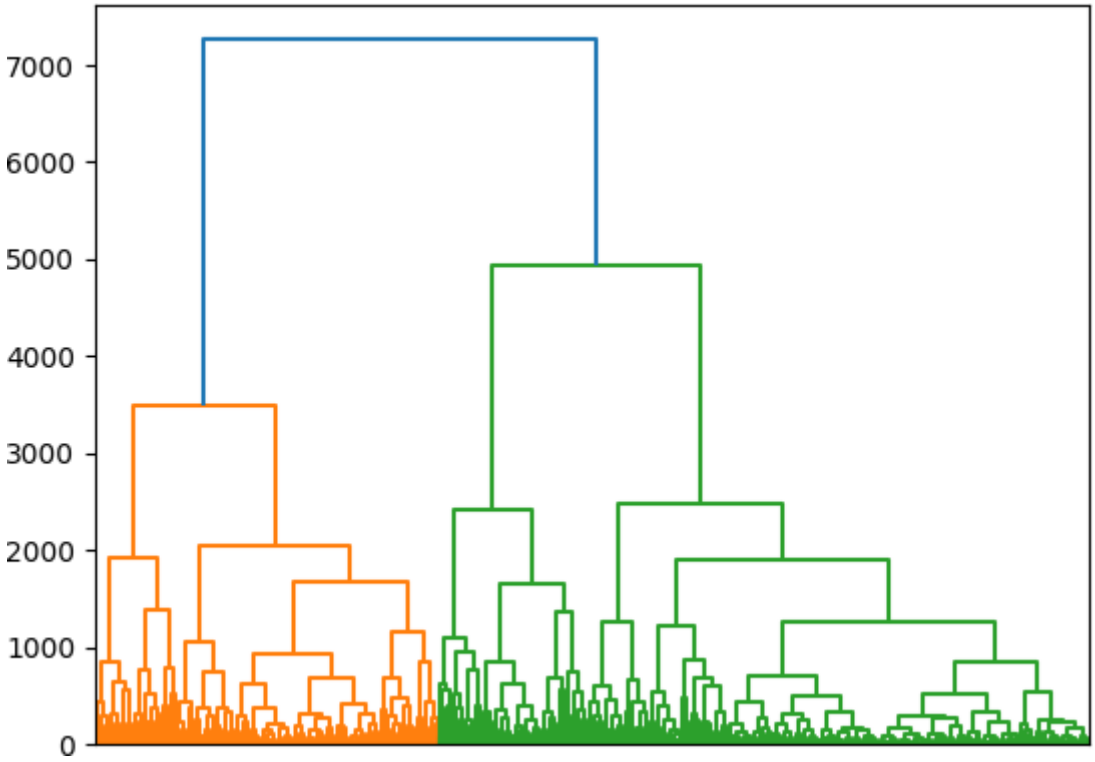
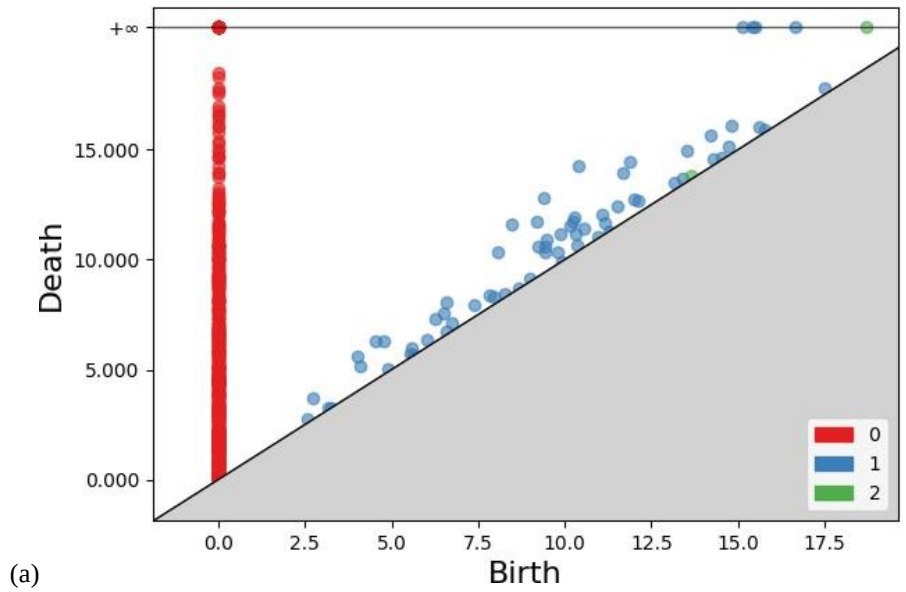
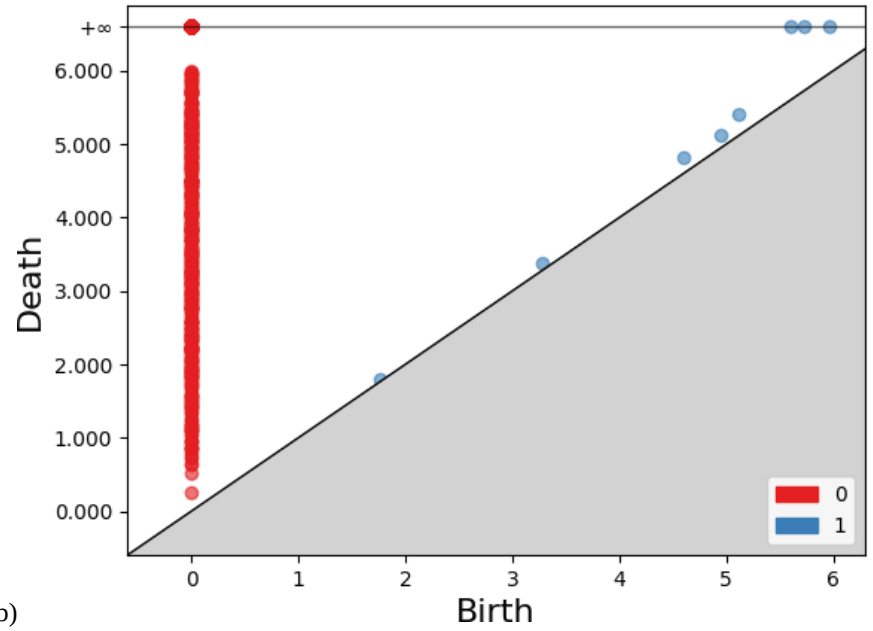


Figure 11. Divisive clustering hierarchy of Human Galanin

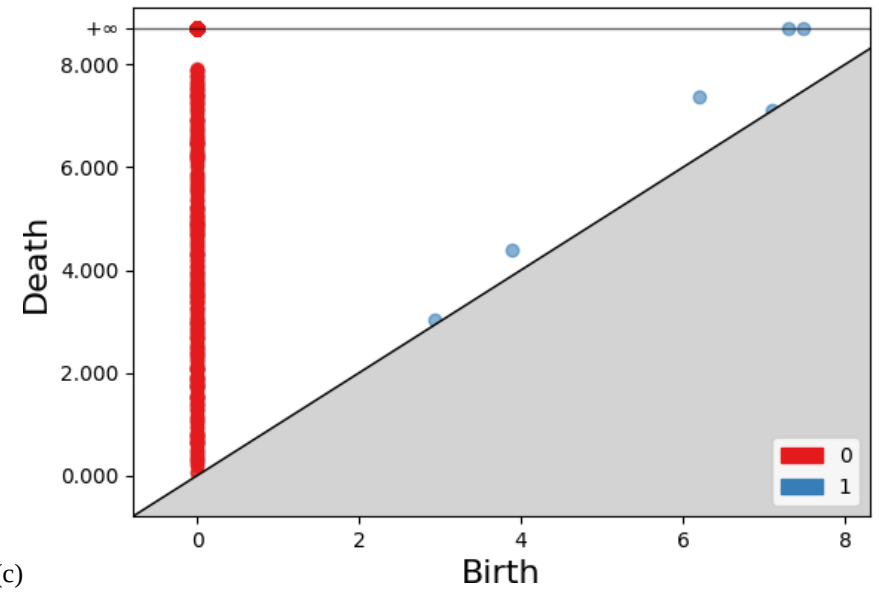
Topologically, the two clusters of hGalanin are quite different from each other, shown in Figures 12(a) and (b), which is in agreement with experimental results produced by Holst et al [66]. The persistence diagram for pGalaninis shown in Figure 12(c). As is evident from the persistence diagrams, the second cluster of hGalanin and the pGalanin cluster are very similar. The study of hGalanin in [66] also shows that there are two molecular forms of hGalanin, one of 30 and another of 19 amino acids. The larger ofthe two peptides has a sequence which is identical to thatof pGalanin except for the following substitutions: Val16, Asn17, Asn26, Thr29 and Ser30. The PDB structures of these clusters is shown in Figure 13, further corroborating that there are two conformations for hGalanin and one of themis structurally similar to that of pGalanin.



(a)

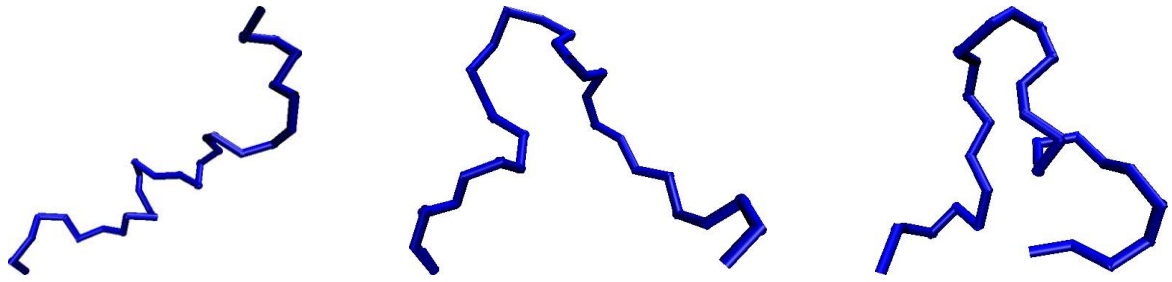


(b)



(c)

Figure 12. Persistence diagrams for the (a)- first cluster of hGalanin, (b)- second cluster of hGalanin, (c)- pGalanin cluster

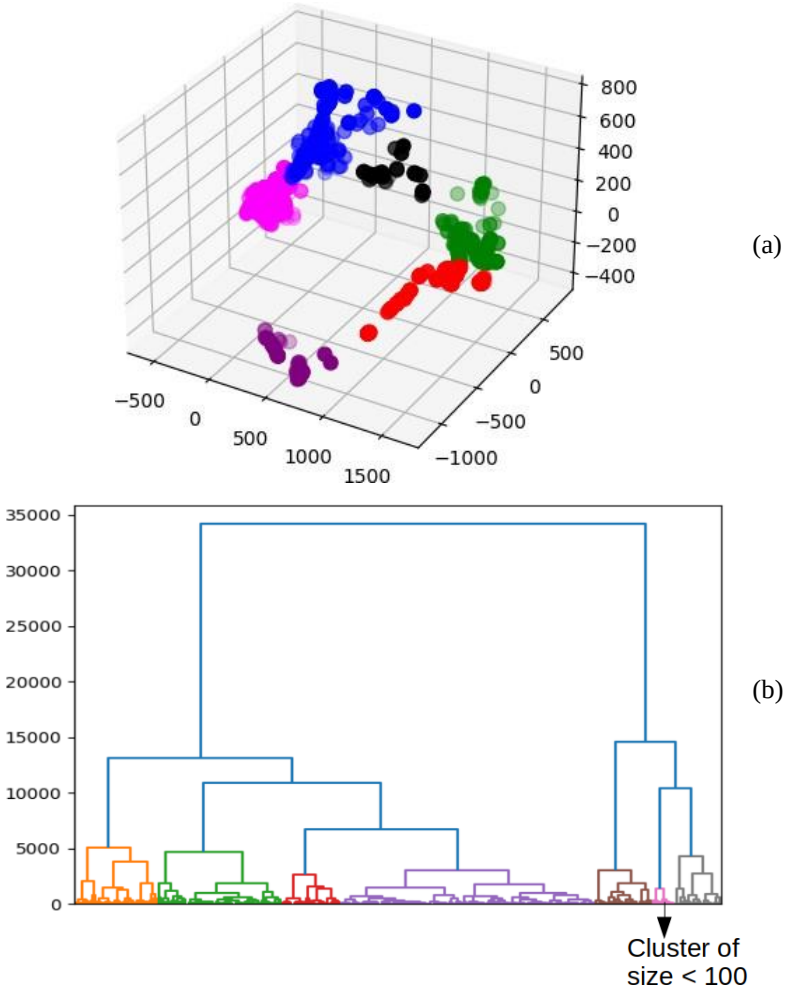


(a) (b) (c)

Figure 13. Conformations of Galanins: (a-b)- Conformations representa- tive of the two clusters of hGalanin, (c) Conformation representative of the single pGalanin cluster

3.4. GroEL

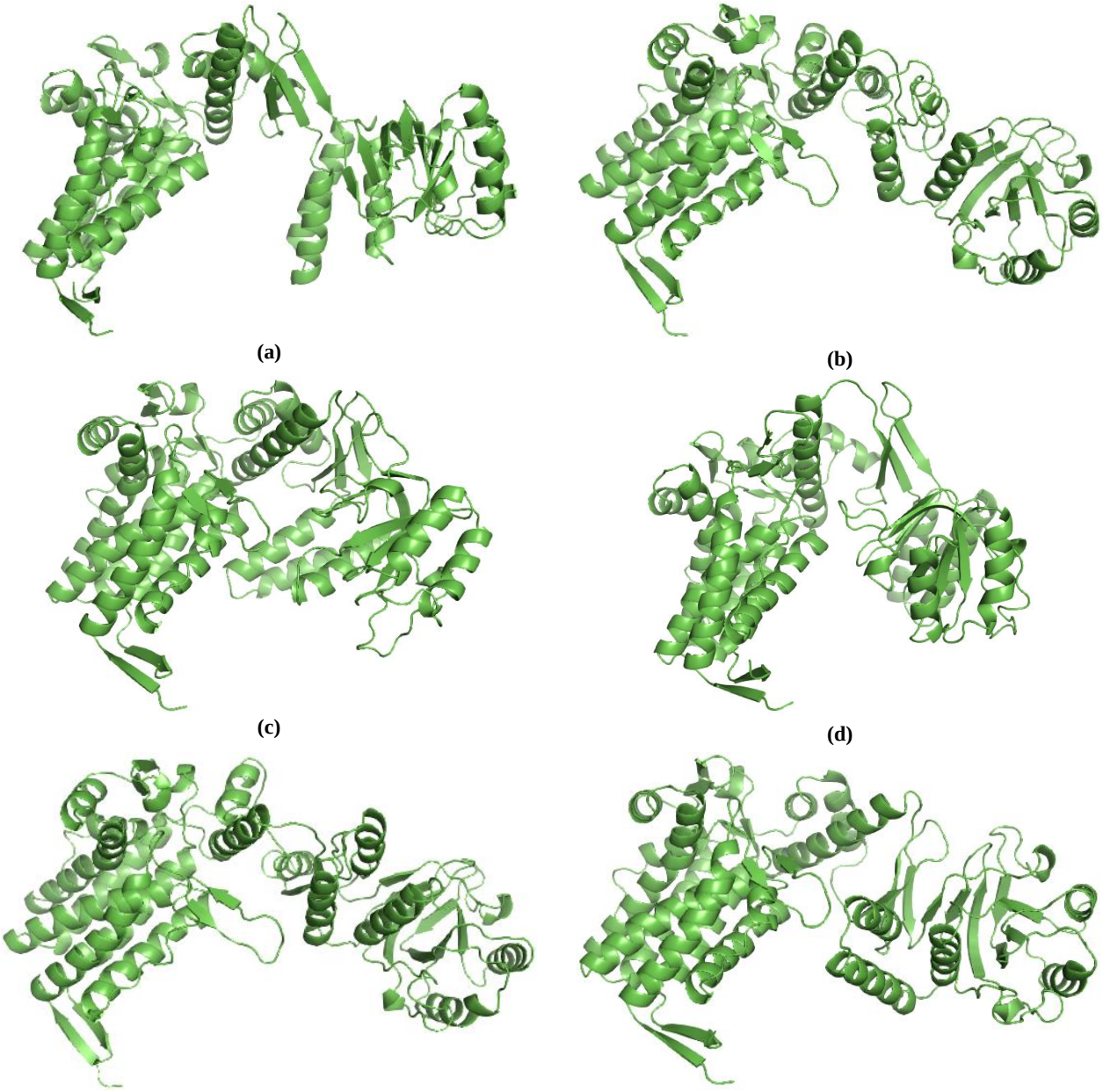
GroEl is a chaperonin protein which is needed by many other proteins for their proper folding. Structurally, GroEL is a dual-ringed tetradecamer, with both the *cis* and *trans* rings consisting of seven subunits each. The conformational changes that occur within the central cavity of GroEL cause for the inside of GroEL to become hydrophilic, rather than hydrophobic, and is likely what facilitates protein folding. GroEL requires the co-chaperonin protein complex, GroES. Binding of substrate protein, in addition to binding of ATP, induces an extensive conformational change that allows association of the binary complex with GroES. It is the heaviest molecule in the study with 524 amino acids. Itis known to undergo large scale conformational changes and hence agglomerative form of hierarchical clustering wasused here. It helped in identifying six different clusters (asis corroborated by [3]), isomap embedding and clustering hierarchy of which is shown in Figure 14(a) and (b) respectively. This is the most diverse molecule in the study and has a complex more perforated conformational space, and hence the isomap embedding is shown in three dimensions to highlight the different clusters. The conformations representative of each of these clusters are shown in Figure 15. Looking at the clusters, one can see the opening and closing along the hinge. The persistence diagrams for thesix individual clusters can be found in Figure 16.



(a)

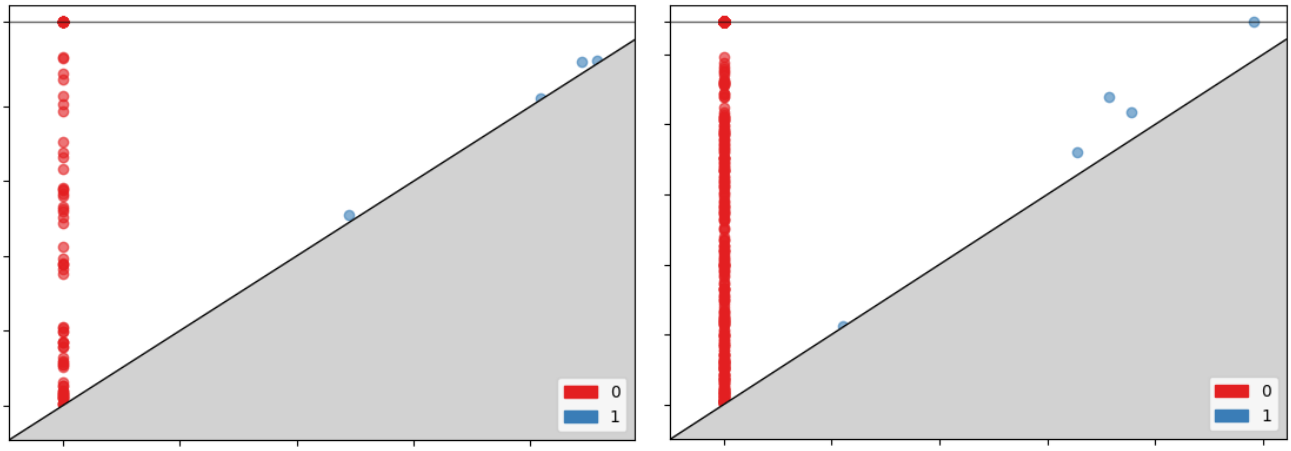
(b)

Figure 14. (a)- Embedding of the six clusters of GroEL. Each cluster is highlighted in a different color. X, Y, Z axes are the respective Isomap coordinates, (b)- Agglomerative clustering hierarchy, the insignificant cluster was taken off the analysis



(a) **(b)** **(c)** **(d)** **(e)** **(f)**

Figure 15. Representatives of the six clusters obtained for the GroEL monomer



(a) (b)

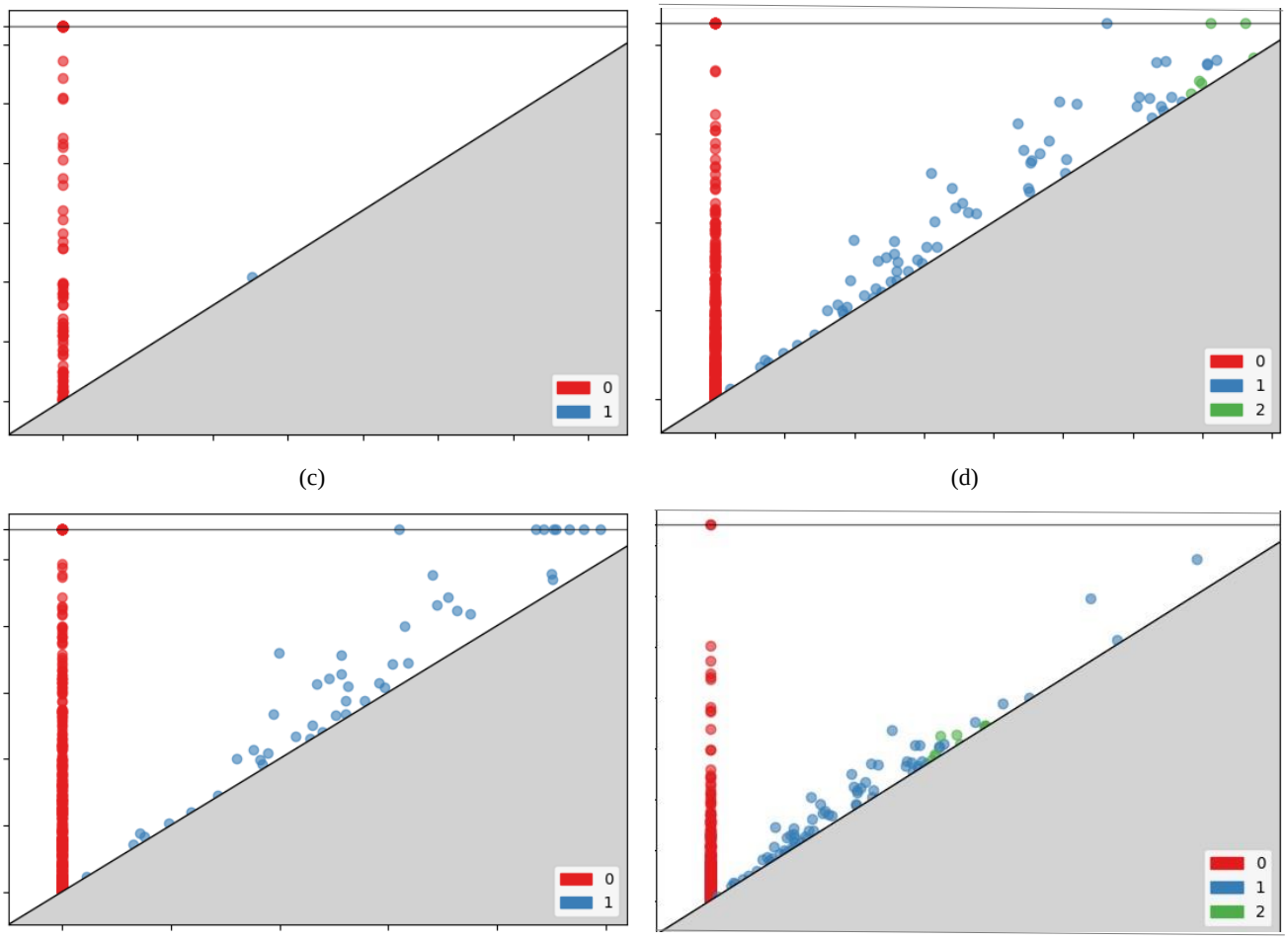
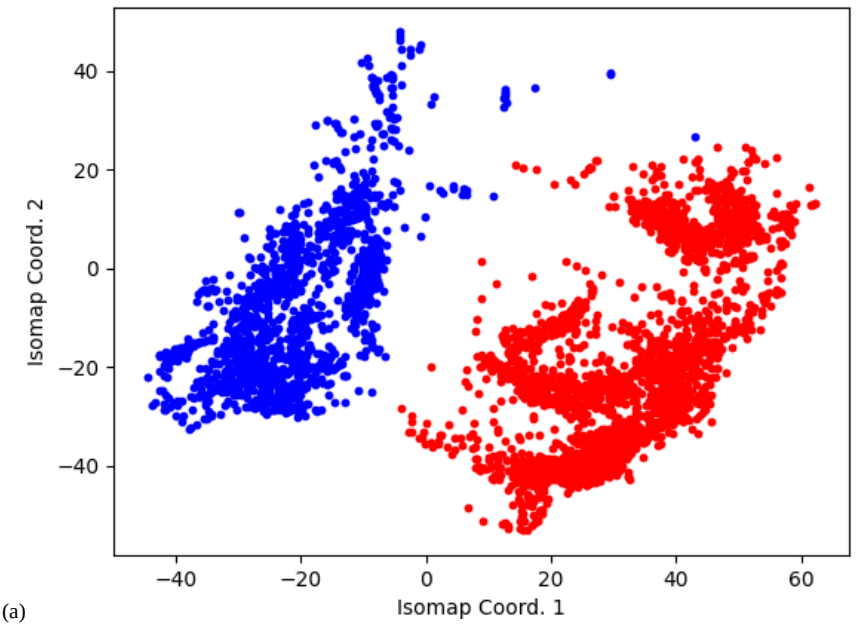


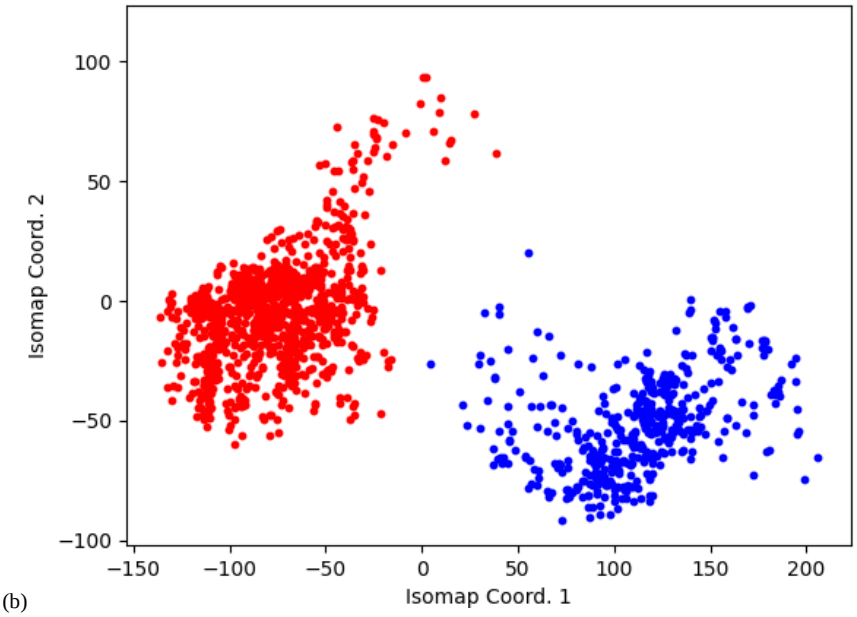
Figure 16. Persistence diagrams for the six clusters of GroEL

4. Analysis of Results

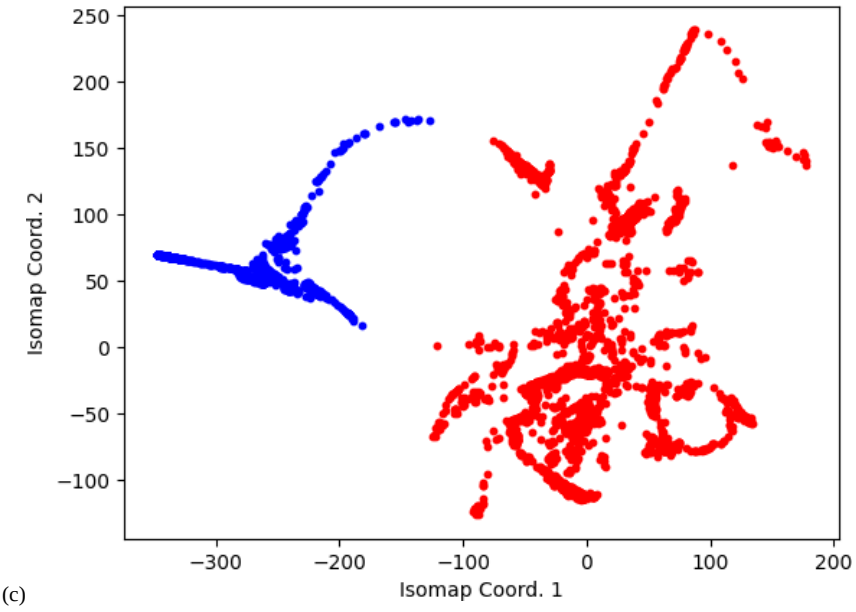
Results above establish that the entire methodology of using dimesionality reduction, hierarchical clustering and topological analysis helps in sampling the conformational landscape of a molecule in a way that truly distinct conformations are identified. As mentioned earlier in the section on generation of data, the data for most molecules here was generated using molecular dynamics simulations; they also yield the potential energy of each simulated conformation, taking into account the relative three dimensional placementof each atom that makes the conformation. Topological analysis of the entire energy space is the same as computing persistent homology of the entire embedding generated after feature reduction (which means without filtering the conformations using instance reduction). To explore the energy landscape of the conformational space we filtered the conformations for three molecules (GDP bound Cdc42, Oxytocin and hGalanin, the ones that were generated using MDsimulations and were hypothesized to have more than one persistent conformation) based on their energy. We retained only the conformations that have energy less than the 80% of conformations, in other words, we performed filtering at 20% filtration of energy. To do this, we simply sorted the conformations based on their energy and picked the first 20%. The embedding of these molecules at this filtration is shown in Figure17. As is clear from the embedding, it can be seen that even at such low levels of energy filtration, the molecules separate into two clusters, as suggested by topological analysis. This indicates that the method described inthis work to characterize conformational space of proteinsis capable of sampling low energy conformations of proteinmolecules that are likely to persist.



(a)



(b)



(c)

Figure 17. Embeddings of 20% lower energy conformations of: (a)- GDP bound Cdc42 (b)- Oxytocin, (c)- Human Galanin

5. Conclusion

Many proteins undergo large-scale conformational changes as part of their function. Characterizing the conformational space of proteins is crucial for understanding their function and dynamics. We present an efficient filtration procedure that works well for sampling the intermediate conformations for protein molecules that undergo large-scale conformational changes as well as for the ones that have a pervasive native state. The method presented is well suited to establish distinctiveness among the clusters generated. Analysis of low energy conformation clusters using dimensionality reduction and algebraic topology and observing its effect with varying energy levels may tell us how the topology of the space evolves as the protein goes through high energy barriers. It can produce interesting results and help in designing of conformational landscapes for targeted drugs. Refining these methods to produce finer samples and study the significance of fleeting high energy conformations is the goal ahead. Hierarchical clustering merges (or splits) two sets of conformations based on the heterogeneity of the entire set. In the future, we aim to develop a method to bias this divide in a way that is more suited for protein datasets. Molecules that undergo rigorous changes in structure are the ones more suited for such work. In particular, having known intermediates would help in guiding the conformational pathway search problem as well. It can divide the search space into smaller instances of the same problem. This is also a portion of the ongoing research.

6. Declarations

6.1. Author Contributions

Conceptualization, A.J., N.H., and E.G.; methodology, A.J., N.H., and E.G.; software, A.J. and N.H.; validation, A.J. and N.H.; formal analysis, A.J. and N.H.; investigation, A.J. and N.H.; resources, A.J., N.H., and EG; data curation, N.H. and A.J.; writing—original draft preparation, A.J., N.H., and E.G.; writing—review and editing, A.J., N.H., and E.G.; visualization, A.J. and N.H.; supervision, N.H. and E.G. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Ethical Approval

Not applicable.

6.5. Declaration of Competing Interest

The authors declare that there is no conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

7. References

[1] Miyashita, O., Wolynes, P. G., & Onuchic, J. N. (2005). Simple energy landscape model for the kinetics of functional transitions in proteins. Journal of Physical Chemistry B, 109(5), 1959–1969. doi:10.1021/jp046736q.

[2] Haspel, N., Moll, M., Baker, M. L., Chiu, W., & Kavraki, L. E. (2010). Tracing conformational changes in proteins. BMC Structural Biology, 10(SUPPL. 1), 1. doi:10.1186/1472-6807-10-S1-S1.

[3] Haspel, N., Luo, D., & González, E. (2017). Detecting intermediate protein conformations using algebraic topology. BMC Bioinformatics, 18(Suppl 15), 502. doi:10.1186/s12859-017-1918-z.

[4] Bryngelson, J. D., Onuchic, J. N., Socci, N. D., & Wolynes, P. G. (1995). Funnels, pathways, and the energy landscape of protein folding: A synthesis. Proteins: Structure, Function, and Bioinformatics, 21(3), 167–195. doi:10.1002/prot.340210302.

[5] Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B., & Woods, R. J. (2005). The Amber biomolecular simulation programs. Journal of Computational Chemistry, 26(16), 1668–1688. doi:10.1002/jcc.20290.

[6] Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. Science, 220(4598), 671–680. doi:10.1126/science.220.4598.671.

[7] Raveh, B., Enosh, A., Schueler-Furman, O., & Halperin, D. (2009). Rapid sampling of molecular motions with prior information constraints. PLoS Computational Biology, 5(2), 1000295. doi:10.1371/journal.pcbi.1000295.

[8] Shehu, A., & Olson, B. (2010). Guiding the search for native-like protein conformations with an Ab-initio tree-based exploration. International Journal of Robotics Research, 29(8), 1106–1127. doi:10.1177/0278364910371527.

[9] Al-Bluwi, I., Vaisset, M., Siméon, T., & Cortés, J. (2013). Modeling protein conformational transitions by a combination of coarse-grained normal mode analysis and robotics-inspired methods. BMC Structural Biology, 13(SUPPL.1), 2. doi:10.1186/1472-6807-13-S1-S2.

[10] Molloy, K., & Shehu, A. (2016). A General, Adaptive, Roadmap-Based Algorithm for Protein Motion Computation. IEEE Transactions on Nanobioscience, 15(2), 160–167. doi:10.1109/TNB.2016.2519246.

[11] Zheng, W., & Brooks, B. (2005). Identification of dynamical correlations within the myosin motor domain by the normal mode analysis of an elastic network model. Journal of Molecular Biology, 346(3), 745–759. doi:10.1016/j.jmb.2004.12.020.

[12] Yang, L., Song, G., & Jernigan, R. L. (2009). Protein elastic network models and the ranges of cooperativity. Proceedings of the National Academy of Sciences of the United States of America, 106(30), 12347–12352. doi:10.1073/pnas.0902159106.

[13] Xia, K., Opron, K., & Wei, G. W. (2015). Multiscale Gaussian network model (mGNM) and multiscale anisotropic network model (mANM). Journal of Chemical Physics, 143(20), 204106. doi:10.1063/1.4936132.

[14] Schröder, G. F., Brunger, A. T., & Levitt, M. (2007). Combining Efficient Conformational Sampling with a Deformable Elastic Network Model Facilitates Structure Refinement at Low Resolution. Structure, 15(12), 1630–1641. doi:10.1016/j.str.2007.09.021.

[15] Frappier, V., Chartier, M., & Najmanovich, R. J. (2015). ENCoM server: Exploring protein conformational space and the effect of mutations on protein function and stability. Nucleic Acids Research, 43(W1), W395–W400. doi:10.1093/nar/gkv343.

[16] Weiss, D. R., & Levitt, M. (2009). Can Morphing Methods Predict Intermediate Structures? Journal of Molecular Biology, 385(2), 665–674. doi:10.1016/j.jmb.2008.10.064.

[17] Castellana, N. E., Lushnikov, A., Rotkiewicz, P., Sefcovic, N., Pevzner, P. A., Godzik, A., & Vyatkina, K. (2013). MORPH-PRO: A novel algorithm and web server for protein morphing. Algorithms for Molecular Biology, 8(1), 19. doi:10.1186/1748-7188-8-19.

[18] Vetro, R., Haspel, N., & Simovici, D. (2013). Characterizing intermediate conformations in protein conformational space. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 7845 LNBI, 70–80. doi:10.1007/978-3-642-38342-7_7.

[19] Chang, H. W., Bacallado, S., Pande, V. S., & Carlsson, G. E. (2013). Persistent Topology and Metastable State in Conformational Dynamics. PLoS ONE, 8(4), 58699. doi:10.1371/journal.pone.0058699.

[20] Gan, G., & Wu, J. (2004). Subspace clustering for high dimensional categorical data. ACM SIGKDD Explorations Newsletter, 6(2), 87–94. doi:10.1145/1046456.1046468.

[21] Karplus, M., & Shakhnovich, E. (1992). Protein Folding: Theoretical Studies of Thermodynamics and Dynamics. In Protein Folding, 127–196, W. H. Freeman and Company, New York, United States.

[22] Bryngelson, J. D., Onuchic, J. N., Socci, N. D., & Wolynes, P. G. (1995). Funnels, pathways, and the energy landscape of protein folding: A synthesis. Proteins: Structure, Function, and Genetics, 21(3), 167–195. doi:10.1002/prot.340210302.

[23] Wilson, D. R., & Martinez, T. R. (2000). Reduction techniques for instance-based learning algorithms. Machine Learning, 38(3), 257–286. doi:10.1023/A:1007626913721.

[24] Arnaiz-González, Á., Díez-Pastor, J. F., Rodríguez, J. J., & García-Osorio, C. (2016). Instance selection of linear complexity for big data. Knowledge-Based Systems, 107, 83–95. doi:10.1016/j.knosys.2016.05.056.

[25] García, S., Derrac, J., Cano, J. R., & Herrera, F. (2012). Prototype selection for nearest neighbor classification: Taxonomy and empirical study. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(3), 417–435. doi:10.1109/TPAMI.2011.142.

[26] Czarnowski, I., & Jędrzejowicz, P. (2006). Instance reduction approach to machine learning and multi-database mining. Annales Universitatis Mariae Curie-Skłodowska, sectio AI–Informatica, 4(1)-60-71.

[27] Son, S.-H., & Kim, J.-Y. (2006). Data Reduction for Instance-Based Learning Using Entropy-Based Partitioning. Lecture Notes in Computer Science, 590–599. doi:10.1007/11751595_63

[28] Boyd, S., Boyd, S. P., & Vandenberghe, L. (2004). Convex optimization. Cambridge University Press, Cambridge, United Kingdom.

[29] Maaten, L., Postma, E., & Herik, J. (2009). Dimensionality reduction: a comparative review. Journal of Machine Learning Research, 10, 1–36.

[30] Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. Science, 290(5500), 2319–2323. doi:10.1126/science.290.5500.2319.

[31] Das, P., Moll, M., Stamati, H., Kavraki, L. E., & Clementi, C. (2006). Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. Proceedings of the National Academy of Sciences, 103(26), 9885–9890. doi:10.1073/pnas.0603553103.

[32] Vajdi, A., Haspel, N., & Banaee, H. (2015). A new DP algorithm for comparing gene expression data using geometrics similarity. Proceedings - 2015 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2015, 1745–1747. doi:10.1109/BIBM.2015.7359948.

[33] Silva, V., & Tenenbaum, J. (2002). Global versus local methods in nonlinear dimensionality reduction. Advances in neural information processing systems, 15 (NIPS 2002), 1-8.

[34] Talwalkar, A., Kumar, S., & Rowley, H. (2008). Large-scale manifold learning. 2008 IEEE Conference on Computer Vision and Pattern Recognition. doi:10.1109/cvpr.2008.4587670.

[35] Adams, H., Tausz, A., & Vejdemo-Johansson, M. (2014). javaPlex: A research software package for persistent (co)homology. In H. Hong & C. Yap (Eds.), Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 8592 LNCS (pp. 129–136). doi:10.1007/978-3-662-44199-2_23.

[36] Watanabe, S., & Yamana, H. (2020). Deep Neural Network Pruning Using Persistent Homology. 2020 IEEE Third International Conference on Artificial Intelligence and Knowledge Engineering (AIKE). doi:10.1109/aike48582.2020.00030.

[37] Dindin, M., Umeda, Y., & Chazal, F. (2020). Topological Data Analysis for Arrhythmia Detection through Modular Neural Networks. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12109 LNAI, 177–188. doi:10.1007/978-3-030-47358-7_17.

[38] The GUDHI Project (2015). GUDHI User and Reference Manual. GUDHI Editorial Board, 2015. Available online: http://gudhi.gforge.inria.fr/doc/latest/ (accessed on March 2022).

[39] Cang, Z., Munch, E., & Wei, G.-W. (2020). Evolutionary homology on coupled dynamical systems with applications to protein flexibility analysis. Journal of Applied and Computational Topology, 4(4), 481–507. doi:10.1007/s41468-020-00057-9

[40] Cámara, P. G. (2017). Topological methods for genomics: Present and future directions. Current Opinion in Systems Biology, 1, 95–101. doi:10.1016/j.coisb.2016.12.007.

[41] Wei, G.-W. (2017). Persistent homology analysis of biomolecular data. Society for Industrial and Applied Mathematics, 2017. Available online: https://sinews.siam.org/Details-Page/persistent-homology-analysis-of-biomolecular-data (accessed on March 2022).

[42] Cang, Z., Mu, L., & Wei, G. W. (2018). Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. PLoS Computational Biology, 14(1), 1005929. doi:10.1371/journal.pcbi.1005929.

[43] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera - A visualization system for exploratory research and analysis. Journal of Computational Chemistry, 25(13), 1605–1612. doi:10.1002/jcc.20084.

[44] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics, 79(2), 926–935. doi:10.1063/1.445869.

[45] Darden, T., York, D., & Pedersen, L. (1993). Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. The Journal of Chemical Physics, 98(12), 10089–10092. doi:10.1063/1.464397.

[46] Kalé, L., Skeel, R., Bhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N., Phillips, J., Shinozaki, A., Varadarajan, K., & Schulten, K. (1999). NAMD2: Greater Scalability for Parallel Molecular Dynamics. Journal of Computational Physics, 151(1), 283–312. doi:10.1006/jcph.1999.6201.

[47] Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J., & Kollman, P. (2003). A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations. Journal of Computational Chemistry, 24(16), 1999–2012. doi:10.1002/jcc.10349.

[48] Haspel, N., Jang, H., & Nussinov, R. (2021). Active and Inactive Cdc42 Differ in Their Insert Region Conformational Dynamics. Biophysical Journal, 120(2), 306–318. doi:10.1016/j.bpj.2020.12.007.

[49] Luo, D., & Haspel, N. (2013). Multi-resolution rigidity-based sampling of protein conformational paths. In 2013 ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics, ACM-BCB 2013 (pp. 786–792). doi:10.1145/2506583.2506710.

[50] Candès, E. J., Li, X., Ma, Y., & Wright, J. (2011). Robust principal component analysis? Journal of the ACM, 58(3), 1–37. doi:10.1145/1970392.1970395.

[51] Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T., Cohen, K. L., … Cohen, K. L. (1999). Robust principal component analysis for functional data. Test, 8(1), 1–73. doi:10.1007/bf02595862.

[52] Fujiki, J. (2007). Spherical PCA with Euclideanization. ACCV'07 Workshop Subspace, November, Tokyo, 61–68.

[53] Joshi, A., & Haspel, N. (2020). A Novel Data Instance Reduction Technique using Linear Feature Reduction. Journal of Artificial Intelligence and Systems, 2, 191–206. doi:10.33969/ais.2020.21012.

[54] Joshi, A. (2019). High Performance Computing Techniques To Better Understand Protein Conformational Space. Ph.D. dissertation, University of Massachusetts, Boston, United State

[55] Joshi, A., & Haspel, N. (2019). Clustering of Protein Conformations Using Parallelized Dimensionality Reduction. Journal of Advances in Information Technology, 10(4), 142–147. doi:10.12720/jait.10.4.142-147.

[56] Wadhwa, R. R., Williamson, D. F., Dhawan, A., & Scott, J. G. (2018). Introduction to persistent homology with tdastats. The Journal of Open Sorce Software. Available online: https://cran.r-project.org/web/packages/TDAstats/vignettes/intro.html (accessed on March 2022).

[57] Valds-Mora, F., Pulgar, T. G., & Lacal, J. C. Translational Oncology Unit CSIC-UAM- La Paz Centro Nacional de Biotecnologia C/Darwin 3, Campus de Cantoblanco, 28049 Madrid, Spain. Available online: http://atlasgeneticsoncology.org/Genes/CDC42ID40012ch1p36.html

[58] Hartman, M. A., & Spudich, J. A. (2012). The myosin superfamily at a glance. Journal of Cell Science, 125(7), 1627–1632. doi:10.1242/jcs.094300.

[59] Del Mar Maldonado, M., & Dharmawardhane, S. (2018). Targeting rac and Cdc42 GT pases in cancer. Cancer Research, 78(12), 3101–3111. doi:10.1158/0008-5472.CAN-18-0619.

[60] Backurs, A., Indyk, P., & Wagner, T. (2019). Space and time efficient kernel density estimation in high dimensions. In H. Wallach, H. Larochelle, A. Beygelzimer, F. Buc, E. Fox, & R. Garnett (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc. 32, 15799–15808.

[61] Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: Visual molecular dynamics. Journal of Molecular Graphics, 14(1), 33–38. doi:10.1016/0263-7855(96)00018-5.

[62] Morris, K. M., Henderson, R., Suresh Kumar, T. K., Heyes, C. D., & Adams, P. D. (2016). Intrinsic GTP hydrolysis is observed for a switch 1 variant of Cdc42 in the presence of a specific GTPase inhibitor. Small GTPases, 7(1), 1–11. doi:10.1080/21541248.2015.1123797.

[63] Melendez, J., Grogg, M., & Zheng, Y. (2011). Signaling role of Cdc42 in regulating mammalian physiology. Journal of Biological Chemistry, 286(4), 2375–2381. doi:10.1074/jbc.R110.200329.

[64] Caldwell, H. K., & Young, W. S. (2006). Oxytocin and Vasopressin: Genetics and Behavioral Implications. Handbook of Neurochemistry and Molecular Neurobiology (3rd Ed), 573–607. doi:10.1007/978-0-387-30381-9_25.

[65] Torres, R., & Polymeropoulos, M. H. (1998). Genomic organization and localization of the human CRMP-1 gene. DNA Research, 5(6), 393–395. doi:10.1093/dnares/5.6.393.

[66] Bersani, M., Johnsen, A. H., Højrup, P., Dunning, B. E., Andreasen, J. J., & Holst, J. J. (1991). Human galanin: Primary structure and identification of two molecular forms. FEBS Letters, 283(2), 189–194. doi:10.1016/0014-5793(91)80585-Q.