UNIVERSITY OF CAPE TOWN

DEPARTMENT OF COMPUTER SCIENCE

CS/IT Honours Final Paper 2020

Title: **Developing a Framework to Analyse Clustering Algorithms for Molecular Dynamics Trajectories**

Author: **Nicholas Alan Limbert**

Project Abbreviation: **ClusterMol**

Supervisor(s): **Associate Professor Michelle Kuttel**

Category	*Min*	*Max*	Chosen
Requirement Analysis and Design	*0*	*20*	
Theoretical Analysis	*0*	*25*	
Experiment Design and Execution	*0*	*20*	**20**
System Development and Implementation	*0*	*20*	**5**
Results, Findings and Conclusions	*10*	*20*	**20**
Aim Formulation and Background Work	*10*	*15*	**15**
Quality of Paper Writing and Presentation	*10*		10
Quality of Deliverables	*10*		10
Overall General Project Evaluation (*this section allowed only with motivation letter from supervisor*)	*0*	*10*	
Total marks		**80**	**80**

Developing a Framework to Analyse Clustering Algorithms for Molecular Dynamics Trajectories

Nicholas A. Limbert
University of Cape Town
Cape Town, South Africa
lmbnic008@myuct.ac.za

ABSTRACT

Molecular Dynamics (MD) is a computer simulation method for generating and analysing the physical behaviours of molecules over time. As simulations increase, the amount of data needing to be analysed also increases. Manual comparisons between visual structures are rarely reliable on such a large scale and extremely time-consuming. Clustering analysis provides a means to classify similar molecular conformations into a limited number of district groups thereby reducing the amount of information to be manually analysed. A wide variety of readily available clustering algorithms have not been implemented and tested on highly flexible molecules such as carbohydrates. Hierarchical clustering provides a well-known approach to classify underlying cluster structures. The Quality Threshold (QT) algorithm guarantees that no cluster exceeds a specified similarity quality threshold. Our findings indicate that QT performs as expected in classifying noise and producing clusters of high similarity while hierarchical clustering is unable to distinguish noise. A major limitation of hierarchical clustering is the high sensitivity to outliers. Although the algorithms performed well, it is important to understand the limitations of clustering. Results are significantly dependent on the selection of atoms for pairwise comparison. Numerous algorithms and validation metrics should be evaluated to determine which clustering algorithm may be most suitable for the specific data set.

CCS CONCEPTS

• **Applied computing** → *Computational biology*; • **Mathematics of computing** → *Cluster analysis*; • **Theory of computation** → *Unsupervised learning and clustering*.

KEYWORDS

clustering analysis, molecular dynamics (MD) simulations, carbohydrate molecules, dominant conformations

1 INTRODUCTION

Molecular Dynamics (MD) simulations allow researchers to analyse the conformations of molecular structures over time to better understand the various macro-molecular structures that may occur. The use of unsupervised learning techniques, specifically clustering analysis has been utilised to deal with the growing amount of data needing to be processed and analysed. Simulation data now exceeds the size where manual analysis of conformations can be reliably performed. Clustering analysis of MD simulation data focuses on grouping similar molecular conformations into well-defined clusters.

Previous clustering analysis of MD simulation data focuses on nucleic acids and proteins, which are not flexible molecules. Highly flexible molecules such as carbohydrates have an extensive domain of possible conformations. While some carbohydrate molecules may remain in stable states with relatively few conformation changes, others may yield many conformations that change multiple times throughout the simulation. The range of possible conformations in flexible molecules will test the capabilities and limitations of different clustering algorithms in dealing with a variety of data. Clustering analysis is inherently data specific and each all results should be evaluated individually.

There are a wide variety of algorithms already adapted to work with MD data, however, there is no single framework which incorporates pre-processing, clustering and post-processing analysis. As a result, a framework was developed for researchers to easily integrate other algorithms, validation measures and useful functions into one package. The framework will allow us to test a variety of algorithms while consolidating additional functions.

Hierarchical clustering is a well-developed clustering technique with numerous resources available. We utilise four different linkage criteria for hierarchical clustering - *single*, *complete*, *average* and *Ward's* method. By exploring a range of linkage criteria we can determine which will be most suitable for MD simulation data. These are standard, widely available methods that provide an intuitive understanding of how clusters are formed. This method produces a hierarchical tree (dendrogram) which allows us to understand possibly underlying structures of the data while also providing control over the number of clusters to produced.

The Quality Threshold (QT) algorithm [12] aims to generate high-quality clusters by iteratively adding observations to a cluster while not exceeding the pre-defined cluster diameter or threshold. Observations are added to minimise the cluster diameter. The Quality Threshold (QT) algorithm has previously been implemented for clustering a range of MD data, however, there are some inconsistencies related to certain implementations [8]. We, therefore, aim to make use of the implementation from González-Alemán et al. which is in line with the originally proposed algorithm from Heyer et al. In addition we make use of a vectorised version by Melvin et al. [19]. This will allow us to determine the effectiveness of the two variations when clustering highly flexible molecules while also illustrating any differences that may arise.

Overall we aim to determine if any of our implementations can effectively cluster highly flexible molecules. Specifically, we aim to determine which hierarchical linkage criteria may be most suitable in clustering highly flexible molecules as well as any limitations that may arise. Additionally, we aim to determine whether the QT algorithms can produce clusters of high quality (high similarity

between conformations). Finally, if there is a significant difference between the original QT implementation by González-Alemán et al. and the QT vector implementation by Melvin et al.

2 BACKGROUND

2.1 Molecular Dynamics Simulations

Molecular Dynamics (MD) is a computer simulation method for generating and analysing the trajectories of atoms and molecules over a period of time [3, 4, 11, 22]. This is a numerical method whereby atoms positions are recalculated over extremely small periods of time to generate an overall motion picture of the molecular system for an extended period of time. Simulations produce extremely large amounts of data due to the increase in computational power [18], size of molecular systems and length of simulations. By simulating these complex interactions of molecules and atoms we can develop a better understanding of their various characteristics and behaviours. These simulations play an important role in determining the characteristics of a molecule, specifically uncovering conformational change in the molecular system.

Differences between frames within a simulation are calculated using the root-mean-square-deviation (RMSD) of selected atom positions. The selection of atoms used in the RMSD calculation is either a backbone structure, residuals or a specific choice of atoms that have certain behaviours. Using all atoms in the RMSD calculation will likely not produce meaningful pairwise differences due to the number of atoms and the complexity of molecules in a simulation. Each frame difference is calculated against every other frame in the simulation which generates a pairwise distance matrix or RMSD matrix. This similarity metric allows for the comparison of molecular structures for each frame which is required for many clustering algorithms. Equation 1 is a generalised version of the calculation between to frames u and v. The distance between atoms is usually returned as length units Angstroms, Å or 10^{-10} meters.

$$RMSD(u, v) = \sqrt{\frac{1}{N}\sum_{i=1}^{n} ||u_i - v_i||} \tag{1}$$

Trajectories of flexible molecules will differ from those of more stable molecules such as nucleic acids and proteins. Highly flexible molecules, such as carbohydrates will usually have many different shapes or conformations. The understanding of these macromolecular structures is important for many applications, specifically vaccinology. These molecules are usually highly flexible and can change conformations multiple times throughout a simulation. We will focus on well know carbohydrate simulations from previous literature. Namely Meningococcal Y (MenY), Meningococcal W (MenW) [16] and variations of Shigella flexneri [13].

2.2 Clustering Molecular Dynamics Simulations

Cluster analysis is a technique for grouping a set of data objects into distinct sets of different dissimilar partitions. A cluster is defined as a collection of data objects where objects within the same cluster are similar to one another while dissimilar to data objects in other cluster formations. Clustering analysis with MD simulations is typically used to group similar molecular conformations into partitions for analysis. Conformations refer to the overall structure of a molecule due to the spatial arrangement of atoms.

2.2.1 *Hierarchical Clustering.*

Hierarchical clustering is a type of agglomerative clustering whereby initially each data object is a single cluster. These clusters are then iteratively merged according to some distance linkage criteria. This process continues until one homogeneous cluster has been formed. Hierarchical clustering produces a dendrogram which illustrates how the cluster formations merge over iterations. This can be used to investigate similar conformations and determine levels of similarity during iterations as clusters are merged.

There are four types of linkage criteria we will investigate when using hierarchical clustering. These linkage criteria outline what is defined as the minimum distance between two clusters. These include - *single*, *complete*, *average* and *Ward's* linkage. It is important to note that *single* and *maximum/complete* linkage only take single points within a cluster into account when determining this minimum distance while *average* and *Ward's* method takes all points within a cluster.

Single-linkage merges two clusters based on the shortest distance between a pair of points within the two clusters, searching for the minimum distance between a pair of elements within two sets before merging. Given clusters u and v, single linkage is defined as follows 2.

$$d(u, v) = \min(dist(u[i], v[j])) \tag{2}$$

Maximum-linkage or *complete-linkage* focuses on searching for the maximum distance between a pair of elements from two sets to define the distance between two clusters. Given clusters u and v, complete linkage is defined as follows 3 .

$$d(u, v) = \max(dist(u[i], v[j])) \tag{3}$$

Average-linkage calculates the minimum distance between two clusters as the average distance between each point in one cluster to another. Given clusters u and v, average linkage is defined as follows.

$$d(u, v) = \sum_{ij} \frac{d(u[i], v[j])}{(|u| * |v|)} \tag{4}$$

Ward's method aims to minimise the total variance between clusters merging. At each iteration it determines which two clusters once merged will have the smallest variance increase. This continues until we have one homogeneous cluster. Ward's variance minimisation algorithm is defined as follows (5). Where u is a newly formed cluster from clusters s and t and v is previously unused cluster and T is defined as 6.

$$d(u, v) = \sqrt{\frac{|v| + |s|}{T} d(v, s)^2 + \frac{|v| + |t|}{T} d(v, t)^2 - \frac{|v|}{T} d(s, t)^2} \tag{5}$$

$$T = |u| + |v| + |t| \tag{6}$$

As we can see the linkage criteria defines the minimum distance between two clusters. Hierarchical clustering allows users to produce clusters based on the k, the number of required clusters. Distance measure d, whereby each cluster does not exceed a

maximum cophenetic distance of d is used to form clusters. A dendrogram illustrates how the clusters are formed and merged over time and aid us in selecting appropriate values for these parameters.

2.2.2 *Quality Threshold Algorithms.*

The Quality Threshold algorithm was original proposed by Heyer et al. [12] to group genes into high-quality clusters. The "quality" of the clusters produced is ensured by finding large clusters where the diameter does not exceed a user-defined similarity threshold. This is also referred to as the cutoff value. The algorithm was originally used to find extremely large clusters of genes while ensuring dissimilar genes are not forced into a cluster as would be the case with other types of clustering algorithms.

The Quality Threshold algorithm requires two users specified parameters, similarity threshold/cutoff value and the minimum amount of elements in each cluster. A candidate cluster is formed from the first data object in the data set, then iteratively every data objects similarity measure is compared, those below the threshold are added to the cluster. Only data objects with similarity a measure differences below the threshold for all current data objects in the cluster will be added. Once this step is complete, we proceed to do this for all data objects in the data set. The is a computationally expensive process. Consequently, each data object forms a cluster. All data objects are candidates for all clusters at this stage.

Once this process has been completed, we take the largest cluster and remove it from the data set. Iteratively removing the next largest cluster from the pool until we reach the minimum user-defined size of a cluster. A data object may only be associated with one cluster, therefore should a data object belong to more than one cluster it will only be in the larger cluster. At this point, there may be data objects that do not belong to any cluster (noise). It is important to note that we may set the minimum size extremely low to generate a significant amount of clusters. This will allow all data objects to be assigned to a specific cluster, albeit we will have many more cluster formations - particularly those of a smaller size that would be deemed insignificant. The minimum size parameters allow for only significant clusters to be returned while labelling the rest as outliers.

There have been multiple implementations of the Quality Threshold algorithm [15, 19]. Many of these already work with a range of MD data. Most notably, the Visual Molecular Dynamics (VMD) package's [14] internal clustering function makes us of the Quality Threshold algorithm. Additionally, clustering plugins for VMD such WMC PhysBio clustering GUI [9] use this internal clustering function. However, there are some inconsistencies related to certain implementations [8]. González-Alemán et al. outline numerous implementations and their shortcomings against the originally proposed algorithm.

2.3 Validation Techniques for Clustering Molecular Dynamics Simulations

We outline some notable validation techniques and indices implemented in previous research with MD simulations.

Shao et al. [21] use several validation metrics such as the pseudo-F statistic, Davies-Bouldin index (DB) [6], SSR/SST ratio and the "critical distance" when the clustering of MD simulations. The DB and pseudo-F statistic are used to determine the overall compactness and separation of all the clusters. Compactness representing how similar items within a cluster are while separation deals with how dissimilar clusters are from one another. The SSR/SST and critical distance are used to determine the ideal cluster count through iterations with different cluster parameters. It is shown that low DB values and high pseudo-F values indicate good partitions of the data. The validation indices behaved as expected, with high pseudo-F statistic and low DB values for good cluster formations. They also observe a constant SSR/SST ratio when the ideal cluster count is reached. Together these metrics provided comprehensive validation of the algorithms implemented for the given data, however, none should be used as a metric individually.

Abramyan et al. [1] make use of three internal cluster validation measures; Calinski-Harabasz (CH), Davies-Bouldin (DB), and Silhouette (S) indices when clustering of MD simulations. These simulations focus on clustering protein adsorption MMD simulation data. High Calinski-Harabasz, Silhouette indices and a low Davies-Bouldin index indicate good cluster formations and separation. All the indices implemented by Abramyan et al. are internal validation measures. These indices are used to determine the compactness and overall separation of clusters formed.

Melvin et al. [18] implement a wide variety of different algorithms however, only make use of the Silhouette index [10] as their basis of comparison. The high Silhouette index indicates ideal cluster formations with high intra-cluster similarity and low inter-cluster similarity.

2.3.1 *Cluster Validation Indices.*

There are numerous cluster validation indices (CVI's) in addition to those mentioned. The variety in clustering validation indices arises as no single measure can capture all aspects of the clustering problem [2, 7]. This reinforces the idea that multiple indices should be used to better understand the cluster analysis output. We make use of the Calinski-Harabasz (CH), Davies-Bouldin (DB) and Silhouette (S) indices to better understand our results and their significance. These indices are already widely used in clustering analysis as well as in MD clustering analysis.

Calinski-Harabasz index is a ratio type index where high values represent ideal cluster formations. The compactness of a cluster is determined by the sum of distances between all objects in the cluster to the local centroid while the cluster separation is calculated by the distances between the local centroids and a globally specified centroid.

Davies-Bouldin index is summation based index where low values represent well-formed clusters. The cluster compactness is calculated as the distance from all the points in a cluster to the centroid. The cluster separation is then determined by the distance between all centroids from the clusters.

Silhouette index is a summation type index that ranges between -1 and 1, with high values representing good cluster formation and separation. The cluster compactness is determined by the sum of the distances between all the points in the same cluster. The separation is estimated by the nearest neighbour distance, this is the distance between the two closest points between the two clusters.

3 DESIGN AND IMPLEMENTATION

We outline the implementations of the Hierarchical and Quality Threshold algorithms. This includes the sources for these algorithms and how they were integrated into the larger framework. We briefly outline the extensive framework developed and some of the key functionality. We present the MD data sets and the methodology we follow in order to obtain results.

3.1 Clustering Algorithms

3.1.1 Hierarchical Clustering.

Implementations of the Hierarchical clustering algorithm with various linkage criteria are available in the python SciPy library [25]. We extended this existing code to allow for clustering of MD simulation data. By reducing the complex trajectories files to the desired RMSD matrix which quantify the similarity/dissimilarity between frames we were able to generate results with this existing clustering package.

Hierarchical clustering has previously been implemented with MD simulations [1]. Hierarchical clustering allows us to grasp the general structure of the data and how these clusters are formed and merged as we iterate through the process. It is important to note that there is no objective way to initially determine the number of clusters present. The number of clusters will depend on where the tree or dendrogram is "cut." By specifying the number of clusters to return, the SciPy library [25] will determine the minimum cophenetic distance threshold value between any two data points within the same cluster. This ensures that the exact number of user-specified clusters will be returned.

Four linkage criteria - *single*, *complete*, *average* and *Ward's* method are used to generate results. The cophenetic distance measures how well the clustering preserves the original data points pairwise distance once a single cluster is formed. Higher cophenetic values are favourable.

An additional advantage of hierarchical clustering is the output of a dendrogram. Dendrograms give the overall structure of the cluster formations and provide an intuitive understanding of possible underlying clusters. However, this should be used cautiously when determining the number of clusters. Hierarchical clustering is also unable to classifying noise into a separate cluster, as all data points belong to a cluster.

3.1.2 Quality Threshold Algorithms.

The Quality Threshold algorithm implemented by González-Alemán et al. [8] and vectorised version by Melvin et al. [19] are publicly available. These algorithms were slightly adapted to work with a pairwise distance matrix from different sources (Molecular Dynamics trajectories, pre-processed data generated by UMAP [17], T-SNE [23] and test data).

González-Alemán et al. outline numerous faults in many available implementations of the QT algorithm. They detail how their in-house implementation is in line with the correct Quality Threshold algorithm originally proposed by Heyer et al. [12]. In addition, they outline how the vectorised version by Melvin et al. [19] is an implementation of an algorithm proposed by Daura et al. [5] rather than the QT algorithm.

The original Quality Threshold algorithm ensures that all data items within a cluster do not exceed the threshold value between every data point within a cluster. The vectorised version only ensures that no data item exceeds the threshold value against the original seed data point for each cluster. The seed is an initial data point as the algorithms iterates. The QT algorithm has more stringent constraints compared to the vectorised version.

3.2 Molecular Dynamics Data Sets and Methods

3.2.1 MenW and MenY.

MenW and MenY are two similar molecules that have different conformations characteristics [16]. MenY represents a molecule that is relatively stable with one single very dominant conformation throughout the simulation. MenW has numerous conformations with varying cluster sizes including the dominant conformation that is present in MenY. MenW is seen as a flexible molecule while MenY is stable with one frequent conformation.

MenW and MenY were initially aligned on two central residues in the 3RU chains before performing clustering. Alignment and select statements are outlined in Table A13. Alignment ensured that each frame is essentially superimposed onto one another to ensure that co-ordinate differences from one frame to another will have meaning. The selection allows us to cluster on only the atoms of importance which are expected to affect the conformation change throughout the simulation. Clustering was performed on ever 10th frame for both simulations, leading to a total of 4003 frames to be clustered.

By using the same alignment and selection statements we can determine how the clustering algorithms handle both stable and flexible molecules with similar parameters.

The only parameter required for Hierarchical clustering (number of clusters) was set to 10 for both MenW and MenY. The Quality Threshold algorithms required two parameters, minimum cluster size and cutoff/threshold value. The minimum cluster size was set as 100 data objects to ensure that only significant clusters are returned. The following cutoff values were selected: 1.5, 2.5, 3.5, 4.5 (Å). The range of values illustrates the different performance characteristics of the two QT variants

3.2.2 Shigella flexneri.

The S. flexneri molecules represent simulations with many different conformations and deemed highly flexible. Clustering of the Shigella flexneri molecules was only performed with the two QT variations. We make use of the S. flexneri Y molecule with three repeating units, selecting everything 10th frame (4002 total frames). Frames are again aligned on a selection of atoms outline in Table A13. Once all frames were aligned clustering was performed on a selection of atoms outlined in Table A13. Using all atoms when clustering will not produce results of significance as these molecules are too complex and the calculated RMSD values would not be representative of the structures in focus. For example, some simulations contain Sodium and Hydrogen atoms which are deemed insignificant when clustering. The Quality Threshold algorithms required two parameters. The minimum cluster size was set at 100 while the following cutoff values were selected: 2.5, 3.5, 4.5 (Å) to illustrate the change in results with S. flexneri Y.

In addition, we cluster S. flexneri 6 with six repeating units. This molecule represents a simulation that has a large number of atoms and residuals, while also being highly flexible. Once again, every 10th frame was selected leading to a total of 3602 frames to cluster. The S. flexneri 6 with six repeating units was clustered with a higher cutoff value ranging between 6.5 and 7. 5 (Å). These cutoffs were selected by looking at the mean and median values in the RMSD matrix as well as recommendations from researches familiar with this molecule. Selection and alignment statements are outlined in Table A13.

3.3 Framework

A framework was developed in order to accommodate the various processing and individual requirements for each clustering job. The framework was developed in python [24] and makes use of standard object-orientated practices. A total of 7 different clustering algorithms were implemented with a range of preprocessing and post-processing functions. Trajectories in the form of a protein data bank file (.pdb) are the main input for the framework in addition to various cluster-specific parameters and preprocessing parameters. Various outputs are produced such as cluster validation indices (CVI's), MD data files of the largest n clusters and cluster labels. Figure 1 gives a general overview of the framework.

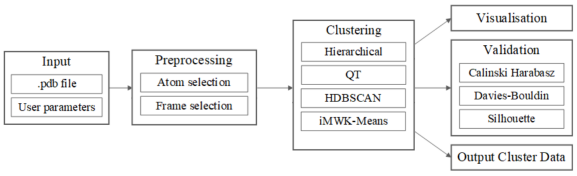

Figure 1: Pipeline of framework design.

4 TESTING AND VALIDATION

Unit testing was performed on all helper functions implemented within the Hierarchical and Quality Threshold algorithm scripts. This ensured that these functions execute as expected. Testing was not performed on graphic output functions such as graphs and plots. The Hierarchical and Quality Threshold algorithm implementations were validated with generated data sets (Figure 2). The data sets have known cluster characteristics and counts (Table A1). We outline the steps taken to validate each algorithm in the sections below while findings are discussed in the Results and Discussion section. Furthermore, we outline steps taken to ensure that the framework performs as expected.

4.1 Framework Validation

The framework has multiple input parameters and requirements for each clustering job. To ensure that all paths produce the expected results a parser was implemented for parameter selection. This enabled us to restrict the selection, type and requirement of parameters while raising errors on incorrect usage. In addition, we implemented error handling when parameters do not satisfy the requirements for specific algorithms - an error will be raised if the entire distance matrix (RMSD matrix) is empty due to extremely low cut-off values. The user will be unable to continue with the clustering or processing jobs unless the required parameters are satisfied.

We encountered some issues when running the framework on different operating systems due to different package requirements and dependencies. The framework requires several Python packages that are not previously installed. We, therefore, implemented a setup script that installs all package dependencies for the framework. Additionally, this deals with any directory issues. Although extensive testing was implemented we were unable to exhaustively test and validate all functions with various types of data sources. Users should validate and test various data sets against the framework to determine its effectiveness and whether it is useful for their application.

4.2 Validation Data Sets

In order to validate the clustering algorithm implementations, we made use of the Python sci-kit-learn library [20] to generate random data sets. Data set 1 **(a)** has distinct clusters of data with a low intra-cluster variation. This enabled us to test whether the algorithms can correctly cluster district separate data in addition to the correct number of clusters. Data set 2 **(b)** consists of noise or random data in order to understand how the algorithms handle highly dissimilar data with no underlying structures. Data set 3 **(c)** contains clusters with a varied and high intra-cluster variation. Many of these clusters are overlapping which demonstrates the performance of the algorithms when there is a high amount of similarity between clusters. Figure 2 illustrates the overall structure of the data sets. Table A1 outlines the parameters used for each data set.

Additionally, data sets from sci-kit-learn such as Iris, Wine, Digits and Breast Cancer data [20] are available in the framework for further testing and validation. We include results from the Iris data set **(d)** which has three clusters, however, the global structure only has two clear partitions. This is real-world data rather than artificially generated data such as data sets 1, 2 and 3.

4.3 Hierarchical Clustering Validation

To ensure that our Hierarchical clustering implementations produce expected results, we made use of data sets with know characteristics and outputs. Firstly, data set 1 with distinctly different clusters allowed us to tests the algorithms ability to produce the correct clusters. Data set 3 has the same amount of clusters but a higher intra-cluster variation. This ensures that the implementations can deal with overlapping clusters. Iris only has two distinct clusters, yet three correct clusters. We expect the implementations to only classify the global structure with outliers as separate clusters. Data set 2 tests the linkage criteria ability to handle unstructured data. The only input parameter required by our Hierarchical clustering implementation is k, the number of clusters. For each data set, we will use the correct number of clusters (Table A1) as the input paraeneters. This allows us to see how the clusters are formed and whether they the same as the correct cluster formations.

4.4 Quality Threshold Clustering Validation

To validate that both QT algorithms can produce acceptable clusters, we made use of the four aforementioned data sets. The data sets

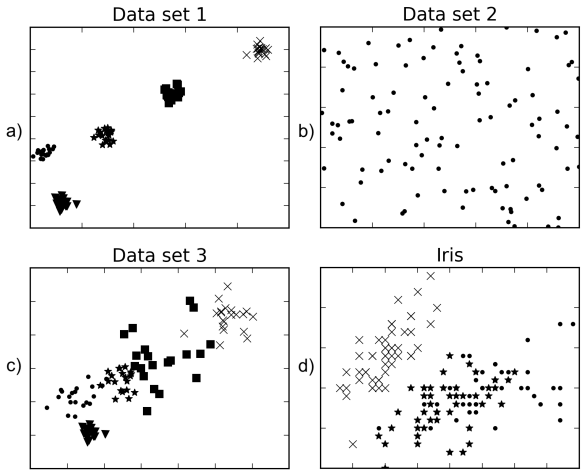

Figure 2: Test Data sets. Data set 1 **(a)** represents distinct compact clusters. Data set 2 **(b)** represents random noise data with no underlying clusters. Data set 3 **(c)** represents clusters with varying variance and overlapping clusters. The Iris data set **(d)** represents three different types of iris flowers based on length and width of the sepals.

aim to test how the algorithms perform with distinct low intra-variance clusters (Figure 2a), highly similar clusters (Figure 2c) and more natural data with overlapping clusters (Figure 2d). In addition, we make use of random data set (Figure 2b) with no underlying characteristics. These implementations require two parameters, a cutoff/threshold value and minimum cluster size (Table A2). As mentioned previously a cutoff value ensures that clusters have an acceptable level of intra-cluster variation that does not exceed when generating a cluster (this ensures that clusters are similar) while only clusters that have a greater member tally than the minimum cluster size are returned. Clusters that are smaller than the cluster size threshold are discarded as noise.

5 RESULTS AND DISCUSSION

We outline the results obtained when clustering the validation data sets and discuss the implications. Most importantly, we outline the various results obtained when performing clustering analysis on MD simulation data. Characteristics of the MD data sets were discussed above. Results for the following MD simulations are shown - Meningococcal Y (MenY), Meningococcal W (MenW) [16] and variations of Shigella flexneri [13].

5.1 Validation

5.1.1 *Hierarchical Clustering.*

Hierarchical clustering requires a linkage criterion, this defines the distance between two clusters. We implemented four linkage criteria - *single*, *average*, *complete* and *Ward's*. The theory and implementation of this algorithm including the various linkage criteria are discussed above.

Tables A3, A4, A5 and A6 in the Supplementary Material summarise the various parameters and results obtained for each linkage criterion. All linkage criteria were able to correctly clusters data set 1 with 5 distinct sets of data. This is evident in the correct cluster counts as well as with positive values for the Silhouette score, Davies-Bouldin index and Calinski-Harabasz index. The *single* linkage criterion produces a single large cluster for data sets 2 and 3, in addition to smaller insignificant clusters. This is expected, as the distance between two clusters when merging is defined as the shortest distance between two clusters in the previous iteration. As a result of the requirement to return a set number of clusters (5), these smaller clusters are merely placeholder elements in order to satisfy this parameter. The Iris data set produces results that are in line with the global structure of the data set, with two significant clusters and a third smaller cluster when using *Ward's* linkage criterion (Figure 3a).

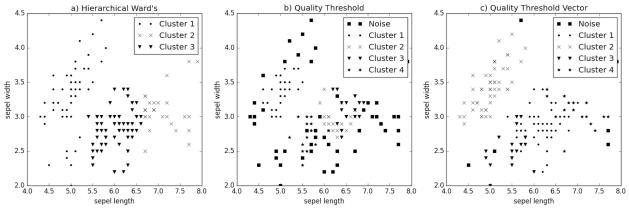

Figure 3: Clusters produced by the Hierarchical clustering - Ward's method **(a)**, Quality Threshold algorithm **(b)** and the Quality Threshold vectorised algorithm **(c)** on the Iris data set.

Ward, *average* and *complete* linkage all produce similar results, but *average* produces better CVI values for all datasets, followed closely by *Ward*. This is expected, as both *average* and *Ward's* linkage criteria take multiple data points in a cluster into account when performing the linkage calculation, while *complete* and *single* only take a single point.

All algorithms produce random clusters from data set 2. These clusters are insignificant due to their high intra-cluster variation. This indicates that the clusters are not compact or similar but rather spread throughout the sample space. The clusters produced are based more on arbitrary points from the data rather than any underlying clusters which are not present in the random data.[1]

5.1.2 *Quality Threshold Clustering.*

Table A2 in the Supplementary Material summaries the testing results obtained for both algorithms. Both algorithms were able to deal with distinctly different clusters even with a low cutoff value. Data set 3 provides clusters with higher intra-cluster-variance. Results indicate that both algorithms were unable to produce all 5 correct clusters. QT original only produced 3 clusters while the vectorised version only indicating 2 clusters were found. Both algorithms were still able to cluster the general structures of the data with an increased cutoff value. Data set 2 or noise is classified as a single cluster by the vectorised version while the original version produced a significant amount of noise and two clusters. The vectorised versions single cluster has a low intra-cluster similarity and could therefore be considered as noise due to the data points being widely spread throughout the cluster. Figure B1 illustrates

[1]Dendrogram, the resulting CVI's and scatter-plots for all testing are available at [http://projects.cs.uct.ac.za/honsproj/2020/].

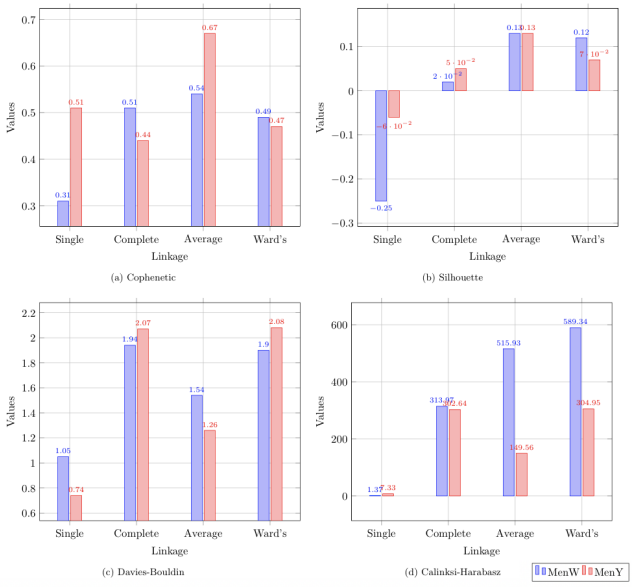

Figure 4: Cluster validation indices for Hierarchical clustering with various linkage criteria. Figure 4**a** shows the Cophenetic distance, while Figure 4**b** shows the Silhouette score for each criterion. Figure 4**d** shows the Calinksi-Harabasz scores. Higher values indicate better clustering for these indices. Figure 4**c** shows the Davies-Bouldin index in which lower values are good.

the global structure and partitioning of the data when using the original Quality Threshold algorithm. It is evident that no meaningful clusters are produced due to the spread and intersection of the clusters.

Figure 3b and 3c show the partitions of the data points for the Iris data set. Interestingly, for the Iris data set, both algorithms produced a total of four clusters. Noise accounted for more than 30 per cent in the original algorithm compared to 5 per cent in the vectorised version. It is evident that both algorithms have partitioned the global clusters even though these may not be the correct clusters. Note that the Quality Threshold algorithms aim to produce clusters where the cut-off of the threshold value for a cluster is not exceeded. This vital characteristic ensures that when clustering MD data no conformations exceeding the root-mean-square-deviation difference between all the frames in a cluster are added.

5.2 Molecular Dynamics Hierarchical Clustering Results

5.2.1 MenW and MenY.

Ten clusters for both MenW and MenY were returned along with their respective CVI's. This value was chosen to allow for a number of clusters to be returned even though some clusters may be insignificant due to their cluster counts. Comparisons of CVI values between MenW and MenY are shown in Figure 4.

Supplementary information for MenW and MenY can be found in Table A7 and A8 respectively. This outlines the various results from Hierarchical clustering performed. Dendrograms for both MenW (B2, B3, B4, B5) and MenY (B7, B6, B8, B9) can be found in the Supplementary Information.

Results show that *single* linkage is unable to cluster both MenW and MenY, producing a single cluster containing almost all data objects. This is further indicated by poor values for both the Silhouette score (positive values indicate good cluster formations) and Calinski-Harabasz index (higher values indicate compact clusters). The *single* linkage also performed worse than all other linkage criteria. We suspect that no meaningful clusters were produced as evident from the results.

The additional linkage criteria all produce similar CVI's for MenW, with *average* and *Ward's* method having more favourable Silhouette scores, Davies-Bouldin and Calinski-Harabasz indices. Cluster counts suggest that there are significantly larger clusters (meaning a conformation is more prevalent). *Complete* has a cluster with 40 per cent of frames, average 32 per cent while *Ward's* largest cluster is approximately 17 per cent of the total clusters. *Ward's* methods produce resulting cluster counts that are more evenly spread, this is in line with the methodology as it attempts to minimise the variance within a cluster when merging and forming new clusters. The 10 conformations using the *average* linkage are shown in Figure 5. *Average* linkage criterion produces clusters with significantly better CVI's than other methods for MenW, however as there is no cluster for noise/outliers, we suspect some clusters to contain conformations that aren't similar to others.

Disregarding the results from *single* linkage, *average* produces one single dominant cluster for MenY. *Ward's* method and *complete* linkage produce results that indicate two dominant conformations. Figure 6 outlines the top conformations produced by *average*, *complete* and *Ward's* method. Once again the *average* linkage has the best validation indices apart from the Calinski-Harabasz index. The value is likely low due to one large cluster with high intra-cluster variance. The significant cluster produced by *complete* linkage in 6 contains an anomaly conformation that does not match the other conformations. This is probably due to *complete* using a single data point when merging clusters rather than multiple data points as is the case with *average* and *Ward's* method. Dissimilar conformations are thus more likely to be grouped into incorrect clusters.

The overall results obtained from the majority of hierarchical clustering for MenW and MenY indicate that no individual linkage criterion produces significantly better results. When inspecting the clusters visually through VMD it is evident that many clusters contain outliers and significantly different conformations. This would negatively impact the CVI's values and indicate why the values are relatively poor. The validation indices should only be used to compare results with an individual varying parameter, in our case this would be the linkage criteria. While this allows us to determine which linkage may be most suitable it does not thoroughly verify or validate our results.

Our results demonstrate that some linkage criteria are more acceptable than others however do not produce clusters to the standard of the previous implementations of the Quality Threshold algorithm. We now take a look at how the QT algorithms can cluster a range of carbohydrate molecules.

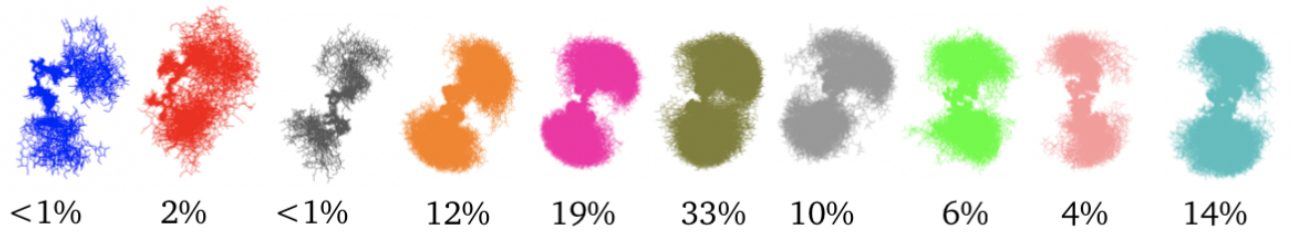

Figure 5: Dominant conformations produced by Hierarchical clustering (average linkage) for MenW

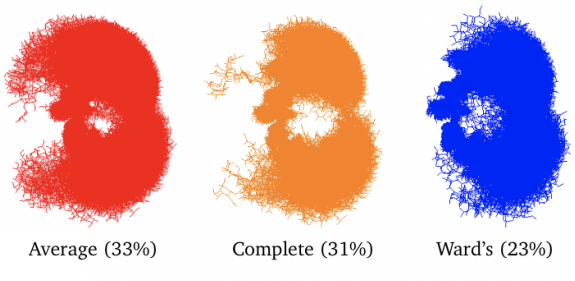

Figure 6: Conformations produced by Hierarchical clustering on MenY.

5.3 Molecular Dynamics Quality Threshold Results

5.3.1 MenW and MenY.

Analysis of MenW and MenY trajectories was performed on every 10th frame, with the trajectory being aligned on two central residuals before clustering. Clustering was then performed using a range of cutoff values for both QT and QT vector. CVI's are generated for all results. Results are given in Table A9, A10 and summarised in Figure 8. Note that CVI calculations do not include the noise cluster which is denoted by a negative value. Including noise would severely affect the resulting CVI's as this cluster does not represent meaningful clustered data.

Results obtained for MenW show that smaller cutoff values have extremely large noise clusters, sometimes more than 50 per cent of the entire data set. When the cutoff value increases, we see cluster counts increase and the noise cluster decrease. However, once the cutoff value is too large, data points tend to fall into only a few clusters. This is shown by the reduced number of clusters for larger cutoff values in Figure 8d. Although lower cutoff values have more favourable CVI's, these values cannot be taken at face value when the majority of data is classified as noise. Noise in simulations can be seen as the change from one stable conformation to another while dominant conformations are relatively stable and hence produce clusters with higher counts. The trade-off between the size of noise data and CVI's should be evaluated carefully when selecting the final parameters for QT clustering.

QT original produces the best results with a cutoff value of 3.5 Å (Figure 7a). It has a high Calinski-Harabasz index, and a relatively low amount of noise compared to lower cutoff values. The vectorised version produces a smaller number of clusters for the same cutoff values. Overall the original version produces more favourable CVI values and cluster counts for all cutoff values when compared to the vectorised version. It is evident that MenW has numerous dominant conformations - this is in line with previous research [16] and validates the effectiveness of QT algorithms with flexible molecules.

MenY represents a more stable molecule where previous research [16] has indicated that only one dominant conformation is evident throughout the simulation.

Analysis performed on MenY shows that there is consistently an individual large cluster with one smaller yet significant cluster also present. Noise is again higher for low cutoff values which is expected with a stricter constraint. This reinforces the idea that as cutoff values increases, noise decreases. Notably, the vectorised version produces only one cluster with a cutoff value of 4.5Å and an extremely large single cluster for 3.5Å. All evidence from both QT and QT vector suggests that there is a single dominant conformation present for MenY. QT vector CVI's show that a cutoff value of 1.5Å and 2.5Å produce more accurate results. This lower cutoff value may be more useful for molecules in a stable state, as evident from the results. Cutoff values can be decreased when the RMSD values are lower due to a high amount of similarity in conformations throughout a simulation.

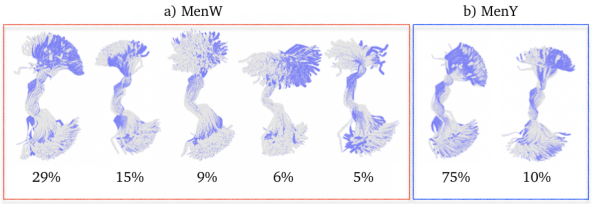

Figure 7: Dominant confrontations produced by QT algorithm for MenW **(a)** and MenY **(b)** with a cutoff of 3.5 Å.

Additionally, the original version produces dominant conformations with some additional less significant conformations. The spread of clusters is higher for the original version. Once again a cutoff value of 3.5Å for the original version produces more favourable

CVI's with a reduced amount of noise. The Calinski-Harabasz index is high with lowered values for the Davies-Bouldin index. The two dominant conformations of MenY are shown in Figure 7b. Visually these conformations appear very similar.

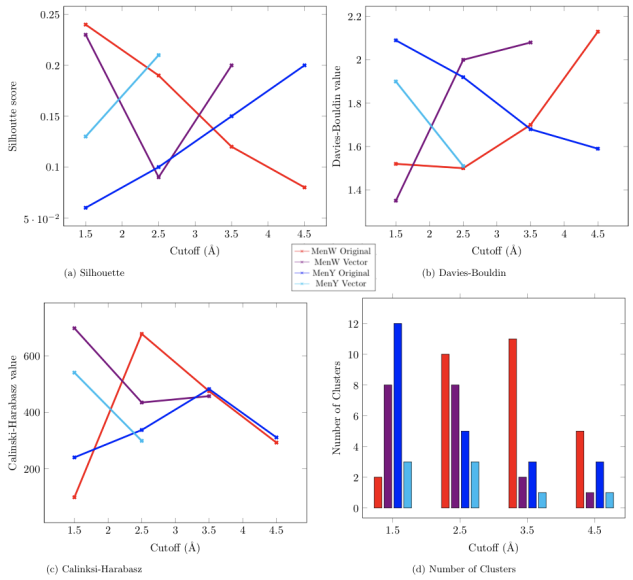

Figure 8: Cluster validation indices for QT and QT Vector clustering with varying cutoff values for MenW and MenY.Figure 8**a** shows the Silhouette score, while Figure 8**c** shows the Calinksi-Harabasz scores. Higher values indicate better clustering for these indices. Figure 8**b** shows the Davies-Bouldin index in which lower values are good. Number of cluster (Figure 8**d**).

5.3.2 *Shigella flexneri Y 3RU and 6 6RU.*

Detailed results for S. flexneri Y 3RU (Table A11) and S. flexneri 6 6RU (Table A12) are available in the Supplementary Material. Results are summarised in Figure 9. S. flexneri Y 3R has higher amounts of noise generated from lower cutoff values for both the original and vectorised version albeit the noise produced by the original algorithm is extremely high being more than 60 per cent of the data. The trade-off between noise and CVI values is evident. When there are fewer, highly correlated clusters, the CVI's are higher even though much of the data is classified as noise. The original version with 4.5Å and the vectorised version with 2.5Å produce extremely similar CVI values, however, the original produces slightly less noise.

The vectorised version produces larger clusters for high cutoff values, possibly grouping non-similar conformations into clusters due to relaxed constraints. The vectorised results obtained with a low cut-off of 2.5Å show promising CVI's and cluster counts. Although, the Silhouette index is low compared to higher cutoff values; the Calinski-Harabasz is substantially higher indicating good cluster formations. The top three conformations produced by the vectorised version with a cutoff value of 2.5Å are outlined in Figure 10a.

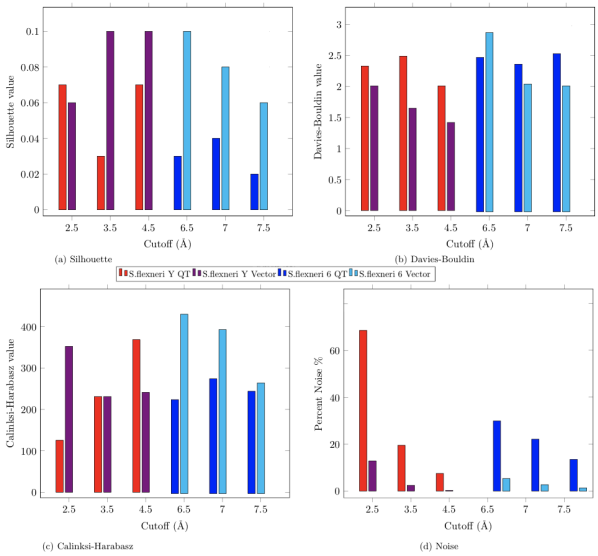

Figure 9: Cluster validation indices for QT and QT Vector with varying cutoff values for S.flexneri Y 3RU and 6 6RU. Figure 9**a** shows the Silhouette score, while, Figure 9**c** shows the Calinksi-Harabasz scores. Higher values indicate better clustering for these indices. Figure 9**b** shows the Davies-Bouldin index in which lower values are good. Noise per cent of data (Figure9**d**).

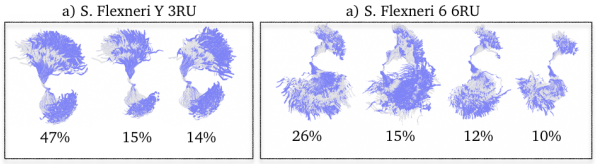

Figure 10: Dominant conformations produced by QT Vector algorithm for Shigella flexneri Y 3RU with a cutoff of 2.5 Å**(a)**. Dominant conformations produced by QT algorithm for Shigella flexneri 6 6RU with a cutoff of 7.5 Å**(b)**.

The results obtained for S. flexneri 6 6R with the vectorised version indicate clusters with higher candidate counts and fewer outliers. Most CVI's indicate that the vectorised version performed slightly better than the original version with similar cutoff values. The reduced number of clusters produced could indicate why better CVI's were achieved as there aren't many clusters to take into account during the calculations. The original QT algorithm with a cutoff value of 7Å produces positive results, achieving lower Davies-Bouldin values compared to other cutoff values. The Calinski-Harabasz index is also slightly higher. There are four dominant conformations that are more than 8 per cent of the frames (Figure 10b). While the original version produced a higher number of clusters with similar parameters many of these clusters are not deemed significant as they are less than 8 per cent of the simulation data.

These molecules are highly flexible with many different conformations. S. flexneri 6 6R has many more atoms to cluster on

compared to S. flexneri Y 3RU. Visual representations indicate that both algorithms can produce acceptable clusters of confirmations provided the correct cutoff values are selected.

5.4 Discussion

Hierarchical clustering is unable to classify noise separately, however, it is still able to produce clusters with relatively good CVI values. A limitation to Hierarchical clustering is that it requires prior knowledge about the number of conformations within a simulation. This information is usually not available unless prior research has been conducted on the specified data set. As expected, *single* linkage is unable to produce clusters of significance and merges most frames into one cluster. *Complete*, however can produce reasonable results given that only a individual data points are taken into account when merging clusters. *Complete* also has better CVI values and a greater cluster count spread to those produced by *single* linkage.

CVI values for *Ward's* method and *average* linkage are much higher in comparison to *single* and *complete* linkages. Visually the larger clusters produced by these methods show distinct conformation differences (Figure 5, 6). Linkage criteria that take multiple data points into account produce significantly better results and should be implemented when using Hierarchical clustering.

Noise is a major issue in hierarchical clustering. A possible solution is to use two-pass clustering analysis, where two different clustering algorithms are implemented. For instance, running the Quality Threshold algorithm on larger clusters produced by hierarchical clustering in order to distinguish and classify noise. We suspected that should these non-similar conformations be removed from the clusters produced by hierarchical clustering the CVI values would improve significantly.

These results confirm that although Hierarchical clustering allows us to generate a good understand of the general structure of underlying clusters, it cannot produce accurate results with highly flexible molecules, exclusively.

It is important to note that we cannot compare CVI values between QT and Hierarchical results as these are internal validation indices (only measuring one common variable change). Cluster formations cannot be compared as we are testing algorithms with different characteristics and parameters.

QT results produce good clusters for highly flexible molecules as is evident front the results obtained for S. flexneri, MenW and to an extent MenY. MenY is not highly flexible, however, both QT and QT vector were able to determine the single conformation in the data. Notably, results are only informative once the correct cutoff value has been found. Lower cutoff values impose a stricter constraint on the QT variations and produce more noise data. A balance between clusters produced and noise should be determined. Minimising noise by increasing the cut-off value leads to larger clusters containing dissimilar conformations.

The original Quality threshold algorithm proposed by Heyer et al. has stricter constraints compared to the version by Daura et al. This is evident from the increased noise clusters in the original version compared to the vectorised version throughout the results for similar cut-off values. The original Quality Threshold algorithm ensures that all data items within a cluster do not exceed the threshold value between every data point within a cluster while vectorised version only ensures that no data item exceeds the threshold value against one data point for each cluster.

It is evident that both QT variants can cluster highly flexible molecules and can produce high-quality clusters with minimal noise albeit the correct cut-off values have to be determined. Selection of a cutoff value should be based on iteratively evaluating different CVI's while changing the cutoff values. One should evaluate the noise and clusters produced to determine the most suitable cutoff value.

6 CONCLUSIONS

Clustering algorithms are inherently unable to consistently produce accurate clusters when data is highly similar. This is mainly due to the random nature of the data points in addition to their being no objective method to correctly clustered unlabeled data. Certain clustering algorithms that can deal with high similarity while also classifying noise separately allow for significantly better results than those which are unable to classify noise.

The Quality Threshold algorithm (original version) produces the best results in comparisons to other algorithms implemented, demonstrating the capability to cluster highly flexible molecules. This further attests to its existing popularity within the domain of clustering MD data. Clustering analysis should always be performed with a range of algorithms and validation indices when evaluating a data set. The framework developed allows for easy integration of existing algorithms in order to evaluate results.

7 FUTURE WORK

This framework could be enhanced through the implementation of additional clustering algorithms as well as a selection of additional cluster validation indices. Importantly, many other formats of Molecular Dynamics simulation files exist and extending the framework to accept these files would benefit a range of researchers. The overall significance of our results could be improved by using longer simulation runs rather than a reduced data set. In addition, results could be compared to other studies involving these simulations, including those performed by other members of this project.

DATA AVAILABILITY

All results obtain are available at http://projects.cs.uct.ac.za/honsproj/2020/ under "clustermol". This includes all outputs from the framework such as resulting clusters, CVI values and any graphics produced.

ACKNOWLEDGEMENTS

I would like to thank my project partners Wen Kang Lu and Robyn McKenzie for their support and commitment to developing this joint framework. Nicole Richardson for providing the protein data bank files for S. flexneri and supplementary information. Most importantly, I would like to thank my supervisor, Associate Professor Michelle Kuttel, for providing guidance and expertise throughout this project. South Africa's National Research Foundation provided funding for this project.

REFERENCES

[1] Tigran M. Abramyan, James A. Snyder, Aby A. Thyparambil, Steven J. Stuart, and Robert A. Latour. 2016. Cluster analysis of molecular simulation trajectories for systems where both conformation and orientation of the sampled states are important. *Journal of Computational Chemistry* 37, 21 (2016), 1973–1982. https://doi.org/10.1002/jcc.24416

[2] J. C. Bezdek and N. R. Pal. 1998. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 28, 3 (1998), 301–315. https://doi.org/10.1109/3477.678624

[3] B. R. Brooks, C. L. Brooks III, A. D. Mackerell Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. 2009. CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry* 30, 10 (2009), 1545–1614. https://doi.org/10.1002/jcc.21287 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.21287

[4] David A. Case, Thomas E. Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M. Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J. Woods. 2005. The Amber biomolecular simulation programs. *Journal of Computational Chemistry* 26, 16 (2005), 1668–1688. https://doi.org/10.1002/jcc.20290

[5] Xavier Daura, Karl Gademann, Bernhard Jaun, Dieter Seebach, Wilfred F van Gunsteren, and Alan E Mark. 1999. Peptide Folding: When Simulation Meets Experiment. *Angewandte Chemie (International ed.)* 38, 1-2 (1999), 236–240.

[6] David L. Davies and Donald W. Bouldin. 1979. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1, 2 (1979), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

[7] Vladimir Estivill-Castro. 2002. Why so Many Clustering Algorithms: A Position Paper. *SIGKDD Explor. Newsl.* 4, 1 (June 2002), 65–75. https://doi.org/10.1145/568574.568575

[8] Roy González-Alemán, David Hernández-Castillo, Julio Caballero, and Luis A Montero-Cabrera. 2020. Quality Threshold Clustering of Molecular Dynamics: A Word of Caution. *Journal of chemical information and modeling* 60, 2 (2020), 467–472.

[9] Luis Gracia. [n. d.]. WMC PhysBio clustering GUI. https://github.com/luisico

[10] M. Halkidi and M. Vazirgiannis. 2001. Clustering validity assessment: finding the optimal partitioning of a data set. (2001), 187–194. https://doi.org/10.1109/ICDM.2001.989517

[11] Berk Hess, Carsten Kutzner, David van Der Spoel, and Erik Lindahl. 2008. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of chemical theory and computation* 4, 3 (2008), 435. https://doi.org/10.1021/ct700301q

[12] L. J Heyer. 1999. Exploring Expression Data: Identification and Analysis of Coexpressed Genes. *Genome research* 9, 11 (1999), 1106–1115.

[13] Jason Hlozek, Neil Ravenscroft, and Michelle M. Kuttel. 2020. Effects of Glucosylation and O-Acetylation on the Conformation of Shigella flexneri Serogroup 2 O-Antigen Vaccine Targets. *The Journal of Physical Chemistry B* 124, 14 (2020), 2806–2814. https://doi.org/10.1021/acs.jpcb.0c01595 arXiv:https://doi.org/10.1021/acs.jpcb.0c01595 PMID: 32204588.

[14] William Humphrey, Andrew Dalke, and Klaus Schulten. 1996. VMD: Visual molecular dynamics. *Journal of molecular graphics* 14, 1 (1996), 33–38.

[15] Xin Jin and Jiawei Han. 2016. *Quality Threshold Clustering*. Springer US, Boston, MA, 1–2. https://doi.org/10.1007/978-1-4899-7502-7_692-1

[16] Michelle M Kuttel, Zaheer Timol, and Neil Ravenscroft. 2017. Cross-protection in Neisseria meningitidis serogroups Y and W polysaccharides: A comparative conformational analysis. *Carbohydrate research* 446-447 (2017), 40–47.

[17] Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (2018).

[18] Ryan Melvin, Ryan Godwin, Jiajie Xiao, William Thompson, Kenneth Berenhaut, and Freddie Salsbury. 2016. Uncovering Large-Scale Conformational Change in Molecular Dynamics without Prior Knowledge. *Journal of Chemical Theory and Computation* 12, 12 (2016), 6130–6146. https://doi.org/10.1021/acs.jctc.6b00757

[19] Ryan Melvin and Freddie Salsbury. 2016. Python Implementation of Quality Threshold Clustering for Molecular Dynamics. https://doi.org/10.6084/m9.figshare.3813930.v2

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[21] Jianyin Shao, Stephen W Tanner, Nephi Thompson, and Thomas E Cheatham. 2007. Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *Journal of chemical theory and computation* 3, 6 (2007), 2312–2334.

[22] Martin Stahl, Harald Mauser, Mark Tsui, and Neil R. Taylor. 2005. A robust clustering method for chemical structures. *Journal of medicinal chemistry* 48, 13 (2005), 4358. https://doi.org/10.1021/jm040213p

[23] Laurens van der Maaten and Geoffrey Hinton. 2008. Viualizing data using t-SNE. *Journal of Machine Learning Research* 9 (11 2008), 2579–2605.

[24] Guido Van Rossum and Fred L. Drake. 2009. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

[25] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Supplementary Material

Table A1: Summary of validation data set characteristics.

Data set	Sample set size	Centres	Cluster size	Variance
1	100	5	20	0.2
2	100	0	0	n/a
3	100	5	20	[1.5, 0.5, 0.9, 0.3, 0.8]
Iris	150	3	50	n/a

Table A2: Summary of validation results for Quality Threshold Algorithms

Data set	Algorithm	Minimum size	Cutoff value	Expected number of clusters	Number of clusters	Cluster counts	Silhouette	Davies-Bouldin	Calinski and Harabasz
1	qt_original	10	1	5	5	[20, 20, 20, 20, 20]	0.87	0.18	4863.54
1	qt_vector	10	1	5	5	[20, 20, 20, 20, 20]	0.87	0.18	4863.54
2	qt_original	10	1	0	2	[-12, 63, 25]	0.20	3.72	17.73
2	qt_vector	10	1	0	1	[100]	n/a	n/a	n/a
3	qt_original	10	4	5	3	[-6, 47, 24, 23]	0.45	2.47	168.12
3	qt_vector	10	4	5	2	[-1, 75, 24]	0.25	0.56	117.80
Iris	qt_original	10	1	3	4	[-54, 35, 21, 21, 19]	0.15	3.64	43.12
Iris	qt_vector	10	1	3	4	[-7, 58, 48, 19, 18]	0.41	3.60	180.94

Table A3: Summary of validation results for Hierarchical clustering (Ward)

Data set	Expected clusters	Number of clusters	Cluster counts	Cophenetic	Silhouette	Davies-Bouldin	Calinski and Harabasz
1	5	5	[20, 20, 20, 20, 20]	0.88	0.87	0.18	4863.55
2	0	5	[26, 24, 25, 6, 19]	0.62	0.25	1.12	34.18
3	5	5	[22, 35, 25, 4, 14]	0.89	0.44	0.87	242.18
Iris	3	3	[50, 37, 63]	0.93	0.55	0.66	554.82

Table A4: Summary of validation results for Hierarchical clustering (average)

Data set	Expected clusters	Number of clusters	Cluster counts	Cophenetic	Silhouette	Davies-Bouldin	Calinski and Harabasz
1	5	5	[20, 20, 20, 20, 20]	0.89	0.87	0.18	4863.55
2	0	5	[29, 31, 9, 9, 22]	0.68	0.30	1.03	39.99
3	5	5	[22, 36, 30, 4, 8]	0.89	0.46	0.77	244.28
Iris	3	3	[50, 37, 63]	0.93	0.55	0.66	554.82

Table A5: Summary of validation results for Hierarchical clustering (complete)

Data set	Expected clusters	Number of clusters	Cluster counts	Cophenetic	Silhouette	Davies-Bouldin	Calinski and Harabasz
1	5	5	[20, 20, 20, 20, 20]	0.89	0.87	0.18	4863.55
2	0	5	[15, 28, 45, 2, 10]	0.65	0.17	1.16	25.06
3	5	5	[4, 20, 41, 25, 10]	0.88	0.42	0.84	221.07
Iris	3	3	[50, 16, 84]	0.90	0.53	0.62	422.24

Table A6: Summary of validation results for Hierarchical clustering (single)

Data set	Expected clusters	Number of clusters	Cluster counts	Cophenetic	Silhouette	Davies-Bouldin	Calinski and Harabasz
1	5	5	[20, 20, 20, 20, 20]	0.90	0.87	0.18	4863.55
2	0	5	[6, 91, 1, 1, 1]	0.45	-0.09	0.83	3.84
3	5	5	[72, 1, 1, 2, 24]	0.82	-0.01	0.76	63.87
Iris	3	3	[50, 4, 96]	0.90	0.47	0.47	299.96

Table A7: Hierarchical clustering results for MenW

Type	Cluster Counts	Cophenetic	Silhouette	Davies-Bouldin	Calinski and Harabasz
Single	[2, 1, 3, 3991, 1, 1, 1, 1, 1, 1]	0.31	-0.25	1.05	1.37
Complete	[11, 151, 189, 1613, 491, 145, 207, 296, 664, 236]	0.51	0.02	1.94	313.97
Average	[25, 59, 26, 485, 767, 1301, 389, 241, 156, 554]	0.54	0.13	1.54	515.93
Ward's	[716, 239, 573, 364, 635, 194, 302, 414, 212, 354]	0.49	0.12	1.90	589.34

Table A8: Hierarchical clustering results for MenY

Type	Cluster Counts	Cophenetic	Silhouette	Davies-Bouldin	Calinski-Harabasz
Single	[2, 4, 3, 3985, 1, 1, 1, 1, 4, 1]	0.51	-0.06	0.74	7.33
Complete	[981, 126, 90, 206, 109, 496, 596, 1265, 54, 80]	0.44	0.05	2.07	302.64
Average	[26, 23, 1, 71, 6, 77, 180, 86, 3443, 90]	0.67	0.13	1.26	149.56
Ward's	[929, 196, 118, 498, 71, 94, 67, 804, 645, 581]	0.47	0.07	2.08	340.95

Table A9: QT cluster results for MenW

QT Type	Cutoff (Å)	Number of clusters	Cluster counts	Silhouette	Davies-Bouldin	Calinski-Harabasz
Original	1.5	2	[-3754, 139, 110]	0.24	1.52	98.84
Original	2.5	10	[-2046, 604, 286, 213, 159, 127, 123, 119, 110, 108, 108]	0.19	1.50	677.54
Original	3.5	11	[-427, 1148, 610, 368, 272, 232, 210, 188, 184 148, 111, 105]	0.12	1.70	474.78
Original	4.5	5	[-139, 2260, 696, 583, 178, 147]	0.08	2.13	292.54
Vector	1.5	8	[-2241, 650, 305, 177, 149, 134, 133, 119, 106]	0.23	1.35	697.78
Vector	2.5	8	[-336, 1599, 658, 547, 242, 198, 169, 144, 110]	0.09	2.00	434.62
Vector	3.5	2	[-148, 3281, 574]	0.20	2.08	457.12
Vector	4.5	1	[-16, 3987]	n/a	n/a	n/a

Table A10: QT cluster results for MenY

QT Type	Cutoff (Å)	Number of clusters	Cluster counts	Silhouette	Davies-Bouldin	Calinski-Harabasz
Original	1.5	12	[-1659, 493, 365, 304, 198, 153, 144, 134, 124, 117, 112, 100, 100]	0.06	2.09	248.00
Original	2.5	5	[-668, 1944, 655, 532, 104, 100]	0.10	1.92	337.24
Original	3.5	3	[-294, 2995, 401, 313]	0.15	1.68	482.18
Original	4.5	3	[-38, 3585, 197, 183]	0.20	1.59	310.84
Vector	1.5	3	[-786, 2185, 602, 430]	0.13	1.90	540.36
Vector	2.5	3	[-157, 3533, 192, 121]	0.21	1.51	298.69
Vector	3.5	1	[-62, 3941]	n/a	n/a	n/a
Vector	4.5	1	[4003]	n/a	n/a	n/a

Table A11: QT cluster results for Shigella flexneri Y 3RU

QT Type	Cutoff (Å)	Number of clusters	Cluster counts	Silhouette	Davies-Bouldin	Calinski-Harabasz
Original	2.5	7	[-2706, 320, 239, 191, 167, 137, 130, 112]	0.07	2.33	125.41
Original	3.5	10	[-781, 1056, 558, 455, 243, 243, 179, 143, 131, 112, 100]	0.03	2.49	231.03
Original	4.5	6	[-301, 1978, 737, 550, 156, 150, 130]	0.07	2.01	368.44
Vector	2.5	6	[-514, 1864, 589, 572, 208, 142, 113]	0.06	2.01	352.10
Vector	3.5	3	[-97, 3182, 419, 304]	0.10	1.65	231.03
Vector	4.5	3	[3748, 139, 115]	0.10	1.42	243.01

Table A12: QT cluster results for Shigella flexneri 6 6RU

QT Type	Cutoff (Å)	Number of clusters	Cluster counts	Silhouette	Davies-Bouldin	Calinski-Harabasz
Original	6.5	10	[-1078, 616, 384, 352, 265, 218, 185, 146, 124, 117, 117]	0.03	2.49	226.59
Original	7	9	[-797, 779, 432, 425, 288, 234, 192, 175, 160, 120]	0.04	2.38	276.90
Original	7.5	10	[-486, 946, 523, 440, 349, 176, 175, 149, 129, 117, 112]	0.02	2.55	246.75
Vector	6.5	3	[-192, 2288, 602, 520]	0.10	2.89	432.72
Vector	7	3	[-97, 2585, 523, 397]	0.08	2.06	395.61
Vector	7.5	3	[-47, 2879, 444, 232]	0.06	2.03	266.76

Table A13: Simulation Additional Information

MD Simulation	Alignment selection statement	Cluster selection statement
MenW	Pre-aligned	type != H and (((resname AGL or resname AGA) and not (name O2 or name O3 or name O4)) or (resname ASI and (name O4 or name C2 or name C3 or name C4 or name C5 or name O6))) and not resid 0 1 10 11
MenY	Pre-aligned	type != H and (((resname AGL or resname AGA) and not (name O2 or name O3 or name O4)) or (resname ASI and (name O4 or name C2 or name C3 or name C4 or name C5 or name O6))) and not resid 0 1 10 11
S. flexneri Y 3RU	resid 8 7 and name C1 C2 C3 C4 C5 O2 O5 and not index 157	name C1 C2 C3 C4 C5 O2 O3 O5 and not name NH N CT C O SOD and not resid 1 12 and not index 38 59 79 125 146 166 212 224 233
S. flexneri 6 6RU	resid 16 15 and name C1 C2 C3 C4 C5 O2 O5 and not index 327 337	name C1 C2 C3 C4 C5 O4 O2 O3 O5 and not name NH N CT C O SOD and not resid 1 2 3 4 21 22 23 24 and not index 108 131 148 144 127 194 164 168 217 213 234 230 250 254 280 299 303 320 340 336 316 366 385 389 402 406 426 422 413

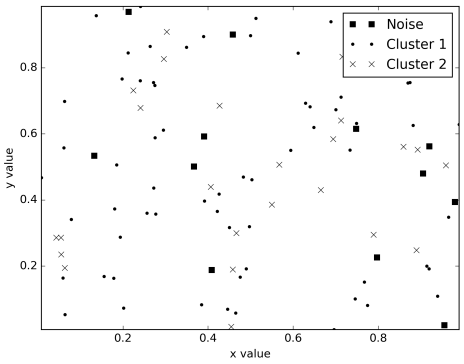

Figure B1: Clusters produced by the original Quality Threshold algorithm on Data set 2.

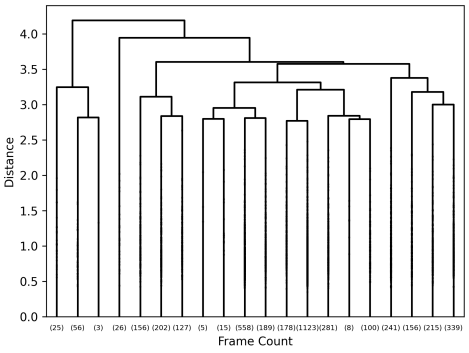

Figure B4: Dendrogram illustrating cluster formations of MenW using average linkage.

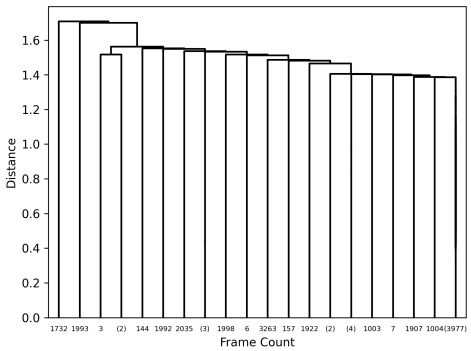

Figure B2: Dendrogram illustrating cluster formations of MenW using single linkage.

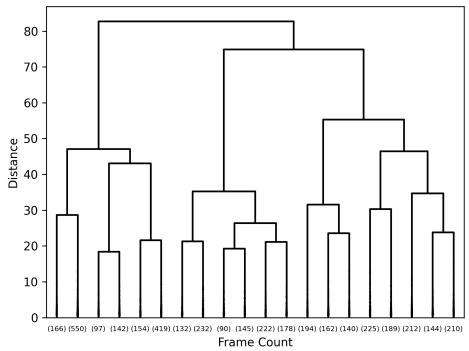

Figure B5: Dendrogram illustrating cluster formations of MenW using Ward's linkage.

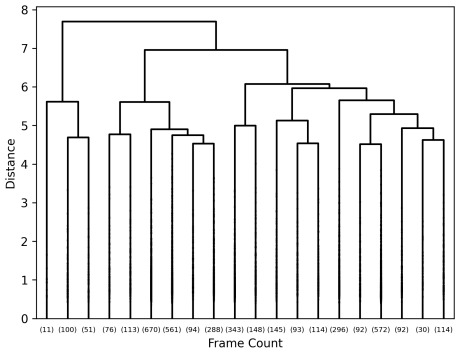

Figure B3: Dendrogram illustrating cluster formations of MenW using complete linkage.

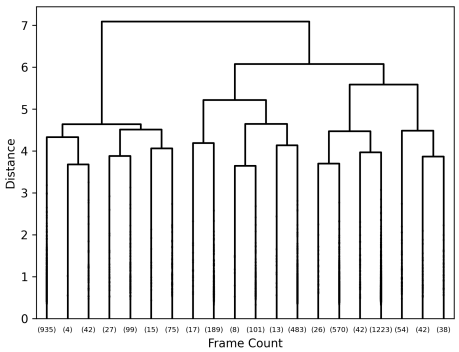

Figure B6: Dendrogram illustrating cluster formations of MenY using complete linkage.

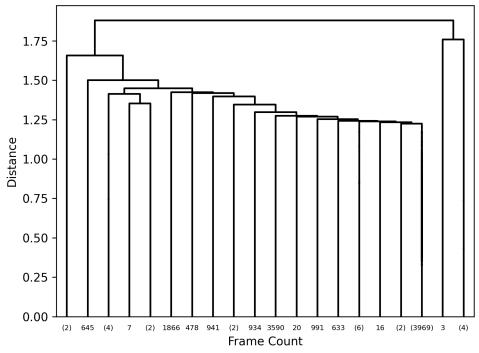

Figure B7: Dendrogram illustrating cluster formations of MenY using single linkage.

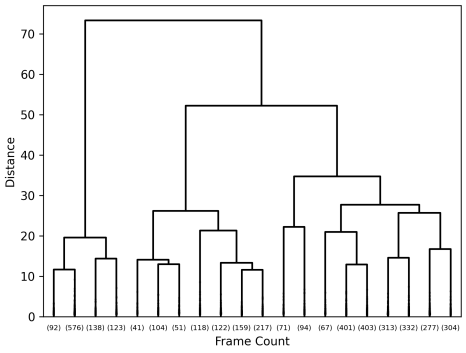

Figure B9: Dendrogram illustrating cluster formations of MenY using Ward's linkage.

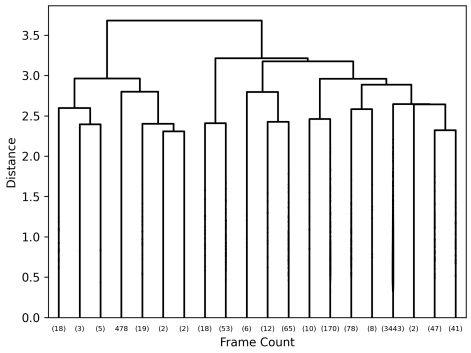

Figure B8: Dendrogram illustrating cluster formations of MenY using average linkage.