Published in final edited form as:

J Comput Chem. 2016 August 5; 37(21): 1973–1982. doi:10.1002/jcc.24416.

Cluster Analysis of Molecular Simulation Trajectories for Systems where Both Conformation and Orientation of the Sampled States are Important

Tigran M. Abramyan¹, James A. Snyder¹, Aby A. Thyparambil¹, Steven J. Stuart², and Robert A. Latour¹

¹Department of Bioengineering, 501 Rhodes Engineering Research Center, Clemson University, Clemson, SC, United States, 29634

²Department of Chemistry, 369 Hunter Laboratory, Clemson University, Clemson University, Clemson, SC, United States, 29634

Abstract

Clustering methods have been widely used to group together similar conformational states from molecular simulations of biomolecules in solution. For applications such as the interaction of a protein with a surface, the orientation of the protein relative to the surface is also an important clustering parameter because of its potential effect on adsorbed-state bioactivity. This study presents cluster analysis methods that are specifically designed for systems where both molecular orientation and conformation are important, and the methods are demonstrated using test cases of adsorbed proteins for validation. Additionally, because cluster analysis can be a very subjective process, an objective procedure for identifying both the optimal number of clusters and the best clustering algorithm to be applied to analyze a given dataset is presented. The method is demonstrated for several agglomerative hierarchical clustering algorithms used in conjunction with three cluster validation techniques.

Graphical Abstract

Cluster analysis of molecular simulation trajectories for molecular systems where both the conformation and orientation of the sampled states are important parameters requires a different approach from the standard clustering methods. These types of systems require cluster analysis methods that discriminate based on both molecular orientation and conformation. New clustering methods that account for both of these parameters are presented and demonstrated by the analysis of trajectories produced in protein-adsorption simulations for validation.

Keywords

Cluster analysis; molecular dynamics; protein adsorption; orientation; conformation

Introduction

Cluster analysis is a statistical data mining tool that seeks to divide data into groups or clusters that share similar qualities.¹ It requires a metric of similarity between the objects in a given dataset upon which a particular clustering algorithm either sorts the objects or partitions the dataset into separate groups. Ideally, the degree of similarity between two objects within the same cluster will be greater than the similarity between two objects that are in different clusters. Cluster analysis is thus used to reveal structure within a given data set, although it doesn't directly provide an explanation for why a particular pattern in the data exists.

For over two decades many research groups have used a variety of clustering algorithms to analyze molecular or system configurations obtained from atomistic simulation trajectories.² For example, Shenkin et al. described a method of cluster analysis in which a pairwise inter-conformational distance matrix in either torsional or Cartesian space was first calculated and then an agglomerative single-link clustering method was used to define clusters. The method was embodied in a program called Xcluster. The folding and unfolding of a three-helix-bundle protein were explored through cluster analysis in the work of Boczek & Brooks. In

their study clustering was performed with hierarchical agglomerative Ward's method, in which the pairwise distance between two structures incorporated the interatomic contact distance between core side-chains, helical hydrogen bond distance, and solvent-accessible surface area. Daura et al. performed MD simulation studies on the folding of two β -peptides, and developed a cluster analysis technique based on backbone root mean square deviation (*RMSD*) using the nearest neighbor algorithm. In a cluster analysis after Brownian dynamics of a protein, Mereghetti et al. used the *RMSD* of all the atoms of the protein as the metric along with agglomerative single-linkage clustering. In a recent rigorous study comparing various cluster analysis methods, the Cheatham group analyzed DNA simulation data using eleven different clustering algorithms, ranging from hierarchical (both divisive and agglomerative) to refinement (means, Bayesian, and self-organizing maps) clustering algorithms. In general, their analysis showed that there is no single ideal algorithm for clustering simulation trajectories. However, they did recommend using the hierarchical average-linkage clustering algorithm if the cluster count was unknown in advance. Importantly, they found that each algorithm has limitations, such as the sensitivity to outliers of some clustering methods, or the tendency of the *k*-means algorithm to generate uniform clusters.

The cluster analysis methods that have previously been developed for analyzing simulation trajectories of biomolecules (e.g., protein, DNA, or lipids), have primarily been developed and used for the analysis of their conformational behavior in solution, in which case the orientation of a given sampled state is not important. However the cluster analysis of datasets where molecular orientation is important (e.g., molecules adsorbed to a material surface) also requires consideration of a molecule's orientation, in addition to its conformational state. This is of particular relevance for protein adsorption behavior, where both the adsorbed orientation and conformation can influence the bioactive state of the protein. For example, the adsorbed orientation of a protein can lead to steric blockage of the bioactive site with subsequent substantial loss in bioactivity, while the same conformation may fully retain its bioactivity if it is adsorbed in an orientation such that the bioactive site is still available for binding to its intended substrate. To address this limitation, we developed a methodology for the cluster analysis of datasets that accounts for both the conformation and orientation of the sampled states of the system while also providing an objective process for identifying the optimum number of clusters and cluster analysis method that should be applied for a given system. We demonstrate the application of these methods by the analysis of collections of sampled states from simulations of protein adsorption on a material surface.

Methods

Structural alignment of adsorbed configurations

In isotropic systems, such as biomolecules in solution, the *RMSD* between a pair of structures is calculated after a translational and orientational alignment, in which one structure is translated and rotated arbitrarily in three dimensions to minimize the *RMSD*. This defines an alignment vector (x, y, θ, ϕ, χ).

For cases like protein adsorption simulations, where the surface imposes both an orientational and translational anisotropy, a full three-dimensional translation and orientation

is not appropriate. We describe two different methods of structural alignment that can be applied in such cases. The discussion here uses the example of proteins adsorbed on a planar material surface for the sake of concreteness, but the methods are equally applicable to any other system where translational and orientational anisotropy is present, because of a surface, substrate, structural matrix or external field.

The first method that we consider is the alignment of the sampled adsorbed states of the protein using only translation in directions parallel to the surface plane (taken as the x - y coordinate plane) with simultaneous rotation about the axis normal to the surface plane (i.e. the z axis) to minimize the *RMSD* with respect to an arbitrarily chosen reference frame (we use the first frame in the trajectory). This alignment procedure thus generates frames of the adsorbed protein which differ from the input trajectory frames by a 3-component alignment vector (x , y , θ_z), as shown in Figure 1 (**Method 1**). This method clusters the sampled states based on their orientation on the material surface, but discriminates between structures that are different distances from the surface. The disadvantage of this alignment method, however, is that it will treat sampled states that have the same orientation and conformation but with their centers of gravity displaced relative to the z coordinate direction as belonging to different clusters. This method would be suitable to help prevent adsorbed and non-adsorbed structures from appearing in the same cluster, even when they are conformationally very similar. It would be unsuitable, however, for adsorption on surfaces with steps or other surface features that would cause identically adsorbed proteins to have different z coordinates, or a system with an external field that induces orientational but not translational anisotropy. In such cases, Method 2 of Figure 1 would be more appropriate.

The second method of structural alignment of the adsorbed protein frames is similar to the first, but includes a translation along the z axis. This type of alignment produces frames which differ from the input trajectory by a 4-component alignment vector of (x , y , z , θ_z), as shown in Figure 1 (**Method 2**). With this method, two structures with similar conformation and orientation relative to the material surface, but with different distances from the defined surface plane, will be combined into the same cluster.

Both of these methods assume molecular adsorption onto a surface with planar isotropy. Alternative methods could be applied for systems where the surface is strongly patterned, or periodic, or non-planar.

Structural alignment of adsorbed configurations

Our clustering approach follows the **typical four basic steps for cluster analysis of any dataset**: feature selection or extraction, cluster algorithm design and selection, cluster validation, and results interpretation.

In the **first step of cluster analysis**, called feature selection, 3-D coordinates of C_α atoms in the protein backbone are extracted as the feature for the clustering. It is important to mention here that the choice of the protein atoms for the pairwise comparison in cluster analysis can greatly influence the clustering results. The highly mobile parts of the protein (e.g., random loop segments) may increase the noise in the structural data. While it may be desirable to ignore these more random segments of a protein when performing cluster analysis for the

analysis of protein behavior in solution, these same segments, which commonly occur along a protein's surface, may play an important role in the adsorption process, may have their structure altered as a consequence of adsorption, or may have their fluctuations damped significantly upon adsorption. Therefore, these segments are not omitted in our procedure for cluster analysis of adsorbed protein frames, because they may contain useful information.

The **second step of cluster analysis** is associated with the selection of the clustering algorithm. Several methods of agglomerative hierarchical clustering, including single-linkage (nearest-neighbor algorithm), complete-linkage (furthest-neighbor algorithm), average-linkage (unweighted paired-group method with arithmetic mean), and Ward's method, as well as partitional clustering methods, such as *k*-means,² were selected for evaluation in our study. These methods were selected to represent some of the most commonly used clustering algorithms.

As the name suggests, any agglomerative hierarchical, or bottom-up, clustering algorithm, starts by assigning each sample in a dataset to its own single cluster. At each following iteration, the closest clusters are merged to form a larger cluster, with this process continued until all of the sampled states are finally combined into a single large cluster in the final iteration. A typical representation of bottom-up clustering is the 'tree' of clusters, or dendrogram, with the 'root' being the largest cluster containing all of the sampled states, and each 'leaf' being a singleton cluster (i.e., each sampled state being its own cluster). The differences between agglomerative hierarchical algorithms are based on the distance measure used to decide which clusters to merge in each iteration.

Single-linkage, for instance, uses the smallest distance between objects in the two clusters as a decision for merging two clusters. In contrast, **complete-linkage** uses the largest distance between objects in the two clusters. The merging criterion for **average-linkage** is the average distance between all pairs of objects in any two clusters. **Ward's method** uses the total sum of squared deviations (*SSD*) from the clusters' centroids. The fusion criterion is based on minimizing the increase in *SSD*. The ***k*-means** algorithm, which is similar to Ward's method in the way that it also minimizes the *SSDs* from the clusters' centroids, first assigns *k* arbitrary initial centroids. Then the algorithm forms clusters by assigning each observation to the nearest centroid. The centroids are then recalculated by minimizing the *SSD* between the centroid and each of the observations in the cluster. These three steps are iterated until the minimum within-cluster *SSDs* from the cluster centroid is achieved.

In many applications involving molecular conformational analysis, the number of clusters is not known ahead of time. This is not a problem for hierarchical agglomerative methods, because the dendrogram provides the optimal clustering for any number of clusters. The *k*-means method, on the other hand, must be run separately for each cluster number, so it becomes computationally very expensive to consider a broad range of cluster numbers. Hence, we decided to drop *k*-means from our evaluation because of its low efficiency.

The main objective of the **third step of cluster analysis**, called cluster validation, is to determine the best partition or the optimal number of clusters that a given dataset should be

grouped into. It is commonly agreed that this is the most challenging step in a clustering procedure. Heuristic evaluation standards and criteria are often used in this step. Among these are: (1) a rule of thumb that uses $\sqrt{N/2}$, clusters, where N is the total number of samples in the dataset; (2) the ‘elbow method’, in which the optimal cluster count is approximated visually by finding the ‘elbow’ region of the objective function versus the number-of-clusters plot; (3) finding the best cutting point of the dendrogram obtained in an agglomerative hierarchical clustering, either visually or using an inconsistency coefficient; or (4) using external and/or internal criteria for cluster validation.

Since *a priori* knowledge about the cluster counts in the ensembles of states obtained from a simulation trajectory is typically unavailable, in this study we employed three commonly used internal validation criteria for this assessment as example methods: Calinski-Harabasz (CH), Davies-Bouldin (DB), and silhouette (S) indices. The CH index assess the clustering performance based on the ratio of between-cluster variance (SSB) to within-cluster variance (SSW). Higher values of this coefficient are associated with the optimal cluster counts. The contribution to the DB index for each cluster is the ratio of the within-cluster variances to the between-cluster distance, maximized over all other clusters. Smaller values of the index therefore indicate better clustering. The S index includes a contribution from each structure that contributes positively if it is closer to every point in its own cluster than any other, and negatively if it is closer to the points in another cluster than its own. Maximizing the index gives the best clusters.

While these techniques generally work well, none of them is necessarily superior to the others, and, as suggested by Xu and Wunch, generally “it is advisable not to depend on a single rule for selecting the number of groups, but to synthesize the results of several techniques.” In light of this statement, in our study we implemented all three cluster validation approaches to evaluate the effectiveness of each of the four clustering agglomerative hierarchical clustering methods.

Identifying the optimal number of clusters that a given dataset should be divided into can be a very subjective process. In order to provide a more objective approach for this process, we propose a methodology in which a given dataset of sampled states is (i) subjected to clustering via four clustering algorithms—single-linkage, complete-linkage, average-linkage, and Ward’s method. Then (ii) the three cluster count validation methods described above—CH, S, and DB—are applied to each algorithm’s clustering results to evaluate the cluster solutions. The resulting 12 graphs of validation index vs. number of clusters are plotted and all local maxima or minima (depending on the validation technique) corresponding to potential cluster solutions are determined. (iii) Cumulative number of observations of cluster solutions from these 12 graphs are then calculated and the optimal number of clusters is chosen as that with the highest occurrence among all 12 clustering / validation techniques. (iv) With this optimal number of clusters, there is then one distinct clustering given by each of the four clustering methods. For each of these clusterings, the distribution of within-cluster $C_a \overline{RMSD}$ is evaluated, with respect to the average structure in that cluster. The clustering algorithm which produces the lowest and most consistent

values of the $C_\alpha \overline{RMSD}$ is then selected, and the clustering it generates for the optimal cluster count is used.

In the **fourth and final step of cluster analysis**, results interpretation, the clusters can be analyzed to determine the properties of interest for the system that is being studied.

We cannot provide a proof that this process identifies the ‘optimal’ number of clusters and analysis algorithm that should be used for a given dataset. (Indeed, such a proof is not possible, since the optimal clustering will vary depending on the distance metric and contains unavoidable subjectivity.) Nonetheless, the results are satisfactory for the range of systems considered, and we propose this procedure as a means to provide an objective basis that can be followed for cluster analysis of molecular conformations.

Model systems and generation of trajectories

For the purpose of demonstrating the above-described procedure for cluster analysis, three different protein-adsorption model systems were generated and used to produce molecular dynamics (MD) simulation trajectories for evaluation. These model systems were used for four separate test cases (TC) 1–4 (**TC 1**, **TC 2**, **TC 3**, and **TC 4**), where the same sampled dataset was used for TC 1 and TC 2. These test cases are illustrated in Figure 2.

TC 1 and TC 2—To evaluate the general performance of the two different methods of structural alignment described above, a common trajectory was designed to test clustering following alignment-methods 1 (**TC 1**) and 2 (**TC 2**). For these test cases, three different orientations of ribonuclease A in its native-state structure (RNase, PDBID 5rsa) in vacuum were placed either directly in contact with or translated 10 Å above a surface, which was represented by the (110) plane of high-density polyethylene (HDPE), thus providing six different configurational states (Figure 2a). Short 200 fs MD simulations of each system at 300 K were then performed to slightly perturb each state, with each protein’s center of mass restrained with a harmonic potential to maintain its general position above the surface plane. Coordinates of the MD simulations were sampled every 5 fs, thus generating an ensemble of 40 states for each protein that were closely associated with its initial orientation, conformation, and position over the surface. The resulting trajectories were then merged into a dataset in which a successful clustering method should identify six distinct clusters when using alignment method 1, or three distinct clusters when using alignment method 2.

TC 3—Our second model system and third test case was designed to assess the ability of the various clustering algorithms to analyze a set of states for which the clusters are slightly overlapping. I.e., it provides a more complex case to test the ability of our proposed objective cluster analysis process to identify the optimum number of clusters, but for a system where the correct answer is known. For this test case, hen-egg white lysozyme (HEWL, PDBID 1gxv) was placed over the HDPE surface with its long axis parallel to the surface and rotated along this axis to produce ten different orientations of the protein on the surface. MD simulations were then conducted for 200 ps in vacuum at 300 K, with trajectories saved every 10 ps. The sampled states were then combined to form a dataset that should be partitioned into ten different clusters, each containing twenty structures (Figure

2b). Since each protein was in close proximity to the surface plane, the analysis was performed using alignment method 1.

TC 4—Our third model system was designed to provide an ensemble of sampled states of an adsorbed protein where the cluster count was unknown. The trajectory for this study was produced from a 19 ns TIGER2A sampling simulation of HEWL protein over a silica glass surface with explicit water to provide an ensemble of sampled states consisting of 405 adsorbed protein configurations on the silica glass surface (Figure 2c). This provides a more strenuous and realistic test of the objective clustering analysis procedure.

Results and Discussion

We used our proposed cluster analysis procedure to first analyze the test cases with known numbers of clusters (i.e., TC 1, TC 2, and TC 3) to assess the ability of our procedure to correctly identify the number of clusters. Per design, the number of clusters expected in these test cases was 6, 3, and 10, respectively. Once we were able to document that this procedure was functioning as intended, we then applied the same procedure to evaluate TC 4, which contained a larger set of sampled states with unknown clustering, to demonstrate its application to analyze a collection of sampled states that were obtained from an actual MD simulation of protein adsorption behavior.

Accordingly, following the alignment of the protein frames for model systems TC 1–3, cluster analysis was performed using our selected set of four agglomerative hierarchical clustering algorithms—single-linkage, complete-linkage, average-linkage, and Ward’s method. Qualitative inspection of the objective function using the ‘elbow method’ as well as dendrogram plots (i.e., a hierarchy tree or tree of clusters) confirm that the datasets contain the expected number of clusters (Figure 3): six clusters in TC 1 following alignment-method 1, three clusters in TC 2 following alignment-method 2, and (at least with some methods) ten clusters for the TC 3. The heuristic approaches begin to be somewhat ambiguous for the more complex TC 3, affirming the need for a more objective method of determining an appropriate number of clusters to be used for system analysis.

Figure 4 demonstrates a more objective approach to determining the optimal number of clusters in a dataset. Validation indices are shown at every cluster size for each of three internal validation techniques (CH, S, and DB) indices, and each of the four clustering methods considered. Local maxima (CH and S) or minima (DB) on these plots correspond to potential cluster numbers. These include several false positives that do not correspond to useful clusterings, but these false positives differ for different methods. The consensus from all 12 combinations is presented in Figure 5, which shows the number of observations with which varying cluster numbers were identified. The largest value in these plots indicates the consensus decision, or cluster number most frequently identified by the 12 clustering/validation combinations.

For models TC 1 and TC 2, the number of observations plots provide the same result as the more heuristic approaches. All 12 combinations successfully identify 6 clusters for TC 1 (Figure 5a) and 3 clusters for TC 2 (Figure 5b). For the more complex model system TC 3,

the number of observations plot (Figure 5c) provides a less ambiguous result than the heuristic approaches. Although the clustering/validation combinations were not unanimous, 10 of the 12 of them correctly identified ten clusters within the dataset. This is a count more than twice as large as any of the false positives.

Having demonstrated the application of the clustering procedure for our three model test cases, we now apply the same procedure to cluster the sampled states from a production-scale MD simulation of a protein adsorbing on a silica surface, system TC 4. The objective function and dendrogram data in Figure 6 provide a general indication that the appropriate number of clusters should be somewhere between about 5–25, although the ‘elbows’ identified by different clustering methods are not distinct, and differ substantially.

Figure 7 shows the internal validation indices for each combination of clustering algorithm and validation technique. Each method identifies several potential clusterings, with no sharp maxima or minima as in model systems TC 1 and TC 2 (and, to a lesser degree, TC 3). The consensus result, shown in Figure 8a, is much less ambiguous. The most frequently identified number of clusters for TC 4 is fairly clearly either 11 or 17, each of which were identified by 7 of the 12 clustering / validation combinations.

Having identified these two optimal numbers of clusters, it still remains to choose the most appropriate clustering. To do so, the within-cluster $C_{\alpha} \overline{RMSD}$ was evaluated for each of the (11 or 17) clusters generated by each of the four clustering algorithms, as shown in Figure 8b. This analysis shows that the Ward method provides the most constant and lowest $C_{\alpha} \overline{RMSD}$ values, especially for the largest clusters. (In configurations generated from equilibrium sampling, the larger clusters will be the most probable and physically important ones, so tight clustering is more important in these states than in smaller clusters representing comparatively rare configurations.) Thus, the Ward method is chosen as the most accurate for this dataset. Figure 8c shows the sizes of each of the 11 and 17 clusters as predicted using the Ward method of clustering. Based on the general principle that fewer clusters are preferred to simplify the overall analysis of a system, we therefore select the lower number of 11 clusters for this system, which can be used for the analysis of desired properties of the molecular system.

Conclusions

For molecular systems where both the orientation and conformation of sampled states are important parameters, conventional cluster analysis methods that cluster based only on conformation are not appropriate. Instead, cluster analysis methods that also discriminate based on molecular orientation are required.

In this work, we propose two methods of cluster analysis to account for molecular orientation, with the difference between the methods being which degrees of freedom are used for alignment of the sampled states (i.e., translation and/or rotation in 3-D Cartesian space). Different methods may be applicable depending on the specific system that is to be analyzed. Also, in an attempt to reduce subjectivity in cluster analysis, we developed a procedure that can be objectively applied for identifying the optimal number of clusters and

selecting the algorithm to be applied for final cluster analysis for a given dataset and given set of clustering methods.

Our proposed methods and procedure were first assessed and validated for three model-system test cases using four selected clustering methods in combination with three validation techniques. We then further demonstrated their use by application to cluster a set of sampled states from a production-scale MD simulation of protein adsorption on a silica surface. While the Ward method was identified as the best performing cluster analysis method in this particular application, the general method preserves the potential for different techniques to be evaluated to select the one that performs the best for a given dataset and application. Regardless of the specific algorithms considered, the proposed procedure should thus be useful for identifying the optimal cluster number and clustering method to be applied for a given system.

Acknowledgments

This project received support from the Defense Threat Reduction Agency-Joint Science and Technology Office for Chemical and Biological Defense (Grant no. HDTRA1-10-1-0028), with partial support also from NIH Grant No. P41 EB001046 and the Department of Bioengineering of Clemson University. Computing resources: Palmetto Linux Cluster, Clemson University. We would also like to acknowledge helpful discussions with Dr. Delphine Dean, Department of Bioengineering of Clemson University.

References

1. Tryon, RC. Cluster analysis: correlation profile and orthometric (factor) analysis for the isolation of unities in mind and personality. Edwards brother, Incorporated, lithoprinters and publishers; 1939.
2. Jain AK, Murty MN, Flynn PJ. Data clustering: a review. *ACM Comput. Surv.* 1999; 31(3):264–323.
3. Rui X, Wunsch D II. Survey of clustering algorithms. *IEEE T. Neural Networ.* 2005; 16(3):645–678.
4. Scheraga HA, Khalili M, Liwo A. Protein-Folding Dynamics: Overview of Molecular Simulation Techniques. *Annu. Rev. Phys. Chem.* 2007; 58(1):57–83. [PubMed: 17034338]
5. Oren EE, Notman R, Kim IW, Evans JS, Walsh TR, Samudrala R, Tamerler C, Sarikaya M. Probing the Molecular Mechanisms of Quartz-Binding Peptides. *Langmuir.* 2010; 26(13):11003–11009. [PubMed: 20499870]
6. Shao J, Tanner SW, Thompson N, Cheatham TE. Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *J. Chem. Theory Comput.* 2007; 3(6):2312–2334. [PubMed: 26636222]
7. Krivov SV, Karplus M. Hidden complexity of free energy surfaces for peptide (protein) folding. *P. Natl. Acad. Sci. USA.* 2004; 101(41):14766–14770.
8. Karpen ME, Tobias DJ, Brooks CL. Statistical clustering techniques for the analysis of long molecular dynamics trajectories: analysis of 2.2-ns trajectories of YPGDV. *Biochemistry-US.* 1993; 32(2):412–420.
9. Shenkin PS, McDonald DQ. Cluster analysis of molecular conformations. *J. Comput. Chem.* 1994; 15(8):899–916.
10. Torda AE, van Gunsteren WF. Algorithms for clustering molecular dynamics configurations. *J. Comput. Chem.* 1994; 15(12):1331–1340.
11. Woolf TB, Roux B. Molecular dynamics simulation of the gramicidin channel in a phospholipid bilayer. *P. Natl. Acad. Sci. USA.* 1994; 91(24):11631–11635.
12. Boczek EM, Brooks C. First-principles calculation of the folding free energy of a three-helix bundle protein. *Science.* 1995; 269(5222):393–396. [PubMed: 7618103]
13. Troyer JM, Cohen FE. Protein conformational landscapes: Energy minimization and clustering of a long molecular dynamics trajectory. *Proteins.* 1995; 23(1):97–110. [PubMed: 8539254]

14. Rhee Y, Pande V. Multiplexed-Replica Exchange Molecular Dynamics Method for Protein Folding Simulation. *Biophys. J.* 2003; 84(2):775–786. [PubMed: 12547762]
15. Brünger AT, Brooks CL, Karplus M. Active site dynamics of ribonuclease. *P. Natl. Acad. Sci. USA.* 1985; 82(24):8458–8462.
16. Unger R, Harel D, Wherland S, Sussman JL. A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins.* 1989; 5(4):355–373. [PubMed: 2798411]
17. Gordon HL, Somorjai RL. Fuzzy cluster analysis of molecular dynamics trajectories. *Proteins.* 1992; 14(2):249–264. [PubMed: 1409572]
18. Michel AG, Jeandenans C. Multiconformational investigations of polypeptidic structures, using clustering methods and principal components analysis. *Comput. Chem.* 1993; 17(1):49–59.
19. Radkiewicz JL, Brooks CL. Protein dynamics in enzymatic catalysis: exploration of dihydrofolate reductase. *J. Am. Chem. Soc.* 2000; 122(2):225–231.
20. Lahiri A, Nilsson L. Molecular dynamics of the anticodon domain of yeast tRNA (Phe): codon-anticodon interaction. *Biophys. J.* 2000; 79(5):2276. [PubMed: 11053108]
21. Mereghetti P, Gabdoulline RR, Wade RC. Brownian Dynamics Simulation of Protein Solutions: Structural and Dynamical Properties. *Biophys. J.* 2010; 99(11):3782–3791. [PubMed: 21112303]
22. Bystroff C, Garde S. Helix propensities of short peptides: Molecular dynamics versus bioinformatics. *Proteins.* 2003; 50(4):552–562. [PubMed: 12577261]
23. Moraitakis G, Goodfellow JM. Simulations of Human Lysozyme: Probing the Conformations Triggering Amyloidosis. *Biophys. J.* 2003; 84(4):2149–2158. [PubMed: 12668424]
24. Rao F, Karplus M. Protein dynamics investigated by inherent structure analysis. *P. Natl. Acad. Sci. USA.* 2010; 107(20):9152–9157.
25. Pandini A, Fornili A, Fraternali F, Kleinjung J. Detection of allosteric signal transmission by information-theoretic analysis of protein dynamics. *FASEB J.* 2012; 26(2):868–881. [PubMed: 22071506]
26. Papaleo E, Mereghetti P, Fantucci P, Grandori R, De Gioia L. Free-energy landscape, principal component analysis, and structural clustering to identify representative conformations from molecular dynamics simulations: The myoglobin case. *J. Mol. Graph. Model.* 2009; 27(8):889–899. [PubMed: 19264523]
27. Kannan S, Zacharias M. Folding simulations of Trp-cage mini protein in explicit solvent using biasing potential replica-exchange molecular dynamics simulations. *Proteins.* 2009; 76(2):448–460. [PubMed: 19173315]
28. Rodrigues JR, Simões CJV, Silva CG, Brito RMM. Potentially amyloidogenic conformational intermediates populate the unfolding landscape of transthyretin: Insights from molecular dynamics simulations. *Protein Sci.* 2010; 19(2):202–219. [PubMed: 19937650]
29. Daura X, Gademann K, Jaun B, Seebach D, van Gunsteren WF, Mark AE. Peptide Folding: When Simulation Meets Experiment. *Angew. Chem. Int. Edit.* 1999; 38(1–2):236–240.
30. Coutsias EA, Seok C, Dill KA. Using quaternions to calculate RMSD. *J. Comput. Chem.* 2004; 25(15):1849–1857. [PubMed: 15376254]
31. Everitt, B.; Landau, S.; Leese, M. Cluster analysis. 2001. Arnold, London: 2001.
32. Gower JC, Ross G. Minimum spanning trees and single linkage cluster analysis. *Appl. Stat.-J. Roy. St. C.* 1969:54–64.
33. Defays D. An efficient algorithm for a complete link method. *Comput. J.* 1977; 20(4):364–366.
34. Sokal RR. A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.* 1958; 38:1409–1438.
35. Ward JH Jr. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 1963; 58(301):236–244.
36. Forgey E. Cluster analysis of multivariate data: Efficiency vs. interpretability of classification. *Biometrics.* 1965; 21(3):768–769.
37. MacQueen, J. Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967; Oakland, CA, USA. 1967. p. 281–297.
38. Lloyd S. Least squares quantization in PCM. *IEEE T. Inform. Theory.* 1982; 28(2):129–137.

39. Hair, JF.; Black, WC.; Babin, BJ.; Anderson, RE.; Tatham, RL. Multivariate data analysis. Vol. 6. Upper Saddle River, NJ: Pearson Prentice Hall; 2006.
40. Mooi, E.; Sarstedt, M. Cluster analysis. Springer; 2011.
41. Milligan GW, Cooper MC. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*. 1985; 50(2):159–179.
42. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. *J. Roy. Stat. Soc. B*. 2001; 63(2):411–423.
43. Jain, AK.; Dubes, RC. Algorithms for clustering data. Prentice-Hall Inc.; 1988.
44. Zahn CT. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE T. Comput.* 1971; 100(1):68–86.
45. Xu R, Wunsch D. Clustering. 2008; 10 Wiley.com.
46. Cali ski T, Harabasz J. A dendrite method for cluster analysis. *Commun. Stat. Theory*. 1974; 3(1): 1–27.
47. Davies DL, Bouldin DW. A cluster separation measure. *IEEE T. Pattern Anal.* 1979; 2:224–227.
48. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 1987; 20:53–65.
49. Wlodawer A, Borkakoti N, Moss D, Howlin B. Comparison of two independently refined models of ribonuclease-A. *Acta Crystallogr. B*. 1986; 42(4):379–387.
50. Refaee M, Tezuka T, Akasaka K, Williamson MP. Pressure-dependent changes in the solution structure of hen egg-white lysozyme. *J. Mol. Biol.* 2003; 327(4):857–865. [PubMed: 12654268]
51. Li X, Snyder JA, Stuart SJ, Latour RA. TIGER2 with solvent energy averaging (TIGER2A): An accelerated sampling method for large molecular systems with explicit representation of solvent. *J. Chem. Phys.* 2015; 143(14):144105. [PubMed: 26472361]

Figure 1.

Illustration of the clustering of frames of sampled states obtained for structured molecules in solution compared to the molecules adsorbed on a surface. Two methods of alignment are displayed for the adsorbed condition. **Method 1**—alignment of adsorbed molecules in which RMSD is minimized through alignment vector (x , y , θ_z); in this method the z coordinates remain unmodified. **Method 2**—alignment of adsorbed molecules in which RMSD is minimized through alignment vector (x , y , z , θ_z).

Figure 2.

Illustration of test cases TC 14. **(a)** Model system for TC 1 and TC 2 is composed of two sets of three different conformational states of RNase over an HDPE surface; one set was positioned close to the HDPE surface and the other translated 10 Å further away from the surface. **(b)** Model system for TC 3 is composed of similar conformational states of HEWL over the HDPE surface, but with 10 different orientations. **(c)** Model system for TC 4 is composed of 405 sampled states of HEWL over a silica glass surface that were obtained from an MD simulation with unknown cluster solutions. All the configurations in the

trajectories produced are shown in all-atom representation (**a–c**), while a representative structure from each cluster is shown in ribbons view (**a** and **b**).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 3.

Objective function at each cluster number (plots on the left) as needed for the ‘elbow method’ of visually identifying the cluster number at which the slope changes abruptly, and dendrograms (plots on the right) generated with each of the agglomerative hierarchical clustering algorithms used in this study for TC 1–3. A suggested threshold line is drawn (blue dashed line) from the elbow region cutting the corresponding dendrogram.

Figure 4.

Validation indices as a function of cluster number for each combination of clustering algorithm (rows) and internal validation techniques (columns) for models TC 1–3 (a–c). Red points indicate the local maxima (CH and S) or minima (DB), which correspond to potential cluster solutions.

(a)

(b)

(c)

Figure 5. Number of observations (labels as Count on the graphs) of the cluster counts obtained from all combinations of clustering methods and validation technique for TC 1–3 (a–c). The highest count corresponds to the optimal solutions for the number of clusters in the trajectory data.

TC 4

Figure 6. Objective function at each cluster number (plots on the left) as needed for the ‘elbow method’ and dendrograms (plots on the right) generated with each of the agglomerative hierarchical clustering algorithms used in this study for TC 4.

Figure 7. Validation indices as a function of cluster number for each combination of clustering algorithm (rows) and internal validation technique (columns) for TC 4. Red points indicate the local maxima (CH, S), and minima (DB), which correspond to potential clustering solutions.

Figure 8.

(a) Number of observations of the cluster counts obtained from all combinations of clustering methods and validation techniques for TC 4. **(b)** C_{α} $\overline{RMSD}$ for each cluster, with respect to the average structure in the cluster. The clusters are numbered in order of decreasing size; i.e. cluster 1 is the largest cluster. **(c)** Sizes of each of the 11 and 17 clusters as identified using the Ward algorithm. The inset shows a characteristic structure from the two largest clusters.