Published in final edited form as:

J Phys Chem B. 2018 December 13; 122(49): 11807–11816. doi:10.1021/acs.jpcb.8b08852.

CATS: A tool for clustering the ensemble of intrinsically disordered peptides on a flat energy landscape

Jacob C. Ezerski^{1,2}, Margaret S. Cheung^{1,2}

¹Department of Physics, University of Houston

²Center for Theoretical Biological Physics, Rice University

Abstract

We introduce the combinatorial averaged transient structure (CATS) clustering method as a means to cluster protein structure ensembles based on the distributions of protein backbone descriptor coordinates. In our study, we use phi and psi dihedral angle coordinates of the protein backbone as descriptors due to their translational and rotational invariance. The CATS method was developed to produce unique structure ensembles that are typically difficult to obtain from flat energy landscapes using a one dimensional separation value (eg. RMSD cut-off). Through the use of higher dimensional descriptor coordinates, we remedy structure resolution shortcomings of standard clustering algorithms due to large RMSD fluctuations between structures. We compare the performance of CATS to an RMSD-based clustering method GROMOS, which may not be the best choice of IDP clustering since separation quality heavily relies on cutoff values instead of energy landscape minima. We demonstrate the performance of CATS and GROMOS by analyzing the all-atom molecular dynamics trajectories of the Tau/R2(273–284) fragment in solution with TMAO and urea osmolytes from prior studies. Our study reveals that the CATS method produces more unique clusters than the GROMOS method as a result of higher dimensional distributions of the descriptor coordinates. The cluster centers produced by CATS correspond to local minima in the multi-dimensional potential mean force, which generates a structure ensemble that adequately samples the energy landscape.

I. Introduction

The datasets from molecular dynamics (MD) simulations can provide valuable insight into protein interactions and dynamics through clustering analysis [1]. There are several main categories of clustering method: hierarchical (eg. single linkage) [2], vector quantization (eg. K-Means)[3, 4], neural network[5, 6], mixture density[7], and fuzzy [8, 9], to name a few [2, 7, 10, 11]. In addition to these, new methods routinely emerge that encompass characteristics from multiple algorithm categories to optimize the clustering of a specific set of data [2, 12–14]. Reduction of dimensionality is sometimes desired to reduce computational cost and necessary sample size; however the reduction of dimensions also can lead to lower separation resolution [15]. RMSD based clustering methods are typically employed to find dominant protein structures due to its simplicity and low dimensionality, however this technique heavily relies on an appropriate cutoff between cluster groups to be effective. In RMSD-based clustering, choosing a separation cutoff that is too large will

produce significant variation within the cluster and increase structure ambiguities, while a small cutoff may over-separate similar structures[16].

Due to the sensitivity of the input parameters and low separation dimensionality, an RMSD-based method may not be the best choice of analysis for proteins that sample a large conformation space. This is a defining characteristic of intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs), which represent a group of proteins that do not form a stable tertiary structure under physiological conditions and possess a large conformation space [17, 18]. A well-studied protein containing a disordered region is Tau[19], whose function regulates microtubule growth in the nervous system [20]. Significant interest in Tau has occurred due to its link with neurodegenerative diseases; notably Alzheimer's disease [21, 22]. Describing the conformational dynamics of IDPs is difficult, therefore several approaches have been developed to characterize them that do not revolve around strict structure resolution (eg. sequence predictions) [23]. A popular experimental method for IDP structure analysis is NMR spectroscopy. Experimental methods like NMR resolve measured data through RMSD fitting to generate structure ensembles, and offer a macroscopic view of IDP dynamics, however these methods are typically unable to fully resolve the conformational substates due to high conformational variation [24–26]. Alternatively, MD simulations offer an atomistic level of resolution to address the problem of IDP interactions[27]. Regardless of method, structure resolution is essential for functional analysis of IDPs. Because of the inherent flexibility and large conformation space exhibited by IDPs [28], generation of dominant structures through the use of clustering is a useful tool in structural analysis.

In this paper, we discuss the development of the Combinatorial Average Transient Structure (CATS) algorithm for determining structural ensembles of IDPs, and use it to analyze the trajectory data for the Tau/R2 fragment in urea and TMAO solutions published by the Shea group [29]. The clusters generated by CATS require no a priori information about the expected structure, which is advantageous for determining significant structures in MD trajectories where experimentally determined structures are unavailable. The choice of backbone phi and psi dihedral angle coordinates as structure descriptors was chosen for clustering due to their rotational and translational invariance. The CATS method generates clusters through each descriptor coordinate distribution, therefore the phi and psi residue dihedral angles have the added benefit of specifying backbone orientation using two values instead of three values in a Cartesian coordinate system. Through our novel technique, we show that CATS is able to generate a robust set of primary structures highlighting the denaturing effect of urea and compacting effect of TMAO on the Tau/R2 fragment. Additionally, we compare our results to the clusters obtained with the GROMOS method and show that CATS produces more structurally unique clusters that correspond to a higher resolution energy landscape as described by a Fano factor of 0.65 and 0.60 for the R2/TMAO and R2/urea trajectories respectively (compared to Fano factors of 0.36 and 0.54 produced by the GROMOS method). The drawback of this higher resolution energy landscape separation is that CATS requires a set of coordinate descriptors with multi-modal distributions that are well defined. CATS was tailored to separate the structures of proteins with assumed meta-stable conformational states, therefore not all datasets are applicable.

II. Methods:

II.1. Combinatorial averaged transient structure (CATS) clustering algorithm

CATS was created as a tool to identify highly probable ensemble structures within a trajectory for proteins that rapidly sample multiple conformations. Generally, any coordinate with a Gaussian-like distribution can be used for clustering. Our interest lies in conformational clustering from atomic coordinates; therefore we designed the process with dihedral coordinates as a descriptor. This descriptor is invariant to the translational or the rotational symmetry from the data set. Initially, the angular distribution of phi and psi dihedral angles for each residue is generated from the trajectory data. Assuming there are N residues, we obtain $2N$ individual coordinate distributions, where each of the $2N$ coordinate distributions is assumed to contain between 1 and 3 Gaussian-like peaks. Each peak is fitted with a Gaussian curve, and the average angle and standard deviation of each curve is used to transform the dihedral coordinates from each trajectory frame into an index value. We obtain a set of $2N$ index values for every frame in the trajectory, which is used to generate clusters based on similarity between each frame's index set. Finally, the clusters are further refined by allowing members of low populated clusters to be absorbed by others. Figure 1 depicts the flow chart of CATS.

II.1.A Dihedral angle distribution analysis—The phi/psi dihedral coordinates for a given molecular dynamics trajectory of an N -residue protein is extracted for each of the N residues of every frame. We histogram these trajectory phi and psi dihedral angles using a bin size of 3.6° on a range of -180° to 180° . The dihedral angle distribution represented by the raw histogram data is smoothed and fitted with Gaussian curves (see figures S3 and S4 for the complete set of R2/TMAO and R2/Urea distributions) of the form:

$$f(\theta) = \sum_{i=1}^N A_i e^{-\frac{(\theta - \mu_i)^2}{2\sigma_i^2}} \quad (1)$$

With amplitude A , angle θ , dihedral average μ and standard deviation σ corresponding to peak i . The maximum number of peaks in the coordinate distribution, N , varies by coordinate and trajectory and generally ranges between 1 and 3. The standard deviation and amplitude in each coordinate distribution is used to calculate an approximate probability value for peak i given by $P_i = \frac{A_i \sigma_i}{\sum_j A_j \sigma_j}$. A maximum of 3 most probable peaks are tabulated, while statistically insignificant distribution peaks (less than 1% probability) are discarded.

II.1.B Trajectory transformation—For a protein with N residues, every frame in the MD trajectory is transformed into a $2N$ -character string of numbers (0, 1, 2 or 3), which represent the specific angular distribution that each dihedral coordinate belongs to. This is accomplished by taking the dihedral coordinates from a given trajectory frame and comparing it to the corresponding coordinate distribution peaks obtained from the previous step. Each of the frame's $2N$ dihedral coordinates will be assigned an indexing value of 0, 1, 2 or 3 if the coordinate value is satisfies

$$|\theta_i - \mu_i^n| < \varepsilon \sigma_i^n \quad (2)$$

where θ_i is the trajectory frame's dihedral coordinate i , n is the coordinate distribution peak index, ε is a user defined constant, μ_i^n and σ_i^n are the distribution peak's average and standard deviation respectively. CATS assumes there is a maximum of 3 possible peaks for a given coordinate distribution, therefore the first peak (starting from -180°) is labeled 0, the second peak is labeled 1 and so on. If a dihedral coordinate from the trajectory frame is unable to be matched to any of the tabulated distribution peaks, it is given an index value of 3.

II.1.C. Cluster formation and refinement

II.1.C.a Initial cluster formation: Clusters are formed by grouping together the transformed trajectory strings. The first frame's indexing string defines the first cluster. The next frame's indexing string is compared to the first cluster by calculating the value of:

$$C = \sum_{i=1}^{2N} \Theta(|\lambda_i^k - \lambda_i|) \quad (3)$$

Where Θ is the Heaviside step function, λ_i^k is the distribution index for the i th coordinate of cluster k and λ_i is the i th distribution index of a given trajectory indexing string. If $C=0$, the frame's indexing string matches the first cluster's indexing string, and the frame joins the cluster. If $C \neq 0$, the indexing string of the frame fails to match with the indexing string of the cluster and it forms a new cluster defined by that unique indexing string. Similarly, every subsequent frame's indexing string is compared to all previous clusters; if there is a match, it joins that cluster, otherwise a new cluster is formed as defined by the newest frame's indexing string. This process is repeated for all trajectory frames until all frames are clustered. There is an option in CATS to specify coordinates in that are to be completely ignored throughout cluster formation and relaxation; this effectively changes the original number of coordinate comparisons ($2N$) to a user dependent value.

II.1.C.b Cluster representation and center structure: For a given cluster, each member is assigned an error factor that describes how close the member's dihedral coordinates are to the tabulated coordinate distributions defining the cluster. This error value, E , is calculated by:

$$E = \sum_{j=1}^{2N} \frac{|\theta_j - \mu_j^n|}{2N} \quad (4)$$

Where θ_j is the cluster member's dihedral angle for coordinate j , μ_j^n is the cluster defined distribution peak average for peak index n and coordinate j . The trajectory frame number of the cluster member with the lowest error value is used to represent the conformation of the cluster.

II.1.C.c. Cluster relaxation: Once a basis set of initial clusters are established, CATS attempts to reduce the number of clusters by combining lower populated clusters that are similar to higher populated clusters. Relaxation is based on two user-defined criteria: cluster

population cutoff, α , and number of ignorable coordinates, β . For clusters with a population under the cutoff value, CATS will compare each member's indexing data to all other clusters, starting from the largest cluster and ending at the smallest cluster, excluding the current cluster. In the initial stage of clustering, all members of a specific cluster must contain the same set of indexing values (given by a string of $2N$ values), however in the relaxation stage, the relaxed cluster member may be absorbed into a different cluster as long as the indexing data between the relaxed member and new cluster does not differ by more than the specified number of ignorable coordinates β . In other words, each frame's indexing data is allowed to fail up to β index comparisons with a given cluster before indicating a failure to match. In the case that a cluster member cannot be absorbed elsewhere, it remains in its original cluster. Cluster centers are not recalculated after relaxation. Clusters that are relaxed are considered to be artifacts of simulation noise or transitions; initially, they are not grouped together with larger clusters due to a small fraction of coordinate index deviations. Essentially, relaxation serves to change the population rankings of major clusters by correcting the deviation.

II.2. Implementation of CATS and GROMOS algorithms on the Tau/R2 fragment

II.2.A Cluster setup using the GROMOS method—Following the original setup, we use the GROMACS GROMOS package as a control [30, 31]. Clusters are formed using the algorithm described in Daura et al [32] using an RMSD cutoff value of 1.4 angstroms.

II.2.B Comparison of CATS clustering method—We focus our comparison between CATS clusters and the clusters obtained by the RMSD-based GROMOS method used in the original R2 fragment study [32]. The potential mean force (PMF) was generated using the distributions of the radius of gyration (R_g) and end-to-end distance (R_{ee}) using 25 evenly-spaced bins to histogram values.

II.3. All atom molecular dynamics simulation of the Tau/R2 fragment

We have taken the data of the all-atom molecular dynamics simulations of the Tau/R2 fragment in TMAO or in urea from the Shea group. Details on the model and simulation setup for Tau/R2(273–284) fragment can be found in the Shea group's publication [29], which will be briefly summarized here. The effect of TMAO and urea on the aggregation-prone (R2) region of the Tau protein is investigated using all-atom molecular dynamics (MD) simulations. Topologies were generated using GROMACS. Trajectories consisting of 6000 frames for the R2/Urea configuration and 7500 frames from the R2/TMAO configuration (totaling 120ns and 150ns respectively) are generated using a leap-frog algorithm to integrate the Newtonian equations of motion in the Nose-Hoover NVT environment at 300K using a time step of 2fs and sampling rate of 20ps.

III. Results

III.1 CATS clusters and their center structures represent local minima in the energy landscape

The dihedral phi and psi angles chosen for clustering produced a series of Gaussian-like peaks in their distributions. Each coordinate's peak describes a probability density, which

associates the potential mean force (PMF) and free energy with the clustering coordinate. CATS transforms the trajectory into a 2N-dimensional energy landscape (where N represents the number of residues) and creates clusters based on the landscape's minima, as shown in figure 2A–D.

CATS method produces a different analysis of the R2 trajectory than GROMOS because GROMOS clusters by average RMSD. CATS uses the probability distribution analysis that is essentially its free energy landscape. Figure 3 illustrates cluster populations and center structures for the four highest populated clusters of CATS and GROMOS compared to the potential mean force (PMF) of the TMAO trajectory. Similarly, figure 4 illustrates the distribution of cluster populations for both methods compared to the PMF of the urea trajectory.

The top four cluster members in the R2/TMAO trajectory produced by CATS and GROMOS are similarly distributed around the center of the lowest point on the PMF contour map shown in figure 4. CATS and GROMOS produce the same center structure for the first cluster, and a beta-hairpin structure for the second cluster. The third cluster of GROMOS is similar to the fourth cluster of CATS, however the third CATS cluster center and fourth GROMOS cluster center have completely different secondary structures. The distribution of cluster members from the third CATS cluster is closer to the center of the PMF map than the members of the fourth GROMOS cluster. Despite the difference in order between the top four GROMOS and CATS clusters, the fourth GROMOS cluster is still captured by the sixth CATS cluster. Similarly, the third CATS cluster is captured by the ninth GROMOS cluster. The top 20 cluster centers for the CATS and GROMOS methods can be found in figures S5–S8.

The R2/urea trajectory produces a more diverse set of clusters. All GROMOS cluster members are distributed around the minimum PMF in the neighborhood of $R_g = 7.5\text{\AA}$, whereas CATS clusters have distributions around both of the PMF local minima. This results in noticeably different center clusters, with only CATS cluster 2 and GROMOS cluster 1 having a similar conformation. GROMOS clusters 2 and 3 appear to be similar, indicating the presence of degeneracies. The fourth GROMOS cluster is a beta-hairpin structure, which is significantly different from the other clustered structures. Despite GROMOS and CATS previously having captured similar structure centers at different ranks, CATS does not capture the hairpin conformation as its center structure throughout the top 20 CATS clusters due largely to the low probability of a positive phi dihedral angle coordinate ($\sim 1\%$) of the 6th residue and poorly defined Gaussian-like peaks in the phi/psi angles of residues 5 and 7, which is responsible for the hairpin conformation (see figure 2C and supporting information for distribution). The low-probability phi peak in residue 6 was not taken into account during the distribution file setup, and therefore all structures containing this peak were flagged as non-clusterable. We re-ran CATS using a modified input file that explicitly includes the low probability phi coordinate of the 6th residue with no observable hairpin conformations (within the first 40 clusters). The frames captured by the 4th GROMOS cluster appeared to be spread over multiple low-ranked CATS clusters due to other dihedral coordinate variations. Despite similar RMSD values, the coordinate index values vary enough to eliminate the hairpin as a top ranking cluster.

III.2 Cluster probabilities based on RMSD derived populations can be misleading

The TMAO and urea PMF both resemble an oblong shape with one local minimum for TMAO and two local minima for urea. The lack of resolution of the PMF contour maps is also clear from the tendency of GROMOS and CATS clusters to not always center on a local minimum, which would occur for highly populated clusters.

Conversely, the 2N-dimensional PMF landscapes shown in figure 2A–D contain higher resolution minima because each CATS cluster occupies a local minimum within the 2N-dimensional energy landscape. Additionally, the center structure produced by CATS for each cluster is the closest to the 2N-dimensional minima. Other members in the cluster vary within the energy landscape around this central point. The RMSD distribution of the GROMOS method (shown in figure S2) illustrates that pairwise RMSD values have a single distribution curve and cannot take into account the PMF at high resolution. The implication of using an RMSD cutoff forces structures that may deviate slightly from the center of a local minimum into another cluster all together. In this case, probabilities derived by GROMOS based on populations of the cluster may not accurately describe stable states of high probability structures.

III.3 CATS clusters structures than GROMOS might have missed

GROMOS clusters have a strict clustering cutoff based on RMSD. Figure 5 illustrates two structures that would have normally been placed in separate clusters due to their large differences in R_{ee} . These two structures are members of the same cluster produced by CATS; one being the center structure, and one being a large deviation in terms of end to end distance. Since CATS produces the center structure at the closest possible minima point of the 2N-dimensional energy landscape for the cluster, any deviations in clustering coordinate would produce a structure at a higher PMF as calculated by the error function in eq. 4. The larger PMF due to backbone deviation does not move the conformation into another 2N-dimensional local minimum (at another cluster), but it does imply that the conformation of the peptide will shift to resemble the center structure under the influence of the PMF. Because CATS clusters form at local energy minima, cluster members can describe how the center structure conformation (given by eq. 4) is perturbed around the local minimum point.

III.4 Clusters produced by CATS have larger structure variations than clusters produced by GROMOS

Figure 6 illustrates the pairwise deviations between center structures produced by CATS and GROMOS methods. In both the TMAO and urea clusters, CATS tends to produce more unique centers than the GROMOS method. In the TMAO clusters, both methods appear to produce similarly varying cluster centers until cluster 12, where CATS begins to generate centers that have large RMSD values compared to previous structures. This trend can be seen by comparing the Fano factor [33] of the two methods: The top 20 TMAO structures for GROMOS and CATS produce a Fano factor of 0.36 and 0.65 respectively. Additionally, CATS produces structures with large comparative RMSD values as shown in figure 6. GROMOS appears to produce 3 centers (cluster 13, 19 and 20) that have large RMSD values compared to other centers. For the top 20 urea structures, GROMOS and CATS produce a Fano factor of 0.54 and 0.60 respectively. The Fano factor is a quantitative measurement of

the noise in a given variable, which in this case is the average RMSD between cluster pairs. Comparing the pairwise RMSD matrix in figure 6 and the Fano factors for the two methods suggests that CATS and GROMOS methods do not share the same top 20 cluster centers. Figures 3 and 4 illustrated how CATS and GROMOS methods produced similar structures with alternate rankings for the top 4 clusters. This trend does not appear to continue according to figure 6A and 6B, which suggests that CATS clusters have larger pairwise RMSD values after the 4th top ranked cluster, meaning CATS and GROMOS methods have less agreement for center structures at lower ranked clusters.

IV. Discussion

The cluster representation (center structure) produced by CATS captures energy landscape minima with better resolution than the center structure of RMSD based algorithms.

In the comparison of the top 20 cluster centers produced by the GROMOS and CATS methods (figure 6), we observed that CATS produces a greater number of unique center structures. Visually, the top 20 CATS cluster centers have a larger averaged RMSD than the top 20 GROMOS structures for both TMAO and urea. We quantify this observation with the Fano factor, where CATS clusters possess a larger Fano factor in both cases. This behavior can be observed in the PMF plots for the urea and TMAO trajectories (figure 3 & 4); the cluster members were shown to sample a larger area on the PMF plot than the members of the GROMOS clusters. Since CATS produces distinct clusters using bimodal dihedral distributions, we expect each cluster group to have distinct features. The majority of the bimodal coordinate distributions have large differences in their average peak values, which are reflected through the large deviations between cluster centers. Since the GROMOS method produces a continuum of different clusters (depending on the cutoff value), there is no guarantee that the conformations of each cluster will be relatively unique. Therefore, the clusters produced by CATS will not sample from the same areas in an energy landscape defined by dihedral angles. This is a feature of RMSD minimization used by the GROMOS method. On the other hand, CATS captures the most probable conformation for IDPs, constrained by the dihedral coordinate energy landscape, where RMSD may not be small. Due to this phenomenon, CATS and GROMOS will assign different ranks/probabilities to similar structures, as observed in figures 3C–D and 4C–D. Thus, IDP-like proteins that have a larger RMSD but exist in the same energy state (as described by the dihedral PMF) may not be grouped together using the GROMOS method.

CATS captures alternate high-probability structures suitable for IDP simulations

The fourth largest cluster produced by the GROMOS method for the R2/urea trajectory, which is a beta-hairpin, was not captured by CATS (within the top 40 structures) despite its relatively large population. CATS appears to capture urea structures that mainly lack conformation. The discrepancy between the CATS and GROMOS top 4 cluster structures comes from the phi/psi coordinate distributions in residues 5–7. The phi dihedral angle PMF for the 6th residue (shown in figure 2C) contains a bifurcation in the distribution curve, where 1% of the phi distribution exists in the 100 degree region, while 99% exist in the –100 degree region (See figure S1). Not only does CATS reject low populated distributions (1% or

lower by default), the construction of the distribution probability inputs also ignored the peak, therefore the hairpin conformation was deemed an “unclustered coordinate” because the dihedral angle values of the coordinate did not fit into any indexed distribution (see II.1.B in methods). CATS was re-run on the R2/urea trajectory with this peak accounted for, however the top 20 structures/clusters still did not contain the beta-hairpin conformation. The coordinate indexes of the CATS transformed trajectory frames that make up the hairpin cluster (GROMOS cluster 4) showed several “unclusterable” coordinates randomly throughout the cluster. We attribute this phenomenon to the hairpin structures belonging to distribution regions that poorly resemble a normal distribution (See figure S4 for the complete set of urea coordinate distributions). We observed several coordinate distributions that do not converge to zero population between distribution peaks, which is where several hairpin coordinates reside. If the hairpin structures were a result of poor simulation sampling, and all distribution peaks obey the law of large numbers, then CATS could successfully have filtered out “noise” since at some limit the population of hairpin conformations would be miniscule compared to others. On the other hand, if the distributions never approach an ideal Gaussian distribution, then CATS would not be as apt for clustering (at least with the choice of dihedral coordinate distributions).

Poorly defined coordinate distributions and relaxation affect accuracy of CATS.

CATS was developed to specifically analyze a trajectory that contains points of variation for the protein backbone. Coordinates with multiple distinct distribution peaks permit clustering. The limit of resolution in this scheme comes from the quality of the distribution peak and the number of coordinates used to describe the structure as a descriptor. The R2 fragment has 24 coordinates total (a phi and psi dihedral angle describing relative orientation for all 12 residues); therefore the maximum possible combination of coordinate distributions (assuming that each coordinate had two peaks in its distribution) is 2^{24} . This combinatorial maximum, however, does not occur due to backbone constraints imposed on the protein during the MD simulation. In the initial analysis of the MD trajectory (II.1.B), the descriptor coordinate distributions are generated. The distributions alone do not characterize the constraints on the number of possible coordinate peak combinations. In phase II.1.C, the backbone constraints, which forbid dihedral coordinate combinations that produce steric clashes or other unrealistic structures, are taken into account by clustering distribution peak combinations that exist in the trajectory. Once this analysis is complete and clusters are formed, the center structure of each cluster is generated based on the cluster member with the lowest error or deviation (see eq. 4) from each coordinate’s distribution peak average (See figures S3 and S4 for a complete set of descriptor coordinate distributions). In this way, the representative center structure of a CATS cluster shares similarities to expectation-maximization (EM) based algorithms [10, 34], while the relaxation routine encompasses the idea of fuzzy clustering [8, 9, 35–39]. The concept behind CATS relaxation is that a protein will still perform a function similar to the cluster representative structure despite a lowered degree of membership. For example, a relaxation factor of 2 ignorable coordinates in the R2 fragment implies that all members of the cluster are at least 91% the same as with the representative structure.

CATS has the ability to cluster any generalized coordinate with separable Gaussian-like distributions. Depending on the data used for cluster analysis, this can be a positive or negative attribute. Coordinates with large standard deviations will produce clusters that reflect a large range of variation in such a coordinate. Conversely, narrow distributions will produce well defined clusters with low variation. These properties can be exacerbated based on the coordinate chosen for clustering. In this study, we used the dihedral angle coordinates to form distributions of the R2 fragment in TMAO and Urea. We chose backbone dihedral angles phi and psi as our clustering coordinate due to their Gaussian-like distributions throughout the trajectory, independence from Cartesian coordinate positions (the R2 trajectories did not need to be aligned prior to clustering), and condensed representation of relative backbone orientation. In our comparison of CATS to GROMOS, we analyzed the top four clusters as functions of R_g and R_{ee} . The clusters from both methods produced a high density population centered about a specific R_g and R_{ee} region. The clusters produced by CATS periodically included members that appear to center about a different point on the density map, which was not observed in the GROMOS results. This feature is most likely due to a larger deviation of dihedral coordinates near the middle residues of the R2 fragment (residues 5, 6, or 7) as observed in figure 5. The distance travelled by a point on the N or C terminus due to a change in a central residue dihedral angle is proportional to the distance between the central residue and the terminus point. Since CATS is not optimized to minimize RMSD distances, clusters produced using dihedral coordinate distributions have the possibility of including members that seem out of place when analyzing member conformations as a function of R_g and R_{ee} . As with all statistical methods, a larger sample size will improve accuracy. The larger R_g and R_{ee} of CATS may not be improved by greater sample size since this is an artifact of the standard deviation exhibited by the coordinate distribution. Coordinates with significantly large standard deviations may not be suited for CATS clustering since structure information cannot be partitioned out.

CATS is better suited for input parameter tuning than the GROMOS method

Determining the correct cutoff/input parameters for clustering proteins occurs on a case-by-case basis. For the GROMOS method, one must supply an RMSD cutoff that is both large enough to cluster a reasonable number of structures together while simultaneously filtering out dissimilar structures. Similarly, CATS requires the user to supply a standard deviation cutoff for the coordinate distributions so that the majority of structures can be grouped together. An important distinction between CATS and GROMOS methods with respect to these user supplied cutoff values is that CATS will generally produce the same center structures regardless of standard deviation used; tuning the inputs for CATS will affect the relative populations of clusters only. On the other hand, tuning the RMSD cutoff values in the GROMOS method will affect both the center structure and populations of clusters. Determination of ignorable coordinates must also be made since volatile coordinates that have little significance on the overall structure (such as the dihedral angles of the first and last residues) should be ignored to avoid over constrained clusters.

A practical consideration of the parameter tuning process is the computational resources required for each method to complete the task. The RMSD based clustering method in GROMOS requires significantly more computational time than CATS (see table S1). Due to

the pairwise RMSD calculations performed by GROMOS, the number of calculations required to generate the observed clusters is much greater than the number of calculations performed by CATS. In this study, we used the ‘backbone’ option for GROMOS, requiring 3 atom pairwise calculations per residue per frame. Each frame comparison requires a symmetric pairwise RMSD matrix to be formed. Therefore, the number of calculations required by GROMOS scales proportional to roughly $O((3N)!)^2$ where N is the number of residues. CATS transforms the trajectory in a simple comparison step proportional to $2N$, then clusters through iterative comparisons between transformed frames, which requires roughly $O(2N)$ if using the dihedral coordinate descriptors from our studies. For larger proteins or larger trajectory sizes, the computational cost will increase dramatically. Although the CATS algorithm produces clusters significantly faster than the GROMOS method, there is a drawback of CATS when producing input parameters. We chose to neglect input parameter timing factors because there are multiple methods for input parameter generation, all of which are dependent on the trajectory. The automated fitting of Gaussian curves to the coordinate distributions based on our in-house numerical solver adds a trivial amount of time to CATS cluster production. This software, however, is in the early development stage; therefore its results can be inaccurate for complex distributions. The user can produce input files using Gaussian fitting software from other sources, as well as a manual fitting tool, which was utilized here.

Conclusions

CATS brings together several aspects of different types of clustering algorithms in a unique way, exploiting distribution trends in the raw data to form clusters. The question we want to address is whether CATS out-performs other clustering algorithms for the purpose of conformational analysis of IDPs. The answer is not simply yes or no, as there are several factors that one must consider when comparing the results from CATS to those of an RMSD-based method, such as GROMOS. The difference between CATS and GROMOS lies in how one defines similarity between structures. RMSD similarity has the disadvantage of being based in Cartesian coordinate space; there is sensitivity to how the structures are aligned beforehand, and ignores energetic deviations within the backbone. CATS defines similar structures as those belonging to the same regions within the PMF space, thus backbone energetics has a significant effect on structure categorization. In recent years, attention has been given to the development of machine learning (ML) algorithms for use in clustering proteins. CATS shares a large similarity with current ML algorithms [40]. With little modification, the CATS algorithm can correct for noise and irregularities discussed earlier that may hamper analysis and be fully capable of self-correction.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement:

We thank Prof. Joan E. Shea and her group for the trajectories from the all-atom molecular simulations used in this work. We thank Prof. Greg Morrison and Jim Briggs for the discussion on CATS. We thank the financial support from the National Science Foundation MCB 1412532 and ACI: 1531814. Supported by a training fellowship from the Gulf Coast Consortia, on the Houston Area Molecular Biophysics Program (NIGMS Grant No. T32GM008280)

References:

- [1]. Shao J, Tanner SW, Thompson N, Cheatham TE, Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms, *J Chem Theory Comput*, 3 (2007) 2312–2334. [PubMed: 26636222]
- [2]. Geva AB, Hierarchical unsupervised fuzzy clustering, *Ieee Transactions on Fuzzy Systems*, 7 (1999) 723–733.
- [3]. Likas A, Vlassis N, Verbeek JJ, The global k-means clustering algorithm, *Pattern Recognition*, 36 (2003) 451–461.
- [4]. Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY, An efficient k-means clustering algorithm: Analysis and implementation, *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 24 (2002) 881–892.
- [5]. Eltoft T, deFigueiredo RJP, A new neural network for cluster-detection-and-labeling, *Ieee Transactions on Neural Networks*, 9 (1998) 1021–1035. [PubMed: 18255785]
- [6]. Karhunen J, Oja E, Wang LY, Vigario R, Joutsensalo J, A class of neural networks for independent component analysis, *Ieee Transactions on Neural Networks*, 8 (1997) 486–504. [PubMed: 18255654]
- [7]. Zhuang XH, Huang Y, Palaniappan K, Zhao YX, Gaussian mixture density modeling, decomposition, and applications, *Ieee Transactions on Image Processing*, 5 (1996) 1293–1302. [PubMed: 18285218]
- [8]. El-Sonbaty Y, Ismail MA, Fuzzy clustering for symbolic data, *Ieee Transactions on Fuzzy Systems*, 6 (1998) 195–204.
- [9]. Eschrich S, Ke JW, Hall LO, Goldgof DB, Fast accurate fuzzy clustering through data reduction, *Ieee Transactions on Fuzzy Systems*, 11 (2003) 262–270.
- [10]. Xu R, Wunsch D 2nd, Survey of clustering algorithms, *IEEE Trans Neural Netw*, 16 (2005) 645–678. [PubMed: 15940994]
- [11]. Koch MA, Wittenberg LO, Basu S, Jeyaraj DA, Gourzoulidou E, Reinecke K, Odermatt A, Waldmann H, Compound library development guided by protein structure similarity clustering and natural product structure, *Proc Natl Acad Sci USA*, 101 (2004) 16721–16726. [PubMed: 15548605]
- [12]. Zhong W, Altun G, Harrison R, Tai PC, Pan Y, Improved K-Means Clustering Algorithm for Exploring Local Protein Sequence Motifs Representing Common Structural Property, *IEEE Transactions on Nanobioscience*, 4 (2005) 255–265. [PubMed: 16220690]
- [13]. Gabrys B, Bargiela A, General fuzzy min-max neural network for clustering and classification, *Ieee Transactions on Neural Networks*, 11 (2000) 769–783. [PubMed: 18249803]
- [14]. Su MS, Chou CH, A modified version of the K-means algorithm with a distance based on cluster symmetry, *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 23 (2001) 674–680.
- [15]. Jain AK, Duin RPW, Jianchang M, Statistical pattern recognition: a review, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 (2000) 4–37.
- [16]. Kufareva I, Abagyan R, Methods of protein structure comparison, *Methods in molecular biology* (Clifton, N.J.), 857 (2012) 231–257. [PubMed: 22323224]
- [17]. Uversky VN, Oldfield CJ, Dunker AK, Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling, *J Mol Recognit*, 18 (2005) 343–384. [PubMed: 16094605]
- [18]. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z, Intrinsically disordered protein, *Journal of Molecular Graphics and Modelling*, 19 (2001) 26–59. [PubMed: 11381529]
- [19]. Eschmann NA, Do TD, LaPointe NE, Shea J-E, Feinstein SC, Bowers MT, Han S, Tau Aggregation Propensity Engrained in Its Solution State, *J. Phys. Chem. B*, 119 (2015) 14421–14432. [PubMed: 26484390]
- [20]. Desai A, Mitchison TJ, Microtubule polymerization dynamics, *Annu Rev Cell Dev Biol*, 13 (1997) 83–117. [PubMed: 9442869]

- [21]. Mudher A, Lovestone S, Alzheimer's disease – do tauists and baptists finally shake hands?, Trends in Neurosciences, 25 (2002) 22–26. [PubMed: 11801334]
- [22]. Grundke-Iqbal I, Iqbal K, Tung YC, Quinlan M, Wisniewski HM, Binder LI, Abnormal phosphorylation of the microtubule-associated protein tau (tau) in Alzheimer cytoskeletal pathology, Proceedings of the National Academy of Sciences, 83 (1986) 4913–4917.
- [23]. Vucetic S, Brown CJ, Dunker AK, Obradovic Z, Flavors of protein disorder, Proteins, 52 (2003) 573–584. [PubMed: 12910457]
- [24]. Oldfield CJ, Dunker AK, Intrinsically disordered proteins and intrinsically disordered protein regions, Annu Rev Biochem, 83 (2014) 553–584. [PubMed: 24606139]
- [25]. Jensen MR, Ruigrok RWH, Blackledge M, Describing intrinsically disordered proteins at atomic resolution by NMR, Current Opinion in Structural Biology, 23 (2013) 426–435. [PubMed: 23545493]
- [26]. Brookes DH, Head-Gordon T, Experimental Inferential Structure Determination of Ensembles for Intrinsically Disordered Proteins, Journal of the American Chemical Society, 138 (2016) 4530–4538. [PubMed: 26967199]
- [27]. Schaeffer RD, Fersht A, Daggett V, Combining experiment and simulation in protein folding: closing the gap for small model systems, Current Opinion in Structural Biology, 18 (2008) 4–9. [PubMed: 18242977]
- [28]. Dyson HJ, Wright PE, Insights into the structure and dynamics of unfolded proteins from nuclear magnetic resonance, Advances in protein chemistry, 62 (2002) 311–340. [PubMed: 12418108]
- [29]. Levine ZA, Larini L, LaPointe NE, Feinstein SC, Shea JE, Regulation and aggregation of intrinsically disordered peptides, Proc Natl Acad Sci U S A, 112 (2015) 2758–2763. [PubMed: 25691742]
- [30]. Pronk S, Pall S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, Shirts MR, Smith JC, Kasson PM, van der Spoel D, Hess B, Lindahl E, GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit, Bioinformatics, 29 (2013) 845–854. [PubMed: 23407358]
- [31]. Berendsen HJC, Vanderspoel D, Vandrunen R, Gromacs - a Message-Passing Parallel Molecular-Dynamics Implementation, Computer Physics Communications, 91 (1995) 43–56.
- [32]. Daura X, Gademann K, Jaun B, Seebach D, van Gunsteren WF, Mark AE, Peptide folding: When simulation meets experiment, Angew Chem Int Edit, 38 (1999) 236–240.
- [33]. Fano U, Ionization Yield of Radiations. II. The Fluctuations of the Number of Ions, Physical Review, 72 (1947) 26–29.
- [34]. Dempster AP, Laird NM, Rubin DB, Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society. Series B (Methodological), 39 (1977) 1–38.
- [35]. Dave RN, Bhaswan K, ADAPTIVE FUZZY-C-SHELLS CLUSTERING AND DETECTION OF ELLIPSES, Ieee Transactions on Neural Networks, 3 (1992) 643–662. [PubMed: 18276465]
- [36]. Carpenter GA, Grossberg S, Rosen DB, FUZZY ART - FAST STABLE LEARNING AND CATEGORIZATION OF ANALOG PATTERNS BY AN ADAPTIVE RESONANCE SYSTEM, Neural Networks, 4 (1991) 759–771.
- [37]. Hathaway RJ, Bezdek JC, Hu YK, Generalized fuzzy c-means clustering strategies using Lp norm distances, Ieee Transactions on Fuzzy Systems, 8 (2000) 576–582.
- [38]. Hathaway RJ, Bezdek JC, Fuzzy c-means clustering of incomplete data, Ieee Transactions on Systems Man and Cybernetics Part B-Cybernetics, 31 (2001) 735–744.
- [39]. Gath I, Gev AB, Unsupervised Optimal Fuzzy Clustering, IEEE Trans. Pattern Anal. Mach. Intell, 11 (1989) 773–780.
- [40]. Perspective: Machine learning potentials for atomistic simulations, The Journal of Chemical Physics, 145 (2016) 170901. [PubMed: 27825224]

Figure 1:

A flow chart of the CATS algorithm. Coordinates that exhibit Gaussian-like distributions are extracted from the trajectory and histogrammed. One or more Gaussian curves are fitted to the distribution of each coordinate, and each peak's average value and standard deviation is generated. These values are compared to the extracted trajectory frame coordinates and transformed into indices that label which Gaussian curve the coordinate belongs to within a given distribution. Clusters are initialized by grouping together trajectory frames that possess the same transformation indices. Refinement allows lower populated clusters to join higher populated ones if their indices have less than a user-defined number of mismatches.

Figure 2:

The R2/TMAO potential mean force in units of $1/kT$ is shown for (A) the phi dihedral angle distribution and (B) the psi dihedral angle distribution. The R2/urea potential mean force in units of $1/kT$ is shown for (C) the phi dihedral angle distribution and (D) the psi dihedral angle distribution. Please refer to figures S3 and S4 for the corresponding distribution curves.

Figure 3:

Comparison between cluster members produced by (A) CATS and (B) GROMOS methods for the top four most populated clusters shown in table 1 in relation to the R2/TMAO potential mean force (in units of $1/kT$) as a function of peptide conformation at room temperature over 100ns. The automated "center" structure (defined by eq. 4 for CATS and by the smallest average RMSD for GROMOS) of each cluster is shown for (C) CATS and (D) GROMOS.

Figure 4:

Comparison between cluster members produced by (A) CATS and (B) GROMOS methods for the top four most populated clusters shown in table 1 in relation to the R2/urea potential mean force (in units of $1/kT$) as a function of peptide conformations at room temperature over 150ns. The automated "center" structure (defined by eq. 4 for CATS and by the smallest average RMSD for GROMOS) of each cluster is shown for (C) CATS and (D) GROMOS.

Figure 5:

Using the first cluster generated by CATS for the R2/TMAO trajectory, the center structure and structure with the largest end-to-end distance is compared through (A) the phi and psi dihedral angles of the center structure (shown in black) and the deviation structure (shown in red) for each coordinate. (B) The end-to-end distance and radius of gyration for the two structures are shown in comparison to the R2/TMAO potential mean force map shown in figure 3 with the two structures overlaid to illustrate conformational differences.

Figure 6:

The pairwise backbone RMSD values (in Angstroms) between center structures of the top 20 ranked clusters are shown for (A) R2/TMAO and (B) R2/urea. Cluster numbers are shown on the x and y axis, with 1 being the most probable cluster. The upper triangle illustrates the RMSD value between a given pair of cluster centers for CATS, and the lower triangle illustrates the RMSD value between a given pair of cluster centers for GROMOS.

Table 1:

Average values for the radius of gyration and end-to-end distance with cluster RMSD for the R2/TMAO and R2/urea trajectories. The values given are for the top four populated clusters of GROMOS and CATS methods.

	cluster #	TMAO			Urea		
		Avg. Rg (Å)	Avg. Ree (Å)	% pop.	Avg. Rg (Å)	Avg. Ree (Å)	% pop.
CATS	1	7.3 ± 0.2	15.0 ± 3.4	3.2	8.0 ± 0.4	22.1 ± 4.5	1.3
	2	7.5 ± 0.2	9.2 ± 3.5	2.0	7.5 ± 0.5	15.2 ± 4.4	1.2
	3	6.9 ± 0.2	10.3 ± 3.6	1.3	8.7 ± 0.4	18.8 ± 3.4	1.0
	4	7.2 ± 0.1	10.5 ± 2.6	1.3	7.3 ± 0.3	10.4 ± 2.4	0.9
GROMOS	1	7.3 ± 0.2	14.5 ± 3.0	4.1	7.3 ± 0.2	13.9 ± 2.0	1.7
	2	7.4 ± 0.2	8.3 ± 2.6	4.1	7.1 ± 0.2	10.0 ± 1.2	1.3
	3	7.2 ± 0.1	11.0 ± 2.3	2.1	7.1 ± 0.1	11.2 ± 1.8	1.1
	4	6.5 ± 0.1	14.0 ± 1.9	2.0	7.0 ± 0.1	4.7 ± 1.2	0.8

Table 2:

Center structure properties from CATS and GROMOS for the TMAO and urea trajectories.

		TMAO		Urea	
	cluster #	Rg (Å)	Ree (Å)	Rg (Å)	Ree (Å)
CATS	1	7.4	10.5	7.5	19.1
	2	7.5	6.4	7.7	17.2
	3	6.9	5.8	8.8	17.8
	4	7.1	10.1	7.6	11.0
GROMOS	1	7.3	14.9	7.4	13.6
	2	7.4	5.4	6.9	9.1
	3	7.4	12.9	7.1	12.2
	4	6.5	13.0	6.9	4.0