A Fast Geometric Clustering Method on Conformation Space of Biomolecules

Jian Sun [*] Yuan Yao [†] Xuhui Huang [‡] Vijay Pande [§] Gunnar Carlsson[†]
Leonidas J. Guibas[*¶]

Abstract

Clustering is a typical approach to study conformation space where close conformations are grouped into the same state. It not only provides a concise representation of the free energy landscape but also is a necessary preprocessing step for building many other more complicated representations such as Markov State Model. However, since the conformation space is often sampled according to Boltzmann distribution which is exponential to the free energy, large amounts of sampled conformations are concentrated at the free energy basins. Typically employed clustering algorithms such as K-means which measure the quality of clustering in terms of variance, tend to split the densely sampled free energy basins into many small clusters and lump the low density regions into the clusters that have quite different structures. In this paper, we consider the clustering from geometric point of view and measure the quality of clustering in terms of geometric property. We introduce a fast efficient clustering algorithm: approximating K-center clustering algorithm. This algorithm has two major advantages (1) It is fast, and can generate thousands of clusters from millions of conformations within several hours on a single PC, which is orders of magnitude faster than K-means clustering algorithm. (2) The output clusters are about of the same radius and thus the population of each cluster represents its relative density determined by the underlying free energy landscape. Moreover, as it efficiently reduces the complexity of the system in a faithful way, it's proved to be very powerful in our experiments to combine it with more deliberate schemes which are otherwise not applicable due to the massive nature of data.

1 Introduction

Elucidation of bio-molecular folding process is critical and fundamental to biology and medicine. However, it is very difficult to experimentally probe the mechanism of the process at atomic resolution. Computer simulations have proved useful for studying biological processes since they can complement experimental tools by providing dynamic information at an atomic level. However, understanding biomoleculuar folding is challenging computationally because it is difficult to sample from the rugged and high-dimensional free energy landscapes. Furthermore, even if this sampling problem is now solved for many systems of interest, there remains the difficulty of representing the free energy surface. Projecting the free energy landscape onto a few order parameters is one common method of depicting the landscape. However, such dimensionality reduction may cause points that are very distant from one another to appear to be close together. In fact, free energy barriers may even be completely obscured [1].

[*] Department of Computer Science, Stanford University, Stanford, CA 94305

[†] Department of Mathematics, Stanford University, Stanford, CA 94305

[‡] Department of Bioengineering, Stanford University, Stanford, CA 94305

[§] Department of Chemistry, Stanford University, Stanford, CA 94305

[¶] To whom correspondence should be addressed. E-mail: guibas@cs.stanford.edu

An alternative approach is to decompose the conformation space into clusters by grouping close conformations into the same state [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. In computer simulation, the conformation space of a biomolecular system is typically sampled by molecular dynamics (MD) or Markov-Chain-Monte-Carlo (MCMC) or their enhanced schemes [13, 14, 15, 16, 17, 18], which produce a sampling of the distribution that is *exponential* to the free energy, i.e., Boltzmann distribution. Therefore, the sampled conformations of high density with a large amount of similar structures nearby account for the regions of low free energy, i.e., free energy basins. A good clustering scheme should preserve such free energy basins by grouping structures within the same basin into the same state, without over-splitting a free energy basin into many clusters [19]. Since estimating the density is equivalent to estimating the free energy, another desired property of a decomposition is to provide the density information as much as possible. However these requirements are not easily reached by most of the typical clustering methods currently in use.

K-means and its variation *K-medoids* are one of the popular methods to cluster the conformation spaces in the literature [12, 19]. However this type of methods measure the quality of clustering in term of variance and minimize the total intra-cluster variances [20]. They tend to divide densely sampled free energy basins into many small clusters to avoid large intra-cluster total variances, whence split conformations with very similar structures into different clusters. As a result, due to the limited quota on the number of clusters k, they lump the low density regions into the clusters corresponding to energy basins nearby and thus fails to identify those intermediate or transition states of interest. Moreover, K-means or K-medoids do not have any control on the volumes of clusters which are often spread out, whence provide poor information about density as well as the free energy landscape. Another set of widely used clustering methods for conformations is *agglomerative methods* [21, 22] including *single-linkage* [10, 11], *average-linkage* [3] and *complete-linkage* [10, 7]. In particular, complete-linkage (also called maximal-distance clustering) controls the diameter of the clusters and may produce a clustering of with good information on density. However, since they are carried out in a bottom-up fashion and much of the computation is wasted on figuring out the grouping order for points within a cluster, these clustering methods are of time complexity at least quadratic order and hence very inefficient especially for large data sets. A closely related clustering method called *leader algorithm* [20] is also widely used [6, 8, 5]. However, the leader algorithm requires a pre-defined threshold that specifies the maximum cluster radius and has no control over the complexity of the resulting clustering. In fact, it may generate many clusters with radius well below the given threshold as it chooses the cluster centers in an ad-hoc fashion.

In this study, we consider the clustering from geometric point of view and measure the quality of clustering in terms of geometric property. Given k, our goal is to find a decomposition of conformation space into k clusters that minimizes the maximum radius of the clusters, instead of the total variance as does in K-means. Such a minimization ensures the obtained clusters are about of the uniform geometric size, whence enables one a better estimation on the density and free energy landscape. This clustering method is called K-center in the literature [23]. Though it is NP-hard, there is a simple and fast 2-approximation algorithm which is of linear time complexity [24, 25]. By utilizing the triangular inequality, the algorithm can be further speeded up to generate thousands of clusters from millions of conformations within several hours on a single PC, which is orders of magnitude faster than K-means. We call this approximation algorithm *AK-center* where "A" stands for approximation.

Furthermore, we show that AK-center clustering can be combined with other more deliberate methods which are otherwise not applicable due to the massive nature of data. As AK-center clustering efficiently reduces the complexity of the system in a faithful way, namely preserving the density and hence free energy information, this *hybrid* strategy of combining AK-center with more deliberate scheme proves to be very powerful. as is demonstrated in the result section where

AK-center is combined with spectral clustering to achieve a concise and clean-cut clustering of conformation space.

2 Method

In K-center clustering problem, the input consists of points in a metric space as well as a preordained number k specifying the number of clusters and the goal is to find a partition of the points into clusters $C_1, \cdots, C_k$ and the cluster centers $\nu_1, \cdots, \nu_k$ from the metric space, so as to minimize the maximum radius of clusters: $\max_j \max_{p \in C_j} d(p, \nu_j)$. This problem is NP-hard but has a simple 2-approximation algorithm [24, 25], meaning that the maximum radius of the outputted clusters is at most twice than that computed by the exact algorithm. It has been proved that 2 is indeed the best approximation factor possible [25]. The algorithm uses a so called *furthest-first traversal* [1] of the data and works as follows. The algorithm first picks any data point as the first cluster center and assigns all data points to the first cluster. Next it chooses the second cluster center as the point furthest from the first one and generates the second cluster by re-assigning to it those data points that are closer to the second cluster center, and then chooses the third cluster center that is furthest from the previous two and generates the third cluster as the data points having it as the closest cluster center, and so on until k cluster centers and thus k clusters are obtained. These k chosen points are often called landmarks. We call this approximation K-center algorithm *AK-center*. Note the above algorithm has one freedom of choosing the first cluster center, which is done in a random fashion in our implementation.

```
**A$K$-center($P$, $k$)**
Pick the first cluster $\nu_1$ arbitrarily from $P$.
Assign all data point to cluster $C_1$
for $i = 2, 3, \cdots, k$ do
    Take as the cluster center $\nu_i$ a point in $P$ furthest from $\{\nu_1, \cdots, \nu_{i-1}\}$, namely $\nu_i$ maximizes $\min_{1 \le j < i} \|p - \nu_j\|$ for any $p \in P$.
    for each data point $p_j \in P$ do
        if $d(p_j, \nu_i) < d(p_j, \nu_l)$ ($\nu_l$ is the current cluster center for $p_j$) then
            Re-assign $p_j$ to the new generated cluster $C_i$.
        end if
    end for
end for
Output $\nu_i$'s and $C_i$'s.
```

Since the maximum radius of the clusters monotonically decreases as the landmarks are added, we can keep adding the landmarks until the maximum radius of the clusters becomes less than a given threshold. Thus, we can obtain a modified clustering algorithm denoted *AK-center(P, δ)* where δ specifies the allowed maximum radius of the clusters.

The complexity of the algorithm is $O(NK)$ in terms of pair-wise distance computation and comparison where N is the number of data points and K is the number of the generated clusters. When we are given millions of conformations, it is not feasible to store all the pair-wise distances and algorithm must compute them on the fly. The distance one often uses is the root mean squared deviation (rmsd), which could be expensive to compute, especially for big molecules. Observe that rmsd is indeed a distance measure, namely it satisfies triangular inequality. If we maintain the

[1] This strategy of choosing landmakrs is used in many data analysis methods such as isomap [26] and manifold learning [27].

distances from the data points to their cluster centers, which needs $O(N)$ spaces, and the pair-wise distances between cluster centers, which needs $O(K^2)$ spaces, by utilizing triangular inequality, we can save a large amount of pair-wise distance computations. Before computing $d(p_j, \nu_i)$ to determine if re-assign the data point p_j the newly generated cluster i in (step 6), we first check if

$$d(p_j, \nu_l) \leq d(\nu_i, \nu_j)/2.$$

If so, by triangular inequality, we have

$$d(p_j, \nu_i) \geq d(\nu_i, \nu_l) - d(p_j, \nu_l) \geq d(p_j, \nu_l)$$

and hence for sure that the data point p_l will not be re-assigned, which saves the computation of the pair-wise distance $d(p_l, \nu_i)$. In the results section, we demonstrate the effectiveness of this strategy.

3 Results

3.1 Systems

We demonstrate the application of AK-center clustering to two model peptide systems in explicit solvent: alanine dipeptide and F_s peptide (Figure 1). For both models, We measure the distance using rmsd involving heavy atoms.

For alanine dipeptide, the conformations are from the trajectories obtained from the 400K replica of a 20 ns/replica parallel tempering simulation described in [19]. There are 975 trajectories, each of which contains 20 ps simulation with conformations stored every 0.1 ps, thus totally 195k conformations. For alanine dipeptide system, it is easy to obtain equilibrium sampling and projection of free energy landscapes onto a pair of torsion angles (ϕ and ψ) is relatively accurate. Therefore we can visualize the resulting clusters and check their various properties on this projection of the free energy landscape. see Figure 3(a).

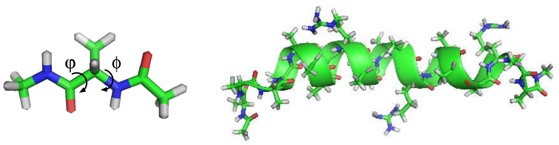


Figure 1.: Left: The terminally blocked alanine dipeptide. Right: The 21-residue helix-forming F_s peptide.

The 21-residue helix-forming F_s peptide is a larger peptide system, MD simulation generates two sets of 1000 trajectories at 302K of varying length of the capped F_s peptide, one set initiated from an ideal helix and another from a random coil [28]. The first 35ns of each trajectory was discarded to make sure the data indeed reach equilibrium, leaving a total of 1975 trajectories, each of which varies in length in 10 to 95 ns with a sampling interval of 100ps, and totally $745,263$ conformations.

3.2 Efficiency

In this section, we demonstrate the efficiency of AK-center clustering method. We compare with the commonly used K-means clustering. Recall that AK-center clustering method has one freedom of choosing the first cluster center ν_1. The experiments on both data sets with different randomly chosen ν_1's show almost the same results in terms of efficiency as well as other properties described in Section 3.3 and Section 3.4 (data not shown). Thus, in the paper, we only discuss the results with the fixed first cluster center that is randomly chosen.

Since rmsd measures the distance between two conformations up to a rigid transformation (translation, rotation or their combination), the mean structure of a cluster is not well defined. The often employed clustering on conformation space is K-medoids, which works iteratively as follows. Randomly choose k points as cluster centers and form k clusters $C_1, \cdots, C_k$ by assigning the rest of points to their closest cluster center, and then update cluster centers, and then form the new k clusters based on the newly chosen cluster centers, and continue until the cluster centers remain unchanged. To update the cluster center of C_i, randomly choose a few trial points and take the one with the least variance.

Table 1 shows the timing of AK-center and K-medoids applying on alanine dipeptide data. AK-center is much more efficient than K-medoids. It also shows that utilizing triangular equality can speed up both methods in more that an order of magnitude. Table 2 shows that AK-center can generate ten thousands of clusters from about a million conformations in a few hours. The timing is collected on a dual core machine with 16G MEM.

k	500	1000	2000	4000
AK-center	29 / 355	46 / 694	76 / 1389	132 / 2834
K-medoids	1262 / 4333	1408 / 7914	1837 / 15067	2657 / 29666

TABLE 1:: Timing (in second) with / without triangle inequality of AK-center and K-medoids on alanine dipeptide data. In K-medoids, we take 100 trial points in updating the cluster center for each cluster and total iterations is 10.

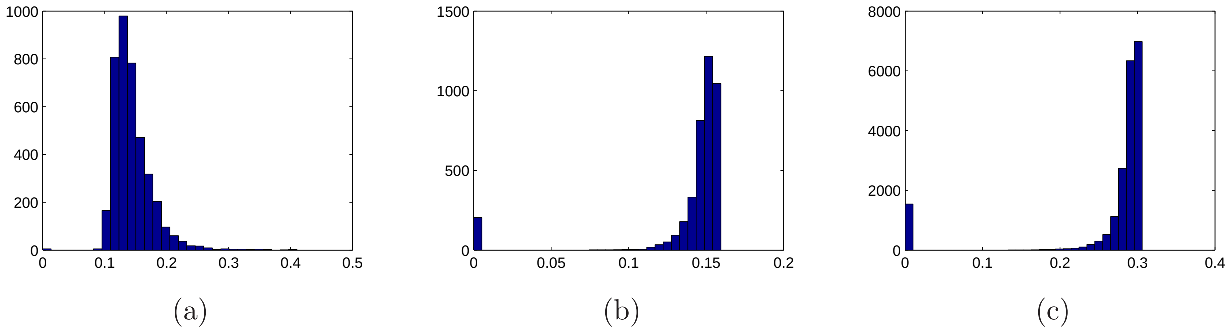
k	1000	5000	10000	20000
AK-center	3815	13528	28595	40915

TABLE 2:: Timing (in seconds) with triangle inequality of AK-center on Fs-peptide data.

3.3 Density information

Due to the good control on radius by AK-center method, below we will show that it preserves density information better than K-medoids method does, which gives important information on the free-energy landscape. This gets verified on both systems. For alanine dipeptide, the histogram of radii of clusters show that most of clusters generated by AK-center are about the same size around 0.14 $\mathring{A}$ (Figure 2(a)), while those by K-medoids have much wider spread (Figure 2(b)). The histogram for Fs-peptide (Figure 2(c)) shows most of clusters generated by AK-center are about the same size around 0.27 $\mathring{A}$. Therefore for the clusters generated by AK-center, the capacity of each cluster well indicates its relative density, and hence the free energy of the state that the cluster covers.

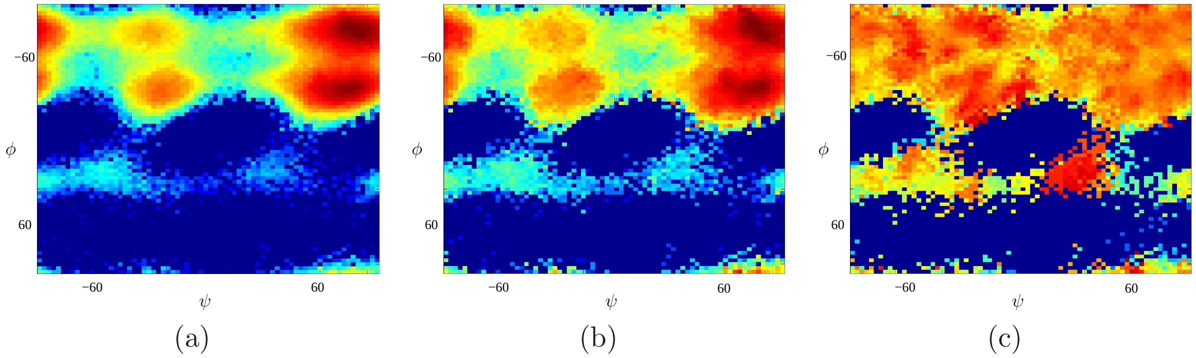
For alanine dipeptide, the mean force potential in $\phi-\psi$ plane can be estimated by the histogram of (ϕ, ψ)'s of all sampled conformations, see Figure 3(a), which is taken as the reference. To demonstrate the density information preserved by clustering, we color a bin in $\phi - \psi$ plane with



(a) (b) (c)

Figure 2.: Histograms of the radii of clusters (a) K-medoids on alanine dipeptide with $k = 4000$.(b) AK-center on alanine dipeptide with $k = 4000$; (c) AK-center on Fs-peptide with $k = 10000$.

the *average* population of those clusters that cover the bin. As shown in Figure 3, AK-center gives much better density estimation than K-medoids does.



(a) (b) (c)

Figure 3.: Mean force potential estimated using density; (a) ground truth; (b) Results from AK-center; (c) Results from K-medoids.

3.4 Improving cluster boundaries

It is known that the clusters obtained by AK-center algorithm may not have clean-cut boundaries [29]. However, after applying AK-center clustering, the complexity of the system is reduced by a lot, often from millions to thousands or tens of thousands, whence more deliberate scheme can be further employed to obtain a clustering with clean-cut boundaries. Below we demonstrate this idea by further applying spectral clustering over the clusters obtained by AK-center, to decompose conformation space into metastable states.

A metastable state consists of many free energy basins where the energy barriers between them are low. Transitions are fast within a metastable state but slow in-between. This separation of time scale enables one to build a Markov State Model (MSM) over the conformation space where each state in MSM is a metastable state. Such MSM can predict the long time behavior of a biological system at the time scale that current computer simulation can not reach [1, 30]. However, it is a challenging task to decompose conformation space into metastable states. One measure on the quality of metastable state decomposition is called the metastability, the sum of self-transition probabilities of all metastable states [19]. The bigger the metastability is, the better the decomposition is since the transition within a metastable state is faster than those in-between. To achieve the maximum metastability, spectral clustering is naturally employed over the transition matrix, as suggested by in [31]. Following [19], we take as microstates the clusters computed by AK-center algorithm ($k = 4000$ for alanine dipeptide and $k = 10000$ for Fs-peptide), and build a transition matrix using the input trajectories, and perform the spectral clustering over the transition matrix to obtain the metastable states. We obtain 6 metastable states with metastability 5.56 for alanine dipeptide data and 20 metastable states with metastability 13.07 for Fs-peptide. Both results are comparable to that (5.64 for alanine dipeptide and 14.05 for Fs-peptide) obtained by a much more complicated scheme in [19] involving 10 iterations of splitting and lumping based on the initial micrcostates generated by K-means. As we discussed before, K-means clustering tends to split the densely sampled free energy basin into many small clusters and lump the low density regions into the clusters of quite different structures. Therefore, it is necessary to perform many iterations of splitting and lumping to obtain good metastable states. On the other hand, AK-center clustering does not lump the conformations with very different structures into the same cluster, thus no extra effect needs to split the initial microstates.

4 Discussion and conclusion

One common concern about K-center clustering is that it tends to generate lots of noise clusters of few population away from the highly populated basin clusters. The algorithm may waste its resource, the quota k as the total number of clusters, on picking up noise. However, due to the special property of conformation space, namely conformations are highly concentrated at free energy basins, such waste is much less severe than that by K-means on splitting densely sampled free energy basins. Moreover, this problem of picking up noise can be leveraged by designing some noise filter and then leaving out those sparsely-populated noise. How to design noise filters systematically is one of our current research direction.

To conclude, we have considered the clustering from geometric point of view and presented a fast efficient clustering algorithm called AK-center for conformation space which preserves the density and energy landscape information. We have also show its efficiency by applications to two data set each with 10^5 conformations: alanine dipeptide and F_s peptide. We have also shown AK-center clustering can be combined with more deliberate schemes, for instance, to produce a concise and clean-cut clustering of conformation space. We believe such hybrid strategy of combining AK-center with other approaches which are otherwise not applicable due to high complexity of the system is not only useful for studying conformation space, but also other systems with the feature that data are highly concentrated at certain regions. We also believe that it is not hard to extend the AK-center clustering method for clustering data sets of billions of conformations, by using the techniques such as hierarchical structure, which is one of our on going research.

References and Notes

(1) Noé, F.; Fischer, S. *Current Opinion in Structural Biology* **2008**, *18*(2), 154–162.

(2) Torda, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **1994**, *15*(12), 1331–1340.

(3) Michel, A. G.; Jeandenans, C. *Computers & Chemistry* **1993**, *17*(1), 49 – 59.

(4) Elmer, S. P.; Pande, V. S. *The Journal of Chemical Physics* **2004**, *121*(24), 12760–12771.

(5) Lyman, E.; Zuckerman, D. M. *Biophys. J.* **2006**, *91*(1), 164–172.

(6) Rao, F.; Settanni, G.; Guarnera, E.; Caflisch, A. *The Journal of Chemical Physics* **2005**, *122*(18), 184901.

(7) Laboulais, C.; Ouali, M.; Bret, M. L.; Gabarro-Arpa, J. *Proteins: Structure, Function, and Genetics* **2002**, *47*(2), 169–179.

(8) Daura, X.; van Gunsteren, W. F.; Mark, A. E. *Proteins: Structure, Function, and Genetics* **1999**, *34*(3), 269–280.

(9) Troyer, J. M.; Cohen, F. E. *Proteins: Structure, Function, and Genetics* **1995**, *23*(1), 97–110.

(10) Torda, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **1994**, *15*(12), 1331–1340.

(11) Shenkin, P. S.; McDonald, D. Q. *J. Comput. Chem.* **1994**, *15*(8), 899–916.

(12) Heather L. Gordon, R. L. S. *Proteins: Structure, Function, and Genetics* **1992**, *14*(2), 249–264.

(13) Swendsen, R. H.; Wang, J. S. *Phys. Rev. Lett.* **1986**, *57*, 2607–2609.

(14) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

(15) Hansmann, U.; Okamoto, Y. *Curr. Opin. Struct. Biol.* **1999**, *9*, 177–183.

(16) Lyubartsev, A. P.; Martsinovski, A. A.; Shevkunov, S. V.; Vorontsov-Velyainov, P. N. *J. Chem. Phys.* **1992**, *96*, 1776–1789.

(17) Marinari, E.; Parisi, G. *Europhysics Letters* **1992**, *19*, 451–458.

(18) Huang, X.; Bowman, G.; Pande, V. S. *Journal of Chemical Physics* **2008**, *128*(20).

(19) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. *Journal of Chemical Physics* **2007**, *126*.

(20) Hartigan, J. A. *Clustering algorithms*; New York: Wiley, 1975.

(21) Hartigan, J. A. *Journal of Classification* **1985**, *2*, 63–76.

(22) Hastie, T.; Tibshirani, R.; Friedman, J. H. *The Elements of Statistical Learning*; Springer, 2001.

(23) Dasgupta, S.; Long, P. M. *J. Comput. Syst. Sci.* **2005**, *70*(4), 555–569.

(24) Gonzalez, T. F. *Theor. Comput. Sci.* **1985**, *38*, 293–306.

(25) Hochbaum, D.; Shmoys, D. *Math of Operations Research 10*(2).

(26) Tenenbaum, J. B.; de Silva, V.; Langford, J. C. *Science* **2000**, *290*, 2319–2313.

(27) Boissonnat, J.-D.; Guibas, L. J.; Oudot, S. Y. Manifold reconstruction in arbitrary dimensions using witness complexes. In *SCG '07: Proceedings of the twenty-third annual symposium on Computational geometry*, pp 194–203, New York, NY, USA, 2007. ACM.

(28) Sorin, E. J.; Pande, V. S. *Biophys J* **2005**, *88*(4), 2472–2493.

(29) J. Shao, S. W. Tanner, N. T.; III, T. E. C. *J. Chem. Theory. Comput* **2007**, *3*, 2312–2334.

(30) Singhal, N.; Snow, C.; Pande, V. S. *Journal of Chemical Physics* **2004**, *121*.

(31) Zha, H.; He, X.; Ding, C.; Simon, H.; Gu, M. Bipartite graph partitioning and data clustering. In *In CIKM*, pp 25–32, 2001.