Characterizing Intermediate Conformations in Protein Conformational Space

Rosanne Vetro, Nurit Haspel, and Dan Simovici

Department of Computer Science, University of Massachusetts Boston
100 Morrissey Blvd. Boston MA 02125 USA
{rvetro,nurith,dsim}@umb.edu

Abstract. In this paper we present a novel parallel coordinate based clustering method using Gaussian mixture distribution models to characterize the conformational space of proteins. We detect highly populated regions which may correspond to intermediate states that are difficult to detect experimentally. The data is represented as feature vectors of N dimensions, which are lower-dimension projections of the protein conformations. Parallel coordinates are a visualization technique that lays out coordinate axes in parallel rather than orthogonal to each other, thereby allowing patterns between pairs of axis as well as outliers to be visually identified in multi-dimensional data. We believe that the size of the resulting clusters may provide information about the likelihood of the corresponding conformations to exist as important intermediates. We tested our method on the conformational space for the enzyme Adenylate Kinase (AdK) which undergoes large scale conformational changes and used our method to detect clusters which may correspond to experimentally known intermediates. Finally, we compare our clusters with the ones generated by the K-Means clustering algorithm and discuss the advantages of our method for the problem of characterizing proteins conformational space.

Keywords: Clustering, Parallel Coordinates, Protein Conformational Search, AdK, Structural Bioinformatics.

1 Introduction

Proteins are flexible molecules that undergo structural (conformational) changes as part of their interactions with other proteins or drug molecules [1]. Changes in torsional angles may induce localized changes or large scale domain motions. Characterizing the conformational space of proteins is crucial for understanding the way they perform their function. There is promise that understanding the connection between protein structure, dynamics and function can contribute a lot to our understanding of how molecular machines function. Therefore, the question of how the structure and dynamics of proteins relate to their function has challenged scientists for several decades but still remains open.

Existing physics-based computational methods that sample the conformational space of proteins include Molecular Dynamics (MD) [2], Monte Carlo (MC) [3] and their variants, as well as approximate methods based on geometric sampling [4–7], Elastic Network Modeling [8], normal mode analysis [9], morphing [10] and more. One of the main challenges in modeling conformational changes in proteins is the difficulty in

L.E. Peterson, F. Masulli, and G. Russo (Eds.): CIBB 2012, LNBI 7845, pp. 70–80, 2013.

detecting intermediate structures that may correspond to transition states. These intermediate states are transient and therefore hard to detect experimentally, but they may be crucial to understanding folding, docking, binding and conformational change processes, as well as for drug design, since many times a drug is targeted as a transition state analog or to block the target molecule from undergoing a structural change. In addition, full-scale conformational search of even a medium sized protein is very demanding computationally and the conformational landscape of proteins is many times rugged and hard to navigate. Therefore, the challenging problem of fully characterizing conformational pathways in proteins still remains open. Recently, we developed a semi-coarse grained conformational search method that conducts a fast, approximate search on the conformational space of proteins undergoing large-scale domain motions [4]. While the method produced feasible conformational pathways, these pathways needed to be clustered and filtered to extract meaningful intermediate conformations.

In this work we use a variant of the above mentioned conformational search algorithm [4] to provide an approximate description of the protein conformational landscape. The algorithm runs a large number of monte-carlo like searches in the conformational space of proteins undergoing large-scale changes. By repeating the procedure a large number of times we produce a set of feasible pathways which provide a good coverage of the space.

In order to find highly populated regions which may correspond to intermediate structures, we introduce a clustering method that takes as input conformational pathways represented by a lower-dimensional projection of the protein conformational space and outputs clusters of data that can give us information about the likelihood of the existence of given structures. The method performs a statistical analysis of multi-dimensional data representing conformations. Each dimension is partitioned in a possibly different number of blocks using model-based clustering with Gaussian mixture models and the data flow between pairs of dimensions is analyzed in order to create disjoint multi-dimensional clusters of conformations and identify structures that are unlikely to be meaningful local minima as outliers.

Experimental results regarding the conformational space for the enzyme Adenylate Kinase (AdK) suggest that the combination of our conformational search and clustering method can help us detect highly populated areas in the conformational space, represented by large clusters, which may indicate the location of important intermediate structures in the protein conformational space, as demonstrated by similarity to known AdK intermediate homologs. In order to evaluate our clustering method, we compare our results with the ones generated by multiple runs of the K-means algorithm [11, 12] and present the advantages of our approach.

The paper is organized as follows. Section 2 presents the methods for protein conformational search and clustering. The methodology is evaluated experimentally in section 3. Finally, the paper is concluded in section 4.

2 Methods

2.1 Protein Conformational Search

We use a semi–coarse grained protein structure representation. The proteins are stripped of their side-chain and hydrogen atoms and represented at the backbone and C-β level

(Glycine is represented by its backbone only). We apply a semi-coarse-grained potential function to approximate the protein energy [13] and an efficient distance measure to estimate the distance between two protein structures based on the positions and angles of their secondary structure elements [4]. This measure represents each protein conformation as a feature vector whose size is the order of magnitude of the number of rigid elements in the protein, thus projecting the structures onto a much lower dimensional space than its full representation and is more "natural" to protein structures than other projections such as PCA, since PCA is limited by its linear nature.

The search algorithm used here is a variant of a Monte Carlo search that leads from one conformational state of the protein (start) to another (goal), applying successive geometric transformations to a randomly selected backbone degree of freedom of the structure while retaining only intermediate structures with an energy below a threshold [4]. We used that method to validate the results at the previous paper. In this paper we used the Monte-Carlo based search rather than a Robotics based search used in the previous work [14], since that method tends to bias the results towards the goal structure and in the present work we wanted to generate as random a sampling of the low-energy conformational space as possible. The reader is encouraged to refer to [4] for more details about the conformational search method.

2.2 Clustering Method

Today, the majority of clustering methods for multi-dimensional data incorporates metric functions that evaluate the distance between feature vectors extracted from a data-set. In this scenario, multiple dimensions are combined and are simultaneously considered according to a metric function in order to create a set of clusters.

In this paper, we propose an alternative clustering method based on parallel coordinates [15] and Gaussian mixture models [16], and argue that it is suited for providing information about the likelihood of the existence of given intermediate conformations in a protein conformational space.

To formally describe our clustering method, we introduce the following notation. The symbol $\mathscr{C}$ stands for a set of conformations represented by feature vectors in the data set, n for the conformational space dimensionality and Σ is the $n \times n$ covariance matrix of the data set where each element along the diagonal, $\Sigma[i,i]$, corresponds to the variance of dimension d_i with $1 \leq i \leq n$. The statistical information provided by Σ is used to create $\mathscr{L}$, an ordered list of dimensions. The threshold used by the algorithm to find outliers or conformations that are unlikely to exist is given by τ where $0 \leq \tau \leq 1$. It corresponds to the minimum fraction of diverging vectors that can form a new cluster, considering the total number of vectors in the original cluster from which the split occurred. Finally, B stands for the matrix containing information about the blocks of each dimension's partition. For instance, $B[i,j]$ corresponds to the block from dimension j in which the corresponding data value from conformation i is located.

Parallel coordinates is a common way of visualizing high-dimensional geometry and analyzing multivariate data. Dimensions or axis are laid out in parallel rather than orthogonal to each other. Each data value of an n-dimensional vector is positioned on the line corresponding to its axis, between the minimum (at the bottom) and the maximum (at the top) values of the axis. Points belonging to the same vector are connected

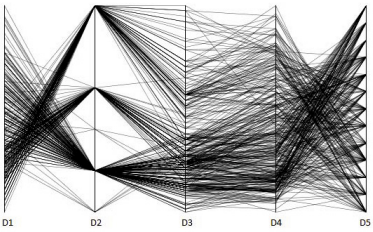


Fig. 1. Example of parallel coordinates for a 5-dimensional data set

by lines, which allows patterns between dimensions and outliers to be visually identified. For example, Figure 1 shows a 5-dimensional data set displayed as a sequence of parallel coordinates. Notice the inverse relationship between $D1$ and $D2$ and the correlation between $D2$ and $D3$: lower values of $D1$ usually imply higher values of $D2$ and vice-versa; higher values for $D2$ usually imply higher values for $D3$ and vice-versa. Likewise, we can visualize the correlation between $D3$ and $D4$ and the cross among the lines between $D4$ and $D5$. Exceptions or outliers corresponding to diverging lines that disrespect the usual behavior between dimensions can also be spotted using this technique.

The real strength of parallel coordinates is in modeling relations between variables, as discussed in [17]. Our method analyzes the variance of each dimension to model those relations. The model is simply represented using $\mathscr{L}$. The purpose of $\mathscr{L}$ is to determine the order in which the algorithm will analyze the data flow between consecutive pairs of dimensions in order to form clusters.

Given a set $\mathscr{C}$ of conformations represented by feature vectors in n dimensions, the covariance matrix Σ of the data set is generated and all dimensions are placed in $\mathscr{L}$ in increasing order of variance. We do not claim that this arrangement is optimal. The optimal ordering of the dimensions is a topic for further study.

In order to assign a unique cluster to each conformation and identify outliers our method first performs a model-based clustering on each dimension separately using Gaussian mixture distribution models to estimate density. A Gaussian or normal mixture model is a parametric probability density function represented as a weighted sum of Gaussian component densities. Gaussian mixture models are commonly used as parametric models of the probability distribution of continuous measurements or features. Model-based clustering [18] is based on a finite mixture of distributions, in which each mixture component corresponds to a different cluster or block. For continuous data, the most common component distribution is a Gaussian distribution. Choosing a suitable number of components gc is essential for creating a useful model of the data and for data partitioning. The authors of [19] state that when a Gaussian mixture model is used for clustering, there might be an overestimation of the number of clusters. This is because a cluster may be better represented by a mixture of Gaussians than by a single Gaussian distribution. In [20] the authors argue that the goal of clustering is not the same as that of estimating the best approximating mixture model. Indeed, our objective in this work is not to find the number of components that best approximates the data,

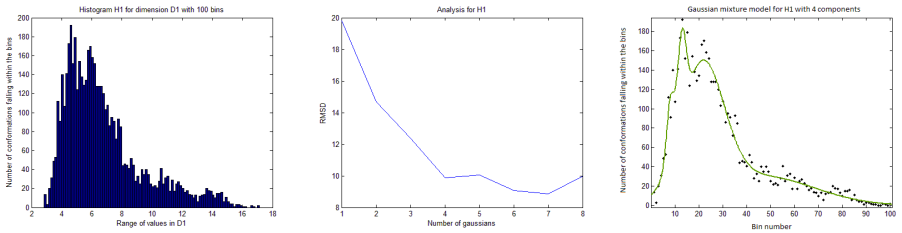
as an estimation for the number of blocks in each dimension's partition. Instead, we determine the minimum number of Gaussian components associated to each dimension's data, whose Root-Mean-Square deviation (RMSD) corresponds to a local minimum or approximates a minimum since the RMSD global minimum usually corresponds to a high number of components.

The process of choosing the number of components gc for the data associated to a dimension from a sample conformation data set is demonstrated in Figure 2. It starts with the generation of a fine histogram with N bins corresponding to a sequence of uniformly spaced single-valued points $\{x_k : k = 1,\ldots,N\}$ with associated data values $\{y_k : k = 1,\ldots,N\}$. Then, a set of m Gaussian models with a number of components varying from 1 to m is used for fitting the histogram's data. The Gaussian model is given by Eq. 1 where a corresponds to the amplitude, b is the centroid location, c is related to the peak width and m is the number of peaks to fit.

$$f(x) = \sum_{i=1}^{m} a_i e^{\left[-\left(\frac{x-b_i}{c_i}\right)^2\right]} \tag{1}$$

We analyze the curve generated for the graph where the x-axis represents the number of components $\{x_i : i = 1,\ldots,m\}$ and the y-axis corresponds to the associated RMSD values. We choose the smallest number of components with an RMSD corresponding to a local minimum or to a value that approximates a minimum.

Once the number of components corresponding to the number of clusters for each dimension is estimated, a Gaussian model-based clustering is used to partition the dimensions. Each element of the conformation's feature vector is assigned to a unique cluster in the corresponding dimension and this information is stored in B.



(a) (b) (c)

Fig. 2. Illustration of data modeling process: (a) histogram of data in a single dimension, (b) plot of RMSD versus number of Gaussian components, (c) data fitting with four Gaussian components

Once $\mathscr{L}$ and B are generated, the initial conformation clusters are created taking into consideration only the data and clusters from dimension $\mathscr{L}[1]$. Then, for each pair of dimensions $(\mathscr{L}[i], \mathscr{L}[i+1])$, we continue to refine our initial set of clusters by grouping the vectors that belong to the same cluster, i.e., those vectors that fall into the same block in $\mathscr{L}[i+1]$. The vectors comprising the cluster must also satisfy the constraint given by τ, whereby any "diverging" set of vectors must have a number of elements greater than a fraction τ of the total number of vectors in the original cluster. Vectors that do not

satisfy this constraint are considered outliers or conformations that are unlikely to exist as meaningful intermediates; such vectors are removed from the clustering process. The final set of clusters is formed after all consecutive pairs of dimensions have been considered according to the order given by $\mathscr{L}$ and the process described above.

The algorithm for the proposed clustering method takes as input $\mathscr{C}$ and τ, and outputs a set of disjoint clusters as well as a set of outliers. The name of each final cluster shows the identification of the corresponding dimension blocks. The pseudocode is presented as Algorithm 1.

Algorithm 1 . Clustering Algorithm

Require: $\mathscr{C}$, τ
Ensure: A set of disjoint clusters and a set of outliers
  Compute Σ
  Compute $\mathscr{L}$
  Compute B
  $i \leftarrow 1$
  Name each initial cluster in $\mathscr{L}[i]$ according to the block from which it originated
  for all consecutive pairs ($\mathscr{L}[i]$,$\mathscr{L}[i+1]$) **do**
    for all clusters up to dimension $\mathscr{L}[i]$ **do** Evaluate $T_\beta = \tau *$cardinality of cluster β
    end for
    Use B to find and group the vectors that belong to same cluster β up to dimension $\mathscr{L}[i]$ and fall into the same block at dimension $\mathscr{L}[i+1]$
    Obtain the cardinality $T_{g\beta}$ of each group g, where β represents the original cluster from which those vectors originated
    if $T_{g\beta} \geq T_\beta$ **then**
      Discard β and create a new cluster with the corresponding vectors
      Add the current dimension block identification (at dimension $\mathscr{L}[i+1]$) to the name of the original cluster β in order to name the new cluster
    else
      Remove corresponding vectors from clustering process and classify them as outliers
    end if
  end for

The implementation of the algorithm includes MATLAB scripts and C code.

3 Experimental Results

3.1 Tested System - Adenylate Kinase (AdK)

The conformational search and subsequent clustering was run on AdK. It is a monomeric phosphotransferase enzyme that catalyzes reversible transfer of a phosphoryl group from ATP to AMP. The structure of AdK, which contains 214 amino acids, is composed of the three main domains, the CORE (residues 1–29, 68–117, and 161–214), the ATP binding domain called the LID (residues 118–167), and the NMP binding domain (residues 30–67). AdK assumes an "open" conformation in the unligated structure

and a "closed"conformation. The RMSD between the two structures is 6.95Å. Supposedly, during the transition from the "open" to "closed" form, the largest conformational change occurs in the LID and NMP domain with the CORE domain being relatively rigid. Our model contains 8 rigid elements where most of the CORE domain was modeled as one large segment and was considered fixed, since it does not undergo a large-scale motion. Hence, the data were represented as feature vectors of 8 dimensions each. We ran the search 30 times in the direction of 1AKE-4AKE and 30 times in the reverse direction. Overall we collected 11,823 intermediate conformations.

3.2 Resulting Clusters

The data set $\mathscr{C}$ generated by our conformational search consists of the 60 pathways containing 11,823 conformations projected onto an 8-dimensional space which represents the protein conformational space. In order to perform a model-clustering of each dimension, the number of blocks in each partition is determined based on the number of components of the Gaussian mixture model chosen. The process for choosing the number of components generates histograms with 100 bins and analyzes Gaussian models with up to 8 components for each dimension. We use MATLAB with default arguments and the Trust-region algorithm [21, 22] to generate each model. We also used MATLAB to run the Gaussian model-based clustering, with the number of mixture components as input argument.

Our experiments use 4 different values of τ, the threshold used for identifying outliers: 0.05, 0.1, 0.2 and 0.3. Among all clusters generated by our method, we are interested in the large clusters (with at least 20 members), which distribute narrowly around their cluster center. While several of the most populated clusters may contain conformations that are close to known intermediates, some of them are narrower than others, regarding their deviation in terms of the centroid location. Since we are evaluating the RMSD between known structures and the resulting clusters centroid location, narrower clusters may produce more desirable results because their standard deviation is lower. We observed that for our sample data set, some clusters having at least 20 conformations contain conformations that are close to known intermediates structures (see Section 3.3). Further study is needed to determine the appropriate size for cluster of interest taking into consideration the standard deviation with respect to the centroid location and RMSD from known structures. Table 1 presents statistics about the resulting clusters according to selected values of τ. Notice that as the value of τ increases, so does the number of outliers detected by the algorithm.

3.3 Comparison with Known Intermediates

In general, knowledge about intermediate conformations is needed in order to provide a case-specific validation, but this knowledge does not always exist. As a matter of fact, intermediate structures are hard to obtain due to their relative high energy with respect to the native structures. With the advances in structural detection and simulation methods, one can expect to have more information about intermediate states in the future. AdK has several known mutant and intermediate structures. In a recent study [23] the energy profile of AdK was produced using elastic network interpolation (ENI).

Table 1. Statistics containing the number of resulting clusters (not considering the cluster containing the set of outliers), number of outliers and number of clusters with at least 20 conformations as well as the number of conformations in the smallest and largest clusters according to selected values of τ

	$\tau = 0.05$	$\tau = 0.1$	$\tau = 0.2$	$\tau = 0.3$
n^o of clusters	329	231	58	10
n^o of outliers	567	1703	4380	7547
n^o of clusters with 20^+ conformations	88	75	44	10
size of smallest cluster	1	1	2	85
size of largest cluster	1312	1312	1312	1312

The method was used to generate the conformational transition pathway between the open and closed form of AdK and back, and compare the intermediates to known structural intermediates. Inspired by that study, we performed a similar test on our results. We focused on five known intermediates: chains A, B, and C of the hetero-trimer Adenylate Kinase from Aquifex Aeolicus (PDB accession code 2RH5), which are conformational change intermediates of the ligand free AdK [24], 1E4Y, which is an AdK mutant having 99% sequence identity with 4AKE and 1AKE and is a closed form of AdK binding with AP5, and 1DVR, which is a mutant that exhibits LID closure [25]. We selected all the clusters that containted at least 20 members and recorded for each cluster center the closest conformation to 1E4Y, 1DVR and to chains A, B and C of 2RH5. Our results are shown in Table 2. For each intermediate, the table shows the lowest RMSD from the closest conformation cluster center and the cluster number. We considered only "well-behaved" clusters, that is - the maximum distance from the cluster average was at most 3Å. Figure 4 shows the distribution of the RMSDs of cluster elements from the cluster center for these clusters. Notice that the same intermediate was the closest to chains B and C of 2RHC. This can be explained by the fact that the two chains are very similar to one another – the RMSD between them is approximately 2Å. The data corresponds to the result of our clustering method with a threshold τ of 0.05. In fact, this value of τ provided the best result obtained by our method. Figure 3 shows the intermediate structures superimposed on the closest cluster center for each intermediate.

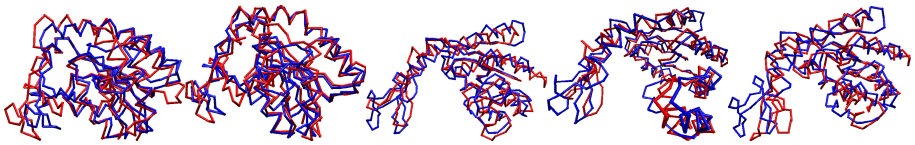
Table 2. RMSDs of cluster centers from five known AdK mutants representing intermediate states. The data were taken from our proposed method, cutoff of 0.05.

Intermediate PDB code	**2RH5(A)**	**2RH5(B)**	**2RH5(C)**	**1E4Y**	**1DVR**
Cluster name	1_2_3_1_1_1_2_3	1_2_1_2_1_1_2_3	1_2_1_2_1_1_2_3	1_2_2_2_2_1_3_3	1_1_3_2_2_1_2_1
Cluster size	33	21	21	84	101
RMSD with cluster average†	2.55	2.49	2.89	2.56	2.77

† The RMSD was calculated with respect to the $C-\alpha$ atoms of the aligned residues between the two proteins

3.4 Comparison of the Proposed Clustering with K-Means

In order to validate our clustering algorithm, we compare our results to others generated by the K-Means algorithm. Weka [26] was the workbench used to run K-Means



(a) 1DVR (b) 1E4Y (c) 2RH5 (chain A) (d) 2RH5 (chain B) (e) 2RH5 (chain C)

Fig. 3. Cluster representatives (blue) superimposed on known intermediates (red). See Table 2 for details.

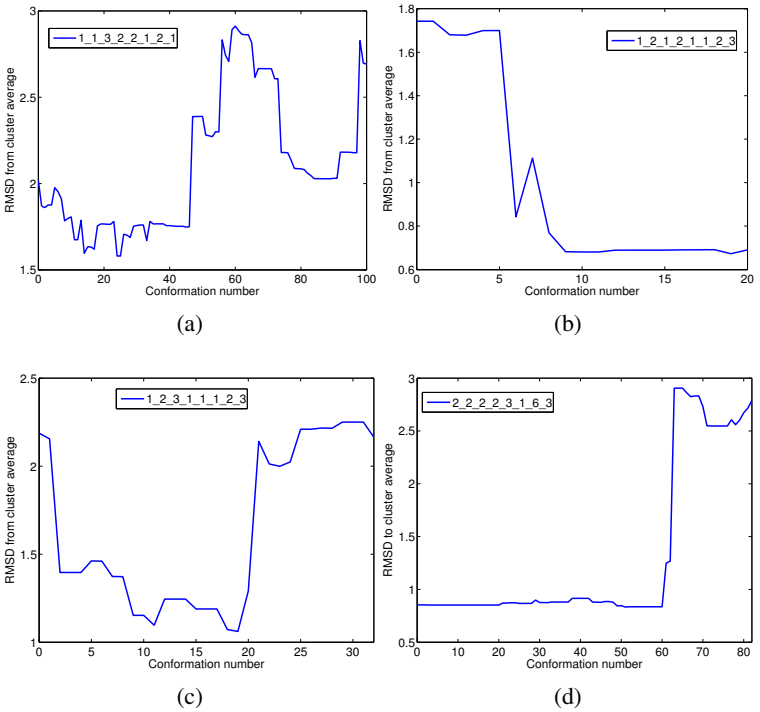


Fig. 4. Distribution of RMSD of cluster elements from cluster average for the clusters representing AdK intermediates (see Figure 3 and Table 2. See inset legend for cluster name.

with Euclidian distance and a maximum of 2000 iterations. We selected 3 well spaced arbitrary values for the number of clusters K within the range of the number of clusters found by our method (see Table 1): 20, 80 and 150. Table 3 shows the results according to the number of clusters K. As seen, with $K = 20$ clusters there was no narrow cluster (with a radius below 3Å) corresponding to intermediates 1DVR and 1E4Y within a reasonable RMSD. Only at $K = 150$ the results were comparable to our method.

Table 3. RMSDs of cluster centers generated by K-Means from five known AdK mutants representing intermediate states

	2RH5(A)		**2RH5(B)**		**2RH5(C)**		**1E4Y**		**1DVR**	
K	RMSD	size	RMSD	size	RMSD	size	RMSD	size	RMSD	size
20	2.69	91	2.73	91	3.35	91	–	–	–	–
80	2.64	28	2.71	112	3.17	36	2.58	55	2.88	20
150	2.52	31	2.43	120	2.95	43	2.52	77	2.82	82

The proteins conformational space may contain many intermediate structures that are unlikely to exist as significant intermediates. Therefore, outlier detection is highly desirable in a clustering algorithm for the characterization of protein conformational space. The K-Means algorithm does not allow for the detection of outliers whereas our clustering method has the advantage of providing a flexible way to detect them. In addition, there is no a-priory way to know K, the number of clusters, in advance, and an educated guess has to be made. Our method provides a more deterministic way to evaluate the number of clusters. As a matter of fact, in this paper K was determined according to the number of clusters discovered by our method (see Table 1).

4 Conclusion

Characterization of protein conformational space is a very challenging problem due to the large amount of calculations required to characterize that complex and multi-dimensional space and due to the scarcity of experimental data regarding intermediate states. In this paper we presented a clustering method based on parallel coordinates and used it to characterize the conformational space of AdK and detect highly populated areas that may correspond to intermediate structures, which are usually hard to detect using experimental methods. In the case of AdK, however, several intermediate homologs exist and we were able to find cluster centers corresponding to these intermediates. The advantage of our method over K-means clustering and other standard clustering methods is that it allows the detection of outliers and does not require the number of final clusters to be given as input. Also, the parameters can be adjusted to gain insight about the optimal number of clusters. Detecting the ideal cutoff for the data and trying to find better ways to merge close clusters is the subject of on-going research.

References

1. Perutz, M.F.: Mechanisms of cooperativity and allosteric regulation in proteins. Quart. Rev. Biophys. 22, 139–236 (1989)
2. Case, D.A., Cheatham, T., Darden, T., Gohlke, H., Luo, R., Merz Jr., K.M., Onufriev, A., Simmerling, C., Wang, B., Woods, R.: The Amber biomolecular simulation programs. J. Computat. Chem. 26, 1668–1688 (2005)
3. Kirkpatrick, S., Gelatt Jr., C.D., Vecchi, M.P.: Optimization by simulated annealing. Science 220, 671–680 (1983)
4. Haspel, N., Moll, M., Baker, M., Chiu, W., Kavraki, L.E.: Tracing conformational changes in proteins. BMC Structural Biology (2010) (in press)

5. Thomas, S., Tang, X., Tapia, L., Amato, N.M.: Simulating protein motions with rigidity analysis. J. Comp. Biol. 14(6), 839–855 (2007)
6. Chiang, T.H., Apaydin, M.S., Brutlag, D.L., Hsu, D., Latombe, J.-C.: Using stochastic roadmap simulation to predict experimental quantities in protein folding kinetics. J. Comp. Biol. 14(5), 578–593 (2007)
7. Raveh, B., Enosh, A., Furman-Schueler, O., Halperin, D.: Rapid sampling of molecular motions with prior information constraints. Plos Comp. Biol. (2009) (in press)
8. Zheng, W., Brooks, B.: Identification of dynamical correlations within the myosin motor domain by the normal mode analysis of an elastic network model. J. Mol. Biol. 346(3), 745–759 (2005)
9. Schroeder, G., Brunger, A.T., Levitt, M.: Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. Structure 15, 1630–1641 (2007)
10. Weiss, D.R., Levitt, M.: Can morphing methods predict intermediate structures? J. Mol. Biol. 385, 665–674 (2009)
11. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice Hall (1988)
12. McQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–296 (1967)
13. Shehu, A., Kavraki, L.E., Clementi, C.: Multiscale characterization of protein conformational ensembles. Proteins: Structure, Function and Bioinformatics (2009)
14. Ladd, A.M.: Motion Planning for Physical Simulation. PhD thesis, Dept. of Computer Science, Rice University, Houston, TX (December 2006)
15. Inselberg, A.: Parallel coordinates: a tool for visualizing multi-dimensional geometry. In: Proceedings of the First IEEE Conference on Visualization, California, USA, pp. 361–378 (1990)
16. McLachlan, G., Peel, D.: Finite Mixture Models. John Wiley and Sons (2000)
17. Inselberg, A.: Visual data mining with parallel coordinates. Computational Statistics 13 (1998)
18. Fraley, C., Raftery, A.E.: Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, 611–631 (June 2002)
19. Baudry, J., Raftery, A.E., Celeux, G., Lo, K., Gottardo, R.: Combining mixture components for clustering. Journal of Computational and Graphical Statistics 19(2), 332–353 (2010)
20. Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 719–725 (2000)
21. Celis, M.R., Dennis, J.E., Tapia, R.A.: A trust region strategy for nonlinear equality constrained optimization. In: Proceedings of the SIAM Conference on Numerical Optimization, pp. 71–82 (1984)
22. Conn, A.R., Gould, N.I.M., Toint, P.L.: Trust-Region Methods. SIAM, PA (2000)
23. Feng, Y., Yang, L., Kloczkowski, A., Jernigan, R.L.: The energy profiles of atomic conformational transition intermediates of adenylate kinase. Proteins 77(3), 551–558 (2009)
24. Henzler-Wildman, K.A., Thai, V., Lei, M., Ott, M., Wolf-Watz, M., Fenn, T., Pozharski, E., Wilson, M.A., Petsko, G.A., Karplus, M., Hübner, C.G., Kern, D.: Intrinsic motions along an enzymatic reaction trajectory. Nature 450(7171), 838–844 (2007)
25. Schlauderer, G.J., Proba, K., Schulz, G.E.: Intrinsic motions along an enzymatic reaction trajectory. J. Mol. Biol. 256, 223–227 (1996)
26. Holmes, G., Donkin, A., Witten, I.H.: Weka: a machine learning workbench. In: Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems, pp. 357–361 (1994)