Published in final edited form as:

J Comput Chem. 2016 May 30; 37(14): 1251–1258. doi:10.1002/jcc.24315.

Inherent Structure versus Geometric Metric for State Space Discretization

Hanzhong Liu, Minghai Li¹, Jue Fan², and Shuanghong Huo*

Gustaf H. Carlson School of Chemistry and Biochemistry, Clark University, 950 Main Street, Worcester, Massachusetts 01610, USA

Abstract

Inherent structure (IS) and geometry-based clustering methods are commonly used for analyzing molecular dynamics trajectories. ISs are obtained by minimizing the sampled conformations into local minima on potential/effective energy surface. The conformations that are minimized into the same energy basin belong to one cluster. We investigate the influence of the applications of these two methods of trajectory decomposition on our understanding of the thermodynamics and kinetics of alanine tetrapeptide. We find that at the micro cluster level, the IS approach and root-mean-square deviation (RMSD) based clustering method give totally different results. Depending on the local features of energy landscape, the conformations with close RMSDs can be minimized into different minima, while the conformations with large RMSDs could be minimized into the same basin. However, the relaxation timescales calculated based on the transition matrices built from the micro clusters are similar. The discrepancy at the micro cluster level leads to different macro clusters. Although the dynamic models established through both clustering methods are validated approximately Markovian, the IS approach seems to give a meaningful state space discretization at the macro cluster level.

Keywords

inherent structure; Markov state model; effective energy surface; clustering method; mean first passage time

Introduction

With the rapid growth of computer power and new developments of algorithms, it is possible for atomic-level molecular dynamics simulations to reach millisecond timescale.^[1,2] Now the question is how to analyze the huge amount of data to obtain structural information as well as thermodynamics and kinetics properties of the system. Frequent approaches include dimensionality reduction, graph-theoretic approach,^[3–8] and Markov state model (MSM).^[9]

*Corresponding authors: shuo@clarku.edu.

¹Present address: Northeastern University, Physics Department, Boston, MA 02115

²Present address: Cellular Research, INC., 4040 Campbell Avenue, Suite 110, Menlo Park, CA 94025

Author Contribution

All the authors substantially contribute to the research design. HL contributes to the acquisition, analysis and interpretation of data, drafting the paper and revising it critically. ML and JF contribute to the acquisition, analysis and interpretation of data. SH contributes to the interpretation of data and revising the draft critically. All authors contribute to approval of the submitted and final versions.

When the dynamics is projected onto one or two dimensions, free energy surface and conformational change pathways can be visualized. However, no matter whether the lower dimensions are user-defined order parameters such as root-mean-square deviation (RMSD), radius of gyration, native contacts, and so on or systematically determined such as principal component analysis,^[10] diffusion maps^[11,12] and Isomap,^[13,14] dimensionality reduction approaches have to face the challenge to prove that they truly preserve the properties in the original space, which is hard in some situations.^[15]

A graph or network can be used to describe the thermodynamics and kinetics of a system.^[16,17] A node on the graph represents a state with certain populations and the edges between nodes can be net reactive flux, reaction rate, or time. Markov state model discretizes a continuous dynamics into states and uses the dynamics of the discrete model to represent the original continuous dynamics. Both graph-theoretical approaches and MSM need to discretize the trajectories into clusters and all the following thermodynamics and kinetics calculations depend on the clustering result.

To capture the kinetic property, a clustering criterion based on kinetic distance is the ideal choice. However, since there is not any simple way to calculate kinetic distance of two conformations, geometric metrics, such as RMSD, are widely used. A small value in RMSD implies fast conversions between the corresponding conformations, thus they are kinetic neighbors. Yet, as pointed out in literature this geometric metric faces challenges.^[18–21] For example, the conversion between two conformations with a relatively small RMSD may be a slow dynamics.^[20] While recent method using transition-based state assignment^[22] leads to a systematic analysis of folding-unfolding trajectories, ranges in conformation space need to be defined first, which is not practical in complex systems.

Inherent structure is a method to map conformations onto potential/effective energy surface by energy minimization, pioneered by Stillinger and Weber.^[23,24] If two conformations are minimized to the same local energy minimum, also called inherent structure (IS), then kinetically they should not be too far away from each other, thus they are kinetics neighbors. In recent years the applications of IS have expanded from liquids to proteins with success.^[21,25–28] It was shown that the IS approach avoided the uncertainties brought by clustering with RMSD cutoff in the detailed characterization of the folded state free energy surface and in the determination of the associated conformation space network.^[21] In this work we investigate how the IS and RMSD-based clustering methods influence our understanding of thermodynamics and kinetics of a model peptide. Unlike Rao's work,^[21] we compare the IS approach with RMSD clustering method at two levels, micro cluster and macro cluster level, for the conformational changes in a wide range, not limited to the folded state.

Methods

Molecular Dynamics Simulations

The dynamics of alanine tetra-peptide, ACE-(ALA)₄-CBX, was simulated using CHARMM^[29] with the polar hydrogen potential function^[30] and the ACE2 implicit solvation model^[31] as reported in our previous work.^[16] Short peptides are good models to

test new methodology. Although alanine dipeptide is widely used as a test system,^[32–34] we need a more complex system than it. The system size of alanine tetra-peptide is still manageable for extensive sampling and it can form a full α -helical turn. A similar size of peptide was used as a test model for computational methods to explore potential energy surfaces and conformational transitions.^[35,36] When a conformation is mapped onto the effective energy (potential energy plus solvation free energy) surface, solvation should be taken into account in the minimization. For consistency and simplicity, the same solvation model is used in both simulation and the subsequent mapping of inherent structures, instead of using an explicit solvation model in simulation and an implicit solvation model in the search of inherent structure. Because the peptide was used as a model system not relevant to any biological functions, a high simulation temperature, 500 K, was set to ensure extensive sampling. The friction coefficient of 64 ps^{-1} was used along with the integration step of 1 fs. Conformations were saved every 1 ps in the production run, resulted in a total of 1 million conformations ($= 1 \text{ }\mu\text{s}$). To test the convergence, we divided the trajectory into four quarters. For each quarter, we plot the Ramachandran plot for each pair of (ϕ, ψ) . As shown in Fig. S1, the Ramachandran plots of each residue for the last two quarters of the trajectory seem almost identical, suggesting that the system has reached equilibrium.

Clustering

There are two stages of clustering: generation of micro clusters and lumping micro clusters to macro clusters. When clustering the conformations into micro clusters, on one hand one wants to ensure that within a single cluster no conformations are separated by large free energy barriers, on the other hand, sufficient statistics of transitions between each pair of micro clusters need to be maintained. Consequently, the number of micro clusters should not be too many nor too few.

Clustering with geometric metric: RMSD—We used combined k-centers clustering and k-medoids clustering implemented the MSMBuild2 package.^[37] Conformations were chosen every 100 ps along the trajectory for the initial clustering of micro clusters, followed by assigning the rest of the conformations to the existing micro clusters. To determine the cutoff of RMSD for the micro clusters, we calculated the distribution of RMSD between every pair of adjacent conformations along the trajectory (Fig. 1), which shows the peak is at $0.56 \text{ }\text{\AA}$ and its tail extends to around $1 \text{ }\text{\AA}$. Thus, the largest RMSD could be reached in the shortest time interval along the trajectory is about $1 \text{ }\text{\AA}$ and the average is $0.56 \text{ }\text{\AA}$. We tried three RMSD cutoffs: $0.56 \text{ }\text{\AA}$, $0.8 \text{ }\text{\AA}$, and $1 \text{ }\text{\AA}$. As shown in Table 1, when the radius of the micro clusters was set to $0.56 \text{ }\text{\AA}$, the fact that 99.8% of the clusters have populations less than 0.1% suggests that the cutoff of $0.56 \text{ }\text{\AA}$ is too small. If the RMSD threshold is raised to $0.8 \text{ }\text{\AA}$, 83.7% of the clusters have populations less than 0.1%. When the cutoff reaches $1 \text{ }\text{\AA}$, the percentage of the clusters that have populations less than 0.1% is lowered to 61.0%. In order to balance the need of sufficient statistics of transitions between clusters and the fine discretization of conformation space, we chose $1 \text{ }\text{\AA}$ cutoff which is the largest RMSD within the shortest time interval along the trajectory, resulted in 733 micro clusters.

IS Analysis—To obtain IS micro clusters, minimization was performed using 100 steps of the steepest descent (SD) followed by the adopted basis Newton-Raphson (ABNR). The

convergence criterion was set to either 20,000 steps of ABNR or the change in energy less than 2×10^{-6} kcal/mol, whichever comes first. The solvation free energy was taken into account during minimization. SD is used to quench a conformation to its nearest effective energy minimum while ABNR assures the convergence of the minimization. A variety of minimization protocols was applied to calculate ISs and it was found that the overall thermodynamic properties of a peptide system were not affected.^[38] To cluster the ensemble of minimized conformations, firstly we sorted the conformations according to their post-minimization energies. Secondly, we employed two cutoffs, energy and all-atom RMSD cutoff, for the clustering. The conformation with the lowest post-minimization energy was chosen as the center of a subset. Conformations within a particular energy threshold to the center were removed from the ensemble to form the subset, then the RMSDs between the conformations in the subset and the center of the subset were computed. Any conformation with an RMSD greater than a given cutoff was returned to the original ensemble. The rest of the conformations in the subset along with its center form an IS micro cluster. The sorting and clustering process was repeated until the ensemble was empty.

We tested different combinations of energy cutoff and RMSD cutoff as shown in Table 2. Since we found that the 733 RMSD micro clusters based on 1 Å-cutoff was reasonable, similar number of IS micro clusters would be desired for comparison of these two clustering approaches. When $E\text{-cutoff} = 0.01$ kcal/mol and $\text{RMSD-cutoff} = 0.05$ Å were applied, 3024 IS micro clusters were generated, among which 2792 clusters contain no more than 0.1% (=1000 conformations) of the total population, suggesting that the cutoffs are too small. When the cutoffs are $E = 0.07$ kcal/mol and $\text{RMSD} = 0.05$ Å, we obtained 771 IS micro clusters, comparable to the number of RMSD micro clusters. We do not believe that $E = 0.07$ kcal/mol and $\text{RMSD} = 0.15$ Å which resulted in 736 IS micro clusters will lead to any appreciable difference at the macro cluster level and the following kinetics calculations from those of the 771 IS micro clusters. We notice that 95 out of 771 clusters contain only one conformation. The maximum difference in post-minimization energy between single-conformation clusters and their nearest energy minimum is 0.32 kcal/mol which is about 5 times of the E cutoff. This indicates that there are rarely visited valleys on the effective energy surface far away from other valleys and they should not be grouped with others.

Lumping

We used PCCA+^[39] implemented in the MSMBuilder2 package to lump the micro clusters into macro clusters. Lag time was set to 1 ps in the construction of transition matrix for lumping. It was shown that the number of RMSD macro clusters between 15 and 48 would satisfy Markovian property for this system.^[16] In this study, we chose 16 after testing different numbers of IS macro clusters.

Results

Clustering Validation

We ran different tests at micro- and macro-cluster level to validate the state space discretization. The validation of RMSD clustering and its following lumping was shown in our previous work.^[16] Here we show the validation of IS clusters. At the micro cluster level,

we calculated the relaxation timescale and plotted it against lag time. The relaxation timescale was calculated by

$$t_i = \frac{-\tau}{\ln \lambda_i},$$

where λ_i is the i th eigenvalue of the micro cluster transition probability matrix and τ is the lag time used to build the matrix. A series of lag time in the range between 1 ps and 150 ps was chosen to build the matrix. One criterion to determine if the dynamics of a system becomes approximately Markovian is that the relaxation timescale is lag time independent. Fig. 2 presents the relaxation timescale as a function of lag time. It shows that after 80 ps all lines are leveled off, which means that the matrices built with lag time greater or equal to 80 ps satisfy Markovian property.

For the purpose of visualization and effective description of the kinetics, the micro clusters were lumped into macro clusters. The number of macro clusters is equal to the number of states with slow relaxation timescales (in the plot) plus one.^[40] Ideally, there would be a large gap between the slow and fast relaxation timescales. In the non-ideal case such as the tetra-peptide system where no such a single strong separation could be identified, the choice of both lag time and number of macro clusters needs to be validated carefully. The Chapman-Kolmogorov test was used to assess whether the dynamics described by the Markov state model can reproduce the continuous dynamics within statistical error. If the equation of

$$[T(\tau)]^k = T(k\tau)$$

is satisfied within certain statistical uncertainty, then the discrete Markov model is consistent with the true dynamics of the system. $T(\tau)$ is the transition matrix based on macro clusters computed using lag time τ , and $T(k\tau)$ is the transition matrix computed from the same macro clusters using a longer lag time $k\tau$, where $k = 1, 2, 3, \dots$. We first fixed the lag time to 80 ps and tested the 16 and 24 macro clusters as shown in Fig. 3A, then we fixed the number of macro clusters to 16 and tested the lag time of 80 ps and 100 ps as presented in Fig. 3B. These two tests show that the Markov state model based on the IS approach is a reasonable approximation for the tested number of macro clusters and lag time. Both Fig. 3A and 3B give the same relaxation time (≈ 500 ps) as Fig. 2. The relaxation timescale is similar to what were found by the RMSD clustering method reported previously.^[16] For the lumping of RMSD clusters, Li *et al.* have shown that the dynamics of the system satisfied the Markovian property if the micro clusters are lumped into 15–48 states.^[16] To compare with the RMSD macro clusters, we chose 16 IS macro clusters because both clustering methods satisfy Markovian with this number of macro clusters. The Chapman-Kolmogorov test is at the macro cluster level under the condition that the interconversions among the micro clusters within a macro cluster are much faster than those among macro clusters. If the condition is not satisfied, other procedure for obtaining Markov state model may be used.^[41]

The method used to calculate the relaxation timescales is equivalent to the one used in Ref.^[22,42] where the authors constructed coarse master equations for peptide folding dynamics. The difference is that when constructing master equations, rate matrix ($\mathbf{K}$) is set up with rate constants as elements, while in the Markov state model, transition matrix $\mathbf{M}$ with transition probability as elements is used. These two matrices have the relation $\mathbf{M} = e^{\mathbf{K}\tau}$, where τ is the lag time. The eigenvalues of these two matrices satisfied

$$\lambda_M = \exp(\lambda_K \tau)$$

$$\lambda_K = \ln \lambda_M / \tau.$$

In Ref.,^[22] the relaxation time was defined as is $-1/\lambda_K$ which is equivalent to $-\tau/\ln \lambda_M$ in this work. When the authors validated their coarse master equations, they check whether the number of kinetic states can capture the slow dynamics by comparing the correlation functions with those obtained directly from the molecular dynamics simulations. This validation is in the same spirit as the Chapman-Kolmogorov test.

Comparisons between RMSD and IS micro clusters

Comparing to the IS clustering result, RMSD gives a relatively even distribution of cluster size (Fig 4A). The population of the largest IS cluster is 14.53% while its counterpart is only 1.18%. Among the 771 IS clusters, 24 clusters have a population greater than 10,000 conformations and 274 clusters have a population less than or equal to 10 conformations. However, among the 733 RMSD micro clusters, there are only 2 micro clusters with populations larger than 10,000 and there is not any micro cluster with populations less than 10 conformations.

Since the IS approach generated many large micro clusters (with population >10,000 conformations) (Fig. 4A), one may expect that a large IS micro cluster combines several RMSD micro clusters. If this is the case, then for a given RMSD micro cluster, all of its conformations can be assigned into a single IS micro cluster. However, we find that on average the conformations of a single RMSD micro cluster can be assigned into about 18 different IS clusters and the maximum energy difference among the IS clusters is 5.61 kcal/mol (Fig 4B and 4C). This result indicates that conformations with close RMSDs are minimized into different basins. In other words, RMSD clustering method groups conformations in different basins on effective energy surface into one cluster given that the numbers of RMSD micro cluster and IS micro cluster are comparable.

We also investigated the conformational difference within one IS micro cluster. We define the radius of an IS micro cluster as the maximum all-atom RMSD between all the conformations in the cluster and the cluster center which is the one with the lowest post-minimization energy. The IS micro clusters whose populations are greater than 100 for a total of 262 clusters (=99% of the population) were used for the calculation of cluster radii. As shown in Fig. 5, the cluster radii range from 3.5 Å to 1.2 Å with the average equal to 2.2 Å which is more than double of the cutoff of RMSD clustering. Are the radii too large so

that the IS approach groups conformations in different Ramachandran regions into one effective energy basin? We plot backbone dihedral angles (ϕ, ψ) of each residue for every conformation in the largest 262 IS micro clusters and find that the (ϕ, ψ) of every residue is localized in a single region in the Ramachandran plot (right-handed α helix, β sheet, or left-handed α helix) (see the movie in the Supporting Information), indicating that no structurally dissimilar states in the Ramachandran plot are grouped into one IS micro cluster. In conclusion, due to the completely different metrics, these two clustering methods gave totally different state discretization at micro cluster level. Conformations with close RMSDs could be minimized into different energy basins, while those with large RMSDs may belong to the same IS micro cluster. Yet, the difference does not lead to inconsistency in the calculated relaxation timescale and the result of Chapman-Kolmogorov test.

RMSD and IS Macro Clusters

At the macro cluster level, the difference in cluster population between RMSD clusters and IS clusters is narrowed with respect to that at the micro cluster level (Fig. 4A and Fig. 6A). Although the largest IS macro cluster is still bigger than its counterpart, the overall population distributions resulted from these two methods tend to be comparable. To investigate the IS and RMSD cluster overlap at the macro cluster level, we compared the IS and RMSD macro clusters whose population is larger than 10,000 conformations; IS macro clusters 1 to 4 (covering 98.1% of the total population) and RMSD macro clusters 1 to 6 (= 98.8% of the total population). Fig. 6B illustrates the mapping of RMSD macro clusters to the IS macro clusters. We find that 97.87%, 85.77%, 95.08%, 99.47% of the conformations of RMSD macro cluster 1, 2, 4, and 5, respectively, constitute 99.6% of the population of IS macro cluster 1. 95.3% population of IS macro cluster 2 is consisted of 95.4% of RMSD macro cluster 3, 41.17% of RMSD macro cluster 6 and 1.34% of RMSD macro cluster 1. Because RMSD macro cluster 6 as well as IS macro cluster 3 and 4 has no predominant one-to-one mapping with their counterparts, only IS macro cluster 1 and 2 as well as RMSD macro cluster 1, 2, 3, 4 and 5 are used in further analyses.

We use sets to represent RMSD macro clusters and the relationships among sets to represent the corresponding relationships among macro clusters. To describe the main conformational feature of each macro cluster, we use a 4-element string to denote the conformation of each cluster. The elements of the string represent the conformation of the four residues of the tetra peptide. The α conformation is defined as -150° ϕ -50° and -75° ψ 0° , and the β conformation is -170° ϕ -60° and 60° ψ 160° . For example, RMSD macro cluster 1 is located in the β -sheet region in the Ramachandran plot (Fig. S2), thus a string of ($\beta, \beta, \beta, \beta$) is given to represent this cluster. If within a single macro cluster a particular residue occupies two regions in the Ramachandran plot, then a percentage is given to show the relative population. As summarized in Table 3, the conformations in RMSD macro cluster 1 adopt β conformation; in RMSD macro cluster 2 only the second residue adopts pure β conformation while the rest of the residues adopt either α or β conformation; RMSD macro cluster 3, 4, and 5 all have a single residue in the α region and the rest are (mainly) in the β region. If the RMSD macro clusters are considered as sets and the intersection of two sets denotes that the clusters have conformations in the common region of Ramachandran plot, then the relationships among the RMSD macro clusters can be represented using a Venn

diagram (Fig. 7): RMSD macro cluster 1 $\cap$ RMSD macro cluster 2; RMSD macro cluster 2 $\cap$ RMSD macro cluster 4; RMSD macro cluster 2 $\cap$ RMSD macro cluster 5; RMSD macro cluster 3 is disjoint with others.

Discussion

As shown above the IS and RMSD clustering methods generate different micro and macro clusters. How does the difference at the micro cluster level lead to the difference at the macro cluster level? To answer this question, we analyzed how an IS micro cluster can be mapped onto RMSD macro clusters. We picked IS micro clusters for analysis according to the following three criteria: 1) cluster size > 100 conformations; 2) 90% of its populations belong to the RMSD macro cluster 1–5; 3) at least 10% of its population belong to RMSD macro cluster 3, or 4, or 5, then these IS micro clusters are defined to contain RMSD macro cluster 3, 4, and 5-component, respectively. As shown in Fig. 8, none of the IS micro clusters contain pure or nearly pure (> 90% population) RMSD macro cluster 4/5 and these two macro clusters exclude each other, in other words, if an IS micro cluster has an RMSD macro cluster 4-component, then it must not have any RMSD macro cluster 5-component, and vice versa. In addition, RMSD macro cluster 4/5-component complements that of RMSD macro cluster 2. Namely, if a IS micro cluster contains conformations from RMSD macro cluster 4/5, it must have a significant component from RMSD macro cluster 2. We also found that RMSD macro cluster 4/5 component seldom coexists with RMSD macro cluster 3 component in the selected IS micro clusters.

We used a string to represent the average composition (weighted by the cluster-population) of IS micro clusters. The elements in the string represent the percentage of the IS micro cluster population that belong to RMSD macro cluster 1–5. For the IS micro clusters that contain RMSD macro cluster 4-component, the composition string is (5.37, 31.41, 0.06, 62.89, 0), indicating that on average 62.89% of the populations of the selected IS micro clusters belong to RMSD macro cluster 4 and 31.41% of the populations belong to RMSD macro cluster 2. The string representation for the IS micro clusters containing RMSD macro cluster 5-component is (4.53, 28.60, 0.23, 0, 66.61), representing on average 66.61% of the populations can be assigned into RMSD macro cluster 5 while 28.60% of the populations can be mapped to RMSD macro cluster 2. Fig. 8 and the composition strings show that the IS approach groups conformations belonging to RMSD macro cluster 4/5 with those from RMSD macro cluster 2 at the IS micro cluster level.

To understand the reason why the IS approach does not mix RMSD macro cluster 1/2 with 3, we performed the same composition analysis. Out of the 42 IS micro clusters that contain RMSD macro cluster 3-component, 25 clusters have pure or nearly pure (> 90% of population) RMSD macro cluster 3-component. The weighted average composition string is (7.79, 1.32, 90.15, 0.01, 0.30), showing that on average 90.15% of the populations can be assigned into RMSD macro cluster 3 and 7.79% of its populations belongs to RMSD macro cluster 1. This result shows that the IS approach mixes conformations belonging to RMSD macro cluster 3 with those from RMSD macro cluster 1 to a low extent at the micro cluster level (see Fig. S3 for details).

We also calculated the mean first passage time (MFPT) from RMSD macro cluster 3, 4, 5 to RMSD macro cluster 1 or 2 using the MSMBuilder2 package. The transition matrix was built using lag time equal to 80 ps. We first set RMSD macro cluster 1 as sink and RMSD macro cluster 2, 3, 4, and 5 as sources to obtain the MFPT and then replaced the sink with RMSD macro cluster 2 and recalculated MFPT from RMSD macro cluster 1, 3, 4, and 5 to the sink. As shown in Fig. 9, the MFPTs from RMSD macro cluster 3 to cluster 1/2 are longer than others to the corresponding sink. Thus, from kinetics point of view it is reasonable that the IS approach lumps RMSD macro cluster 1, 2, 4, and 5 while separates them from RMSD macro cluster 3.

In summary, we investigated how the applications of two different clustering approaches, inherent structure and geometry-based clustering, influence our understanding of the thermodynamics and kinetics of tetra alanine peptide. We have found that the two clustering methods gave distinct clustering results at micro cluster level. Some conformations with close RMSDs may be quenched into different ISs, while other conformations with large RMSDs could be minimized into the same IS. However, the two methods give similar relaxation timescale which is based on the state space discretization at the micro cluster level. This result is in line with the finding in literature that Markov state models were robust when various combinations of metrics and clustering methods were applied.^[43] When the micro clusters were lumped into macro clusters, the state space discretization is still different by different methods. PCCA+ is a lumping procedure based on kinetic relations between the micro clusters. Thus, some RMSD macro clusters contain conformations located in different regions in Ramachandran plot with fast interconversions. The largest IS macro cluster merges several large RMSD macro clusters. The discrepancies at the macro cluster level originate from the differences in the micro cluster level, but they can be rationalized from the kinetics and structural point of view. Although different state space discretization at macro cluster level may all satisfy Chapman-Kolmogorov test, it is worthwhile to explore the meaningful and relevant state space discretization for an individual system of interest.

Even though RMSD is not a good criterion to assign kinetic states, it has the advantage of capturing large conformational changes as in protein folding.^[44,45] The optimal application for IS is for the energy landscape that has a few deep basins separated by high barriers. Our tetrapeptide system is representative of proteins whose energy landscape characterized by many minima separated by relatively low barriers. It was suggested that for proteins the IS approach is ideal for more restricted regions, such as the vicinity of native state.^[21] However, as shown in our test case, IS gives meaningful macro clusters for conformational changes in a wide range.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work is funded by National Institutes of Health (R01-GM088326).

References

1. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. Science. 2011; 334:517. [PubMed: 22034434]
2. Voelz VA, Bowman GR, Beauchamp K, Pande VS. J. Am. Chem. Soc. 2010; 132:1526. [PubMed: 20070076]
3. Fan J, Duan M, Li DW, Wu H, Yang H, Han L, Huo S. Biophys. J. 2011; 100:2457. [PubMed: 21575580]
4. Krivov SV, Karplus M. Proc. Natl. Acad. Sci. U. S. A. 2004; 101:14766. [PubMed: 15466711]
5. Evans DA, Wales DJ. J. Chem. Phys. 2003; 118:3891.
6. Rao F, Caflisch A. J. Mol. Biol. 2004; 342:299. [PubMed: 15313625]
7. Andrec M, Felts AK, Gallicchio E, Levy RM. Proc. Natl. Acad. Sci. U.S. A. 2005; 102:6801. [PubMed: 15800044]
8. Hubner IA, Deeds EJ, Shakhnovich EI. Proc. Natl. Acad. Sci. U.S.A. 2006; 103:17747. [PubMed: 17095606]
9. Bowman, G.; Pande, V.; Noe, F.; GR, B.; VS, P.; F, N., editors. New York London: Springer; 2014.
10. Jolliffe, IT. Principal Components Analysis. New York: Springer; 1986.
11. Ferguson AL, Panagiotopoulos AZ, Debenedetti PG, Kevrekidis IG. Proc. Natl. Acad. Sci. U.S.A. 2010; 107:13597. [PubMed: 20643962]
12. Coifman RR, Lafon S, Lee AB, Maggioni M, Nadler B, Warner F, Zucker SW. Proc. Natl. Acad. Sci. U.S.A. 2005; 102:7426. [PubMed: 15899970]
13. Tenenbaum JB, de Silva V, Langford JC. Science. 2000; 290:2319. [PubMed: 11125149]
14. Das P, Moll M, Stamati H, Kavraki LE, Clementi C. Proc. Natl. Acad. Sci. U.S.A. 2006; 103:9885. [PubMed: 16785435]
15. Duan M, Fan J, Li M, Han L, Huo S. J. Chem. Theory Comput. 2013; 9:2490. [PubMed: 23772182]
16. Li M, Duan M, Fan J, Han L, Huo S. J. Chem. Phys. 2013; 139:185101. [PubMed: 24320303]
17. Duan M, Liu H, Li M, Huo S. J. Chem. Phys. 2015; 143:135101. [PubMed: 26450332]
18. Malmstrom RD, Lee CT, Van Wart A, Amaro RE. J. Chem. Theory Comput. 2014; 10:2648. [PubMed: 25473382]
19. Kellogg EH, Lange OF, Baker D. J. Phys. Chem. B. 2012; 116:11405. [PubMed: 22958200]
20. McGibbon RT, Pande VS. J. Chem. Theory Comput. 2013; 9:2900. [PubMed: 26583974]
21. Rao F, Karplus M. Proc. Natl. Acad. Sci. U.S.A. 2010; 107:9152. [PubMed: 20435910]
22. Buchete N-V, Hummer G. J. Phys. Chem. B. 2008; 112:6057. [PubMed: 18232681]
23. Stillinger FH, Weber TA. Phys. Rev. A At. Mol. Opt. Phys. 1983; 28:2408.
24. Stillinger FH, Weber TA. Science. 1984; 225:983. [PubMed: 17783020]
25. Sangha AK, Keyes T. J. Phys. Chem. B. 2010; 114:16908. [PubMed: 21114309]
26. Kim J, Straub JE, Keyes T. Phys. Rev. Lett. 2006; 97:050601. [PubMed: 17026089]
27. Nakagawa N, Peyrard M. Proc. Natl. Acad. Sci. U.S.A. 2006; 103:5279. [PubMed: 16567628]
28. Kim J, Keyes T. J. Phys. Chem. B. 2007; 111:2647. [PubMed: 17311447]
29. Brooks BR, Brooks CL 3rd, Mackerell AD Jr, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M. J. Comput. Chem. 2009; 30:1545. [PubMed: 19444816]
30. Neria E, Fischer S, Karplus M. J. Chem. Phys. 1996; 105:1902.
31. Schaefer M, Bartels C, Leclerc F, Karplus M. J. Comput. Chem. 2001; 22:1857. [PubMed: 12116417]
32. Hummer G, Kevrekidis IG. J. Chem. Phys. 2003; 118:10762.
33. Chodera JD, Singhal N, Pande VS, Dill KA, Swope WC. J. Chem. Phys. 2007; 126:155101. [PubMed: 17461665]
34. Huo S, Straub JE. J. Chem. Phys. 1997; 107:5000.

35. Krivov SV, Chekmarev SF, Karplus M. Phys. Rev. Lett. 2002; 88:038101. [PubMed: 11801090]
36. Czerminski R, Elber R. J. Chem. Phys. 1990; 92:5580.
37. Beauchamp KA, Bowman GR, Lane TJ, Maibaum L, Haque IS, Pande VS. J. Chem. Theory Comput. 2011; 7:3412. [PubMed: 22125474]
38. Rao F. J. Comput. Chem. 2011; 32:1113. [PubMed: 21387337]
39. Deuffhard P, Weber M. Linear Algebra Appl. 2005; 398:161.
40. Bowman GR, Huang X, Pande VS. Methods. 2009; 49:197. [PubMed: 19410002]
41. Hummer G, Szabo A. J. Phys. Chem. B. 2015; 119:9029. [PubMed: 25296279]
42. Buchner GS, Murphy RD, Buchete NV, Kubelka J. Biochim. Biophys. Acta. 2011; 1814:1001. [PubMed: 20883829]
43. Prinz JH, Wu H, Sarich M, Keller B, Senne M, Held M, Chodera JD, Schutte C, Noe F. J. Chem. Phys. 2011; 134:174105. [PubMed: 21548671]
44. Rao F, Caflisch A. J. Mol. Biol. 2004; 342:299. [PubMed: 15313625]
45. Rao F, Caflisch A. J. Chem. Phys. 2003; 119:4035.
46. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Genome Res. 2003; 13:2498. [PubMed: 14597658]

Figure 1.
Distribution of RMSD between every pair of adjacent conformations along the 1- μ s trajectory.

Figure 2. Relaxation timescale as a function of lag time computed from micro cluster transition matrices for the IS approach.

Figure 3.

Chapman-Kolmogorov test of the Markov state model built with the IS approach. A) When the lag time is fixed to 80 ps, the 16 and 24 macro clusters are tested. B) When the number of macro clusters is fixed to 16 and the lag time of 80 ps and 100 ps is tested. The error bars give the statistical uncertainty as defined in Ref.^[9]

Figure 4. Comparisons between RMSD and IS micro clusters. (A) Micro cluster population. (B) Number of IS micro clusters to which conformations in a given RMSD micro cluster belong. (C) Maximum IS energy difference within a given RMSD micro cluster. Clusters are sorted from the largest size to the smallest size.

Figure 5.
RMSD radius of the largest 262 IS micro clusters.

Figure 6. Comparisons between RMSD and IS Macro clusters. (A) Macro cluster population. (B) Mapping of RMSD macro clusters to the IS macro clusters. If more than 95% conformations of an RMSD macro cluster can be mapped into a single IS macro cluster, then a single arrow is shown. Otherwise, arrows are shown how an RMSD macro cluster is mapped onto different IS macro clusters. The arrows corresponding to < 5% mapping are not shown.

Figure 7.

Relationships among the large RMSD macro clusters. RMSD macro cluster 1 $\cap$ RMSD macro cluster 2; RMSD macro cluster 2 $\cap$ RMSD macro cluster 4; RMSD macro cluster 2 $\cap$ RMSD macro cluster 5; RMSD macro cluster 3 is disjoint with others. The size of circle represents the population of cluster in a linear scale. The populations of RMSD macro cluster 1 to 5 are: 57.7 %, 20.9 %, 9.2%, 5.7%, and 4.1 %.

Figure 8.

Composition of IS micro clusters that have RMSD macro cluster 4-component (A) and RMSD macro cluster 5-component (B). The IS micro clusters are in descending order of cluster size and the magnitude of populations are shown on the top. Color Scheme: Violet=RMSD macro cluster 1, Cyan=RMSD macro cluster 2, Black=RMSD macro cluster 3, Green=RMSD macro cluster 4, Red=RMSD macro cluster 5.

Figure 9.

Mean first passage time (MFPT) from RMSD macro cluster 2, 3, 4, and 5 to RMSD macro cluster 1 and from RMSD macro cluster 1, 3, 4, and 5 to RMSD macro cluster 2 in units of ps. MFPT(2→1) = 190.86 ps and MFPT(1→2) = 479.79 ps. The size of circle represents the population of cluster in a linear scale. The populations of RMSD macro cluster 1 to 5 are: 57.7 %, 20.9 %, 9.2%, 5.7%, and 4.1 %. The graph is drawn using Cytoscape.^[46]

Table 1

Clustering results of different RMSD cutoffs.

RMSD cutoff (Å)	Number of clusters	Population of the largest cluster	% of clusters with populations <0.1%
0.56	5594	1446	99.8
0.8	1777	4552	83.7
1.0	733	11793	61.0

Table 2

Effective energy threshold and all-atom RMSD cutoff in IS calculations

E cutoff (kcal/mol)	RMSD cutoff (Å)	Number of Clusters	Number of clusters with size 10	Number of clusters with size = 1	Size of the largest cluster
0.01	0.05	3024	1146	313	21771
0.02	0.05	1965	720	198	43013
0.03	0.05	1534	535	152	63956
0.03	0.15	1477	481	133	63956
0.035	0.05	1301	478	141	74573
0.04	0.05	1267	468	136	85611
0.05	0.05	1215	445	134	106507
0.05	0.07	1198	431	126	106507
0.06	0.05	1065	379	117	127749
0.06	0.15	1023	339	100	127749
0.065	0.05	927	318	104	137830
0.07	0.05	771	274	95	145349
0.07	0.15	736	243	80	145349

Table 3

Conformational features of large macro clusters (population >10,000 conformations)

Macro Cluster [*]	Residue 1	Residue 2	Residue 3	Residue 4
RMSD 1	β	β	β	β
RMSD 2	α/β (70%)	β	α/β (80%)	α/β (36%)
RMSD 3	β	α	β	α/β (83%)
RMSD 4	α	β	β	β
RMSD 5	β	β	α	β

^{*} Note that IS macro cluster 1 combines RMSD macro cluster 1, 4, 5 and 85.8% RMSD macro cluster 2. IS macro cluster 2 is almost identical to RMSD macro cluster 3.