JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Distribution of Reciprocal of Interatomic Distances: A Fast Structural Metric

Ting Zhou and Amedeo Caflisch*

Department of Biochemistry, University of Zurich, CH-8057 Zurich, Switzerland

Supporting Information

ABSTRACT: We present a structural metric based on the Distribution of Reciprocal of Interatomic Distances (DRID) for evaluating geometrical similarity between two conformations of a molecule. A molecular conformation is described by a vector of $3N$ orientation-independent DRID descriptors where N is the number of molecular centroids, for example, the non-hydrogen atoms in all nonsymmetric groups of a peptide. For two real-world applications (pairwise comparison of snapshots from an explicit solvent simulation of a protease/peptide substrate complex and implicit solvent simulations of reversible folding of a 20-residue β-sheet peptide), the DRID-based metric is shown to be about 5 and 15 times faster than coordinate root-mean-square deviation (cRMSD) and distance root-mean-square deviation (dRMSD), respectively. A single core of a mainstream processor can perform about 10^8 DRID comparisons in one-half a minute. Importantly, the DRID metric has closer similarity to kinetic distance than does either cRMSD or dRMSD. Therefore, DRID is suitable for clustering molecular dynamics trajectories for kinetic analysis, for example, by Markov state models. Moreover, conformational space networks and free energy profiles derived by DRID-based clustering preserve the kinetic information.

INTRODUCTION

Molecular dynamics (MD) simulations generate information of protein motion at atomic level of detail and femtosecond time resolution. Clustering of snapshots saved along MD trajectories is almost always required for structural analysis and/or to extract thermodynamics and kinetics. Usually, the snapshots are clustered first according to geometric similarity, and the clusters are then grouped into physically relevant states using kinetics-preserving procedures, for example, the cut-based free energy profile (cFEP) method[1,2] or spectral clustering methods.[3−5] Most often than not, the performance of the structural metric will fundamentally affect the efficacy and efficiency of the geometric and subsequent kinetic clustering.[6]

Coordinate root-mean-square deviation (cRMSD) is a natural metric for quantifying the similarity of two conformations.[7,8] The cRMSD compares the Cartesian coordinates of corresponding atomic nuclei in two conformations of a molecule ($[[(1/N)\sum_{i=1}^{N}((x_i - \tilde{x}_i)^2 + (y_i - \tilde{y}_i)^2 + (z_i - \tilde{z}_i)^2)]]^{1/2}$, where N is the number of the atoms). However, it does not explicitly consider the distances between pairs of atoms in the same conformation. Thus, cRMSD does not necessarily correlate with energy, which is a function of interatomic distances.[9] Another disadvantage of cRMSD is that the result relies on the optimal superposition of the conformations being compared, which is more time-consuming than the calculation of the root-mean-square deviation itself. To speed up the process of superposition, Theobald recently developed the quaternion-based characteristic polynomial (QCP) method,[10] which is 30−70 times faster than the eigen decomposition methods.[11−14] Another disadvantage of cRMSD is that the value of the coordinates depends on the orientation of the molecule. Therefore, it is inefficient to use Cartesian coordinates in conformational searching across a database.

Another frequently used geometry-based metric is the distance root-mean-square deviation (dRMSD). The dRMSD ($[[(2/N(N-1))\sum_{i=1}^{N}\sum_{j=1}^{i-1}(d_{ij} - \tilde{d}_{ij})^2]]^{1/2}$, where d_{ij} is the distance between atom i and atom j) and its variants[9,15−17] do not need structural alignment, and show higher correlation with energy[9] than cRMSD because dRMSD is a function of interatomic distances. The original dRMSD strongly weighs large values of atomic distance. To reduce the overweight, in the contact map method only the distances smaller than a threshold are compared.[16] Hole and Sander have proposed the use of the percentage of change in distance to evaluate the similarity.[15] However, these variants may introduce new problems, such as discretization for the contact map and unsatisfactory triangular inequality for Hole and Sander's method.[9] Moreover, dRMSD uses a larger amount of floating numbers than do Cartesian coordinates (a total of $N(N-1)/2$ and $3N$, respectively, where N is the number of atoms used for the comparison). In the majority of clustering algorithms, the geometrical similarity between each new conformation and the representative of individual clusters needs to be calculated. For clustering a complex system whose kinetic network[18] has a large amount of clusters, dRMSD requires that the clustering program uses a larger amount of memory for storing the representatives of the clusters (in terms of distance matrices) than cRMSD (in terms of Cartesian coordinates), or to recalculate distance matrices from the Cartesian coordinates when needed. Overall, both cRMSD and dRMSD have their shortcomings. Evaluating the similarity among conformations is often the bottleneck of performance in coarse-graining of MD snapshots.

Received: April 18, 2012
Published: July 9, 2012

dx.doi.org/10.1021/ct3003145 | J. Chem. Theory Comput. 2012, 8, 2930−2937

In this Article, we propose an efficient method for measuring the geometrical similarity between two conformations of a macromolecule based on a one-dimensional array (vector) of moments of Distribution of Reciprocal of Interatomic distances (DRID). Each conformation is encoded in an orientation-independent DRID vector, which has the same length as the vector of Cartesian coordinates. The difference between two conformations can be efficiently evaluated as the root-mean-square deviation between the two DRID vectors. Because DRID does not need superposition of two conformations, it is about 5 times faster than the cRMSD version with the fast QCP method. Because a shorter vector is used to represent a conformation, DRID is about 15 times faster than dRMSD. Furthermore, DRID uses an amount of memory similar to that of cRMSD, which is much smaller than dRMSD for large molecules. Moreover, the geometric distance measured by DRID is a better estimate of kinetic distance than either cRMSD or dRMSD, which indicates that DRID is an effective structural metric for coarse-graining snapshots of MD trajectories into geometrically similar clusters for kinetic analysis.

METHODS

One important difference between DRID and previously reported metrics is the use of the multiplicative inverse (reciprocal) of the distance rather than the distance itself. If two atoms are far away from each other, that is, separated by several covalent bonds, their distance is strongly affected by thermal motion because of the leverage effect (Figure 1). In contrast, such thermal "noise" does not influence the very similar interaction energy of pairs of distant atoms (which is close to zero). To reduce but not completely neglect the effect of large separations, reciprocals of distances are used to encode the DRID descriptor vector. Note that variations of the covalent bond lengths have little effect on the conformational changes, but the fluctuations of their reciprocal would have a large influence. Thus, the distances between pairs of atoms separated by a single covalent bond are excluded from DRID calculation. Two "sets" of atoms have to be defined for DRID evaluation: a set of n centroids and a set of N atoms for distance evaluation. Note that these two sets can be identical as in both examples below. The centroids are the (sub)set of atoms in nonsymmetric groups. In the process of measuring the similarity of topologically different molecules,[19] the selection of centroids is based on values of discontinuous maximum/minimum functions so that the centroids can "hop" from one atom to the other in two similar conformations.[20] The alteration of centroids will lead to significantly different values in the elements of the vector for two similar conformations. A fixed set of atoms are chosen as centroids in DRID, which is thus not affected by the centroid-hopping problem. The atoms in symmetric groups (in proteins, e.g., hydrogen atoms in methyl groups, the ortho-/meta- carbon atoms of the benzyl and phenolic ring in Phe and Tyr side chains, respectively, as well as the oxygen atoms in carboxyl groups) are neglected because they introduce more noise than signal. An illustrative example of DRID evaluation is shown in Figure 2. Three moments of the distribution of reciprocal distances are employed to characterize the essential features of atomic-distance sets based on individual centroids. The first centroid-based descriptor is the mean of the reciprocal of the distances $\mu_i \equiv [\sum_j(1/d_{ij})]/(N - 1 - nb_i)$, where d_{ij} is the distance of the atom j from centroid atom i, N is the number of atoms, and nb_i is the number of atoms bonded to centroid i. The centroid i and the atoms covalently bound to it are excluded from the sum. The second descriptor is the square root of the second central moment of the reciprocal of the same set of atomic distances $\nu_i \equiv \{[\sum_j(1/d_{ij} - \mu_i)^2]/(N - 1 - nb_i)\}^{1/2}$. The third descriptor is the cubic root of the third central moment of the same distribution $\xi_i \equiv \{[\sum_j(1/d_{ij} - \mu_i)^3]/(N - 1 - nb_i)\}^{1/3}$.

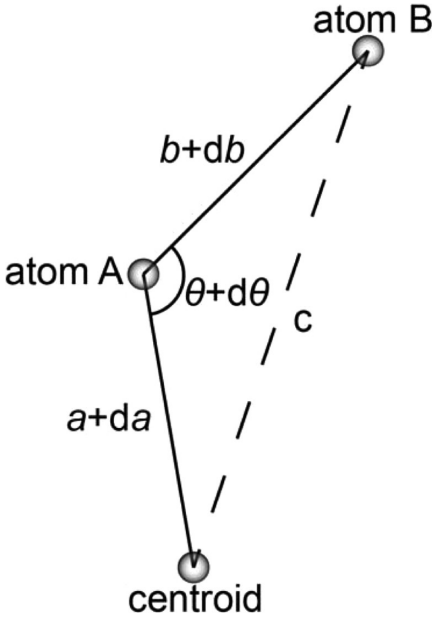

Figure 1. Leverage effect for large atomic distances. The distance between the centroid and the atom A is a + da, where da is the perturbation of the distance due to thermal motion. Accordingly, the db and dθ are the perturbation of the distance b and the angle θ, respectively. The perturbation of the distance c between the centroid and the atom B, which is two bonds away, can be calculated as $dc = [(a - b\cos\theta)da + (b - a\cos\theta)db + ab\sin\theta\, d\theta]/(a^2 + b^2 - 2ab\cos\theta)^{1/2}$. The coefficient of the last term of uncertainty of dc has the same order of magnitude as the size of the system and the distance from the centroid. Therefore, the uncertainty of the distance increases with the distance itself.

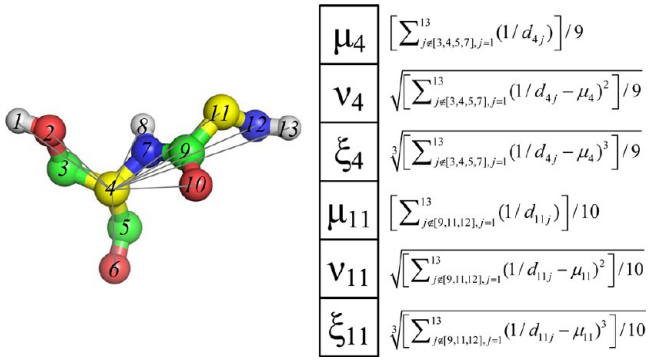

Figure 2. Schematic view of the DRID encoding process. The left panel shows the three-dimensional model of the glycine−serine dipeptide. The sequential identification of the atom is denoted on top of each atom. The α carbon atoms (4th and 11th atoms) were chosen as centroids (yellow spheres). The d_{ij} is the distance from atom i to atom j. The formulas for calculating the components of the six-dimensional DRID vectors are shown on the right panel.

dx.doi.org/10.1021/ct3003145 | J. Chem. Theory Comput. 2012, 8, 2930−2937

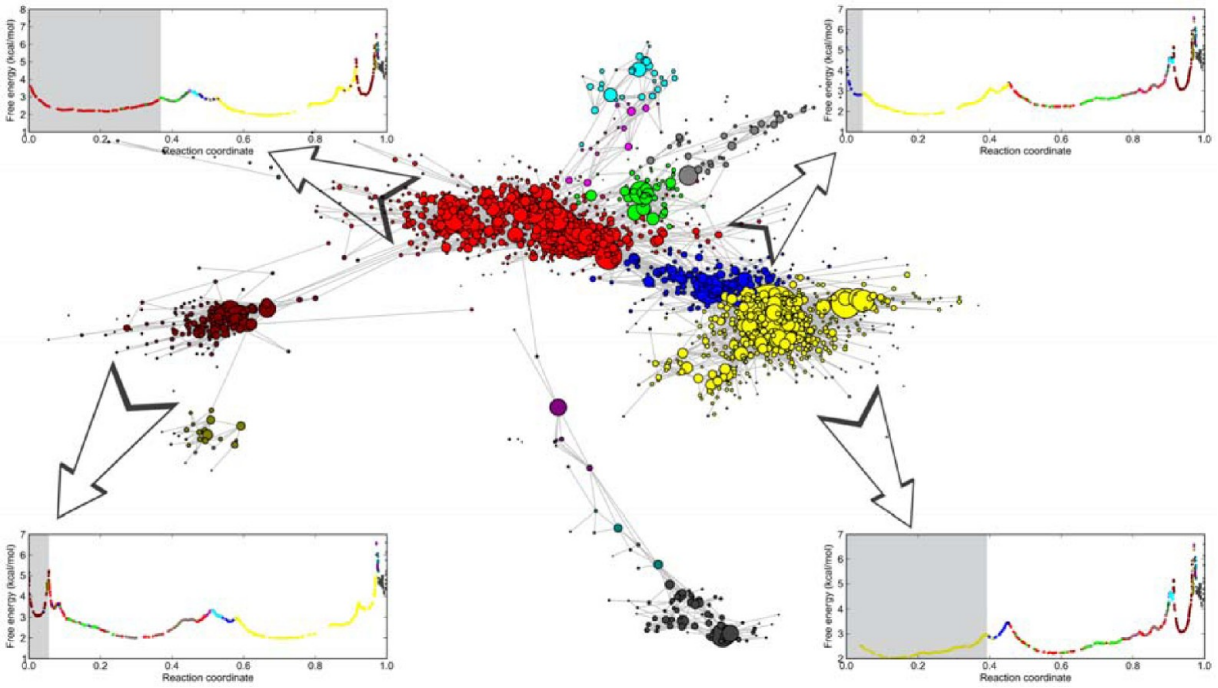

Figure 3. Network analysis of the BACE−substrate complex. The basins on the free energy landscape of the BACE−substrate complex identified by cFEPs are shown by different colors. The cFEPs plotted using the representatives of each of the four largest basins (red, blue, brown, and yellow) are shown as insets in the four corners. In each cFEP, the basin used as the reference state is marked by the gray rectangle. In the conformational space network (middle of the figure), nodes and links are conformations (i.e., clusters obtained by DRID metric) and direct transitions in the MD trajectory, respectively. The size of each node is proportional to its statistic weight. The forced directed layout of nodes is determined by the Fruchterman−Reingold algorithm.[34] The overlapped basins in the cFEP have similar kinetic distance from the reference state; for example, in the two insets on the right, red and green basins overlap because their kinetic distances from either blue or yellow basins are similar.

Thus, given a set of n atoms selected as centroids, each conformation of the (macro)molecule is described by a DRID vector of $3 \times n$ orientation-independent descriptors. The geometrical distance between two conformations of a (macro)-molecule described by (μ_i, ν_i, ξ_i) and $(\tilde{\mu}_i, \tilde{\nu}_i, \tilde{\xi}_i)$, respectively, is evaluated as the root-mean-square deviation of two DRID vectors, which is $\{(1/3n)\sum_{i=1}^{n}[(\mu_i - \tilde{\mu}_i)^2 + (\nu_i - \tilde{\nu}_i)^2 + (\xi_i - \tilde{\xi}_i)^2]\}^{1/2}$ and has units of 1/distance.

RESULTS AND DISCUSSION

To evaluate the performance of DRID, we compared it to cRMSD and dRMSD for two representative systems: (1) the explicit solvent MD simulations of the complex of the aspartic protease β-secretase (BACE) and an octapeptide substrate,[21,22] and (2) the implicit solvent[23] MD simulations of reversible folding of a 20-residue antiparallel β-sheet peptide (Beta3s).[24,25]

Explicit Solvent Simulations of the BACE−Substrate Complex. The MD trajectories (136 710 snapshots) from the recently published simulations of the complex of BACE with an octapeptide substrate[22] are used as a first test of DRID. The set of centroids included the 144 C, N, O, and S atoms in nonsymmetric groups in the BACE active site (residues 32−35, 71−73, 76, 198−199, 227−228, and 231−232) and substrate (except for Phe(P4′), which undergoes large fluctuations[21] and does not interact with BACE). The set of atoms for distance evaluation consisted of the same 144 non-hydrogen atoms. Each MD snapshot was first encoded in a 432-dimensional DRID vector. The WORDOM implementation of the leader-like clustering algorithm[26,27] based on DRID metric with a cutoff 1.8×10^{-3} Å^{-1} yielded 1603 clusters, and 50 283 transitions between them were observed in the MD sampling. These clusters and transitions are depicted as nodes and links of the conformational space network (Figure 3), respectively. The topology of this network is consistent with the one based on dRMSD clustering (Figure 4 in ref 22). Note that the DRID cutoff was chosen to obtain a number of nodes and links similar to that in the previous study. To show that the clusterings obtained by DRID and dRMSD are quantitatively comparable, the lengths of hydrogen bonds critical for catalysis are plotted along the natural reaction coordinate of the cFEP (Figure 4). The conformations in individual basin identified by cFEPs based on either DRID or dRMSD show similar distributions of hydrogen-bond lengths. As an example, for both DRID and dRMSD clustering, all snapshots in the yellow and blue basins in Figure 3 exhibit a hydrogen bond between the side chains of Ser35 and Asp32 (upper-left panel of Figure 4), while in about 20% of the snapshots of red and green basins the hydrogen bond is lost. The consistency in conformational space networks and cFEPs obtained by DRID- or dRMSD-based clustering indicates that DRID can be used for structurally clustering snapshots into mesostates for further kinetic analysis.

To test the robustness of DRID, different clustering thresholds were used for building the conformational space network and plotting the cFEP. This analysis shows that DRID has good self-consistency as the kinetic ordering of basins is robust with respect to the choice of threshold (Figure S1 in the Supporting Information).

Implicit Solvent Simulations of the Structured Peptide Beta3s. Folding of the 20-residue antiparallel β-sheet peptide Beta3s (with amino acid sequence

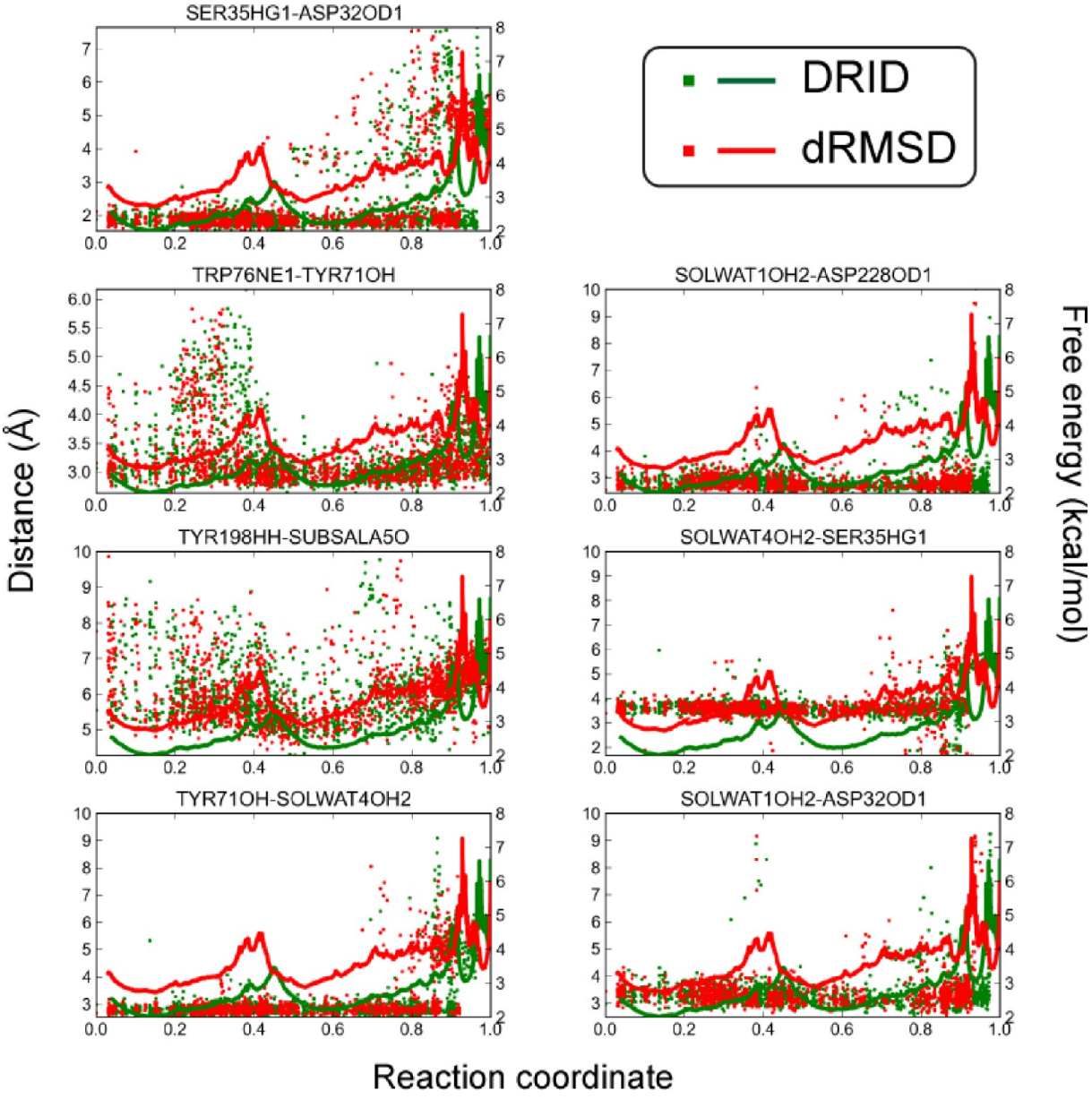

Figure 4. Structural characterization of cFEP basins of the BACE−substrate complex. The cFEP and the length of seven hydrogen bonds in the representative snapshots of individual clusters are depicted in solid lines and dots, respectively. The names of atoms, between which the distances are measured, are denoted on top of each figure; for example, the dots in the top-left figure present the distances between atoms HG1 in Ser35 and OD1 in Asp32. The reference state of the cFEP is the most populated node, which is inside the yellow basin in Figure 3. The cFEPs were determined by clustering using either the dRMSD or the DRID. In both cases, conformations with similar distributions of hydrogen-bond distance were grouped together.

TWIQNGSTKWYQNGSTKIYT in single-letter code) has been extensively investigated by MD simulations.[2,18,25,28−33] By using an implicit solvent model,[23] Beta3s folds reversibly from a heterogeneous denatured ensemble to its native structure, a three-stranded antiparallel β-sheet.[24] The free energy profile of Beta3s calculated upon clustering by DRID is compared to the profiles calculated upon clustering by cRMSD[2] or dRMSD (Figure 5). The 159 unsymmetrical heavy atoms of Beta3s (out of 215 atoms) were used for calculating cRMSD, dRMSD, and DRID. A trajectory of Beta3s, saved every 0.2 ps for 1×10^8 snapshots, was used as input for coarse-graining. The kinetic networks obtained by cRMSD clustering (with 2.5 Å cutoff) and DRID clustering (with 5.5×10^{-3} Å^{-1} cutoff) show a high agreement. The difference in the height of the barrier that separates the native basin from other parts of the free energy landscape is 0.1 kcal/mol, and the difference in the population of the native state is 0.1%. The cFEP calculated by dRMSD (with 2.0 Å cutoff) is qualitatively consistent with the cRMSD-cFEP and DRID-cFEP, but the height of the barrier separating the native basin from other parts of the free energy landscape is 0.43 kcal/mol lower. As for the BACE−substrate complex, the cFEP of Beta3s calculated upon DRID clustering is robust with respect to the choice of the cutoff in the range 0.005−0.0055 Å^{-1}, while a cutoff value of 0.006 Å^{-1} yields a similar shape of the cFEP but does not preserve the barrier to exit from the folded state (Figure S2 in the Supporting Information). As in the case of cRMSD and dRMSD, it is difficult to suggest a rule of thumb for the cutoff. For practical

dx.doi.org/10.1021/ct3003145 | J. Chem. Theory Comput. 2012, 8, 2930−2937

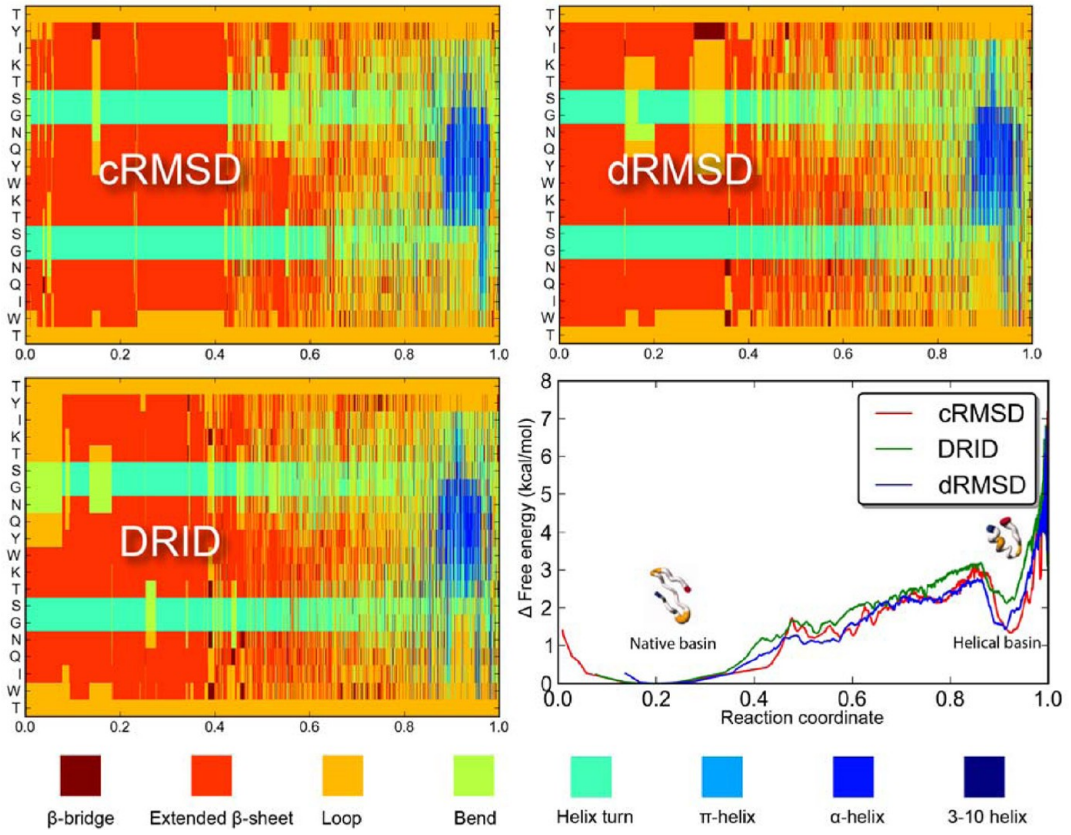

Figure 5. Comparison of cFEPs of Beta3s derived from cRMSD, dRMSD, and DRID coarse-graining. The top panels and bottom left panel show the colored DSSP[35] strings of the cluster representatives, which are arranged according to the reaction coordinate of the corresponding cFEP (bottom right panel). The *y*-axis label is the Beta3s sequence, and the legend of colors for different secondary structure elements in the traces is shown in the bottom. For the three cFEPs based on different clustering methods, the native state and the helical state are the two largest basins, which are denoted by cartoon models.

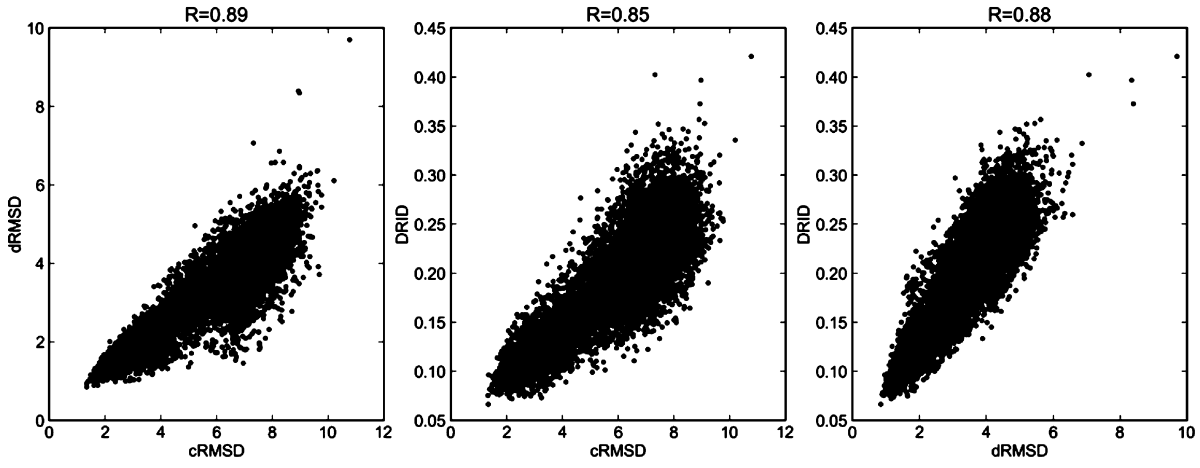

Figure 6. Correlation among geometrical distances. A set of 10^4 pairs of snapshots were randomly selected from a 20 μs trajectory of Beta3s for calculating the three geometrical distances. The Pearson correlation coefficients of pairwise metrics are denoted on top of each panel. The units of cRMSD and dRMSD are Å, while the units of DRID are $Å^{-1}$.

applications, it is necessary to cluster several times using different cutoffs, and directly compare the kinetic properties derived from the network with those derived from the simulation trajectory.[2] This "trial and error" procedure can take advantage of an efficient metric like DRID, as well as a fast clustering algorithm.[6]

Correlation among Geometrical Distances. The scatter plots of geometrical distances measured by cRMSD, dRMSD, and DRID are shown in Figure 6. The Pearson correlation coefficients among each of the three pairs of distances are larger than 0.85, which indicates these three distances are highly correlated. Pairwise distances calculated by DRID correlate slightly more with those calculated by dRMSD than cRMSD as both DRID and dRMSD use intrastructure distances, while cRMSD requires structural alignment and employs distances between pairs of atoms in different structures.

dx.doi.org/10.1021/ct3003145 | J. Chem. Theory Comput. 2012, 8, 2930−2937

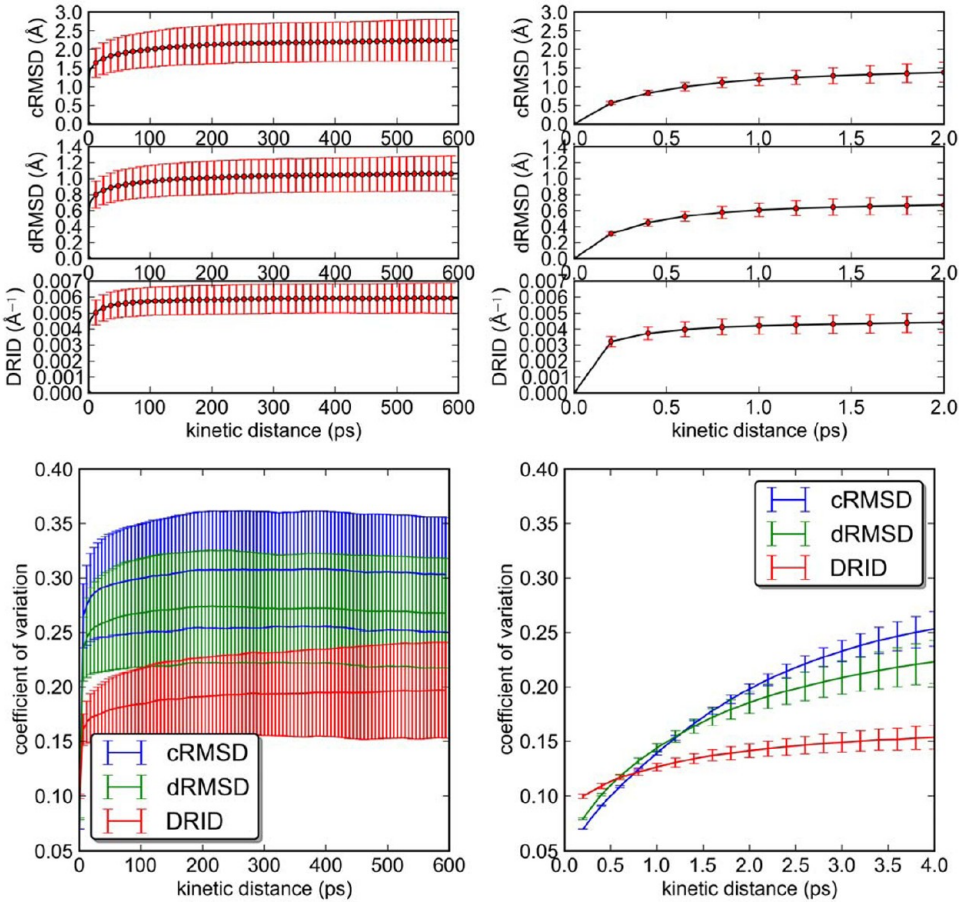

Figure 7. Relation between geometrical distance and kinetic distance for Beta3s conformations. All heavy atoms in nonsymmetric groups were used to calculate cRMSD, dRMSD, and DRID. (Top) Geometrical distance averaged over 10^5 pairs of conformations at each value of the kinetic distance. The conformations were extracted from a 20 ns trajectory, which was randomly selected from a 20 μs MD trajectory of Beta3s. Quantitatively similar results were obtained for other segments of the trajectory (data not shown). In the top-right panel, the x-axis is zoomed in at the kinetically ballistic regime (i.e., low picosecond kinetic distance between two conformations). The error bars (red) denote the standard deviations of the geometrical distances of pairs of conformations at a given kinetic distance. (Bottom) Mean value of the coefficient of variation of the geometric distance at individual kinetic distances. The error bars denote the standard deviation of the coefficient of variation across 10 different segments of the trajectory. The x-axis is zoomed in at the kinetically ballistic regime in the bottom-right panel. The smaller coefficient of variation observed for DRID (at time intervals larger than 0.8 ps) indicates that the DRID-based distance is closer to the kinetic distance than cRMSD or dRMSD metrics. Similar results were obtained for the BACE−-substrate complex (Figure S3 in the Supporting Information).

Relation between Kinetic Distance and Geometrical Distances. The kinetic distance between two MD snapshots saved along a continuous trajectory was estimated as the simulation time needed to evolve from one to the other. Geometrical distances measured by cRMSD, dRMSD, and DRID show a similar relation to kinetic distance for Beta3s (Figure 7) and the BACE−substrate complex (Figure S3 in the Supporting Information). The value of the geometric distance increases fast within about 1 ps (a time interval along which the dynamics is still in the ballistic regime), and reaches a plateau when the kinetic distance is larger than about 100 ps. Because the absolute values of the three structural metrics are different, the coefficient of variation, which is defined as the ratio of the standard deviation to the mean, is used to compare the metrics. A large coefficient of variation implies large uncertainty in the relation between the corresponding geometrical distance and the kinetic distance. For kinetic distances above 1 ps, the coefficient of variation of DRID is about 30% lower than dRMSD and 40% lower than cRMSD (Figure 7), which shows that for conformation pairs with a fixed kinetic distance, the dispersion of the DRID-based distances is smaller than in the case of the other two structural metrics.

It is interesting to evaluate how the convergence of a kinetic property, for example, the folding time, depends on the metric used for clustering. For each of the three metrics, the mean first passage time from the helical basin of Beta3s to the folded state was calculated for each of a set of conformational space networks generated by clustering simulation trajectories of increasing length. The plots of mean first passage time as a function of simulation length show similar convergence properties for the three metrics with the values obtained by DRID always closer to the converged value than the other two metrics (Figure S4 in the Supporting Information).

Timing and Memory Usage. Both dRMSD and DRID do not need to overlap the conformations; they just require the calculation of the root-mean-square deviation of two descriptor vectors to estimate the geometric difference. Therefore, before evaluating the structural similarity, for dRMSD and DRID the descriptor vector of each conformation has to be calculated. In the benchmark test cases, pairwise structural distances of 13 671 conformations of the BACE−substrate complex and

dx.doi.org/10.1021/ct3003145 | J. Chem. Theory Comput. 2012, 8, 2930−2937

10 000 conformations of the Beta3s are calculated. The aforementioned atomic subsets of both systems (144 atoms for the BACE−substrate complex, 159 atoms for Beta3s) were used in all of the three metrics. The fast QCP method,[10] which is the fastest approach in public literature, was used for calculating cRMSD. On a single core of the Intel i7-950 CPU, for both test cases, DRID is about 4.5 times faster than QCP-based cRMSD and is about 17 times faster than dRMSD (Table 1). The dRMSD needs to load more floats into memory than

Table 1. Benchmarks

	BACE−substrate complex[a]			Beta3s[b]		
	cRMSD	dRMSD	DRID	cRMSD	dRMSD	DRID
vector calculation (s)[c]	0	0.9	1.0	0	0.8	1.0
total for pairwise comparison (s)[d]	125	467	28	73	302	17
speedup of DRID	4.5	16.9	1	4.3	17.7	1

[a]The 144 heavy atoms in nonsymmetric groups of the binding site and substrate were used for all three metrics. [b]The 159 heavy atoms in nonsymmetric groups were used for all three metrics. [c]In cRMSD, the Cartesian coordinates do not need to be converted into descriptor vectors. In DRID, not only interatomic distances are calculated, but also distributions thereof; therefore, more time is needed than dRMSD for calculating descriptor vectors. [d]A total of 93 441 285 pairwise comparisons of 13 671 conformations was carried out for BACE and 49 995 000 pairwise comparisons of 10 000 conformations for Beta3s. The time needed for cRMSD is longer than DRID because of calculation of the optimal overlap only in the former. The time needed for dRMSD is longer than DRID because the descriptor vector has length $n\ (n-1)/2$ in the former and $3n$ in the latter (where $n = 144$ for BACE and $n = 159$ for Beta3s).

cRMSD and DRID and is slower than the other two metrics because both cRMSD and DRID need to compare the vector containing $3 \times n$ floats, whereas dRMSD needs to compare the vector of $n(n-1)/2$ floats, where n is the number of unsymmetrical atoms.

For dRMSD, the number of atoms involved in clustering has to be limited because a high number of atoms, as in big proteins or macromolecular assemblies, make calculation of all interatomic distances prohibitive. For DRID, an approximation similar to "non-bonded interaction cutoff" used in MD energy calculation can be applied to speed up the interatomic distance calculation: any reciprocals of the distance larger than a cutoff can be approximated to zero. In both of our test systems, which contain about 150 atoms, this cutoff is not necessary.

CONCLUSIONS

We have presented a one-dimensional array of moments of DRID, which is useful for the efficient evaluation of the geometrical similarity between pairs of conformations of a (macro)molecule or molecular complex. The DRID vector is suitable for coarse-graining MD snapshots into geometrically similar clusters for kinetic analysis because the root-mean-square deviation of two DRID vectors better reflects the kinetic distance (i.e., the temporal separation along the trajectory) than either cRMSD or dRMSD. Because DRID does not need structural alignment and encodes a conformation in a much shorter vector than dRMSD, the comparison of conformations via DRID metric is about 4.5 times faster than the fastest cRMSD method in previous literature and 17 times faster than dRMSD. The evaluation of the DRID deviation of 10^8 pairs of conformations of a peptide (or enzyme binding site) of about 150 atoms can be carried out in one-half a minute on a single core of a mainstream processor. Importantly, for large molecules, a neighbor list similar to the nonbonding list in molecular mechanics programs could be used to restrict the calculations of distance reciprocals to those pairs of atoms within a predefined cutoff value. The exclusion of distant pairs of atoms is appropriate for DRID as distance reciprocals are negligible at large separation.

The orientation-independent DRID vector can be straightforwardly implemented into a recently published tree-based algorithm,[6] which is superior to leader-like algorithm for constructing conformational space network. The DRID vector can also be efficiently used in database searching algorithms (e.g., balanced trees, hashing) for the efficient identification of conformations that are similar to a query conformation.

ASSOCIATED CONTENT

Supporting Information

Comparison of cut-based free energy profile of BACE−substrate complex coarse-grained by DRID under different thresholds (Figure S1). Comparison of cut-based free energy profile of Beta3s coarse-grained by DRID under different thresholds (Figure S2). Relation between geometrical distance and kinetic distance for conformations of BACE−substrate complex (Figure S3). Relation between the simulation time and mean first passage time from the helical to the native basin of Beta3s (Figure S4). This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Phone: (+41 44) 635 55 21. Fax: (+41 44) 635 68 62. E-mail: caflisch@bioc.uzh.ch.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Dr. Sabyashachi Mishra for sharing the MD simulation data of BACE−substrate complex. This work was supported by a grant of the Swiss National Science Foundation to A.C.

ABBREVIATIONS

DRID, Distribution of Reciprocal of Interatomic Distances; cRMSD, coordinate root-mean-square deviation; dRMSD, distance root-mean-square deviation; MD, molecular dynamics; cFEP, cut-based free energy profile; QCP, quaternion-based characteristic polynomial; BACE, aspartic protease β-secretase

REFERENCES

(1) Krivov, S. V.; Karplus, M. *J. Phys. Chem. B* **2006**, *110*, 12689.
(2) Krivov, S. V.; Muff, S.; Caflisch, A.; Karplus, M. *J. Phys. Chem. B* **2008**, *112*, 8701.
(3) Schutte, C.; Fischer, A.; Huisinga, W.; Deuflhard, P. *J. Comput. Phys.* **1999**, *151*, 146.
(4) Deuflhard, P.; Huisinga, W.; Fischer, A.; Schutte, C. *Linear Algebra Appl.* **2000**, *315*, 39.
(5) Deuflhard, P.; Weber, M. *Linear Algebra Appl.* **2005**, *398*, 161.
(6) Vitalis, A.; Caflisch, A. *J. Chem. Theory Comput.* **2012**, *8*, 1108.
(7) Unger, R.; Harel, D.; Wherland, S.; Sussman, J. L. *Proteins* **1989**, *5*, 355.

(8) Gordon, H. L.; Somorjai, R. L. *Proteins: Struct., Funct., Genet.* **1992**, *14*, 249.
(9) Wallin, S.; Farwer, J.; Bastolla, U. *Proteins: Struct., Funct., Genet.* **2003**, *50*, 144.
(10) Theobald, D. L. *Acta Crystallogr., Sect. A* **2005**, *61*, 478.
(11) Kabsch, W. *Acta Crystallogr., Sect. A* **1978**, *34*, 827.
(12) Horn, B. K. P. *J. Opt. Soc. Am. A* **1987**, *4*, 629.
(13) Diamond, R. *Acta Crystallogr., Sect. A* **1988**, *44*, 211.
(14) Kearsley, S. K. *Acta Crystallogr., Sect. A* **1989**, *45*, 208.
(15) Holm, L.; Sander, C. *J. Mol. Biol.* **1993**, *233*, 123.
(16) Vendruscolo, M.; Subramanian, B.; Kanter, I.; Domany, E.; Lebowitz, J. *Phys. Rev. E* **1999**, *59*, 977.
(17) Koehl, P. *Curr. Opin. Struct. Biol.* **2001**, *11*, 348.
(18) Rao, F.; Caflisch, A. *J. Mol. Biol.* **2004**, *342*, 299.
(19) Ballester, P. J.; Richards, W. G. *J. Comput. Chem.* **2007**, *28*, 1711.
(20) Zhou, T.; Lafleur, K.; Caflisch, A. *J. Mol. Graphics Modell.* **2010**, *29*, 443.
(21) Hong, L.; Koelsch, G.; Lin, X. L.; Wu, S. L.; Terzyan, S.; Ghosh, A. K.; Zhang, X. C.; Tang, J. *Science* **2000**, *290*, 150.
(22) Mishra, S.; Caflisch, A. *Biochemistry* **2011**, *50*, 9328.
(23) Ferrara, P.; Apostolakis, J.; Caflisch, A. *Proteins: Struct., Funct., Genet.* **2002**, *46*, 24.
(24) De Alba, E.; Santoro, J.; Rico, M.; Jimenez, M. A. *Protein Sci.* **1999**, *8*, 854.
(25) Ferrara, P.; Caflisch, A. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 10780.
(26) Seeber, M.; Cecchini, M.; Rao, F.; Settanni, G.; Caflisch, A. *Bioinformatics* **2007**, *23*, 2625.
(27) Seeber, M.; Felline, A.; Raimondi, F.; Muff, S.; Friedman, R.; Rao, F.; Caflisch, A.; Fanelli, F. *J. Comput. Chem.* **2011**, *32*, 1183.
(28) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 13841.
(29) Muff, S.; Caflisch, A. *Proteins* **2008**, *70*, 1185.
(30) Muff, S.; Caflisch, A. *J. Phys. Chem. B* **2009**, *113*, 3218.
(31) Muff, S.; Caflisch, A. *J. Chem. Phys.* **2009**, *130*, 125104.
(32) Qi, B.; Muff, S.; Caflisch, A.; Dinner, A. R. *J. Phys. Chem. B* **2010**, *114*, 6979.
(33) Zhou, T.; Caflisch, A. *J. Chem. Theory Comput.* **2012**, *8*, 2134.
(34) Fruchterman, T. M. J.; Reingold, E. M. *Software Pract. Exp.* **1991**, *21*, 1129.
(35) Anderson, C. A. F.; Palmer, A. G.; Brunak, S.; Rost, B. *Structure* **2002**, *10*, 175.