

SCUD: Fast Structure Clustering of Decoys Using Reference State to Remove Overall Rotation

HONGZHI LI,¹ YAOQI ZHOU^{1,2}

¹ Department of Physiology & Biophysics, Howard Hughes Medical Institute Center for Single Molecule Biophysics, State University of New York at Buffalo, 124 Sherman Hall, Buffalo, NY 14214

² Department of Macromolecular Science, The Key Laboratory of Molecular Engineering of Polymers, Fudan University, People's Republic of China

Received 25 January 2005; Accepted 27 March 2005

DOI 10.1002/jcc.20251

Published online in Wiley InterScience (www.interscience.wiley.com).

Abstract: We developed a method for fast decoy clustering by using reference root-mean-squared distance (rRMSD) rather than commonly used pairwise RMSD (pRMSD) values. For 41 proteins with 2000 decoys each, the computing efficiency increases nine times without a significant change in the accuracy of near-native selections. Tests on additional protein decoys based on different reference conformations confirmed this result. Further analysis indicates that the pRMSD and rRMSD values are highly correlated (with an average correlation coefficient of 0.82) and the clusters obtained from pRMSD and rRMSD values are highly similar (the representative structures of the top five largest clusters from the two methods are 74% identical). SCUD (Structure Clustering of Decoys) with an automatic cutoff value is available at <http://theory.med.buffalo.edu>.

© 2005 Wiley Periodicals, Inc. J Comput Chem 26: 1189–1192, 2005

Key words: decoy clustering; near-native detection; root-mean-squared distance

Introduction

The current state-of-the-art methods for *ab initio* protein–structure prediction rely on energy-guided generation of a large number of decoys and clustering of the decoys for near-native identification.^{1,2} Clustering simplifies the data analysis by reducing the number of decoys resulted from the large-scale conformational search. More importantly, it provides the population of the structures in the conformational space sampled by a given energy function. Near-native structures are often assumed to be the structures with the highest populations because native structures are often located in a large free-energy basin in the free energy landscape.³ A reasonable energy function should sample near-native conformations more frequently even if it is unable to rank them among the lowest energy conformations.

There are many measures of structural similarity that can be used for clustering. Examples are comparison of distance matrices,⁴ the Fourier method,⁵ geometric hashing,⁶ and contact-map overlap.^{7,8} The most commonly used method is based on the pairwise root-mean-squared distance (pRMSD).^{1,9,10} The method based on pairwise RMSD, however, is time-consuming because clustering requires $N(N - 1)/2$ RMSD calculations for N structures and N often ranges from 10^3 to 10^6 for near-native detections.

In this work, instead of performing accurate $N(N - 1)/2$ pRMSD calculations for clustering, we evaluate the RMSD value using a

randomly selected reference conformation. The reference conformation was used to remove overall translation and rotational motion for all the decoys and RMSD between any two conformations is calculated without further reorientation. Using a decoy set of 41 proteins, we found that clustering based on this reference RMSD (rRMSD) produces essentially the same near-native structures as clustering based on the pairwise RMSD values. This finding was confirmed by tests on additional protein decoy sets with different reference conformations. The new method, which is called SCUD (Structure Clustering of Decoys), is about nine times faster on average for clustering of 2000 decoys of 41 proteins. More significant saving of computing time is expected for larger number of decoys.

Correspondence to: Yaoqi Zhou; E-mail: yqzhou@buffalo.edu

Contract/grant sponsor: NIH; contract/grant numbers: R01 GM 966049 and R01 GM 068530

Contract/grant sponsor: HHMI to SUNY Buffalo

Contract/grant sponsor: the Center for Computational Research

Contract/grant sponsor: the Keck Center for Computational Biology at SUNY Buffalo

Contract/grant sponsor: National Science Foundation of China; contract/grant number: 20340420391 (to Y.Z.)

Methods

The Decoy Sets

A total of 41 α -helical proteins whose number of residues ranges from 40 to 124 were selected. The initial structures of the proteins were constructed with random dihedral angles for the residues in nonhelical regions and native dihedral angles for the residues in helical regions. The initial structures were then minimized in dihedral space. The energy function for minimization is a combination of the DFIRE energy function,¹¹ an improper torsion energy and a simple repulsive potential. Each protein contains 2000 decoys. Details about decoy generation will be published elsewhere. This set was limited to α -helical proteins because helical proteins are easier to fold than other types of proteins.

We also used an independent Rosetta decoy set of 12 β proteins.¹² They are 1bdo, 1csp, 1gvp, 1tit, 1wiu, 1ark, 1tul, 1who, 2ncm, 4fgf, 1ksr, and 1sro, with 1000 decoys for each protein.

Clustering by pRMSD

Decoys are clusterized based on pairwise C_α -atom-based RMSD values. One common method to cluster structures is to use a predetermined cutoff RMSD value.¹³ We found that it is difficult to set one cutoff value for all 41 proteins because this will lead to too few clusters for some proteins and too many clusters for the others. Thus, we use a protein-dependent cutoff value at which the top three largest clusters contain 5% of all structures to be clustered (i.e., 100 structures out of a total of 2000 structures, other percentages are also tested). This method was used so that (1) the top three largest clusters have statistically meaningful number of structures, and (2) the number of clusters is large enough to reflect the diversity of the structures. The RMSD cutoff values range from 0.5–8 Å, depending on the sizes of the nonhelical portions of proteins. Other similar methods for automatically choosing cutoff values were also proposed.¹⁰ Once the RMSD cutoff value is determined, two structures are defined in the same cluster when their pRMSD is less than the cutoff value. A cluster is made of the structures whose pRMSD between them are all less than the cutoff value. In the cluster, the structure with the most neighboring structures is the representative structure of the cluster. Representative structures are ranked by the sizes of clusters.

One implicit assumption made here is that the structures of decoys are populated in clusters. This is usually true if the energy function used to generate the conformations is reasonable.

Clustering by rRMSD

A randomly selected structure from decoys is used as a reference structure. All decoys are reoriented to minimize RMSD between each decoy and the reference structure (N reorientations for N decoys). After the reorientation, the RMSD value between any two decoys is calculated without any further reorientation. The method for clustering based on reference RMSD (rRMSD) is the same as described above for the method based on pRMSD. For comparison, more than 20 additional reference conformations including the

Table 1. Results of Decoy Clustering by pRMSD and rRMSD for 41 All- α Proteins.

PDB ^a	N_{res} ^b	RMSD _{min} ^c (Å)	Best RMSD (Å) ^d	
			pRMSD	rRMSD
1G6U	48	0.4	0.76 (1)	0.76 (1)
2ERL	40	2.1	2.7 (4)	2.9 (4)
1LP1	55	2.5	2.1 (3)	2.9 (1)
1EZ3	124	3.2	5.0 (5)	5.1 (1)
1LVFA	106	1.7	2.8 (1)	3.2 (1)
1BW6	56	1.8	3.6 (1)	3.6 (1)
1DV0	45	2.5	4.4 (2)	3.6 (4)
1EDK	56	1.8	2.5 (1)	8.5 (1)
1EF4	55	4.6	5.8 (1)	5.1 (4)
1IDY	54	4.6	5.9 (1)	5.9 (1)
1BDD	60	4.0	7.1 (3)	6.6 (4)
1MBE	53	3.3	4.6 (1)	8.5 (4)
1PRB _{10–53}	44	1.8	2.6 (2)	2.6 (2)
1PRU	56	5.1	5.7 (5)	5.7 (3)
2SPZ	58	1.6	3.2 (1)	1.9 (5)
2HOA	68	6.9	8.3 (2)	12.2 (4)
1CKT	71	4.2	10.4 (2)	10.4 (1)
1DV5	80	3.8	7.4 (4)	7.4 (3)
1GAB _{10–51}	42	1.9	2.7 (2)	2.7 (1)
1LBU _{17–76}	60	4.4	8.1 (4)	7.8 (3)
1LEA _{6–52}	47	2.3	4.5 (1)	4.1 (5)
1LRE	81	4.1	5.8 (4)	5.3 (2)
2OCC	79	5.9	7.7 (2)	7.7 (2)
4helix	106	4.7	6.4 (1)	7.1 (1)
1A04A _{158–213}	56	2.3	6.8 (3)	2.5 (2)
1A6S	87	4.6	7.7 (5)	7.7 (1)
1CSA	65	2.9	4.6 (1)	8.0 (3)
1FFH _{2–88}	87	2.7	3.1 (1)	3.1 (2)
1NKL	78	2.7	4.4 (1)	4.2 (4)
2ABD	86	5.2	8.1 (5)	8.1 (1)
1AISB _{1109–1196}	88	2.7	6.5 (5)	5.1 (2)
1BONA _{1–68}	68	4.2	5.4 (4)	5.4 (3)
1BOXA _{916–977}	62	2.3	3.9 (2)	3.1 (1)
1UNKA	87	5.2	6.5 (4)	10.5 (1)
1CTJ	89	5.4	7.9 (4)	8.6 (4)
1KDXA	81	4.5	7.1 (1)	7.1 (1)
1BMTA _{651–740}	90	3.2	3.6 (3)	5.1 (3)
1QC7 _{235–320}	86	2.9	4.9 (4)	4.9 (1)
1BXM	98	6.3	8.6 (2)	8.8 (3)
1NGR	85	5.7	9.8 (5)	9.8 (4)
1RZL	91	5.6	7.9 (1)	7.9 (5)
Average	71.4	3.6	5.5 (2.6)	5.9 (2.4)

^aThe PDB id code.

^bNumber of residues of a given protein.

^cThe smallest RMSD value from native (based on C_α atoms only) in 2000 decoys.

^dThe smallest RMSD value from native among the top five largest clusters clusterized by the pRMSD or the rRMSD method. The number in parentheses is the rank of the best near-native structure.

native conformation are used for clustering and analysis of 41 helical proteins. Another 11 reference conformations are used to cluster 12 β -proteins.

Table 2. Near-Native Selection of 41 All- α Proteins by Clustering with pRMSD and rRMSD Based on Different Reference Conformations.

Reference ^a	RMSD ^b (Å)	Best RMSD (Å) ^c		
		Top 1	Top 5	Top 10
pRMSD ^d	N/A	7.6	5.5	5.3
20 Decoys	10.7 ± 0.5	7.94 ± 0.32 (25.0 ± 3.1/41)	5.83 ± 0.18 (26.1 ± 3.0/41)	5.34 ± 0.17 (28.8 ± 3.1/41)
Initial 1	15.9	8.1 (21/41)	5.7 (21/41)	5.3 (25/41)
Initial 2	15.6	8.2 (22/41)	5.6 (27/41)	5.2 (29/41)
Initial 3	17.6	7.7 (24/41)	5.8 (29/41)	5.4 (29/41)
Native	0	5.4 (38/41)	5.1 (38/41)	4.8 (37/41)

^aReference conformations are those from randomly selected 20 decoys, randomly selected three initial structures (before minimization), and native, respectively.

^bAverage RMSD from the native for the reference conformation.

^cThe average best RMSD value from native among top 1, top 5, and top 10 clusters (averaged over 41 proteins). The number in parentheses is number of proteins whose near-native structures predicted by rRMSD values are at least as accurate as or more accurate than those from the pRMSD method. Error bars are one standard deviation.

^dThe results from clustering of 2000 decoys by pRMSD.

Results and Discussion

Table 1 lists the names of proteins along with number of residues, the minimum RMSD values from the native in the 2000 decoys, the smallest RMSD value (from the native) among the top five structures selected by pRMSD and rRMSD methods, respectively. Here, the top five structures mean the top five representative structures for the five largest clusters. The rRMSD values were obtained from a randomly selected decoy conformation. In 27 out of 41 cases (66%), the rRMSD method yields near-native structures that are either more accurate than or at least equivalent to those from the pRMSD method. The average best RMSD values from native for all 41 proteins given by pRMSD and rRMSD methods are essentially the same (5.5 and 5.9 Å, respectively), considering large fluctuation of best RMSD values among different protein decoys.

Table 2 examines the effect of different reference conformations on the results of clustering by rRMSD. Twenty reference

conformations were used. The average and standard deviation of RMSD values from native are obtained from the top 1 structure, the best in top 5, and the best in top 10. The pRMSD method and rRMSD yielded essentially the same quality near-native structures based on the average RMSD of the selected near-native structures from the native (within statistical uncertainty). Moreover, the average number of proteins whose near-native structures predicted by rRMSD are at least as accurate as or more accurate than those from the pRMSD method are greater than 25 (i.e., >60%) in all cases. To enlarge the possible effect due to the use of a particular reference conformation, we further use the native conformation and three conformations randomly selected from initial structures used in decoy generations (i.e., before minimization). The use of native conformation indeed leads to a better near-native selection from top 1, in particular. The use of initial structures far from the native (average RMSD of 16 Å from the native) does not make any significant change in the results of clustering for top 1, top 5, and top 10 selections.

Table 3. Near-Native Selection of the Rosetta Decoy Set of 12 β -Proteins by Clustering with pRMSD and rRMSD with Different Reference Conformations.

Reference ^a	Best RMSD (Å) ^b		
	Top 1	Top 5	Top 10
pRMSD ^c	11.9	10.2	9.7
11 Decoys	11.6 ± 0.4 (8.3 ± 1.6/12)	10.2 ± 0.2 (7.5 ± 1.4/12)	9.7 ± 0.2 (7.6 ± 1.3/12)
Native	10.2 (11/12)	9.4 (11/12)	9.1 (11/12)

^aReference conformations are those from randomly selected 11 decoys and the native conformation, respectively.

^bThe average best RMSD value from native among top 1, top 5, and top 10 clusters (averaged over 12 proteins). The number in parentheses is number of proteins whose near-native structures predicted by rRMSD values are at least as accurate as or more accurate than those from the pRMSD method. Error bars are one standard deviation.

^cThe results from clustering of 1000 decoys by pRMSD. The average of the minimum RMSD from native in 1000 decoys of the 12 β -proteins is 7.4 Å.

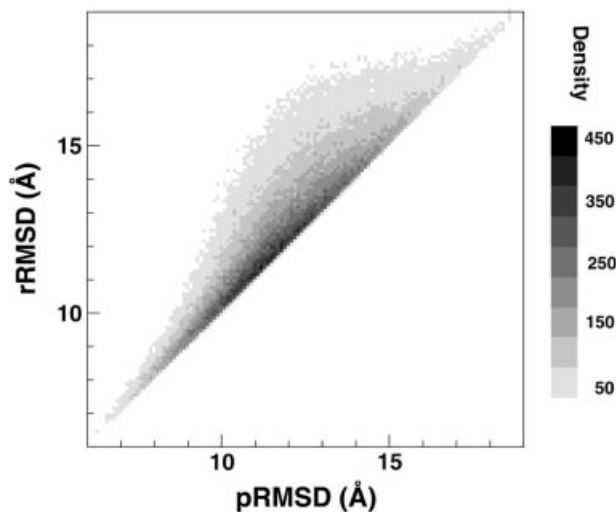


Figure 1. The density plot of rRMSD vs. pRMSD value for protein 1KDXA. (A grid of 0.1 Å is used to collect the density.) Most of data points are clustered above and close to the $x = y$ line. The correlation coefficient is 0.82.

To further confirm the above results on α proteins, we also used an independent Rosetta decoy set of 12 all- β proteins. The results of clustering based on rRMSD and that based on pRMSD are compared in Table 3. The results on all- β proteins confirm that clustering by pRMSD or by rRMSD using different reference conformations all produce essentially the same near-native conformations in average.

The similar results on near native selections based on pRMSD and rRMSD are somewhat surprising. To provide a further understanding, we calculated the correlation coefficients between pRMSD and rRMSD. The correlation coefficients vary from protein to protein, ranging from 0.7 to 0.99. The average correlation coefficient for all 41 proteins and for 20 reference decoy conformations is 0.82 ± 0.01 (one standard deviation). Obviously, the high correlation between pRMSD and rRMSD explains the similarity on near-native selections. One example with a correlation coefficient of 0.82 (the same as the average correlation coefficient) is shown in Figure 1. Most of the data points are clustered above and close to the $rRMSD = pRMSD$ line.

We have also examined the similarity between clusters obtained by pRMSD and by rRMSD, respectively. We define that two clusters are identical if the representative structures of the two clusters are within the cutoff distance defined by top 3 largest decoys having 5% decoy structures. We found that 54% ($\pm 7\%$) of the top 1 cluster, 74% ($\pm 2\%$) of the top 5 clusters, and 80% ($\pm 1\%$) of the top 10 clusters are the same between the clusters obtained from pRMSD and those from rRMSD. These results were

obtained from the average over all 41 proteins with 20 different decoy conformations as the reference conformation. More than 50% top 1 cluster shared by pRMSD and rRMSD clustering further explains the usefulness of rRMSD in near-native selections.

The use of rRMSD offers a significant time savings because reorientation needs to be performed only once. Even for 2000 decoys of 41 proteins, the average time for pRMSD clustering for one protein is 648 s, compared to 70 s for rRMSD clustering. That is, SCUD based on rRMSD is about nine times faster than traditional clustering based on pRMSD. The time saving is expected to be much greater when the number of decoys is significantly larger than 2000.

One interesting question is how the definition of clusters affects the overall results. For the above results, we used top three largest clusters having 5% of decoys as the criterion to determine the RMSD cutoff for a cluster. We found that the average best RMSD from the top five for 41 helical proteins changes from 6.1, 5.8, 5.8, 5.9, 6.1, and 6.3 as 1, 3, 5, 10, 15, and 20% of all decoy structures contained in the top three clusters are used instead. This result is the average from using 20 decoys as reference structures. Although the change is small, there appears to exist a minimum between 3 and 5%.

Acknowledgments

We thank Dr. Hongyi Zhou, Dr. Chi Zhang, and Mr. Song Liu for many helpful discussions.

References

- Shortle, D.; Simons, K. T.; Baker, D. *Proc Natl Acad Sci USA* 1998, 95, 11158.
- Zhang, Y.; Kolinski, A.; Skolnick, J. *Biophys J* 2003, 84, 1145.
- Dobson, C. M.; Šali, A.; Karplus, M. *Angew Chem Int Ed* 1998, 37, 868.
- Holm, L.; Sander, C. *J Mol Biol* 1993, 233, 123.
- Chechetkin, V. R.; Lobzin, V. V. *J Theor Biol* 1999, 198, 219.
- Fischer, D.; Bachar, O.; Nussinov, R.; Wolfson, H. *J Biomolec Struct Dynam* 1992, 9, 769.
- Godzik, A.; Skolnick, J.; Kolinski, A. *J Mol Biol* 1992, 227, 227.
- Caprara, A.; Carr, R.; Istrail, S.; Lancia, G.; Walenz, B. *J Comput Biol* 2004, 11, 27.
- Betancourt, M. R.; Skolnick, J. *J Comput Chem* 2001, 22, 339.
- Zhang, Y.; Skolnick, J. *J Comput Chem* 2004, 25, 865.
- Zhou, H.; Zhou, Y. *Protein Sci* 2002, 11, 2714 [Corrections 2003, *Protein Sci* 12:2121].
- Simons, K.; Bonneau, R.; Ruczinski, I.; Baker, D. *Proteins* 1999, 37, 171.
- Karpen, M. E.; Tobias, D. J.; Brooks, C. L., III. *Biochemistry* 1993, 32, 412.