SPICKER: A Clustering Approach to Identify Near-Native Protein Folds

YANG ZHANG, JEFFREY SKOLNICK
Center of Excellence in Bioinformatics, University at Buffalo, 901 Washington St., Buffalo, New York 14203

Received 11 October 2003; Accepted 20 November 2003

Abstract: We have developed ***SPICKER***, a simple and efficient strategy to identify near-native folds by clustering protein structures generated during computer simulations. In general, the most populated clusters tend to be closer to the native conformation than the lowest energy structures. To assess the generality of the approach, we applied ***SPICKER*** to 1489 representative benchmark proteins ≤200 residues that cover the PDB at the level of 35% sequence identity; each contains up to 280,000 structure decoys generated using the recently developed ***TASSER*** (Threading ASSembly Refinement) algorithm. The best of the top five identified folds has a root-mean-square deviation from native (RMSD) in the top 1.4% of all decoys. For 78% of the proteins, the difference in RMSD from native to the identified models and RMSD from native to the absolutely best individual decoy is below 1 Å; the majority belong to the targets with converged conformational distributions. Although native fold identification from divergent decoy structures remains a challenge, our overall results show significant improvement over our previous clustering algorithms.

 J Comput Chem 25: 865–871, 2004

Key words: ***SPICKER***; near-native folds; ***TASSER***

Introduction

Given a perfect energy function (e.g., the Go-like potential[1]), the native conformation of a protein can be identified on the basis of its energy (i.e., the native state is the minimum energy conformation). Furthermore, of all the energetically and structurally well-defined clusters, the native state should be the most populated at low temperatures. In the more realistic situation encountered in protein structure prediction, the lowest energy state is usually not the conformation nearest to native because of imperfections in the force field. But given a reasonable energy function, these near-native states should be still among the most populated states at low temperature.[2] Based on this hypothesis, approaches based on structure clustering have been used for fold selection.[2,3]

There are two issues involved in cluster-based approaches to native fold identification. First of all, for a set of n structure decoys, an $n \times n$ matrix of RMSD distances for all decoy pairs needs to be constructed; for very large numbers of structures, the size of this matrix could easily exceed the memory of contemporary computers. The problem becomes more acute when parallel sampling methods are used[4–6] and a large number of decoy structures from different replicas need to be clustered. In previous approaches using ***SCAR***,[3] Betancourt and Skolnick implemented clustering in two steps: the structures in each replica are clustered, and then the resulting cluster centroids are clustered to obtain the final models. Because the matrix size is reduced compared with the matrix of the entire decoy set, this two-step clustering process requires much less computer memory. However, the cluster population information present in the first pass is neglected in the second pass; therefore, the relative cluster population is not appropriately accounted for.

Second, the distribution of decoy conformations depends on the energy landscape, which can be very different depending on the level of prediction difficulty. For a comparative modeling target, the homologous templates provide consensus spatial restraints;[7] thus, the resulting structures will be tightly distributed near their templates. On the other hand, for a "New Fold" target, the *ab initio* based simulations,[8–10] generate much more divergent structures. Therefore, the definition of appropriate cutoff values is important for the correct identification of representative folds.

To address these issues, we describe a new clustering program, ***SPICKER***, in which clustering is performed in a one-step procedure using a shrunken but representative set of decoy conformations and the pairwise RMSD cutoff is determined by self-adjusting iteration. To assess its generality, we apply ***SPICKER*** to a large-scale benchmark set of 1489 nonhomologous proteins that

Correspondence to: J. Skolnick; e-mail: skolnick@buffalo.edu

Contract/grant sponsor: Division of General Sciences of the National Institutes of Health; Contract/grant number: GM-37408

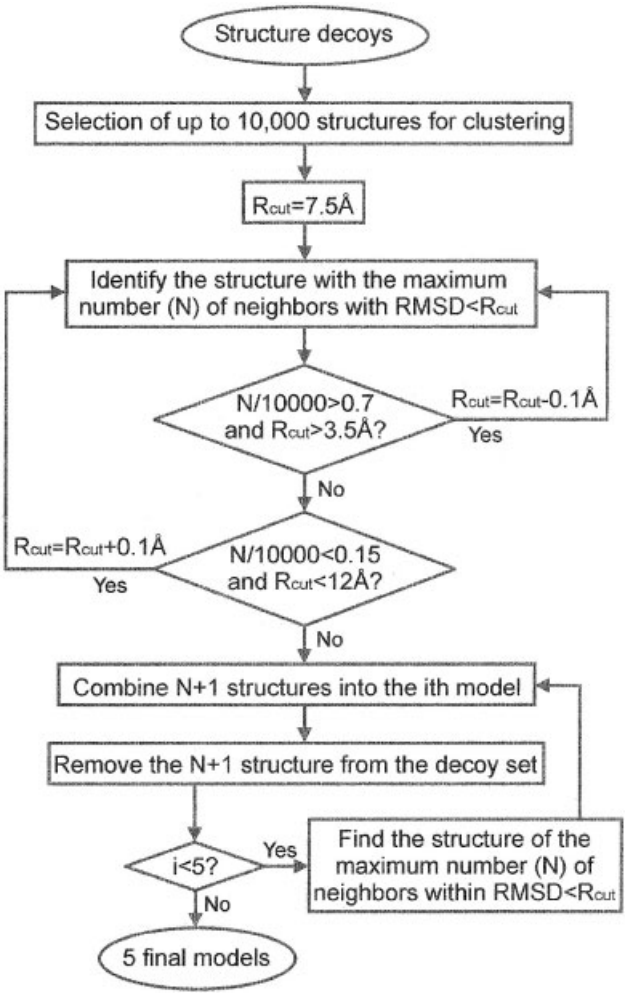

Figure 1. Flow chart of the ***SPICKER*** clustering algorithm.

represent all protein structures in the PDB ≤200 residues; each protein has up to 280,000 decoy structures obtained from low-temperature replicas generated by our threading/assembly/refinement program ***TASSER***.[11]

Methods

The ***SPICKER*** fold identification algorithm consists of decoy shrinking, cutoff iteration, cluster identification, and final model combination, a flow chart. A flow chart of the procedure is presented in Figure 1.

Decoy Selection

Decoy structures are taken from computer simulations, which could be, for example, Monte Carlo or Molecular Dynamics trajectories.[10–12] In principle, to find the most representative structures, we should take as many structures as possible. However, because of computer memory limitations, we cannot use all the structures when their number exceeds 10^4. Here, we shrink the decoy set by taking the lowest energy conformation in each subset of n/S decoys, with the shrinkage factor $S = 10^4$, when n is too large for the pairwise RMSD matrix to fit into memory. When the final time interval between the selected snapshots is shorter than the self-correlation time to crossenergy barriers, this reduction in the number of decoys does not influence the decoy distribution in the important regions of phase space. To confirm this, we show in Figure 2 the average clustering result of 100 randomly selected targets, when different values of the shrinkage factor S are used to select structures from the 280,000 decoys. When the number of clustered structures is larger than about 4000, well within computer memory capacity, there is no obvious dependence of the clustering results on the number of structures used, although the average RMSD from native to the absolute best individual decoy decreases slightly as the number of selected structures increases.

RMSD Cutoff Used for Clustering

The pairwise RMSD cutoff R_{cut}, under which two structures are considered as clustered neighbors, is iteratively decided by the interplay of the cutoff and the ratio of number of decoys in the most populated cluster to the total number of decoy structures. The initial R_{cut} is set to 7.5 Å. If the structures are too tightly distributed, R_{cut} will gradually decrease until the first cluster includes less than 70% of the total number of structures or until R_{cut} is 3.5 Å. On the other hand, if the structures are too divergent, R_{cut} will gradually increase until the first cluster includes more than 15% of structures or until R_{cut} = 12 Å.

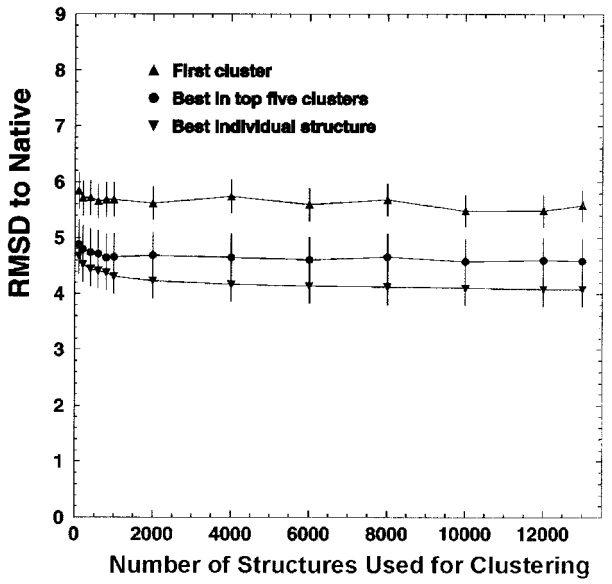

Figure 2. RMSD to native of cluster models and the best individual structure in a shrunken decoy set vs. the number of structures of the compressed decoy set used in ***SPICKER*** clustering. The shown data are averaged over 100 target proteins, each having originally up to 280,000 decoy conformations generated by ***TASSER***.[11] The lines connecting the points serve to guide the eye.

Cluster Identification

The members of each cluster are constructed as follows: for a given R_{cut}, the first cluster contains the structure with the most neighbors (this structure comprises the "cluster center" structure), as well as the structures of all its neighbors. To identify the second cluster, all the structures in the first cluster are excluded and the second cluster contains the structure with the most number of neighbors in the remaining structure decoys, as well as the structures of all its neighbors. This process is iterated so the lth cluster identified contains the structure with the most neighbors in the remaining decoys after excluding all members of the preceding $l - 1$ clusters, plus their associated neighbor structures.

Final Model Construction

The decoy structures in each cluster are used to construct a representative model for the cluster. There were two ways to select these representative models:[2,3] i.e., either selecting the individual "cluster center" structure that was initially used to define the cluster, or superimposing all members of the cluster on the "cluster center" structure and then averaging all the superimposed structures to define the cluster centroid. Although the averaged centroid structure usually has a lower RMSD to native than the "cluster center" structure, global averaging can result in irregular local structures when the cluster cutoff is high. To address this issue, we divide the protein into up to five subchains, each containing more than 20 residues, and average the subchains separately after the superposition of the subchains onto the "cluster center" structure. The resulting local structures of the combined model are less irregular than the centroid but are of lower RMSD to native than the individual "cluster center" structure. Here it should be noted that, in this process, all decoy structures (including previously clustered structures) with an RMSD to the "cluster center" structure below R_{cut} are used to build the final model.

Results and Discussions

Benchmark Set of Decoy Structures

To comprehensively evaluate the methodology, we applied the ***SPICKER*** algorithm to a representative set of PDB proteins comprised of 1489 targets. Here, for each protein the decoys are generated by ***TASSER***,[11] a recently developed protein structure assembly algorithm, designed to build full-length protein models by rearranging the continuous fragments excised from threading templates under the guide of an optimized force field.[10] The conformational phase space is searched using a parallel hyperbolic Monte Carlo simulation.[6] This benchmark set includes 877 targets that have templates identified with high confidence by our threading program ***PROSPECTOR_3***,[13] called "Easy" targets. In this set, the structure decoys are relatively tightly distributed, because the templates have high alignment coverage and consensus alignments, which impose consistent restraints onto the Monte Carlo structure assembly processes. The benchmark

Table 1. Summary of Clustering Results from 1489 Proteins by ***SPICKER*** and ***SCAR.***[a]

	Easy targets (877 cases)						Medium/hard targets (612 cases)					
	SPICKER			***SCAR***			***SPICKER***			***SCAR***		
	BT[b]	T5[c]	T1[d]	BT[b]	T5[c]	T1[d]	BT[b]	T5[c]	T1[d]	BT[b]	T5[c]	T1[d]
⟨RMSD⟩[e]	**3.35**	**3.75**	**4.79**	*3.29*	*4.28*	*5.58*	**7.06**	**8.01**	**10.49**	*6.92*	*8.83*	*11.53*
N_{winner}[f]	**0**	**767**	**653**	*406*	*92*	*213*	**0**	**485**	**396**	*318*	*121*	*213*
$N_{RMSD<6.5}$[g]	**810**	**769**	**682**	*813*	*734*	*611*	**296**	**221**	**121**	*306*	*177*	*87*
$N_{RMSD<6.0}$	**785**	**756**	**652**	*789*	*711*	*590*	**268**	**191**	**107**	*275*	*155*	*74*
$N_{RMSD<5.5}$	**763**	**737**	**615**	*764*	*680*	*555*	**223**	**167**	**90**	*236*	*127*	*66*
$N_{RMSD<5.0}$	**740**	**700**	**592**	*743*	*644*	*519*	**195**	**132**	**74**	*201*	*104*	*56*
$N_{RMSD<4.5}$	**706**	**658**	**541**	*714*	*588*	*467*	**159**	**102**	**58**	*167*	*84*	*45*
$N_{RMSD<4.0}$	**669**	**608**	**487**	*673*	*533*	*403*	**120**	**76**	**45**	*124*	*65*	*34*
$N_{RMSD<3.5}$	**599**	**520**	**411**	*609*	*440*	*321*	**94**	**61**	**34**	*96*	*48*	*23*
$N_{RMSD<3.0}$	**493**	**413**	**320**	*508*	*342*	*241*	**64**	**38**	**24**	*66*	*37*	*17*
$N_{RMSD<2.5}$	**359**	**291**	**223**	*368*	*233*	*146*	**42**	**24**	**17**	*44*	*25*	*10*
$N_{RMSD<2.0}$	**212**	**176**	**124**	*223*	*125*	*73*	**29**	**17**	**8**	*28*	*11*	*3*
$N_{RMSD<1.5}$	**85**	**58**	**43**	*90*	*39*	*22*	**13**	**8**	**6**	*12*	*4*	*1*
$N_{RMSD<1.0}$	**17**	**10**	**9**	*17*	*5*	*3*	**7**	**5**	**4**	*6*	*3*	*1*

[a]The data from ***SPICKER*** and ***SCAR*** are denoted by bold and italic fonts, respectively. All RMSD numbers are in Angstroms.
[b]The best structure of lowest RMSD to native among all the decoys used in clustering process.
[c]The best model in top five clusters.
[d]The first ranked cluster according to the cluster density in ***SPICKER*** or according to the average energy in ***SCAR***.
[e]Average RMSD to native over 877 targets in Easy set and over 612 targets in Medium/hard set.
[f]The number of targets that have a lower RMSD to native between ***SPICKER*** and ***SCAR*** models.
[g]The number of targets that have a RMSD to native below some threshold value.

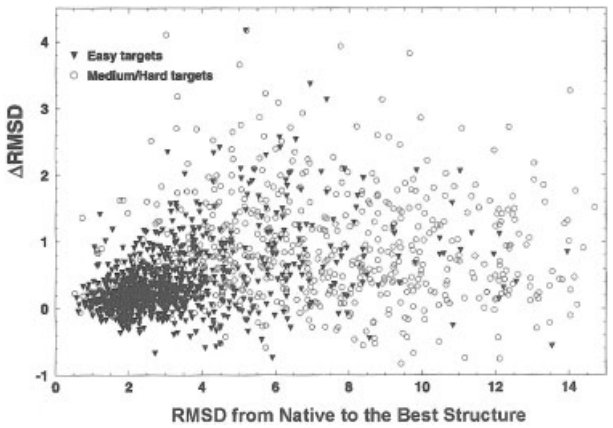

Figure 3. RMSD distribution of the best in top five models with respect to the best individual decoy structure. ΔRMSD is the difference between RMSDs from native to the model and from native to the decoy, as defined in eq. (1).

set also includes another 612 "Medium/Hard" protein targets that ***PROSPECTOR_3*** assigns templates with lower confidence. These templates typically have good structural alignments (i.e., in more than 90% of cases, the RMSD to native in structure alignments is below 6.5 Å); but in only 1/3 of the cases is the RMSD to native in the threading-based alignments below 6.5 Å.[13] The alignments are, on average, also of lower coverage than for the "Easy" set. Thus, the structure decoys are more divergent as compared to the "Easy" targets. All the structure decoys, together with the clustered models, are available on our Web site: http://bioinformatics.buffalo.edu/abinitio/1489.

Summary of Clustering Results

In Table 1, we list a summary of clustering results from ***SPICKER*** and ***SCAR***. Because ***SCAR*** uses all the generated structure decoys, the RMSD of the best individual structures is lower than the shrunken decoys by ***SPICKER***. However, the RMSD of final clustered models is lower in ***SPICKER***. For example, for the "Easy" targets, we have 608/877 cases that have at least one model in top five clusters with a RMSD to native below 4 Å using ***SPICKER***; when ***SCAR*** is used, this number is 533/877. For the "Medium/Hard" targets, the RMSD of the best of the top five clusters generated by ***SPICKER*** is below 6.5 Å in 221/612 cases; using ***SCAR***, this number is 177/612. Overall, if we look at the RMSD of the best of the top five clusters, in 1252 cases, ***SPICKER*** does better; and in 213 cases, ***SCAR*** does better. If we only focus on the first ranked cluster, in 1049 cases ***SPICKER*** does better and in 426 cases ***SCAR*** does better.

Comparison of Models with the Best Individual Structures

For a given set of decoy structures, the goal of clustering is to provide models that are as good as the best individual structures found in the decoy set. Because the final models in ***SPICKER*** are obtained by local averaging of clustered decoys, occasionally the combined model can be even closer to native than the best individual structures. For example, in 194 of the 1489 protein targets, one of top five ***SPICKER*** models has an RMSD to native smaller (in a the range up to 0.9 Å) than any individual structure; 145 of them belong to the "Easy" set and 49 to the "Medium/Hard" set.

However, as shown in the Table 1, the overall quality of the structure provided by clustering is still worse than the best individual structure. For the "Easy" targets, the difference between the RMSD from native of the best of the top five clusters and the RMSD from native of the best individual structure is about 0.4 Å on average, while this difference in "Medium/Hard" cases is around 1 Å on average. If we compare the first ranked cluster with the best individual structures, the former is about 1.5 Å farther from native for the "Easy set" and 3.5 Å farther from native for the "Medium/Hard" set. In Figure 3 and Table 2, we present the distribution of the difference between the RMSD errors of the best of the top five clusters and that of the best individual structures, i.e.,

$$\Delta\text{RMSD} = \text{"RMSD from native to the best of top five cluster models"} - \text{"RMSD from native to the absolutely best decoy."} \quad (1)$$

Table 2. Comparison of Top Five ***SPICKER*** Models with the Best Individual Decoy Structure.[a]

	Easy targets	Medium/hard targets	In total
$\Delta\text{RMSD} \leq 0.5$ Å	615 (70.1%)	222 (36.3%)	837 (56.2%)
0.5 Å $< \Delta\text{RMSD} \leq 1.0$ Å	165 (18.8%)	164 (26.8%)	329 (22.1%)
1.0 Å $< \Delta\text{RMSD} \leq 1.5$ Å	56 (6.4%)	103 (16.8%)	159 (10.7%)
1.5 Å $< \Delta\text{RMSD} \leq 2.0$ Å	22 (2.5%)	67 (10.9%)	89 (6.0%)
2.0 Å $< \Delta\text{RMSD} \leq 2.5$ Å	14 (1.6%)	22 (3.6%)	36 (2.4%)
2.5 Å $< \Delta\text{RMSD} \leq 3.0$ Å	2 (0.2%)	17 (2.8%)	19 (1.3%)
3.0 Å $< \Delta\text{RMSD}$	3 (0.3%)	17 (2.8%)	20 (1.3%)
In total	877 (100%)	612 (100%)	1489 (100%)

[a]ΔRMSD denotes the difference of the RMSD errors of identified models and the best individual decoys and is defined in eq. (1). The numbers in parentheses are the percentage of protein targets in the specified protein sets.

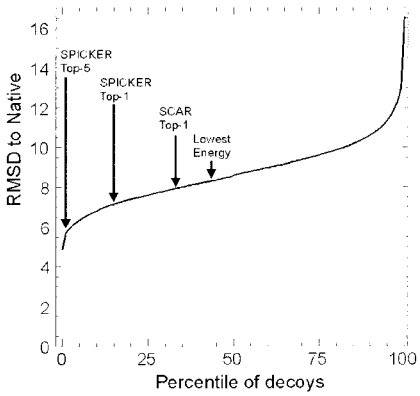

Figure 4. RMSD from native vs. the percentile rank of decoys. The arrows mark the corresponding position of the models identified by different methods.

For "Easy" targets, for around 90% of the cases, ***SPICKER*** identifies models ≤1 Å worse from native than the best decoy; for "Medium/Hard targets, this happens in only 63% of the cases. The dependence of fold identification ability upon the category of targets is understandable because for "Medium/Hard" targets, the decoy structures are much more divergent, and there is more opportunity to pick up wrong structures. This result also highlights the necessity of further improving the current clustering strategy to identify the best folds for divergently distributed decoys.

Rank of Identified Model among Entire Decoy Sets

In Figure 4, we sort the decoy conformations according to their RMSD to the native structure and obtain the average rank of cluster models among all the decoys. On average, the best of top five clusters is in the top 1.4% of all decoy structures and the RMSD of the first cluster is in the top 14.9%. For ***SCAR*** clusters, the average rank of best of the top five clusters and that of the first cluster are in the top 5.8 and 32.7%, respectively. If we select the individual decoy structure of lowest energy, the average rank is in the top 43.1% of all decoys. In Figure 5, we also show the histogram of first clustered models and the lowest energy structure along with the percentile of structure decoys sorted by their RMSD to native. The RMSD of the lowest energy structures distributes almost evenly among all quality structures, while the ***SPICKER*** cluster models are obviously more populated for higher quality decoys; this demonstrates the significant advantage of structure identification by clustering.

Indicator of Model Quality

A sensitive and objective indicator of the quality of predicted models is of vital importance for successful blind protein structure

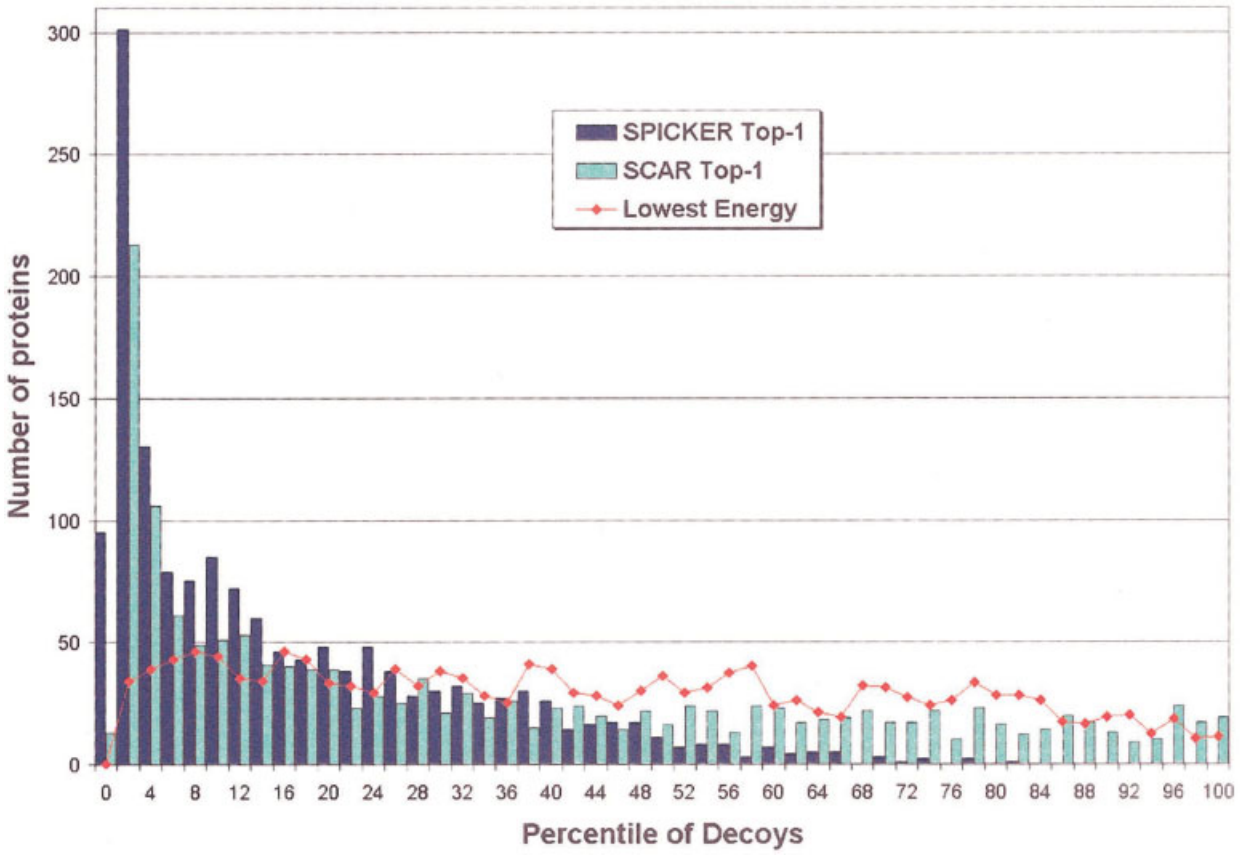

Figure 5. Histogram of the RMSD to native of selected models by different methods with respect to their percentile rank.

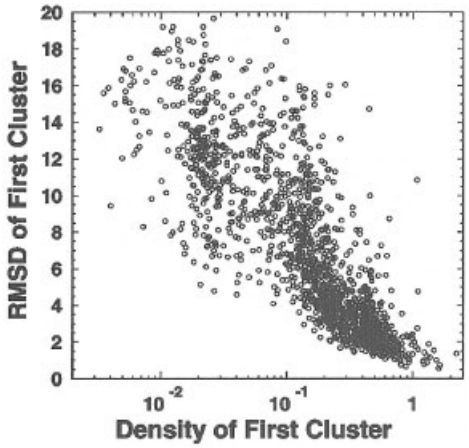

Figure 6. RMSD to native of the first cluster for 1489 targets vs. the normalized cluster density that is defined in eq. (2).

prediction. Because the distributions of decoy conformations in the Monte Carlo simulation are closely related with the alignment coverage of templates and the consistency between restraints and the inherent energy terms of the basic force field,[10] structure convergence can be considered as an indicator of fold quality. Here, we define a normalized structure cluster density as:

$$D = \frac{M}{\langle \mathrm{RMSD} \rangle M_{\mathrm{tot}}} \tag{2}$$

where M is the multiplicity of conformations in the cluster, M_{tot} is the total number of decoy conformations submitted to ***SPICKER*** for clustering, and $\langle \mathrm{RMSD} \rangle$ denotes the average RMSD of the conformations to the combined model of the cluster.

In Figure 6, we show the RMSD from native to the first combined model vs. the normalized structure density. There is a strong correlation between the cluster quality and the degree of cluster convergence. Most dense clusters have a RMSD from native below 4 Å when $D > 0.2$; this corresponds to a cluster having more than 60% of its structures enclosed in a conformational space with a 3 Å of pairwise RMSD. On the other hand, when $D < 0.02$, a value that corresponds to a cluster enclosing less than 15% of structures within 7.5 Å, there is never a model with a RMSD from native that is less than 6 Å.

In Figure 7, we also show the distribution of the highest cluster density of each target. The gray area denotes the targets that have an RMSD to native above 6 Å for the best of the top five clusters, while the dark area shows those with a RMSD below 6 Å. If we define a successful fold prediction when one of top five structures has an RMSD to native below 6 Å, and where we assess the likelihood of folding success based on the cluster density cutoff at $D_{\mathrm{cut}} = 0.1$, the false positive and false negative rate is 10 and 12% according to Figure 7. This cutoff value is equivalent to the case

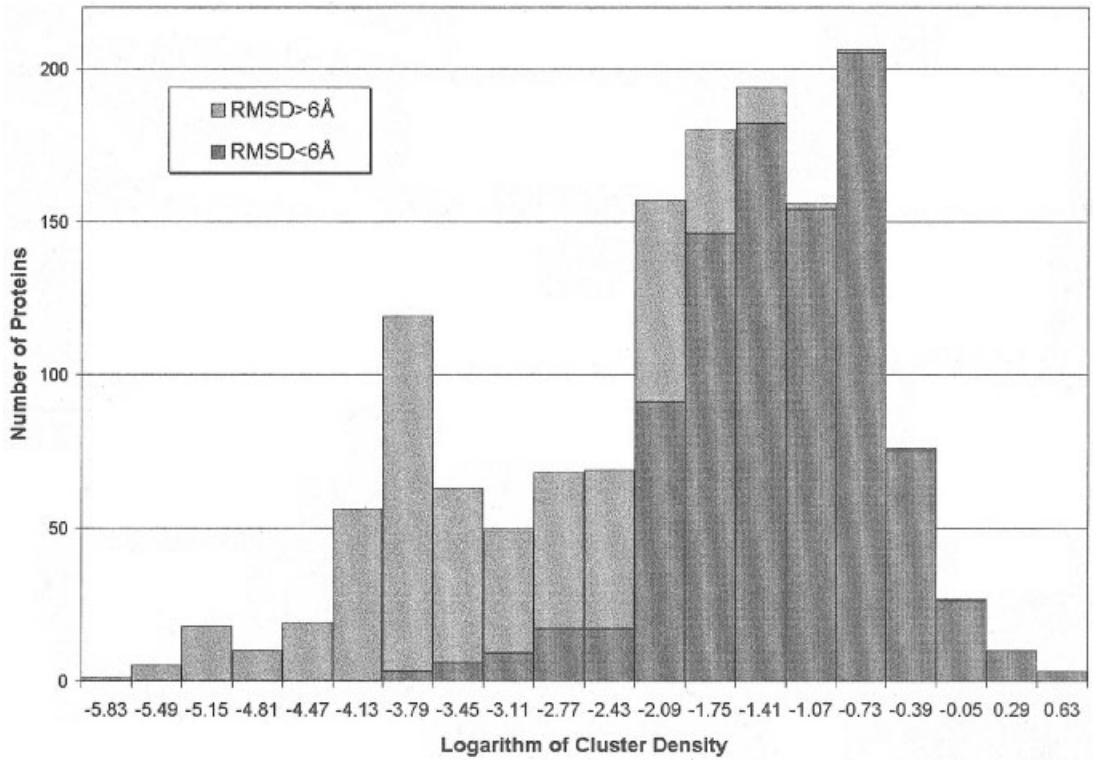

Figure 7. Histogram distribution of structure density of the first cluster for 1489 benchmark targets. The targets of best RMSD in top five clusters below and above 6 Å are shown in different color.

where more than half of structures are enclosed within 5 Å of the highest density area.

Cluster Population in Parallel Exchange Monte Carlo Simulations

One of the main differences between ***SCAR*** and ***SPICKER*** is at their methods of dealing with structure decoys in parallel Monte Carlo sampling simulations.[4–6] Let's take the replica exchange Monte Carlo simulation[5] as an illustrative example.

Suppose we have N copies of the simulated molecule, each at a different temperature T_i. A state of the composite system is specified by $X = \{x_1, x_2, \ldots, x_N\}$, where x_i is the conformation of the molecule in the ith replica. The goal of the replica transition (i.e., $p_{i\leftrightarrow j} = \exp(\beta_i - \beta_j)(E(x_i) - E(x_j)))$ in the simulation is to have an equilibrium distribution of X:

$$P(X) = \prod_{i=1}^{N} \frac{e^{-\beta_i E(x_i)}}{Z(T_i)}$$

where $\beta_i = 1/k_B T_i$, and $Z(T_i) = \Sigma\ e^{-\beta_i E(x_j)}\Omega(x_i)dx_i$ is the overall partition function of the molecule at temperature T_i, with $\Omega(x_i)$ the conformational density of x_i.

The population of structures in one ***SPICKER*** cluster n is proportional to the summation of the probability for the molecule to adopt the conformation x_0 in all replicas, i.e.,

$$\begin{aligned} n &\sim \sum_i \int P(X)\delta(x_i - x_0)\Omega(x_1)dx_1\Omega(x_2)dx_2\cdot\cdot\cdot\Omega(x_N)dx_N \\ &= \sum_i \frac{\int e^{-\beta_i E(x_i)}\Omega(x_1)dx_1}{Z(T_1)} \frac{\int e^{-\beta_2 E(x_2)}\Omega(x_2)dx_2}{Z(T_2)} \\ &\quad \cdots \frac{\int e^{-\beta_i E(x_i)}\delta(x_i - x_0)\Omega(x_i)dx_i}{Z(T_i)} \cdots \frac{\int e^{-\beta_N E(x_N)}\Omega(x_N)dx_N}{Z(T_N)} \\ &= \sum_i \frac{e^{-\beta_i E(x_0)}\Omega(x_0)}{Z(T_i)} \\ &= \sum_i \frac{z(T_i, x_0)}{Z(T_i)} \end{aligned}$$

where $z(T_i, x_0)$ is the partition function of the ith replica with conformation x_0 at the temperature T_i, which is related to its conformational free energy by $F(T_i, x_0) = -k_B T \ln z(T_i, x_0)$.

The difference between ***SCAR*** and ***SPICKER*** is that, for each cluster in an individual replica, only the centroid structure of the cluster was used in the final clustering in ***SCAR***. This is equivalent to treating the partition function $z(T_i, x_0)$ as being a constant independent of T_i (i.e., all clusters are equally populated). Therefore, the relative cluster population is not appropriately accounted for in ***SCAR***.[3]

Conclusions

We have developed ***SPICKER***, a simple and efficient approach to identify near-native folds by clustering structure decoys. This algorithm is used in a large-scale fold selection experiment on a representative benchmark set comprised of 1489 benchmark targets, each including up to 280,000 structure decoys generated by the threading assembly refinement program ***TASSER***.[11] The combined models of the highest structure density are significantly closer to the native structures than the lowest energy structure. On average, the RMSD to native of the highest density cluster is in the top 14.9% of all decoy structures, while the lowest energy state is in top 43.1% of decoys, and is almost evenly distributed among all quality conformations. For the divergently distributed decoys found for the "Medium/Hard" target proteins, the relative rank of the cluster identified model is worse than that for the convergent decoys in the "Easy" target set; this demonstrates that the fold identification of nonclustered decoys still remains a challenge to current clustering approaches.

The overall result shows improvement over our previous two-step clustering algorithm ***SCAR***.[3] For the same set of decoy structures, ***SPICKER*** identifies around 10% more models with a RMSD < 6.5 Å to native than ***SCAR***. The reason may be that information of the cluster populations in the first step is neglected in the second step of clustering in ***SCAR***. Because a uniform shrinking of the Monte Carlo generated trajectories does not change the relative cluster density, we perform ***SPICKER*** clustering in a one-step process. Another difference that may contribute to the improvement is that the clusters in ***SPICKER*** are ranked according to cluster density, while the clusters in ***SCAR*** are ranked by energy, which has been shown to be less sensitive to the model quality.

The strong correlation between cluster density and the RMSD to native shows that this quantity can be considered as a reliable indicator of the likelihood of folding success. For the 1489 representative targets, if the highest cluster density is greater than 0.1 (equivalent to a folded state including more than half of conformations clustered within 5 Å of each other), one of the top five clusters has an RMSD to native below 6 Å in around 90% of the cases. These results provide a quantitative reference for the assessment of future blind genome scale protein structure predictions.

References

1. Go, N.; Taketomi, H. Proc Natl Acad Sci USA 1978, 75, 559.
2. Shortle, D.; Simons, K. T.; Baker, D. Proc Natl Acad Sci USA 1998, 95, 11158.
3. Betancourt, M. R.; Skolnick, J. J Comp Chem 2001, 22, 339.
4. Hansmann, U. H. E. Chem Phys Lett 1997, 281, 140.
5. Swendsen, R. H.; Wang, J. S. Phys Rev Lett 1986, 57, 2607.
6. Zhang, Y.; Kihara, D.; Skolnick, J. Proteins 2002, 48, 192.
7. Sali, A.; Blundell, T. L. J Mol Biol 1993, 23, 779.
8. Pillardy, J.; Czaplewski, C.; Liwo, A.; Lee, J.; Ripoll, D. R.; Kazmierkiewicz, R.; Oldziej, S.; Wedermeyer, W. J.; Gibson, K. D.; Arnautova, Y. A.; Saunders, J.; Ye, Y. J.; Scheraga, H. A. Proc Nat Acad Sci USA 2001, 98, 2329.
9. Simons, K. T.; Strauss, C.; Baker, D. J Mol Biol 2001, 306, 1191.
10. Zhang, Y.; Kolinski, A.; Skolnick, J. Biophysics J 2003, 85, 1145.
11. Zhang, Y.; Skolnick, J. Proc Natl Acad Sci USA 2003, submitted.
12. Duan, Y.; Kollman, P. A. Science 1998, 282, 740.
13. Skolnick, J.; Kihara, D.; Zhang, Y. Protein 2004, in press.