Finding the Needle in a Haystack: Educing Native Folds from Ambiguous *Ab Initio* Protein Structure Predictions

MARCOS R. BETANCOURT, JEFFREY SKOLNICK

Laboratory of Computational Genomics, The Donald Danforth Plant Science Center, 893 N. Warson Rd., Creve Coeur, Missouri 63141

Received 22 May 2000; accepted 18 August 2000

ABSTRACT: Current *ab initio* structure-prediction methods are sometimes able to generate families of folds, one of which is native, but are unable to single out the native one due to imperfections in the folding potentials and an inability to conduct thorough explorations of the conformational space. To address this issue, here we describe a method for the detection of statistically significant folds from a pool of predicted structures. Our approach consists of clustering and averaging the structures into representative fold families. Using a metric derived from the root-mean-square distance (RMSD) that is less sensitive to protein size, we determine whether the simulated structures are clustered in relation to a group of random structures. The clustering method searches for cluster centers and iteratively calculates the clusters and their respective centroids. The centroid interresidue distances are adjusted by minimizing a potential constructed from the corresponding average distances of the cluster structures. Application of this method to selected proteins shows that it can detect the best fold family that is closest to native, along with several other misfolded families. We also describe a method to obtain substructures. This is useful when the folding simulation fails to give a total topology prediction but produces common subelements among the structures. We have created a web server that clusters user submitted structures, which can be found at http://bioinformatics.danforthcenter.org/services/scar.
 J Comput Chem 22: 339–353, 2001

Keywords: protein folding; structure prediction; substructure prediction; clustering methods; relative RMSD

Correspondence to: J. Skolnick; e-mail: skolnick@danforthcenter.org

Contract/grant sponsor: NIH, Division of General Medical Sciences; contract/grant number: GM-37408

Introduction

The prediction of protein structures with *ab initio* methods depends on two main factors. The first is the availability of a reliable potential that singles out the native structure of a given sequence. The second is the efficient exploration of the conformational space guided by the energy landscape. In the ideal structure prediction scheme, the native structure corresponds to the potential global minima. Conformational space can be explored by using a method such as simulated annealing,[1] and the conformation with the lowest energy should be the native state. Unfortunately, the problem is that neither of these factors is fully achievable at the moment.

Most potentials, ranging from the ones that describe protein interactions in full atomic detail to those based on reduced geometries with knowledge-based interactions, can lead to a global energy minimum that does not correspond to the native state.[2] Additionally, the exploration of the conformational space becomes increasingly difficult as the protein size increases, resulting in incorrect global minima predictions. In general, an *ab initio* folding method can lead to a series of structures that may include partially folded structures, folded structures not precisely native, or native structures not at the global energy minimum. This seems to be the case for *ab initio* methods such as the ones presented at the Third Meeting on Critical Assessment of Techniques for Structure Prediction (CASP3).[3–7] It is also the case for a more recently developed method called the Side Chain Only model (SICHO),[8] on which we focus our attention for convenience. As its name suggests, this is a reduced model that represents each residue by a single bead in a high-resolution lattice, with interactions corresponding to those between the side chain centers of mass. The potential is mostly knowledge based and includes local, pairwise, and multibody interactions. Folding simulations with this model can yield hundreds of structures, each of which is a low-energy structure arising from independent folding trajectories. From these structures, one cannot reliably predict the native fold by choosing the structure with the lowest energy because the difference in energies might not be significant enough. Furthermore, it is not immediately evident that the generated structures are representative, which could be indicated by the convergence of a significant number of trajectories to similar topologies. Therefore, a reliable method that can identify the common folds is required. We note that these problems are common to all folding algorithms and are not unique to the SICHO model.

This problem, on a different low-resolution *ab initio* model, was approached using distance geometry techniques.[6,9] In distance geometry, a consensus structure is built from the distances between residues in the given set of predicted protein structures. When all of the structures are used to determine the distances, a large number of incorrect conformations can overwhelm the distance geometry method, resulting in erroneous predictions. Furthermore, the distance geometry method can fail if the predicted structures significantly cluster around more than one topology.

Our approach to this problem consists first in determining if there are significant similarities between structures, next finding the similar structure groups, and then obtaining consensus structures for the groups. To accomplish this, we use clustering methods. Consensus folds are obtained for each cluster by averaging the structures and then, instead of using distance geometry in the traditional way, we adjust the interresidue distances by a simple potential minimization. The potential is built from the average interresidue distances of the structures in a cluster. When no significant clusters are found among the structures, we search for possible substructure clusters. We describe this approach in the following sections and apply it to structures generated by the SICHO model using proteins with known native conformations.

Random Structures: The Reference State

The first step towards analyzing the resulting structures of an *ab initio* folding simulation is to determine whether there are similarities among groups of structures. If there are enough similarities then we can expect structure clusters. Otherwise, if the structures "look" like a collection of random conformations, then clusters of correlated structure are unlikely, and the folding simulation probably failed. The latter can happen either because the conformational search is not thorough enough or because the potential is flawed, or both. Therefore, to be able to tell when a group of structures cluster, we must compare them to the reference state of random structures.

To this end, we would like to utilize a distance measure between structures that indicates whether

these structures look similar or random. The root-mean-square distance [or RMSD, see eq. (1)] is a clear and convenient way of measuring the similarity between structures; however, the significance of its value depends on the size of the structure.[10] For example, a particular RMSD obtained for two large protein structures that indicates similarity may indicate dissimilarity between two smaller ones. Rather, it would be preferable to use a universal similarity measure that is zero between two identical structures and one between random structures, independent of chain length. One possibility is to scale the RMSD by its average value between pairs of randomly selected protein structures of a given length. This requires a precise expression for the corresponding RMSD as a function of chain length. Instead, we define a metric related to the ratio of the RMSD between two structures to a quantity that naturally scales with the size of the structures. This quantity is chosen to have the same average RMSD as the one for random structures in the long chain limit.

More specifically, let $\mathrm{RMSD}_{\alpha\beta}$ be the RMSD between two conformations α and β, both with N residues. Also, $\vec{r}_{\alpha,i}$ let and $\vec{r}_{\beta,i}$ be the respective coordinates of the residues at position i, for $i = 1, \ldots, N$, and with zero average. Without loss of generality, one set of coordinates per residue (e.g., the α carbon coordinates) is used for structure comparisons. Then, $\mathrm{RMSD}_{\alpha\beta}$ can be written as

$$\begin{aligned}\mathrm{RMSD}^2_{\alpha\beta} &= \frac{1}{N}\sum_{i=1}^{N}\left(\vec{r}_{\alpha,i} - Q\vec{r}_{\beta,i}\right)^2 \\ &= R^2_{g\alpha} + R^2_{g\beta} \\ &\quad - 2\left(\frac{\sum_{i=1}^{N}\vec{r}_{\alpha,i}\cdot Q\vec{r}_{\beta,i}}{\sqrt{\sum_{i=1}^{N} r^2_{\alpha,i}}\sqrt{\sum_{i=1}^{N} r^2_{\beta,i}}}\right)R_{g\alpha}R_{g\beta},\end{aligned} \tag{1}$$

where Q is the rotation matrix that optimally aligns the vectors, and $R_{g\alpha}$ and $R_{g\beta}$ are the radius of gyration for structures α and β, respectively. The term in parenthesis is the correlation coefficient between the aligned structures. The correlation coefficient contains the alignment information and for a random group of structures, its average should be independent of structure size, for long enough chains beyond any persistence length. This is, in fact, what we find from actual protein structures. If $\langle\ \rangle_{\alpha\beta}$ is the average over a random ensemble of conformations, then the correlation coefficient for aligned conformations is asymptotically given by

$$\left\langle \frac{\sum_{i=1}^{N}\vec{r}_{\alpha,i}\cdot Q\vec{r}_{\beta,i}}{\sqrt{\sum_{i=1}^{N} r^2_{\alpha,i}}\sqrt{\sum_{i=1}^{N} r^2_{\beta,i}}}\right\rangle_{\alpha\beta} \approx c. \tag{2}$$

From calculations on almost 1300 nonhomologous (with less than 30% sequence identity) random structures from the protein data bank (PDB), we have determined the constant to be $c \approx 2.4^{-1}$. The fact that this correlation does not decay to zero as the chain length increases is the result of the alignment process. If we replace the correlation coefficient in eq. (1) by its average asymptotic value, the equation gives an estimate of the average RMSD between two arbitrary structures of given sizes. From this observation, we define the relative RMSD (RRMSD) as

$$\mathrm{RRMSD}_{\alpha\beta} = \frac{\mathrm{RMSD}_{\alpha\beta}}{\sqrt{R^2_{g\alpha}R^2_{g\beta} - 2cR_{g\alpha}R_{g\beta}}}. \tag{3}$$

We have found that between similar structures of equal lengths the RRMSD is less sensitive to small differences in structure size, but is more discriminant when the size differences are large. This allows the RRMSD to differentiate better between similar and dissimilar structures, in contrast to the RMSD.

The average RRMSD for random structures is not completely size independent. For smaller random proteins, the average RRMSD deviates significantly from unity, as shown in Figure 1. In this plot, the average and standard deviation of the RRMSD were obtained for the PDB random structure set. For a particular sequence length N, a segment of this length was selected at random from each structure, given that the total chain length is larger than or equal to N. The figure shows that the dispersion and average RRMSD converges shortly after 100 residues. Also shown are multiple-exponential fits to these curves. The fits show that the slowest decay length is between 30 and 40 residues, indicating a characteristic length in protein structures caused by strong biases to form secondary structures. This conclusion arises after comparing these curves to similar curves using the ideal and freely jointed chains for which this characteristic length almost disappears.

The distribution of RRMSD values is obtained from the RRMSD between all pairs of conformations with a given chain length. Figure 2 shows RRMSD distribution examples for short and long chain lengths. Notice that for short chains, there is a small maximum at small RRMSD values, indicating possible clustering between α-helices or between β-strands.

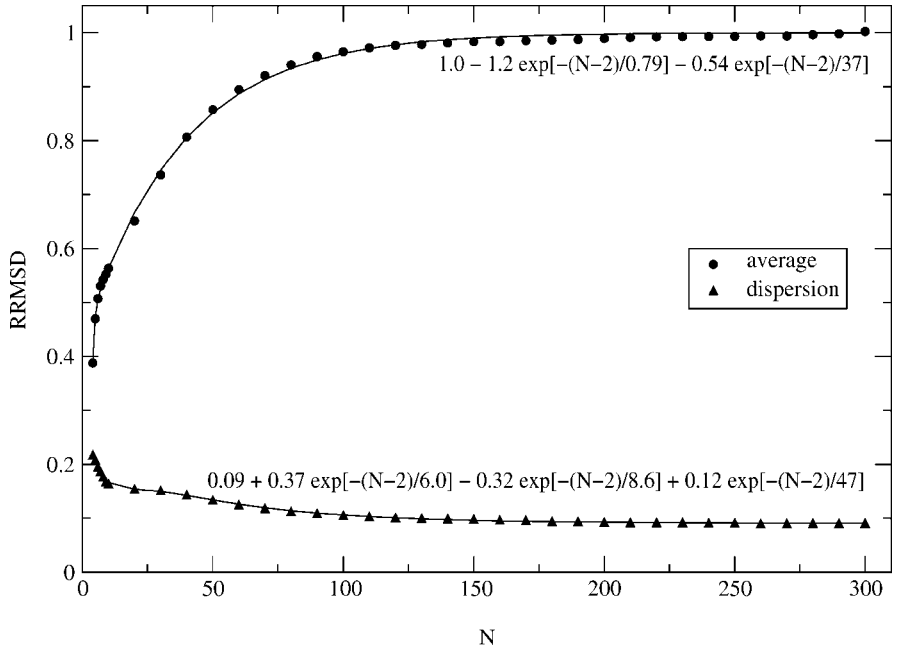

FIGURE 1. Average RRMSD and deviation for polypeptide random structures. The data were obtained from over 1200 nonhomologous protein structures. The solid curves are the fits shown by the formulas.

The RRMSD value allows us to determine when the similarity between two structures is more significant than random. For long enough chains ($N > 100$), if the RRMSD between two structures is near or greater than 1, then they are unlikely to be correlated. For smaller chains, the average RRMSD value for random polypeptides decreases, and the exponential fit shown in Figure 1 must be used.

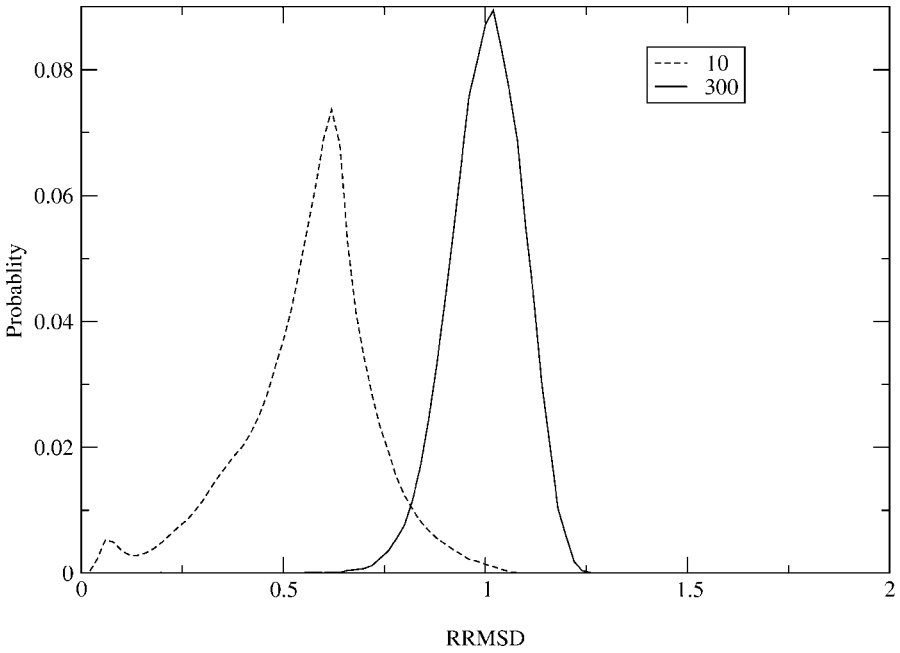

FIGURE 2. Examples of RRMSD distributions for random polypeptides. The solid line corresponds to polypeptide chains with 300 residues while the dashed line corresponds to chains with 10 residues. The probability was obtained by dividing the RRMSD in bins of 0.02. The distribution results from the comparisons between all pairs of structures.

The level of clustering in a group of structures can be determined by comparing its RRMSD distribution to that of a random group of structures. This is estimated by calculating the RRMSD value that gives a particular Z-score in each distribution. The Z-score is defined as the difference between a particular RRMSD value and the mean divided by the standard deviation. As a default, we use a Z-score of -1 for distribution comparisons. For the random case, the RRMSD value corresponding to a Z-score of -1 approaches 0.91 as the chain length increases. If for a given group of structures this RRMSD value is smaller than the one for the random case, then cluster formation is likely.

Structure Clustering Method

The detection of structure groups with similar conformations requires clustering techniques from the theory of multivariate analysis.[11] There are numerous clustering methods that vary in applicability, complexity, and requirements. Here we develop a clustering algorithm of the partitioning class tailored to our particular problem. The basic elements are similar to the K-means clustering method.[11] K-means is a partitioning method in which the average of each cluster is computed to optimize the clusters iteratively. The clusters are optimized by maximizing the correlation between the members of each cluster and minimizing the correlation between members of different clusters.

Overall, our method consists of detecting higher concentrations of similar structures in a space defined by the RRMSD metric. The clusters are determined from these high concentration regions, and then the average structure, or centroid, of each cluster is calculated. The clusters along with their centroids are refined in an iterative process that maximizes a cluster packing measure. When two or more clusters significantly overlap, the best cluster is selected according to a compactness criterion, and then redundant clusters are eliminated. Finally, the centroids are optimized by adjusting the average distance between residues, determined from the clusters where they come from. Figure 3 shows the general flowchart of the algorithm. The following steps describe the algorithm in detail.

1. Compute the RRMSD matrix for all structure pairs and compute the global cluster cutoff. The RRMSD values are stored in a matrix with elements $d_{\alpha\beta}$. A global cluster cutoff, κ, is introduced to represent the distance above which two structures are not likely to be in the same cluster. By default, it is defined as the average minus one standard deviation of the RRMSD distribution for random protein structures with the same chain length. RRMSD values smaller than the cutoff represent structures that are likely to be related. In some folding simulation strategies, the structures may be biased towards a selected group of template structures. The resulting fold predictions may involve variations of some or all of the structure around the template with relatively smaller changes in RRMSD. The default selection of κ must be adjusted to the RRMSD distribution of the structures being clustered. By trial and error, we found that the optimal selection of the default κ is the RRMSD value for two standard deviation *above* the average for the distribution of structures being in the

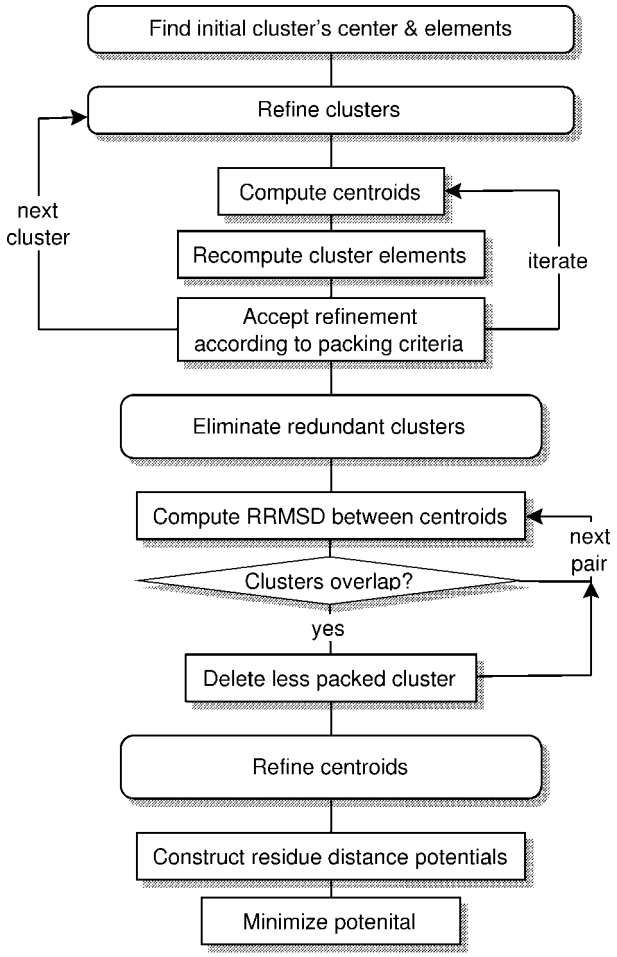

FIGURE 3. Flow chart for the structure clustering method.

cluster, whenever this value is smaller than the initial (from the random structures) κ default value.

2. Compute the structure packing number for all structures. The packing number for structure α is defined as

$$\zeta_\alpha = \sum_{\beta}^{S} e^{-2(d_{\alpha\beta}/\kappa)^2} \qquad \text{for } d_{\alpha\beta} \leq \kappa, \qquad (4)$$

where S is the total number of structures. It conveys the notion of number and compactness by measuring how many structures β are close to α, and how close they are. The contribution of a structure β to the sum is 1.0 if β and α are identical and drops to $\approx$0.35 if $d_{\alpha\beta} = \kappa$. Structures with a large packing number are assumed to be near the center of a cluster.

3. Select the cluster center and members (loop). The structures are initially arranged in order of decreasing packing number, and the number of clusters is set to zero. Then, (a) select the next center as the structure with the highest structure packing number that does not belong to any previous cluster. This selection allows the initial identification of the cluster members from which the cluster centroid can be calculated. By selecting a center structure not in a previous cluster, the new cluster should be significantly different from the previous ones. (b) Compute the local cluster cutoff. This step takes into account the variability in the cluster's internal dispersion. It is computed from

$$\kappa_i^2 = \frac{\sum_{\alpha}^{S_i} \sum_{\beta>\alpha}^{S_i} d_{\alpha\beta}^2 e^{-2(d_{\alpha\beta}/\kappa)^2}}{\sum_{\alpha}^{S_i} \sum_{\beta>\alpha}^{S_i} e^{-2(d_{\alpha\beta}/\kappa)^2}}, \qquad (5)$$

where S_i is the number of structures in cluster i. The Gaussian weight focuses the attention on those structures around the center that are less likely to be randomly related. It can be thought of as a cluster quality control factor. (c) Select the cluster members according to the cluster cutoff. While the global cutoff eliminates structures that are not likely to be in the same cluster, the local cutoff selects structures that are likely to be in the same cluster. Structures with an RRMSD with respect to the center structure below the local cutoff are selected as cluster members. (d) Accept the cluster if there are enough members. A lower limit to the number of members in a cluster is set for efficiency. This number varies with the total number of available structures. Typically, we set its value to about 2% of the total number of structures. (e) Exit the loop if the cluster is too small or there are no more structures to cluster. Because the remaining structures (if any) have smaller packing numbers, it is safe to assume that they all yield small clusters. The number of clusters at this point has reached an upper limit. The remainder of the algorithm could eliminate but will not generate any new clusters.

4. Refine clusters (loop). The structure selected as the cluster center might not be at the center of the cluster, so an iterative process of centroid calculation and cluster member selection is carried out. (a) Obtain the centroid by averaging the cluster structures (loop). The average is done by aligning the structures using the singlular value decomposition method (SVD),[12] as in the RRMSD calculation. The following steps are taken: (i) select an initial structure as the centroid; (ii) align a new structure to the centroid using SVD; (iii) add the aligned structures in a separate sum and compute the new centroid from the sum of structures; and (iv) repeat for all cluster structures. The centroid is independent (up to rotations) of the order in which the structures are added. (b) Recalculate the cluster members. As the position of the centroids varies, some new structures can join the cluster while current ones can leave. In particular, this takes into account clusters for which their centroid is not near to any of the structures. The selection of the cluster members is done as follows: (i) compute the RRMSD between the centroid and all structures; and (ii) select cluster structures with an RRMSD below the cluster cutoff. (c) Calculate new cluster cutoff. The local cluster cutoff is updated for the new cluster members using eq. (5). (d) Compute the cluster packing number. The cluster packing number is a similar measure to the structure packing number, but it takes all cluster structure pairs into account. It is defined as

$$\eta_i = \frac{2}{S_i - 1} \sum_{\alpha}^{S_i} \sum_{\beta>\alpha}^{S_i} e^{-2(d_{\alpha\beta}/\kappa_i)^2}. \qquad (6)$$

The packing number is used as a measure of cluster size and tightness. (e) Update the cluster if it is better packed. Accept the refined cluster if its packing number is larger than before its refinement; otherwise, keep the pre-

vious cluster. When the packing number is maximized, continue to the next cluster.

5. Eliminate redundant clusters. During the cluster refinement process, some clusters become identical or almost identical. The clusters are compared and eliminated as follows: (a) compute the RRMSD between the centroids of two clusters. (b) Decide if the clusters overlap. Two clusters are defined to overlap if their centroids fall within each other's cutoffs. That is, two clusters overlap if the RRMSD_{ij} between the centroids of clusters i and j satisfies the conditions $\text{RRMSD}_{ij} < \kappa_i$ and $\text{RRMSD}_{ij} < \kappa_j$. This criterion is consistent with the initial selection of cluster centers (see step 3a). However, it allows for one cluster to be contained within another, as long as the smaller cluster does not include the centroid of the larger one. (c) If two clusters overlap, delete the cluster with the smallest packing density. From the packing number, we define the packing density as

$$\rho_i \equiv \frac{\eta_i}{S_i}. \tag{7}$$

The packing density is exclusively a measure of cluster tightness. Note that while in step 4e we optimized the cluster size and tightness, here we select the tighter clusters only. (d) Repeat for every pair of clusters.

6. Refine centroids. The centroids obtained from the clustering process capture the global geometry of the clustered structures. However, due to variations in the structure alignments, the details can be averaged out. To recover structural fidelity, we adjust the distance between residues to approach the average values in the cluster structures. This is achieved through the following potential minimization procedure. (a) For all the structures in a cluster, compute the average and the dispersion of the distances between each pair of residues. We let the average distances between residues i and j be Δ_{ij} and the dispersion be σ_{ij}. (b) Construct harmonic potentials for each pair of residues from the average distances and dispersions. The potential between residues i and j is defined as

$$V_{ij} \equiv \frac{1}{2}\frac{(|\vec{r}_i - \vec{r}_j| - \Delta_{ij})^2}{\sigma_{ij}^2}. \tag{8}$$

Within each cluster, highly conserved distances have a larger "spring" constant ($1/\sigma_{ij}^2$) than more variable ones. (c) Minimize the potential using the centroid as the initial condition. The total distance constraint potential

$$V = \sum_{i,\, j>i} V_{ij} \tag{9}$$

is minimized using the conjugate gradients method.[12] The resulting structure is the desired representative structure of the cluster.

At the algorithm output, one obtains the refined centroid for each cluster and the structures for each cluster. Cluster quality can be assessed from the average RRMSD among structures, the cluster cutoff, and their packing density. To ensure consistency with physical constraints, a final step can be included that consists of the reconstruction of the atomic details. To the distance constraint potential, several terms describing the bonded constraints and excluded volume can be added. A final centroid refinement can be achieved by minimizing this potential. We have found that this additional step does not significant modify the centroid's geometry, although it eliminates obvious structural errors (such as residue overlaps) produced by the structure averaging procedure. We do not discuss this final step here.

Examples of Structure Clustering

As an example, we consider the folding simulations carried out with the SICHO folding algorithm[8] on the ribosomal binding protein (1ctf). This is a short monomeric protein (68 residues) for which the structure can be predicted reasonably well with the SICHO model. The folding simulations considered in this example resulted in 430 minimum energy structures from independent annealing trajectories.

The prediction results can be summarized in a plot describing the correlation between the structure's energy (in units of the model special potential) and the RMSD to the native structure described in the protein data bank (PDB). This plot is shown in Figure 4. The simulations yield an optimal prediction at 3.58 Å from the actual native conformation. However, it is evident that this structure cannot be identified from the lowest energy structure. In fact, there are 66 structures with lower energy than the one with the optimal RMSD. The lack of a strong correlation in these simulations, particularly at small RMSD values, is indicative of the folding model's inadequacy to identify the native conformation as the lowest energy structure. The difficulty

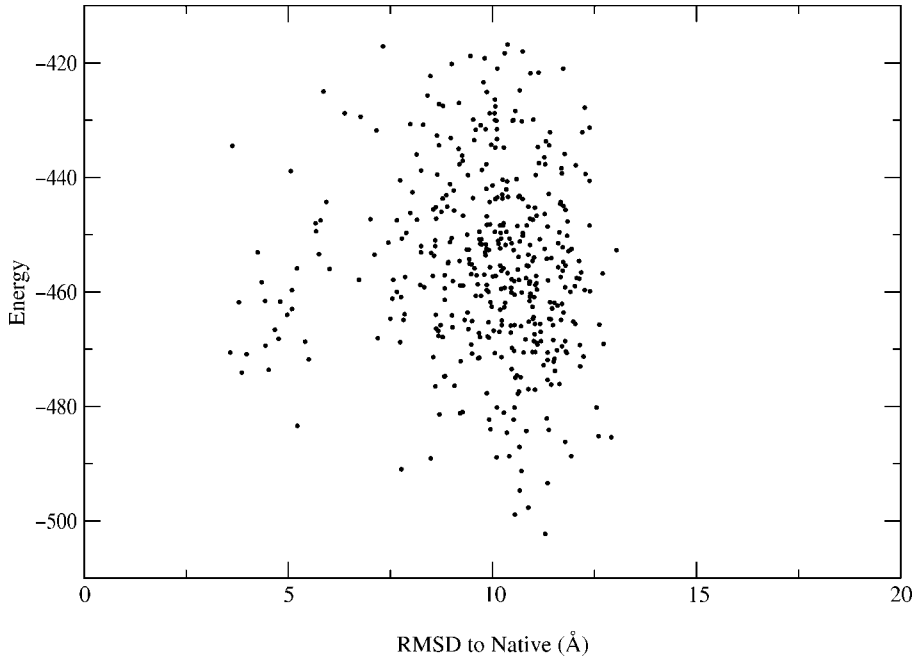

FIGURE 4. Energy and native RMSD correlation for the 1ctf folding simulations. Each point represents one of 430 structures.

in this case arises both from defects in the potential as well as the search scheme. Nevertheless, the appearance of small RMSD structures indicates that native-like conformations are being generated, although at a somewhat higher energy.

Representative conformations, including the native-like one, can be identified if there is significant clustering around them. In Figure 5, we show the RRMSD distribution for the 1ctf structures under consideration and the distribution of a random group of polypeptide segments with the same length as 1ctf. It is clear that there is significant clustering between the structures, as indicated by the lower RRMSD values in relation to the random ones. The RRMSD for 1ctf at the average value of the distribution minus one standard deviation is 0.62, which is less than that of a random structure with 68 residues, or 0.79. Therefore, we can be confident that the folding simulations significantly converged to at least one representative structure, and the clustering analysis can produce meaningful results.

The clustering analysis of the example results in three distinctive clusters. The main properties of these clusters are listed in Table I. For the three clusters, the average RRMSD is significantly smaller than the global cutoff (a difference of about 0.3 smaller), indicating that the structures within them are well correlated. Not indicated in the table is that none of the clusters have structures in common. There is one dominant cluster with more structures (125) than the other two (55 and 33). This cluster has the lowest energy of the three clusters.

The correspondence of the centroids to the native structure is described in Table II. The native structure corresponds to centroid 2, which has the second highest packing density. In terms of the RMSD, centroid 2 is as close to the native structure (3.55 Å) as the best-input structure (3.58 Å). The cluster corresponding to centroid 1 contains most of the lowest energy structures. The centroid conformations are shown in Figure 6. Note that they consist mainly of different global arrangements of mostly similar secondary structure elements.

In addition to 1ctf, we applied the clustering algorithm to simulations of 25 other monomeric proteins (1aba, 1bbhA, 1c5a, 1cewI, 1cis, 1ego, 1fas, 1fc2C, 1ftz, 1gb1, 1gpt, 1hom, 1ife, 1lea, 1mba, 1poh, 1pou, 1shaA, 1stfI, 1tlk, 256bA, 2azaA, 2pcy, 2sarA, 5fd1). In an attempt to make a fair comparison, we compare the best of the three lowest energy centroids against the best of the three lowest energy (unrelated) structures. On average, the best centroids are almost 1 Å closer to the native structure than the best lowest energy structure. In three cases (1ego, 1poh, 1lea) the best lowest energy struc-

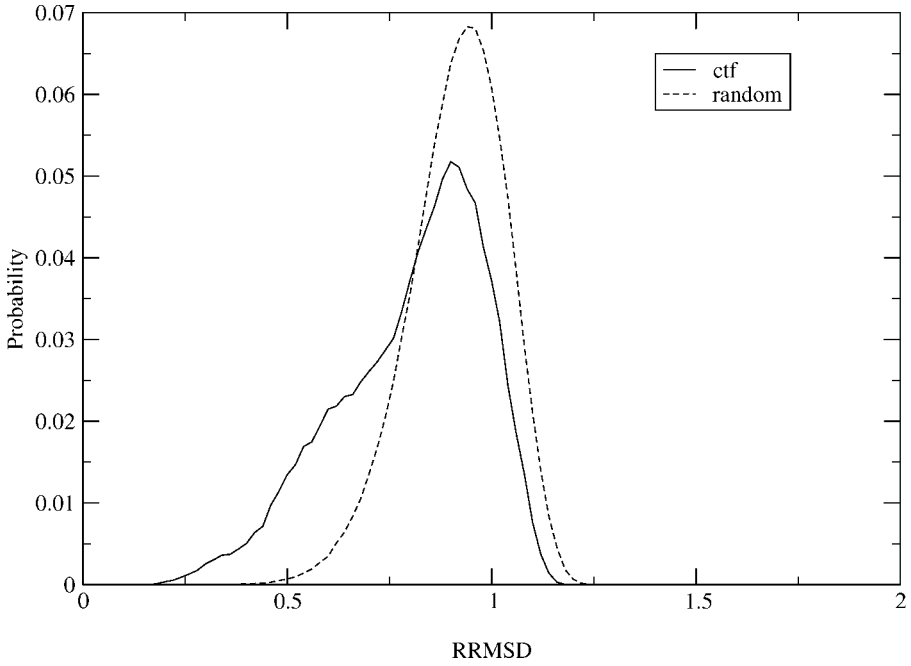

FIGURE 5. RRMSD distributions for the 1ctf folding simulation. The dashed line corresponds to random polypeptides of the same length as 1ctf.

tures were between 0.5 to 1.0 Å better than the best centroids. In one case (1ego), the best RMSD structure corresponds to the lowest energy structure, and is significantly different from the rest of the structures. Therefore, clustering for this structure is bound to fail. On the other hand, four centroids (1fc2C, 2azaA, 1ife, 1gpt) were better by approximately 3 Å than the best low-energy structures, and seven others between 0.5 and 1.5 Å better than the best low-energy structures.

For comparison, we repeated the centroid determination using the popular complete-linkage hierarchical clustering method.[11] In this case, the same metric (RRMSD) was used, and the centroids were calculated after the clusters were determined. One property of the hierarchical approach is that the number of clusters is more sensitive to the selection of the RRMSD cutoff value. In this case, we found optimal to use a cutoff corresponding to one standard deviation above the mean of the RRMSD distribution for the structures being clustered, which roughly corresponded to a value of ≈1.1. In the complete-linkage case, this cutoff means that no two structures in a cluster are farther apart than about 1.1 RRMSD. For many of the proteins considered, the centroids obtained from this approach were significantly similar to the ones obtained from our partitioning approach. This is indicative of well-defined clusters that can be obtained from various methods. As far as the RMSD to the native struc-

TABLE I.
Cluster Properties for the 1ctf Simulated Structures.

i	⟨Energy⟩	S_i	ρ_i	κ_i	⟨RRMSD⟩	⟨RMSD⟩
1	−463.14	125	0.44	0.46	0.51	5.66 Å
2	−455.58	33	0.44	0.53	0.51	5.68 Å
3	−449.77	55	0.40	0.54	0.54	6.05 Å

⟨Energy⟩ is the average energy of all the structures in the cluster. S_i is the number of structures in a cluster. ρ_i is the packing density. κ_i is the cluster cutoff. ⟨RRMSD⟩ and ⟨RMSD⟩ are the average RRMSD and RMSD, respectively, between each cluster structures.

TABLE II.
Comparisons of 1ctf Cluster Centroids with the PDB Structure.

i	RRMSD	RMSD	DRMSD
1	0.95	10.60 Å	5.06 Å
2	0.32	3.55 Å	2.69 Å
3	0.82	9.21 Å	4.96 Å

The columns correspond to the relative, standard, and distance RMSD.

ture determines the adequacy of the method, we again compare the best three centroids by using both clustering methods. On average, the results for the hierarchical method yielded structures almost 1 Å farther (6.9 vs. 6.0 Å) away from the native structure than our partitioning approach. Only in a few cases were the centroids from the hierarchical approach significantly better than the centroids from the partitioning approach. We conclude that the cluster refinement process in our partitioning approach improves the quality of the clusters and centroids, and yields more native like centroids.

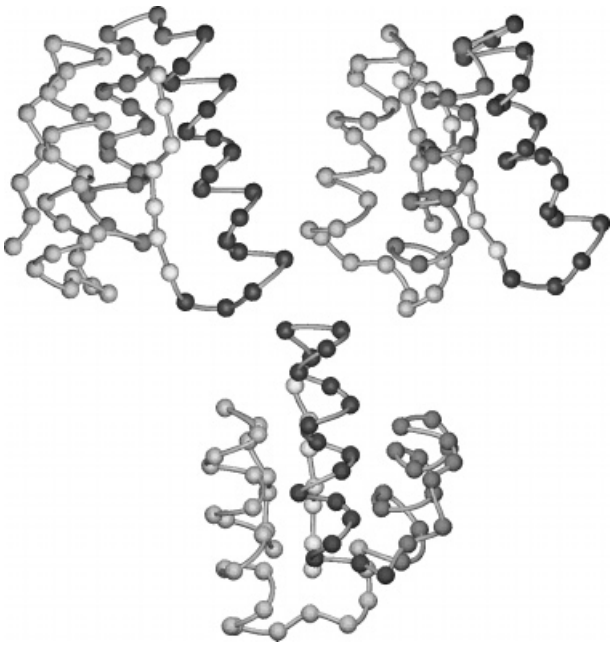

FIGURE 6. Centroids for 1ctf folding simulations. Structure (2) is the one closer to the actual native structure.

Substructure Determination

When the folding simulations fail to yield structures that cluster, it is possible that some common (and perhaps correct) folded substructural elements are present. In some cases, correct substructures fold before the global structure folds and the rate-determining step is the assembly of the substructural elements. It is also possible that the native state of a protein contains highly mobile segments that fail to fold to a definite structure, but it also contains other stable parts that do fold correctly. In such cases, it would be useful to detect folded substructures in the form of secondary or super-secondary structural motifs.

Ideally, we would like to find the residues that form consistent substructures and their resulting substructures by using clustering techniques. In general, this can be a very complicated problem because we would need to compare all the structures formed by all possible combinations of residues. Instead of searching for substructures in structural space, we search for them in residue space. In this space, the metric is a measure of how consistent the residue positions are in relation to each other. With this approach, a substructure can be identified by clustering the residues with relative positions that are significantly conserved.

More specifically, the metric that we use in residue space is defined as

$$D_{ij} \equiv \sqrt{\sum_{\alpha} \left(|\vec{r}_{\alpha i} - \vec{r}_{\alpha j}| - \Delta_{ij}\right)^2} \qquad (10)$$

for two residues i and j, where the sum runs over all structures in a given group of structures and Δ_{ij} is the corresponding average distance. In this space, two residues i and j are considered to be in the same position (at zero distance D_{ij}) when their structure space distance $|\vec{r}_{\alpha i} - \vec{r}_{\alpha j}|$ is constant within the structure group.

In principle, some substructures could be common to all structures, while others could be present only in some total-structure clusters. Note that, to avoid confusion, we are now referring to the structure clusters described in the previous section as being total-structure clusters. Therefore, the structures used in residue clustering are chosen either to detect global substructures or local ones. For detecting global substructures, a "super" cluster is created by combining all total-structure clusters. The structures that do not cluster are discarded as statistical noise. Each total-structure cluster (global or local)

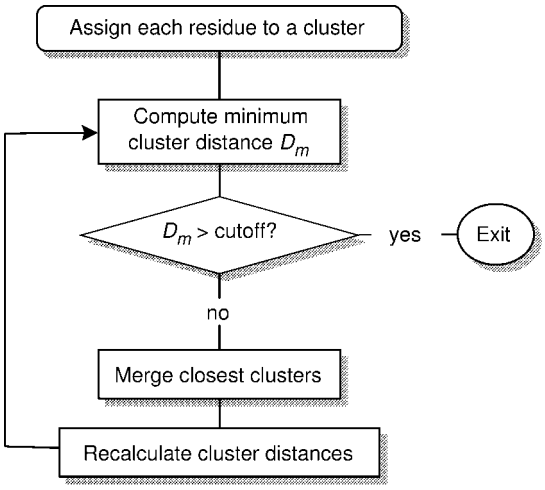

FIGURE 7. Flow chart for the substructure clustering.

defines a particular metric D_{ij} for which the substructures are obtained.

The clustering method derived previously is not easily applicable in residue space. Therefore, we use the more generally applicable method of average-linkage hierarchical clustering.[11] A flow chart with the algorithm general steps is shown in Figure 7. In a hierarchical clustering method, larger clusters are gradually formed by merging smaller clusters. The process starts with as many clusters as there are residues, i.e., each residue is initially a cluster. The initial distance between each cluster is set to the distance between their corresponding residues (i.e., D_{ij}). This distance is used as the criteria for merging clusters in a series of steps. In each step, the closest pair of clusters is determined and combined into a single larger cluster, resulting in a reduction of the total number of clusters. For this process to be repeated, the distance between the newly formed cluster and the other clusters must be determined. In average-linkage clustering, the distance between two clusters is calculated by averaging the distance D_{ij} between all the members of one cluster with all the members of the other. The process of merging clusters is repeated until the minimum distance between clusters reaches a desired cutoff value. This method is relatively easy to implement, but the clusters formed can be more sensitive to metric fluctuations. The average-linkage method produces substructure clusters for which the average distance between any two residues in a cluster is smaller than or equal to a given threshold. This threshold can be set to correspond to the approximate fluctuating error one would allow for the position of residues within a structure. For example, if one wishes to determine the substructures to residue resolution, the threshold may be in the neighborhood of 3.8 Å. Alternately, the threshold can be set to vary according to the distribution of D_{ij} values between all residues. In particular, setting the threshold to the distribution average results in substructures of significant size. This choice is convenient in that it self-adjusts to the fluctuations of the cluster structures, although it can generate substructures with significant fluctuations.

The resulting residue clusters are used to construct the substructures by analogy to the centroid calculations of the total-structure clusters. The coordinates of these residues are aligned using SVD, independently of the remaining residues, and their average is computed. When the total structure fails to cluster significantly, the substructure averages should give a better representation of the substructures. The advantage is that only residues with consistent positions are being aligned and the fluctuations are reduced.

To test the substructure determination method, we analyze the folding simulations of one monomer of the cochaperonin GroES. This monomer consists of 97 residues and has a mobile loop between residues 17 and 32.[13] The structure of this loop is stabilized in the presence of the GroES companion, GroEL. Therefore, at best, only part of this structure can be predicted by folding simulations of the monomer in isolation.

The folding simulations generated 858 structures from which two major clusters were obtained. The average RRMSD for both clusters is approximately 0.85, indicating that the global structures are close to random. The RRMSD between the centroids and the native GroES structure in the GroES/GroEL complex is 0.99 and 1.01 for the centroids. Evidently, there is no global correspondence between the predicted structures and the known GroES structure.

To determine the substructures, the cluster cutoff value was set according to the distribution of distances. For the global cluster, the default cutoff resulted in a value of 4.7 Å, indicating significant fluctuations. The properties of the resulting substructures for the combined total-structure clusters are summarized in Table III. The substructures contain between 11 and 36 consecutive residues. Note that in general, the residues are not necessarily consecutive. The second column shows the average RRMSD between substructures in the clusters. Because of the relatively small number of residues, we show the average RRMSD divided by its random value in column 3. Despite the large fluctuations among the structures in total-structure clus-

TABLE III.
Substructure Properties for the Combined Clusters in the GroES Simulations.

Residues	⟨RRMSD⟩	⟨RRMSD⟩/RRMSD^o	RRMSD	RMSD	DRMSD
1–16	0.58	0.88	0.34	3.69 Å	2.40 Å
17–33	0.47	0.70	0.75	8.63 Å	9.41 Å
34–50	0.48	0.72	0.35	4.66 Å	2.41 Å
51–61	0.41	0.68	0.40	4.00 Å	1.13 Å
62–97	0.65	0.80	0.86	11.17 Å	7.37 Å

The combined clusters were obtained by combining the structures of the two total-structure clusters (184 total structures). The average of the distance distribution criteria was used in the selection of the substructure clusters cutoff, resulting in a cutoff value of 4.71 Å. The average RRMSD values are computed between the substructures in each substructure cluster. The last three columns are in relation to the known GroES structure, in the GroES/GroEL complex.

ters, most substructure clusters contain correlated structures.

The last three columns of Table III show the comparison to the corresponding portions of the known GroES structure. The geometries for both cases are shown in Figures 8 and 9. Figure 8 corresponds to the structures with the smallest average RRMSD values, and Figure 9 to the ones with the largest average RRMSD values. Interestingly, the mobile loop is singled out by the substructure analysis, as shown in Figure 8a. Not surprisingly, the mobile loop prediction is unrelated to the native one. However, the folding simulations did not show significant structure fluctuations for the loop in comparison to the other substructures, as indicated by the average RRMSD. Nevertheless, the other two substructures with small average RRMSD (Figs. 8b and c) are generally similar to their native counterparts. The predicted substructures with the largest average RRMSD values (Fig. 9) are also similar, to some extent, to the native ones. Figure 9a corresponds to the substructure with the smallest RRMSD despite coming from a cluster with considerable fluctuations. The substructures shown in Figure 9b resemble a β-hairpin. The predicted substructure is comparable to the mirror image of the native one, with the exception that the contacting residues between strands are shifted in relation to each other.

The substructures analysis applied to the structures of a particular total-structure cluster results

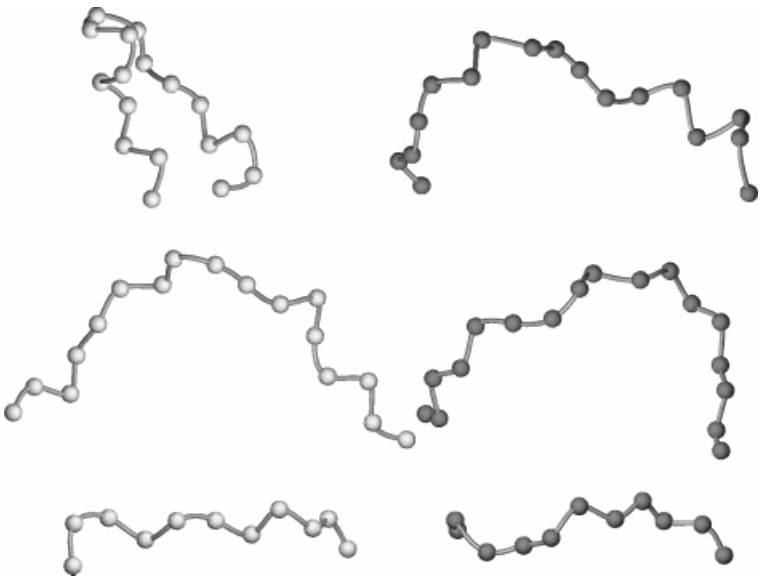

FIGURE 8. Substructures of lower ⟨RRMSD⟩. The light (left) and dark (right) substructures correspond to the native GroES and clustered substructures, respectively.

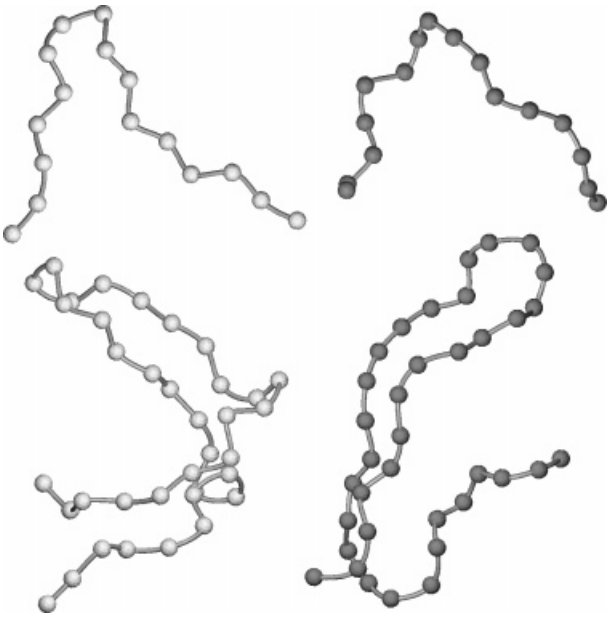

FIGURE 9. Substructures of higher ⟨RRMSD⟩. The light (left) and dark (right) substructures correspond to the native and clustered substructures, respectively.

in similar structures to the ones obtained by using the combined total-structure cluster. In some total-structure clusters, the substructures are better correlated, as indicated by smaller average RRMSD values between them. In others, the substructures are just a combination of the substructures obtained from the combined clusters. In general, the substructures in total-structure clusters can be different from the one in the combined clusters.

World Wide Web Structure Clustering Server

We have created a World Wide Web (WWW) structure-clustering server that can be presently reached at http://bioinformatics.danforthcenter.org/services/scar. There, a user can submit a group of structures and obtain, via e-mail, the results of the structure and substructure clustering analysis. The calculation of the clustering algorithm is carried out in our local machines dedicated as public WWW servers.

The algorithm requires a few parameters, most of which are adequately selected from the default values. The first set of input parameters describes the structure file formats. The structure coordinates can be given either in PDB format or in the SICHO format[8] (see the Web page for details). Examples for both formats are shown in the server. Only a single atom representation of the residues is used for clustering calculations. The second set of input parameters allows changing the clustering default options. When the structure files consist of a series of structures from independent folding simulation trajectories, the user can specify how to select the structures from these files. The options are the minimum energy structure, the structures obtained from clustering the trajectory file (see the concluding remarks sections), or the particular structure indicated by the user. The user must also specify the global cluster cutoff, and for substructures analysis, the residue cluster cutoff. The last input section consists of the submission of the structures. These can be submitted in a number of different file compression formats.

Concluding Remarks

In this work, we have developed a method for analyzing the results of *ab initio* structure predictions and for extracting the significant folds, from where the native state could be obtained. The procedure consists in three parts. First, the variety of structures resulting from the folding simulations are compared to a reference state of random structures to determine if the simulations were successful in generating significant folds. To this end we introduced the relative RMSD, which provides a general measure of similarity between structures. Second, we cluster the simulated structures and find their average structures, representing the significant folds. Finally, we analyze the common substructures (or protein domains) appearing in the folded structures, which can be useful when the global fold fails to cluster.

The relative RMSD allow us to compare structures by an almost universal (size independent) scale. In this scale, two structures with an RRMSD of one are uncorrelated, regardless of structure size. The measure is also useful for detecting significant correlations among a group of predicted protein structures obtained from *ab initio* folding simulations. The RRMSD distribution for random polypeptides converges to a universal curve with a mean of 1.0 and a standard deviation of approximately 0.09. This convergence is significantly attained for chains of more than 100 residues. By comparing the distribution of RRMSD values of the predicted structures to those of a random group of polypeptides, we can determine whether the simulation was adequate or if it requires a better sampling of the conformational space. If the simulation is adequate, the structures can be grouped into representative clusters.

There are several benefits to clustering simulated structures. One is that it narrows down the predictions, from hundreds or thousands, to just a few characteristic structures. We have found that there are much better chances of finding the optimal fold between this small group of structures than from a similar number of the lowest energy structures. Optionally, other types of scoring functions or potentials can be used among this set of structures to identify the native one. Another benefit of clustering is that, whenever the native structure is known, it allows for the detection of typical incorrect folds that can be used to analyze and improve the potential.

The clustering method we have developed proves to be very successful in detecting significant structures. The use of structure averaging in the clustering algorithm serves a dual purpose. First, it allows for a better and more robust determination of the clusters. This is in contrast to popular hierarchical clustering methods, which, to maintain generality, may not take advantage of a system's special properties. Second, the centroids produce the global consensus topology of a cluster that is used as the initial structure in the distance refinement optimization. A feature of our method is that it allows clusters to overlap. In this way, structures that are a combination of more than one characteristic structure can belong to several clusters. Therefore, the selection of the structures in a cluster does not depend on other clusters and their construction is more independent and robust. Another feature of the clustering method is that the cutoff that determines the cluster size is locally adjustable for each cluster. If the cutoff was determined globally instead, there is a possibility that a large cluster (in the RRMSD sense) is truncated while, on the other hand, more than one small cluster can be combined into a single cluster. The results show that the cluster cutoffs vary from one cluster to another, possibly indicating size fluctuations of the basin boundaries around the clusters in the energy landscape. The packing density ρ, defined by eqs. (6) and (7), seems to be a useful quantity in classifying the quality of a cluster, as long as the cluster is significant in size. For the cases studied, the native structure is among the clusters with the highest ρ values, although this result probably depends on the fidelity of the potential and the prediction methods. Note that it is important to cluster structures that come from statistically independent simulations. Otherwise, the clustering method will only detect artificial correlations. In some minimization techniques, the folding trajectories will include both artificial and relevant correlations. For this case, we cluster the structures in each trajectory and then cluster the resulting centroids of all trajectories.

As in many partitioning clustering methods, our method maximizes the correlation between elements in a cluster and minimizes the correlation between elements of different clusters. The correlation between elements in a cluster is captured by the packing number and packing density, eqs. (4) and (7), respectively. These are direct measures of the dispersion errors in each cluster. The minimization of correlations between different clusters is obtained from the overlap condition. It states that the distance between two centroids must be larger than the dispersion for each of the corresponding clusters.

Using substructure cluster analysis, we are able to detect well-defined substructures even when the total structures fail to cluster. The substructures are in the form of secondary and some time supersecondary structures, depending on the quality of the total cluster. One of the purposes of the substructure cluster analysis was to detect the substructures of proteins with globally fluctuating or unstable domains. As an example, we analyzed the substructures of the folding simulations of GroES, which has been determined by NMR studies to have an unstable domain.[13] In this case, our clustering analysis was able to detect the unstable domain even though the folding algorithm did not label the domain as unstable, as indicated by the consistency of the substructure (low average RRMSD) in relation to the other substructures. In our analysis, we studied the substructures for the structures belonging to each individual total-structure clusters, and to the combined cluster formed by the union of the individual clusters. We found in this case that there were no significant differences between the individual and combined cluster substructures. This is a reflection of the folding simulations, showing that the differences of the various substructures lie mostly in the rearrangements of substructures. That is, the global topology has relatively small effects on the substructures. In general, whenever the total structures fail to cluster, the quality of the substructures averaged individually is superior to the one appearing in the total-structure centroids.

Clustering in residue space, with the constraint of selecting structures from total-structure clusters, is an effective way of finding the substructures. The average-linkage hierarchical method seems adequate enough for this purpose. Nevertheless, the boundaries and size of the substructures are somewhat arbitrary, and a refinement of the substructure clusters could improve the results.

Acknowledgments

We would like to thank Andrzej Kolinski for useful insights into this problem. We are also grateful to Daisuke Kihara for carrying out the folding simulations.

References

1. Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Science 1983, 220, 671.
2. Ortiz, A. R.; Kolinski, A.; Skolnick, J. Proteins 1998, 30, 287.
3. Simons, K. T.; Bonneau, R.; Ruczinski, I.; Baker, D. Proteins Suppl 1999, 3, 171.
4. Ortiz, A.; Kolinski, A.; Rotkiewicz, P.; Ilkowski, B.; Skolnick, J. Proteins Suppl 1999, 3, 177.
5. Osguthorpe, D. J. Proteins Suppl 1999, 3, 183.
6. Samudrala, R.; Xia, Y.; Huang, E.; Levitt, M. Proteins 1999, 3, 194.
7. Lee, J.; Liwo, A.; Ripoll, D. R.; Pillardy, J.; Scheraga, H. A. Proteins Suppl 1999, 3, 204.
8. Kolinski, A.; Skolnick, J. Proteins 1998, 32, 475.
9. Huang, E.; Samudrala, R.; Ponder, J. Protein Sci 1998, 7, 1998.
10. Reva, B. A.; Finkelstein, A. V.; Skolnick, J. Fold Design 1998, 3, 141.
11. Gnanadesikan, R. Methods for Statistical Data Analysis of Multivariate Observations; John Wiley & Sons, Inc.: New York, 1997, 2nd ed., Chap. 4.
12. Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. Numerical Recipes: The Art of Scientific Computing; Cambridge University Press: New York, 1986, Chap. 2.
13. Landry, S. J.; Zeilstra–Ryalls, J.; Fayet, O.; Georgopoulos, C.; Gierasch, L. M. Nature 1993, 368, 255.