Selecting High Quality Protein Structures from Diverse Conformational Ensembles

Ashwin Subramani, Peter A. DiMaggio Jr., and Christodoulos A. Floudas*
Department of Chemical Engineering, Princeton University, Princeton, New Jersey

ABSTRACT Protein structure prediction encompasses two major challenges: 1), the generation of a large ensemble of high resolution structures for a given amino-acid sequence; and 2), the identification of the structure closest to the native structure for a blind prediction. In this article, we address the second challenge, by proposing what is, to our knowledge, a novel iterative traveling-salesman problem-based clustering method to identify the structures of a protein, in a given ensemble, which are closest to the native structure. The method consists of an iterative procedure, which aims at eliminating clusters of structures at each iteration, which are unlikely to be of similar fold to the native, based on a statistical analysis of cluster density and average spherical radius. The method, denoted as ICON, has been tested on four data sets: 1), 1400 proteins with high resolution decoys; 2), medium-to-low resolution decoys from Decoys 'R' Us; 3), medium-to-low resolution decoys from the first-principles approach, ASTRO-FOLD; and 4), selected targets from CASP8. The extensive tests demonstrate that ICON can identify high-quality structures in each ensemble, regardless of the resolution of conformers. In a total of 1454 proteins, with an average of 1051 conformers per protein, the conformers selected by ICON are, on an average, in the top 3.5% of the conformers in the ensemble.

INTRODUCTION

The Protein Structure Prediction problem is one of the most challenging problems in molecular and systems biology. The main aim is to predict the final three-dimensional structure of a protein, given only its amino-acid sequence. To address this problem, various methods encompassing a wide array of techniques are used. Protein structure prediction methods are broadly classified as homology-based methods, fold recognition techniques, and first-principles based methods. Recent detailed overviews of these methods are available elsewhere (1–3).

A number of techniques, spanning a wide variety of fields, have been used for the identification of near native folds. These can be broadly classified as force-field based techniques and clustering techniques. Force-field based techniques aim at capturing the energetic interactions that occur in proteins either through physics-based energy functions, or through knowledge-based potentials. CHARMM (4), AMBER (5), ECEPP (6), ECEPP/3 (7), UNRES (8), and ECEPP-05 (9) are examples of some physics-based potentials. A significant amount of research has been dedicated toward optimizing the weight parameters of the physics-based force fields, in order to increase the correlation between the potential energy of the protein and the nearness to the native fold (10,11). Knowledge-based force fields are usually calculated using two different approaches. One approach uses the Boltzmann equation, which is based on the idea that lower energy states are more frequently observed. The second approach is based on parameter estimation, which aims to represent the amino acids of a protein either as a single atom, or as a group of atoms. Parameters are estimated either based on distance between specific atom pairs or on the identity of amino acids, which are trained to ensure that the native structure has an energy value much lower than the decoy structures. Distance-based force fields, based on C_α-C_α, Centroid-Centroid (12–16), or all-atoms (17,18), have been shown to be successful in identifying the native structure from a large ensemble of near-native structures.

A second approach to the identification of near-native folds is clustering. Problems of data clustering and organization are pervasive over a number of disciplines. The most common approaches can be classified as either hierarchical (19) or partitioning (20) clustering. A number of other frameworks for clustering have also been proposed, including model-based clustering (21), neural networks (22), simulated annealing (23), genetic algorithms (24), and data classification (25). Most algorithms use heuristics for their searching procedures, which may result in suboptimal clustering because of analysis of only local comparisons. Recent works have presented a novel clustering approach based on global optimum search (26), which includes a procedure to determine the optimal number of clusters to be used (27–29). Clustering methods have been previously also used as a part of loop structure prediction algorithms (30), where the aim is to eliminate loop structures, which are unlikely to be close to the native structure in an iterative manner.

The field of rearrangement clustering has emerged as a very effective technique for optimally minimizing the dissimilarity metric between the data points in large distance matrices. Recently, a rigorous global optimization method for biclustering biological data was introduced (31). This method, denoted as Optimal RE-Ordering of rows and columns (OREO), is based on optimal reordering of the rows and columns of a data matrix to globally minimize the

Submitted April 13, 2009, and accepted for publication June 30, 2009.

*Correspondence: floudas@titan.princeton.edu

Editor: Costas D. Maranas.

0006-3495/09/09/1728/9 $2.00

doi: 10.1016/j.bpj.2009.06.046

dissimilarity metric, by formulating the physical permutations of rows and columns as either a network flow problem or the traveling salesman problem (TSP) (32). Highly favorable results were presented when the method was tested on several sets of biological and image reconstruction data.

A number of methods have been proposed, which cluster the decoys of a protein based on some mutual distance metric, and then aim to find the structure that has the highest number of similar structures. Shortle et al. (33) proposed a pairwise root mean-square deviation (RMSD)-based clustering method to this effect. The authors presented a scoring function to rank the quality of the decoys. This scoring function aims to predict the probability that the given sequence would fold into the decoy structure. The prior probability of the structure is derived from excluded volume and packing terms. The likelihood term in the Bayesian expression is derived from hydrophobic and pairwise interactions such as salt bridges and disulfide bonds. Based on such an elimination criterion, the method reduces the working ensemble to the 1000 best scoring structures. Based on a (1000 × 1000) pairwise RMSD matrix, the structure with the most neighboring structures is selected. SPICKER (34), a state-of-the-art simple and efficient method to identify near-native folds, also follows a similar idea. This method takes into consideration the fact that depending on whether it is a new fold, or an existing one, most structure prediction techniques are likely to produce a wide or narrow ensemble of structures, respectively. Hence, SPICKER modifies the radius of cutoff for the definition of a cluster. The top five clusters in terms of size are selected, and the cluster centroids and medoids are suggested as the structures closest to the native.

The idea of dihedral-angle based clustering of protein structures has also been investigated by researchers. Dihedral angles provide a good representation of the protein structure itself, since in the dihedral angle space; we can assume two degrees of freedom for each amino acid. Circular clustering is the most effective way of handling dihedral angles (35). The idea of circular clustering is to identify the fact that for a dihedral angle, +180° and –180° are the same. Hence, objective functions defining the dissimilarity metric should reflect this property.

A potential source of error in structure prediction algorithms comes from the misprediction of the topology of the target protein. This is especially a problem for structure prediction using homology-based algorithms, which rely on sequence and structural homologs found by metaservers. In such cases, any clustering method that uses the predicted structural ensemble as the starting point, without any prior knowledge of the native structure, is likely to concentrate, and hence predict, an incorrect decoy structure as the one most likely to be near-native. Situations such as these are especially detrimental to iterative algorithms, as they rely on the assumption that the previous stages of the algorithm would have initiated the search technique in the correct set of directions. For an incorrect topology-based ensemble, the concept of correct set of directions fails to hold significant meaning, and can hence end up driving the algorithm toward a poor prediction.

In this article, we present an Iterative Clustering approach for Optimal selection of Near-native structures (ICON). We use the idea of rigorous global rearrangement clustering as presented by DiMaggio et al. (31) to cluster ensembles of protein structures in a blind case manner, that is, without the knowledge of the native structure. We introduce an objective function that reflects the dihedral angle properties. Furthermore, we use a combination of statistical and analytical techniques to eliminate structures that are unlikely to be close to the native structure. This is presented as an iterative framework, and appropriate termination criteria are introduced. The main thesis behind ICON is that if two conformers of a protein are very similar to the native structure, they are likely to be similar to each other as well. However, if two protein structures are very dissimilar to the native structure, it is not necessary that they would be similar to each other. We implement this thesis by eliminating clusters of protein structures that are very dissimilar to each other. The algorithm and its implementation are presented in detail in the following sections. The method has been tested on an extensive data set of 1400 proteins containing high resolution decoys. The proteins in this data set have a pairwise sequence similarity of <35%. It has further been tested on a number of medium-to-low resolution conformer sets. The first data set in the medium resolution data set involves structures from the Decoys 'R' Us dataset (36). The second dataset in this regime is generated from the first-principles protein folding framework ASTRO-FOLD (37). Finally, the method has been tested on select targets of the recently concluded CASP8 experiment.

METHODS

In this section, we introduce the novel iterative clustering method, ICON. A flow diagram for the algorithm is shown later in Fig. 1.

At each stage, all the dihedral angles of all the conformers of the protein in the working set are put into the evaluation matrix $M(i, j)$, where i represents the row number of the conformer, and j represents the particular dihedral

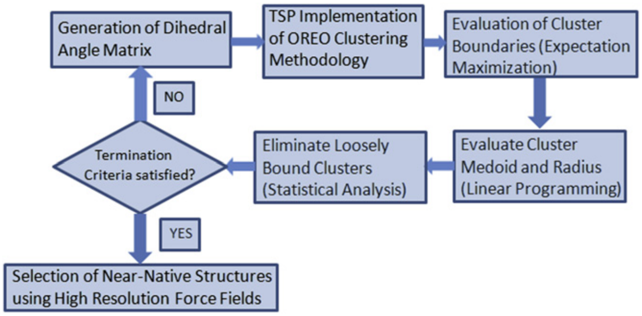
Generation of Dihedral Angle Matrix
TSP Implementation of OREO Clustering Methodology
Evaluation of Cluster Boundaries (Expectation Maximization)
Evaluate Cluster Medoid and Radius (Linear Programming)
Eliminate Loosely Bound Clusters (Statistical Analysis)
Termination Criteria satisfied?
NO
YES
Selection of Near-Native Structures using High Resolution Force Fields

FIGURE 1 Flow sheet representing the ICON algorithm.

angle. Since the end residues of the protein do not define dihedral angles, we end up with a matrix of dimensions $N*K$, where N is the number of conformers in the working set at this stage, and $K = 2*(N_p - 2)$, where N_p is the length of the protein. Based on this cost matrix, we implement the traveling salesman problem (TSP) formulation of the novel biclustering method OREO (31), which is described below in brief.

TSP implementation of the OREO approach

The aim in this section is to provide a detailed description of the variables and the objective function of the TSP model, which provides the optimal rearrangement of the rows of the cost matrix $M(i, j)$. The index pair (i, j) represents the particular row i and the particular column j of the cost matrix, whose individual element shall be denoted by $m_{i,j}$. Two rows i and i' are identified as adjacent rows in the final arrangement of the matrix, where row i' lies immediately below row i. This would mean that in the final arrangement, a binary variable $y_{i,i'}$ can be defined as

$$y_{i,i'} = \begin{cases} 1 & : \text{ if row } i' \text{ immediately precedes row } i \\ 0 & : \text{ otherwise} \end{cases}. \quad (1)$$

To finally place a particular row next to another one, the objective function is to minimize the dissimilarity between the two rows. A number of metrics of similarity can be used to define the objective function. The most commonly used objective functions are symmetric in nature (31). However, for specific data sets, the objective function can be tailored according to the nature of the problem. For example, if it is known, a priori, that the neighboring rows of the final matrix would be such that for one row, the trend is monotonic, then the objective function can be forced to penalize only those cases when this trend is violated.

For our problem here, we introduce the objective function

$$\text{Objective} = \sum_i \sum_{i'} y_{i,i'} \phi\left(m_{i,j}, m_{i',j}\right), \quad (2)$$

where $\phi(m_{i,j}, m_{i',j})$ is given by

$$\phi\left(m_{i,j}, m_{i',j}\right) = \sum_j \min\left(m_{i,j} - m_{i',j}, 360 - \left(m_{i,j} - m_{i',j}\right)\right)^2. \quad (3)$$

The objective function should reflect the circular nature of the dihedral angles. In particular, a squared difference potential cannot be used, because it would ensure that a dihedral angle of +180° and –180° are furthest away from each other, when, in fact, they are identical. Hence, the objective function in Eq. 2 selects between the minimum of the difference in dihedral angles and their difference from 360°. This way, if the difference between the dihedral angles is >180°, the value chosen in the objective function is the correct one.

The traveling salesman problem (TSP) (32) is one of the most studied problems in combinatorial optimization. The main objective is to visit a list of N cities and return to the starting city via the lowest cost route. In the TSP formulation, each row of our cost matrix represents a node (or a city). If an edge connects two such nodes, then the two rows (i.e., the two conformers of the protein) are placed next to each other in the final arrangement. Therefore, the objective of the TSP can be reformulated as visiting each conformer of the working set exactly once via these edges, while incurring the minimum cost, and to return to the first conformer. The cost of traveling from one node to the next is the objective function as expressed above.

Since the problem definition requires a circular tour that starts and ends at the same conformer, we introduce a dummy conformer to connect the first and the last structures. The cost of traveling from this dummy conformer to the top one is zero. The TSP formulation of this problem can be expressed mathematically, using a series of constraints to ensure that each row has exactly one neighbor above and below it. It is represented as

$$\min \sum_{i,i'} c_{i,i'} \times y_{i,i'}, \quad (4)$$

$$\sum_{i'} y_{i,i'} = 1 \ \forall i, \quad (5)$$

$$\sum_{i} y_{i,i'} = 1 \ \forall i'. \quad (6)$$

Here $c_{i,i'}$ represents the cost of creating a final ordered list such that rows i and i' are placed adjacent to each other. For ICON, the cost function, and hence the objective function, is given as in Eq. 2. It should be noted that cyclic tours satisfy the constraints above. Hence, additional constraints are implemented to eliminate these subtours. Such constraints are very efficiently incorporated into TSP solvers such as Concorde, via cutting plane methods.

Cluster boundary definition and analysis

Once the final order of rows is determined, we have a path from the first conformer of the final matrix to the last one.

An important step in the methodology is determining the number of clusters for a given reordering. After optimally reordering a set of features, the clusters are determined in a hierarchical manner. Let us define the final ordering to range over the index $i = 1,\ldots,|I|$. First, the pairwise distances, $d(i, i + 1)\ \forall i < |I|$, between all neighboring elements in the final ordering are computed and stored on a sorted list, which we will define as SL, from lowest to highest. The most similar pair of elements (i.e., $\{(i, i + 1): d(i, i + 1) \leq d(i', i' + 1)\ \forall (i, i' \neq i)\}$) is merged to form the first cluster, $c_{i, i+1}$, and the distances $d(i - 1, i)$, $d(i, i + 1)$, and $d(i + 1, i + 2)$ are removed from SL. The distances between this new cluster, $c_{i, i+1}$, and the elements immediately below and above it in the final ordering are then computed and these distances are added to the sorted list (i.e., $d(i - 1, c_{i, i+1})$ and $d(c_{i, i+1}, i + 2)$ are computed and added to SL). One should note that the distance between an element and a cluster, $d(i, c)$, is based upon the average distance of the element i to all members of the cluster c. The merging of two elements, an element and a cluster, or two clusters decrements the size of SL by one, and this process is repeated until SL reaches some specified value.

The minimum distance found in SL is generally an increasing function of $|I| - |SL|$. When initially creating new clusters, the most similar elements in the final ordering are merged and the corresponding minimum distance found in SL, say $d^{\min}$, would increase slowly as the number of elements in SL decreases. After all the most similar elements have been properly grouped into clusters, we inevitably encounter the situation where only dissimilar clusters and/or elements are candidates for forming a new cluster, which should result in a noticeable increase in $d^{\min}$. Thus, if we can confidently determine where this distance begins to change substantially, we can quantify when to terminate the merging of clusters.

Conceptually, this amounts to finding the "knee" in curve of $d^{\min}$ as a function of $|SL|$. To illustrate, consider the black circles in Fig. S5.1 in the Supporting Material, which represents the values of $d^{\min}$ as $|SL|$ decreases. It is easy to geometrically approximate where the knee in this plot occurs. As shown by the dashed green and blue lines in the figure, this curve can be represented by two distinct linear segments, and where these two segments intersect identifies the knee in the curve. This amounts to solving two separate linear regression problems, where the slope and intercept of each line segment, say line 1 and line 2, is a function of the points used to fit that line segment. Therefore, we need a robust way of determining which points belong to which line segment. Previous attempts (38) employed a complete enumeration approach for assigning points to the two line segments, and the assignment resulting in the minimum weighted RMSD was selected. In this section, we present an efficient and automated strategy for assigning the points to the two line segments and determining the knee in the distance curve as a function of the number of clusters.

The algorithm begins by selecting one-fifth of the points of highest $|SL|$ value (e.g., the points on the right in the figure) and fitting a line segment through these points by solving a least-squares regression problem (note that this corresponds to line segment 1). We then compute the vertical distances from all the points to this line segment and examine the corresponding distribution of these distances. The resulting distances assume a bimodal distribution, with a large, narrow distribution shouldering zero that corresponds to the points that are close to this line segment (which we will denote as class 1, since they belong to line segment 1), and a smaller, broad distribution extending to larger distances (which we will denote as class 2, as they belong to the second line segment). The average, variance, and mixture proportions of these two distributions can be computed by solving a mixture model, where here we assume that these distances approximately follow Gaussian distributions. To solve this Gaussian mixture model, we use the method of expectation maximization to maximize the log-likelihood function (39) that each of the points belongs to the first line segment (by default, the remaining points will belong to the other line segment). This provides us with the posterior probability distribution of the points belonging to either line segment, and we refit the line segment using those points that have a posterior probability >0.5 for belonging to class 1. This procedure is iterated until the slope of this line segment converges to some value.

We also implement convergence strategies to avoid singularities and pathological behavior (39). For instance, from a statistical point of view, the standard deviation (SD) of the first line segment should be at least one-third of the largest distance for any outlier (i.e., $3\ \sigma_1 = d_{\mathrm{outlier}}^{\max}$, where an outlier is a point assigned to class 2 whose neighboring points on either side belong to class 1). Also, if the class 2 distribution begins to collapse (that is, $<90\%$ of the points belong to class 2), then we restart the expectation-maximization algorithm with a lower SD for class 1. It was also observed in previous work (38) that sometimes the initial points on the far right-hand side can skew the fit of the line segment. To address this issue, we check that the majority of the points in class 2 (e.g., at least 75% of the points in class 2) lie above the hyperplane defined by the first line segment. If they do not, then we eliminate the 10% of points with the largest $|SN|$ value and reiterate the aforementioned procedure until the above criterion is satisfied.

Evaluating cluster medoids and average spherical radii

Once we have the set of conformers partitioned into the individual clusters, we would like to eliminate all the clusters that are sparse and/or include structures that are outliers. This is done by evaluating the cluster centers for each of the clusters. The cluster average radius (modeling the cluster as a hypersphere) is then calculated by averaging the pairwise RMSD of the cluster medoid with all other conformers of that cluster. The cluster medoid is the closest node to the cluster centroid. The evaluation of the cluster medoid can be modeled as an integer linear optimization problem. The objective is to minimize the distance of the cluster medoid to each of the elements of the cluster, while making sure that only one such point exists. Let us define binary variables $y(i)$ that are assigned to 1, if conformer i is the medoid of its cluster and zero otherwise. We define parameters $exist(i)$ to have the value 1, if conformer i lies in the current cluster and 0 otherwise. The model can be formulated as

$$\min \sum_{\mathrm{i}} \sum_{\mathrm{i}} b\left(i, i'\right) exist\left(i'\right) y(i), \tag{7}$$

such that

$$y(i) \leq exist(i), \tag{8}$$

$$\begin{array}{l} \sum_{\mathrm{i}} y(i) = 1 \\ y(i) = 0 - 1 \end{array}, \tag{9}$$

where $b(i, i')$ is an element of the matrix B, a square matrix of dimension $N \times N$, which represents the pairwise RMSDs between each pair of conformers in the working set. Once the cluster medoid is identified, the average cluster radius is calculated by evaluating the average pairwise RMSD of the conformers of the cluster to the medoid. The aforementioned model is implemented individually for each cluster.

Eliminating clusters based on cluster properties

Based on the number of elements in a cluster, N_{j}, and its cluster average RMSD, denoted as $\overline{\mathrm{RMSD}}$, we define the cluster concentration CC_{j} as

$$CC_{\mathrm{j}} = \frac{N_{\mathrm{j}}}{\overline{\mathrm{RMSD}_{\mathrm{j}}}}. \tag{10}$$

Based on the definition of the cluster concentration, we would like this term to be as large as possible. Larger number of elements in the cluster shows the possibility of multiple local minima surrounding this region in the energy landscape, whereas a low average RMSD shows the tightness of the cluster (based on the very similar backbone dihedral angles of the conformers in the cluster). This is highly desirable, as it is likely that the clusters that contain outlier structures would have lower number of conformers and/or a high average RMSD.

The value of cluster concentrations for each of the clusters defines a distribution, which can be modeled as a Gaussian distribution. It is desirable to have a metric, which would ensure that we end up selecting as many of the good clusters from this distribution as possible, while ensuring that we do our best to eliminate the poorer ones. To achieve this, we propose the following iterative sequence of steps. We first check whether, for this given distribution, there are any clusters that have their concentration value >3 SD above mean. If there are clusters that satisfy this condition, then we remove them from the current list and store them for the next stage of clustering. These clusters are the ones most likely to contain structures closest to the native. For the remainder of the list, we reevaluate the mean and the SD. It is then again checked whether there are clusters >3 SD above mean. This procedure is carried out until there are no more clusters >3 SD of the mean of the existing distribution.

If, initially, we do not find any clusters >3 SD of the mean, we check whether there are clusters which are >2 SD, and if not, then 1 SD above the mean of this distribution. Once these are found, a similar procedure of removal and analysis of the shortened list of clusters is done until there are no more clusters above this new threshold.

Finally, we select all clusters that were previously removed and stored, along with all clusters that are greater than or equal to the median of this modified distribution, as these would correspond to the clusters with the highest concentration. The median is selected because it is more resistant to the existence of outliers in the distribution of cluster concentrations. All the conformers in the selected clusters form the working set of the next iteration.

This entire procedure is carried out for 10 iterations, or until the number of conformers in the working set is not $<50\%$ of the number of conformers in the original ensemble.

Selection of near-native structures

At the end of the iterative clustering procedure, we select the most likely structures that are going to be closest the putative native structure. To do this, we collect the medoids of all the clusters at the end of the final stage, which have their concentration value above or equal to the median of this distribution. For each of these cluster medoids, we implement the novel C_α-C_α and Centroid-Centroid distance-dependent high-resolution force fields, proposed by the literature (12,13). These force fields aim to isolate native and near-native folds of a protein as lower energy structures, compared to structures further away from the native structure. Finally, the five cluster medoids with the lowest energies are picked as the selected structures.

Automated implementation

ICON has been implemented as an automated procedure, which collects the dihedral angle matrix and pairwise RMSD data and implements the entire iterative algorithm. The TSP implementation of the OREO clustering approach has been implemented as a C++ program in interaction with CPLEX 11.0 (ILOG, Cintech Iii, Singapore) and Concorde (William Cook, Georgia Institute of Technology, Atlanta, GA). Steps that comprise of collection of dihedral angle data, evaluation of cluster boundaries, evaluation of cluster medoids, and the elimination of loosely bound clusters, are implemented in C language. Finally, the high resolution force fields have been implemented in C++, and are available for download (12,13). We plan to make ICON freely available to the scientific community, by releasing it as a single executable, which can be run with CPLEX and Concorde licenses.

ICON has been tested for computational time on a single processor machine using a single processor CPLEX implementation, as well as with a multiprocessor CPLEX implementation on a computer cluster. The parallel version of ICON was run on Quad core Intel Xeon 2.83 GHz processors. For a run on a test protein (PDB: 1elr chain A) with 762 conformers, the parallel version of ICON took five CPU minutes for a run. ICON was also tested on a single Intel Pentium (Santa Clara, CA) (4) 3.2 GHz processor. The same run is completed in ~25 min.

COMPUTATIONAL RESULTS

ICON was applied on a large number of proteins. The test sets were divided into three distinct categories: high resolution data set, medium-to-low resolution data set, and CASP8 targets. The medium-to-low resolution data sets is further subcategorized into ensembles from Decoys ‘R’ Us (36) and ASTRO-FOLD (37). Fig. 2 shows the brief results of the application of the method. As can be seen from the histogram plot, the method performs consistently at >90% in almost all cases, by using both the C_α-C_α and Centroid-Centroid force fields.

The following subsections present details on the generation of the datasets, and the results generated from the application of the novel clustering method.

High resolution data set

For generating the decoys for the high resolution data set, a well-represented collection of 1400 proteins developed by Zhang and Skolnick (40) was used. All the proteins of this set are nonhomologous, single domain proteins with a maximum pairwise sequence similarity of 35%. The length of the proteins varies from 41 to 200 amino acids. It also has a mixed representation of α-, β-, and α/β-proteins.

The generation of decoys for each of these proteins was carried out using a torsion angle dynamics approach, DYANA (41). The main premise of the decoy generation framework is the idea of retaining distance information among the residues within the hydrophobic core of the protein. Once the hydrophobic core has been defined, distance bounds are introduced among the hydrophobic residues to relax the native distance between them. Further details on the generation and quality of the decoys generated can be found elsewhere (13).

For all decoys of the 1400 proteins, the proposed novel iterative clustering method, ICON, was applied. At the final stage, to select the final five conformers from the list, both the C_α-C_α (12) and the Centroid-Centroid (13) distance-dependent force fields were applied, and the results compared to the state-of-the-art SPICKER method (34). Fig. 3 shows the rank of the selected conformer using both the criteria. The rank reflects the number of structures in the individual ensemble that are ahead of the picked structure. To give a better representation of how many conformers are better or worse than the selected structure, Fig. 4 presents a percentile graph of the number of structures that are worse than the selected structure for each of the force fields. As shown in the figure, for 84.7% proteins out of the high resolution data set, ICON selects structures that are above the 95 percentile in terms of quality of the structure.

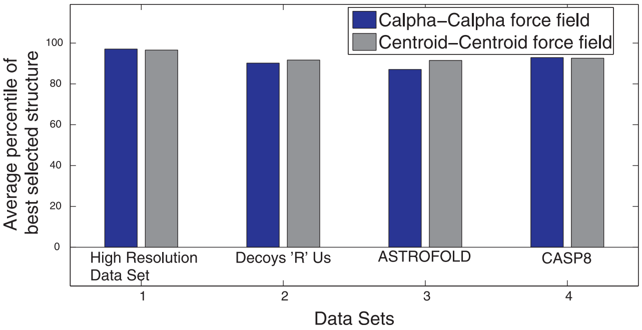

FIGURE 2 Histogram presenting overall results of ICON algorithms on individual test sets.

As a comparison to the SPICKER method, we ran SPICKER on the same dataset. Fig. 3 presents a graph where the proteins have been sorted based on the ranks of the structures selected by SPICKER method. Further, ranks that were selected by ICON are also presented there. As can be seen, for a large number of cases, ICON performs better than the SPICKER method. Furthermore, Fig. 5 presents a graph

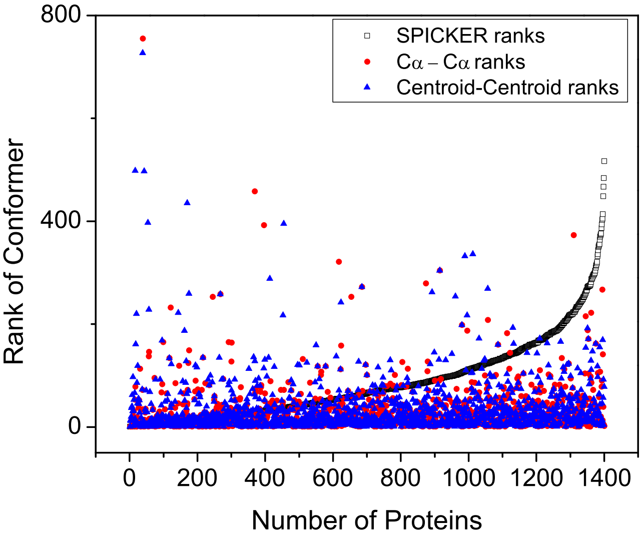

FIGURE 3 Graph representing ranks of structures selected by SPICKER and ICON. Approximately 83% of points from the ICON algorithm fall below the monotonic curve represented by SPICKER.

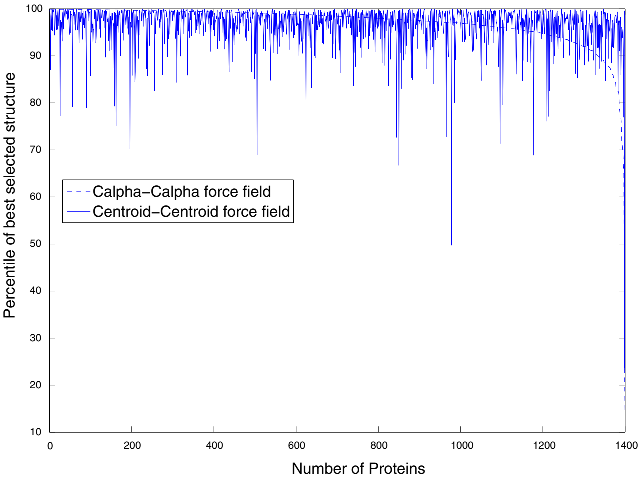

FIGURE 4 Graph representing percentile of selected structure with respect to the best structure of the ensemble.

showing the difference in RMSDs between the best structures selected by ICON, and the corresponding structures picked by SPICKER. In 81.5% of the cases, ICON performs better in selecting near-native structures from the ensembles, whereas in 86.2% of the cases, ICON performs at least as well as SPICKER. This suggests that ICON can select near-native structures from a given high resolution ensemble of protein structures. A detailed presentation of the results for this dataset is in the Supporting Material.

To compare the relative contributions of the clustering algorithm and the force field toward the performance of the ICON algorithm, the C_α-C_α and Centroid-Centroid energies of all decoys of a set of 150 proteins from the high resolution data set were evaluated. These 150 proteins were not used in the generation of either of these force fields. We compiled the five lowest energy structures in these ensembles, and the lowest RMSD structures among these were analyzed. The average percentile of the selected structures by using just the C_α-C_α and Centroid-Centroid force fields were 92.7% and 91.9%, respectively, whereas the use of the proposed approach resulted in the average percentiles to be 97.1% and 96.6%, respectively. Comparisons were also made with the statistical full-atom Rosetta potential. By using just the Rosetta potential, the average percentile of best selected structure was 70.8%. This shows that the algorithm improves the structures selected by the force fields. The force fields themselves perform quite well on this test set, since they were trained on a large, high resolution training set of 1250 proteins, and are likely to handle high-resolution ensembles well.

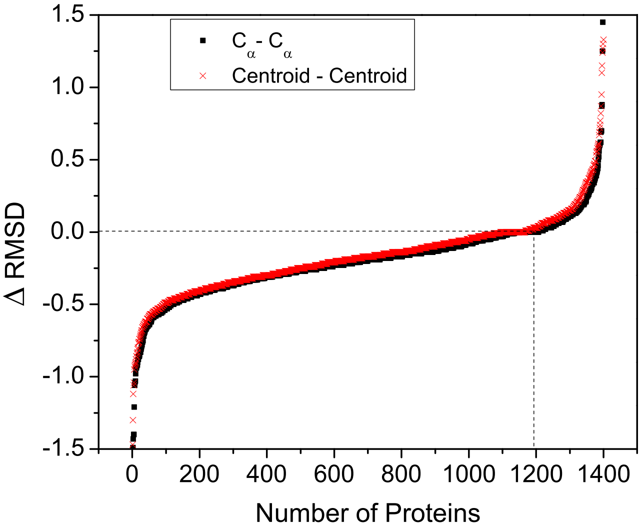

FIGURE 5 Graph representing difference in RMSD values of structures selected by ICON and SPICKER, $\Delta\text{RMSD} = rmsd_{\text{ICON}} - rmsd_{\text{SPICKER}}$. Note that 1193 points fall below zero, representing the number of cases where the structure selected by ICON has a lower RMSD than the one selected by SPICKER.

Medium-to-low resolution data set: Decoys ‘R’ Us

To generate the medium-to-low resolution data set, two distinct procedures were used. Firstly, the algorithm was tested on five decoy sets of the Decoys ‘R’ Us database (36), which are identified as the most challenging datasets based on the results of Rajgaria et al. (13). These included the test sets FISA and FISA-CASP3 (42), LMDS (43), LATTICE-SSFIT (44,45), and SEMFOLD (46). The FISA and FISA-CASP3 datasets were generated by a fragment-insertion simulated annealing procedure, where the procedure was used to assemble nativelike fragments from unrelated proteins using Bayesian scoring functions. Each conformer of the LMDS dataset is a local minima structure obtained using the ENCAD function, which contains a penalty term for steric clashes and a favorable contribution term for compactness and native similarity. The LATTICE dataset was generated by firstly generating all possible conformations using a tetrahedral lattice. A scoring function was used to rank these structures. Some of the best structures were minimized locally using a different energy function, while maintaining the secondary structure features.

As can be seen from Fig. S2.1, ICON performs well in selecting structures in the top 10th of the ensemble of structures in most cases. It is of particular importance to note that the sets of decoys used in this test set have been generated by different methods. Further, the RMSD ranges of the individual decoys for a significant number of proteins have a majority of their ensembles in the medium or low resolution decoy range. For this data set, using the C_α-C_α force field, the average percentile of structures selected by ICON is 91.2%. Using the Centroid-Centroid force field, ICON selects structures with an average percentile of 92.0%.

Comparisons were also made between the use of the proposed clustering algorithm, and the use of the force fields directly. By directly using the C_α-C_α and Centroid-Centroid force fields, we get an average percentile of best structure to be 80.2% and 82.1%, respectively. By using the all-atom Rosetta potential, we get an average selected percentile of 83.8%. Note that ICON exhibits a superior performance (i.e., 91.2% and 92.0%) when compared to the exclusive use of C_{α}-C_{α}, Centroid-Centroid, and the all-atom Rosetta potentials (80.2%,82.1%, and 83.8%, respectively).

A disparity is seen in the results obtained using either the different force fields. This may be attributed to the nature of the ensembles of structures produced. Since the C_{α}-C_{α} force field does not account for side-chain information, a protein structure where side chains are too close or too far will not be accounted for. On the other hand, the Centroid-Centroid force field accounts for the side chains of the structures. Hence, misplaced side chains would cause an increase in the Centroid-Centroid energy of the structure, while keeping the backbone-based energy constant.

Medium-to-low resolution data set: ASTRO-FOLD

A first-principles based method was also used for the generation of medium resolution ensembles of a small set of proteins. The fourth stage of the ASTRO-FOLD algorithm uses torsion angle dynamics, deterministic global optimization, and a stochastic computational space annealing procedure to predict the tertiary structure of a protein given its amino-acid sequence (37,47–51). As can be seen from the results shown in Table S3.1, the average percentiles of the best structure selected by ICON are 87.1% and 91.5%, using the C_{α}-C_{α} and Centroid-Centroid force fields, respectively. Further, the method is seen to be reasonably independent of the quality of the ensemble provided to it. This is particularly important, since for a protein with a completely new fold, it is possible that structure prediction techniques may not be able to produce structures of high resolution quality.

Selected CASP8 targets

A selection of the CASP8 Targets was also used as an additional test set for evaluating the effectiveness of the proposed iterative clustering method, ICON. The results for the CASP8 targets are presented in Table S3.2. As can be seen from the results, for the range of RMSDs of the structures in the ensemble, the structures selected by ICON are of good quality. This is particularly important to note in this case, as the range of RMSDs lies within the medium-to-low resolution regime.

The average percentile of selected structures using the C_{α}-C_{α} force field is 92.9%, whereas it is 92.6% when the Centroid-Centroid force field is used. The CASP8 data set provides the most realistic, and up-to-date test set of the ICON formulation. Based on its nature, the CASP8 target structures are not known a priori. This is especially relevant for target structures with low sequence and structural homology to databases. As can be seen, ICON performs very well by selecting structures that are, on an average, within the top 7.5% of predicted structures in the respective ensembles.

Stagewise enrichment of ICON

To demonstrate the benefit of the proposed iterative clustering method ICON, we present a histogram for each individual stage of the iterative procedure, which shows the distribution of the conformers in the working set of the method at the particular stage for an example protein (PDB: 1elrA). As can be seen from Fig. S4.1, at each stage, the better structures are retained more than the comparatively worse structures. Table 1 shows the enrichment factor for each bin in the histogram for the various stages. The enrichment factor for a bin at a stage is given by

$$\text{Enrichment} = \frac{N_{\text{bin,stage}}/N_{\text{bin,start}}}{N_{\text{total,stage}}/N_{\text{total,start}}}. \qquad (11)$$

As can be seen from Table 1, the RMSD regions that we are interested in (the ones closer to the native structure) are enriched in a favorable manner (>1). The top two regions are enriched significantly. Clearly, the conformers that are most likely to be falsely selected would lie in the middle of this table. As can be seen, a very large number of them are eliminated at the individual stages. Furthermore, for RMSD ranges 1.5–4.0, the trend is monotonic, which is highly favorable.

DISCUSSION

A novel iterative clustering method, ICON, is introduced to identify the near-native folds for a protein from an ensemble of given structures. The method uses a clustering in the dihedral angle space via a TSP-based framework and eliminates loose and widely spread clusters at each iteration to reach the final solution. ICON was tested on a set of 1400 nonhomologous proteins. The method identified structures within the top 10% of conformers in 97% of cases. The average percentile of the selected conformer was 2.9%; that is, on average, the selected conformer was in the top 2.82% of the conformers in the ensemble.

The method was also tested on medium resolution data sets taken from external sources, and performed very well. The fact that the accuracy of the method does not deteriorate significantly when considering a variety of high resolution, medium resolution, and low resolution datasets suggests that the method is robust to diverse conformational ensembles.

TABLE 1 Table showing enrichment factor for different RMSD regions over number of stages

RMSD range	N	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5	Stage 6
0.5–1.0	49	1.1091	1.1905	1.2843	1.3412	1.4547	1.6518
1.0–1.5	236	1.0843	1.1020	1.171	1.2531	1.3232	1.3969
1.5–2.0	206	1.0113	0.9734	0.9497	0.9950	1.0216	0.9870
2.0–2.5	103	0.9674	0.9675	0.9032	0.7596	0.6261	0.6325
2.5–3.0	104	0.8492	0.8647	0.8024	0.7071	0.6527	0.5125
3.0–3.5	54	0.8387	0.7877	0.7347	0.6085	0.5343	0.4752
3.5–4.0	9	1.0064	1.0803	1.064	1.0431	0.5657	0.2193

As was discussed in the Introduction, the overall effectiveness of the method will depend to a certain degree on the prior distribution of a given ensemble. For a given target protein, if the ensemble of predicted structures are of a different topology or fold to the native structure, using clustering techniques to identify the near-native structure can be limited. In such cases, even the correct selection of structures with lowest RMSDs to the native would hold little significance. Such situations could introduce sources of errors in the iterative process of the algorithm, and the algorithm may result in a direction that is not favorable.

SUPPORTING MATERIAL

Three tables and three figures are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(09)01210-7.

C.A.F. gratefully acknowledges financial support from the National Science Foundation, National Institutes of Health (grant Nos. R01 GM52032 and R24 GM069736), and the U.S. Environmental Protection Agency (EPA) (grant No. GAD R 832721-010). Although the research described in the article has been funded in part by the U.S. EPA's Science To Achieve Results program through grant No. R 832721-010, it has not been subjected to any EPA review and does not necessarily reflect the views of the Agency, and no official endorsement should be inferred.

REFERENCES

1. Zhang, Y. 2008. Progress and challenges in protein structure prediction. *Curr. Opin. Struct. Biol.* 18:342–348.
2. Floudas, C. A., H. K. Fung, S. R. McAllister, M. Monnigmann, and R. Rajgaria. 2006. Advances in protein structure prediction and de novo protein design: a review. *Chem. Eng. Sci.* 16:966–988.
3. Floudas, C. A. 2007. Computational methods in protein structure prediction. *Biotechnol. Bioeng.* 97:207–213.
4. MacKerell, A. D., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, et al. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B.* 102:3586–3616.
5. Cornell, W., P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, et al. 1995. A second generation force field for the stimulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* 117:5179–5197.
6. Momany, F. A., R. F. McGuire, A. W. Burgess, and H. A. Scheraga. 1975. Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. *J. Phys. Chem.* 79:2361–2381.
7. Némethy, G., K. D. Gibson, K. A. Palmer, C. N. Yoon, G. Paterlini, et al. 1992. Energy parameters in polypeptides. X. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *J. Phys. Chem.* 96:6472–6484.
8. Liwo, A., S. Oldziej, S. Czaplewski, K. Urszula, and H. A. Scheraga. 2004. Parameterization of backbone-electrostatic and multibody contributions to the UNRES force field for protein-structure prediction from ab-initio energy surfaces of model systems. *J. Phys. Chem. B.* 108:9421–9438.
9. Arnautova, Y. A., A. Jagielska, and H. A. Scheraga. 2006. A new force field (ECEPP-05) for peptides, proteins and organic molecules. *J. Phys. Chem. B.* 110:5025–5044.
10. Wroblewska, L., A. Jagielska, and J. Skolnick. 2008. Development of a physics-based force field for the scoring and refinement of protein models. *Biophys. J.* 94:3227–3240.
11. Arnautova, Y. A., Y. N. Vorobjev, J. A. Vila, and H. A. Scheraga. 2009. Identifying native-like protein structures with scoring functions based on all-atom ECEPP force fields, implicit solvent models and structure relaxation. *Proteins.* 65:726–741.
12. Rajgaria, R., S. R. McAllister, and C. A. Floudas. 2006. A novel high resolution C_α C_α distance-dependent force field based on a high quality decoy set. *Proteins.* 65:726–741.
13. Rajgaria, R., S. R. McAllister, and C. A. Floudas. 2007. Distance dependent centroid to centroid force fields using high resolution decoys. *Proteins.* 70:950–970.
14. Tobi, D., and R. Elber. 2000. Distance-dependent, pair potentials for protein folding: results from linear optimization. *Proteins.* 41:40–46.
15. Wagner, M., J. Meller, and R. Elber. 2004. Large-scale linear programming techniques for the design of protein folding potentials. *Math. Program.* 101:301–318.
16. Meller, J., M. Wagner, and R. Elber. 2002. Maximum feasibility guideline in the design and analysis of protein folding potentials. *J. Comput. Chem.* 23:111–118.
17. Qiu, J., and R. Elber. 2005. Atomically detailed potentials to recognize native and approximate protein structures. *Proteins.* 61:44–55.
18. Lu, H., and J. Skolnick. 2001. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins.* 44:223–232.
19. Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA.* 95:14863–14868.
20. Hartigan, J. A., and M. A. Wong. 1979. Algorithm AS 136: a K-means clustering algorithm. *Appl. Stat.* 28:100–108.
21. Wolfe, J. H. 1970. Pattern clustering by multivariate mixture analysis. *Multivariate Behav. Res.* 5:329–350.
22. Jain, A. K., and J. Mao. 1996. Artificial neural networks: a tutorial. *IEEE Comput.* 29:31–44.
23. Klein, R. W., and R. C. Dubes. 1989. Experiments in projection and clustering by simulated annealing. *Pattern Recognit.* 22:213–220.
24. Bhuyan, J. N., V. V. Raghavan, and K. E. Venkatesh. 1991. Genetic algorithm for clustering with an ordered representation. *Proc. 4th Int. Conf. Gen. Alg.* 408–415.
25. Busygin, S., O. Prokopyev, and P. Pardalos. 2007. An optimization-based approach for data classification. *Opt. Meth. Soft.* 22:3–9.
26. Floudas, C. A., A. Aggarwal, and A. R. Ciric. 1989. Global optimum search for nonconvex NLP and MINLP problems. *Comput. Chem. Eng.* 13:1117–1132.
27. Tan, M. P., J. R. Broach, and C. A. Floudas. 2007. A novel clustering approach and prediction of optimal number of clusters: global optimum search with enhanced positioning. *J. Glob. Optim.* 39:323–346.
28. Tan, M. P., J. R. Broach, and C. A. Floudas. 2007. Evaluation of normalization and pre-clustering issues in a novel clustering approach: global optimum search with enhanced positioning. *J. Bioinform. Comput. Biol.* 5:895–913.
29. Tan, M. P., E. N. Smith, J. R. Broach, and C. A. Floudas. 2008. Microarray data mining: a novel optimization-based approach to uncover biologically coherent structures. *BMC Bioinformatics.* 9:268–288.
30. Monnigmann, M., and C. A. Floudas. 2005. Protein loop structure prediction with flexible stem geometries. *Proteins.* 61:748–762.
31. DiMaggio, P. A., S. R. McAllister, C. A. Floudas, X.-J. Fend, J. D. Rabinowitz, et al. 2008. Biclustering via optimal re-ordering of data matrices in systems biology: rigorous methods and comparative studies. *BMC Bioinformatics.* 9:458.
32. Applegate, D., R. Bixby, V. Chvatal, and W. Cook. 2007. The Traveling Salesman Problem: A Computational Study. Princeton University Press, Princeton, NJ.
33. Shortle, D., K. T. Simons, and D. Baker. 1998. Clustering of low-energy conformations near the native structures of small proteins. *Proc. Natl. Acad. Sci. USA.* 95:11158–11162.

34. Zhang, Y., and J. Skolnick. 2004. SPICKER: a clustering approach to identify near-native protein folds. *J. Comput. Chem.* 25:865–871.

35. Dowe, D. L., L. Allison, T. I. Dix, L. Hunter, C. S. Wallace, et al. 1996. Circular clustering of protein dihedral angles by minimum message length. *Proc. 1st Pacific Symp. Biocomput.*

36. Samudrala, R., and M. Levitt. 2000. Decoys 'R' Us: a database of incorrect protein conformations to improve protein structure prediction. *Protein Sci.* 9:1399–1401.

37. Klepeis, J. L., and C. A. Floudas. 2003. ASTRO-FOLD: a combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophys. J.* 85:2119–2146.

38. Salvador, S., and P. Chan. 2004. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. *Proc. IEEE Int. Conf. Tools Artif. Intel.* 576–584.

39. Bishop, C. M. 2006. Pattern Recognition and Machine Learning. Springer, New York.

40. Zhang, Y., and J. Skolnick. 2004. Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc. Natl. Acad. Sci. USA.* 101:7594–7599.

41. Guntert, P., C. Mumenthaler, and K. Wuthrich. 1997. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.* 273:283–298.

42. Simons, K. T., C. Kooperberg, C. Huang, and D. Baker. 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* 268:209–225.

43. Keasar, C., and M. Levitt. 2003. A novel approach to decoy set generation: designing a physical energy function having local minima with native structure characteristics. *J. Mol. Biol.* 329:159–174.

44. Samudrala, R., Y. Xia, M. Levitt, and E. S. Huang. 1999. A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. *Pac. Symp. Biocomput.* 4:505–516.

45. Xia, Y., E. S. Huang, M. Levitt, and R. Samudrala. 2000. Ab initio construction of protein tertiary structure using a hierarchical approach. *J. Mol. Biol.* 300:171–185.

46. Samudrala, R., and M. Levitt. 2002. A comprehensive analysis of 40 blind protein structure predictions. *BMC Struct. Biol.* 2:3–18.

47. Klepeis, J. L., and C. A. Floudas. 1999. A comparative study of global minimum energy conformations of hydrated peptides. *J. Comput. Chem.* 20:636–654.

48. Klepeis, J. L., C. A. Floudas, D. Morikis, and J. D. Lambris. 1999. Predicting peptide structures using NMR data and deterministic global optimization. *J. Comput. Chem.* 20:1354–1370.

49. Klepeis, J. L., and C. A. Floudas. 2002. Ab initio prediction of helical segments of polypeptides. *J. Comput. Chem.* 23:246–266.

50. Klepeis, J. L., and C. A. Floudas. 2003. Prediction of β-sheet topology and disulfide bridges in polypeptides. *J. Comput. Chem.* 24:191–208.

51. Klepeis, J. L., and C. A. Floudas. 2003. Ab initio tertiary structure prediction of proteins. *J. Glob. Optim.* 25:113–140.