NIH Public Access
Author Manuscript
Proteins. Author manuscript; available in PMC 2010 June 15.

Published in final edited form as:
Proteins. 2010 April ; 78(5): 1137–1152. doi:10.1002/prot.22634.

MUFOLD: A new solution for protein 3D structure prediction

Jingfen Zhang[1,2], **Qingguo Wang**[1], **Bogdan Barz**[3], **Zhiquan He**[1,2], **Ioan Kosztin**[3], **Yi Shang**[1], and **Dong Xu**[1,2,*]

[1] Department of Computer Science, University of Missouri, Columbia, Missouri

[2] Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, Missouri

[3] Department of Physics and Astronomy, University of Missouri, Columbia, Missouri

Abstract

There have been steady improvements in protein structure prediction during the past 2 decades. However, current methods are still far from consistently predicting structural models accurately with computing power accessible to common users. Toward achieving more accurate and efficient structure prediction, we developed a number of novel methods and integrated them into a software package, MUFOLD. First, a systematic protocol was developed to identify useful templates and fragments from Protein Data Bank for a given target protein. Then, an efficient process was applied for iterative coarse-grain model generation and evaluation at the Cα or backbone level. In this process, we construct models using interresidue spatial restraints derived from alignments by multidimensional scaling, evaluate and select models through clustering and static scoring functions, and iteratively improve the selected models by integrating spatial restraints and previous models. Finally, the full-atom models were evaluated using molecular dynamics simulations based on structural changes under simulated heating. We have continuously improved the performance of MUFOLD by using a benchmark of 200 proteins from the Astral database, where no template with >25% sequence identity to any target protein is included. The average root-mean-square deviation of the best models from the native structures is 4.28 Å, which shows significant and systematic improvement over our previous methods. The computing time of MUFOLD is much shorter than many other tools, such as Rosetta. MUFOLD demonstrated some success in the 2008 community-wide experiment for protein structure prediction CASP8.

Keywords

protein structure prediction; CASP; multidimensional scaling; scoring function; clustering; molecular dynamics simulation

INTRODUCTION

Knowledge of the three-dimensional structure of a protein often gives a basis for understanding its function. However, the gap between numbers of known protein sequences and structures has been dramatically increasing. As a comparison, Release 40.6 of July 28, 2009 of UniProtKB/TrEMBL[1] contained 8,926,016 nonredundant sequence entries, whereas there were 59,790 structures (representing 34,480 nonredundant sequences) in Protein Data Bank (PDB)[2] as of August 25, 2009, two orders of magnitude fewer. One important tool to bridge this gap is computational prediction, which has become more and more important because of a huge number of biological sequences being generated by new sequencing technologies.

*Correspondence to: Dong Xu, 201 Engineering Building West, University of Missouri, Columbia, MO 65211. xudong@missouri.edu.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Compared with experimental approaches, computational methods are cheap and fast, and they can be applied at a much larger scale.

Computational protein structure prediction methods can be classified into three categories: (1) comparative modeling (CM),[3–7] (2) threading,[8–16] and (3) *ab initio* prediction.[17–21] Both CM and threading are template-based, whereas *ab initio* methods do not require templates. CM methods, based on the fact that evolutionarily related proteins typically share a similar structure, build models for the target protein by aligning the target sequence to evolutionarily related (i.e., homologous) template structures. Although *ab initio* methods, not relying on any template, predict a protein structure by optimizing some scoring functions, which describe the physical/statistical properties of a protein. In contrast to CM and *ab initio* predictions, threading methods are designed to match target sequence to templates of similar structural folds, where the target and template sequences are not required to be evolutionarily related.

Although significant progress has been made in the protein structure prediction over the past 2 decades, current computational methods are still far from consistently providing accurate structures with reasonable computing time. For proteins sharing more than 30% sequence identity to their homologous templates, models built by CM or threading methods are typically comparable to low-resolution experimental structures. However, when the sequence identity drops below 30%, modeling accuracy sharply decreases because of substantial alignment errors. *Ab initio* methods require extensive computing resources because searching space is huge and optimizing process has very high computational complexity, whereas the prediction results are generally unreliable because of limitations of scoring functions and conformational search. Currently, heuristic optimization methods such as genetic algorithms and Monte Carlo simulations are time-consuming so as to generate structural models often far from the global optimal solution of a scoring function. In addition, widely used scoring functions are generally not accurate enough to identify the best structure from the generated structure pool. Hence, although a number of prediction servers, such as Modeller,[5] HHpred,[7] I-TASSA,[16] and Robetta,[21] have been developed, protein structure prediction has not been widely applied in molecular biology studies other than homology modeling with structural templates of high-sequence identity, because of low prediction accuracies and long computing time.

To address the above two issues, we propose a hybrid methodology of using whole and partial template information, MUFOLD, to cover both template-based and *ab initio* predictions using the same framework toward achieving improved accuracies and fast computing in automated predictions. The key advantages of MUFOLD are as follows: (1) a fast model generation method based on exploring different folds and alignments for the same target in short computing time, and (2) an effective molecular dynamics (MD) simulation method to identify the best structure from generated models. MUFOLD uses template information, but does not assume that the correct fold is uniquely identified. Theoretically, for any given protein up to 100 residues, there is at least one already solved structure in the PDB that has a root-mean-square deviation (RMSD) < 4 Å for 90% of its residues regardless of whether it is evolutionarily related to other proteins with solved structures,[15] which suggests that the protein structure prediction problem can be solved in principle by using template-based methods. The main issue is to identify the correct fold and alignment, which is often a daunting task for CM or threading methods in most cases. In contrast to the conventional CM and threading methods, which mainly rely on one fold and one alignment, we use a set of possible folds, and for each fold, multiple templates (full-size and partial fragments) and various alignments are applied. This is made possible by a much faster graph-based method, multidimensional scaling (MDS)[22] in coarse-grain model generation than the time-consuming optimization methods such as genetic algorithms and Monte Carlo optimizations by current popular servers. Furthermore, we rank models using their dynamics properties in addition to static scoring functions. We have

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

demonstrated that the behavior of simulated unfolding under heating in MD simulation can reveal the quality of structural model more effectively than static scoring functions.

We have implemented our methods into the software package MUFOLD, which is available upon request. As a new framework for protein structure prediction, MUFOLD is at its early stage of development. Nevertheless, MUFOLD demonstrated proof of principles in the 2008 community-wide experiment for protein structure prediction CASP[23] (critical assessment of techniques for protein structure prediction), which provides an assessment of the state-of-the art in this field.

MATERIALS AND METHODS

The overview of MUFOLD is presented in Figure 1, which includes three main parts: (1) template selection and alignments, that is, recognizing potentially useful templates/fragments in PDB for the target sequence and finding alignments; (2) coarse-grain model generations and evaluations, including fast model generations using MDS at the Cα or backbone atoms level based on the spatial restraints derived from alignments, evaluations of models through clustering and static scoring functions, and iterative improvement of selected models by integrating spatial restraints and previous models; and (3) full-atom model evaluation through MD simulations.

Template selection and alignment

MUFOLD is designed to cover the CM, threading, and *ab initio* structure predictions using a unified framework. The first step of MUFOLD is to find suitable templates and alignments. Here, the "template" is a general concept including the global homologous templates, nonevolutionarily related (analogous) templates, and the locally compatible protein fragments for the targets from PDB. MUFOLD adaptively applies different strategies for various targets. In general, target sequences are classified into three categories as follows:

1. "Easy" targets that have significant hits by applying sequence-profile alignment tool such as PSI-BLAST[24] against the PDB database, that is, there is at least one alignment hit (reported by PSI-BLAST) that can cover more than 70% of the target sequence and with *E*-value 1e−3 or less. As homologous templates with high confidence alignments can be easily found for this case, it is intuitive that the sequence alignments can be used to obtain high-quality distance restraints directly.
2. "Medium" targets that have remote homologies to templates obtained by using profile–profile alignment (e.g., HHSearch[25]), that is, there is at least one alignment hit (reported by PSI-BLAST or HHsearch) with *E*-value less than 1e−2 (excluding "easy" targets defined earlier). These targets probably have the correctly identified fold information, but the alignments may be incorrect. Therefore, we try to obtain various alignments by applying different tools and parameters for the correct fold. For example, we apply PSI-BLAST with the parameters of "−j 3 −e 0.001 and nr90" for profile and "−e 11000" for alignment. We also apply HHsearch with default parameters. Coupled with the optimization of MDS, we sample distance restraints and improve the restraints iteratively.
3. "Hard" targets that cannot be assigned to any template even by profile–profile alignment, that is, targets other than the above two cases. We use a simple fast threading approach to search for possible templates by aligning secondary structures. Here, we apply BLAST to obtain alignments between the predicted secondary structures of the target sequence and the actual secondary structures of templates. More specifically, we predict secondary structures of a target protein using in-house tool MUPRED,[26] PSIPRED,[27] and SSPro,[28] and we convert the predicted secondary

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

structures of H, E, C to 9 or 21 alphabets according to their predicted secondary structure types and associated confidence values. We then train score matrixes (similar to the Blosum matrix) for either 9 or 21 alphabets. In such fast alignments using BLAST, although the top-1 hit may not represent the correct fold, the compatible protein fragments of top-*n* (20–100) folds usually include the correct fold.

As we often obtain multiple alignments for one template protein sequence, we select top alignment hits for the model generation using the following score:

$$\text{score}=\sum_i (X_i - \overline{X_i})/\text{dev}(X_i),$$

Here X_i represents the *i*-th feature value in terms of the likelihood of the alignment to be correct. For example, X_1 is the sequence similarity of the hit, X_2 the secondary structure similarity of the hit, X_3 the solvent accessibility similarity of the hit, and so forth. Note, both the secondary structure and the solvent accessibility for a target are predicted based on its sequence, and the corresponding values for a template are calculated based on its 3D structure. $\bar{X}_i$ is the average value of X_i from all alignment hits, and dev(X_i) is the standard deviation of X_i from $\bar{X}_i$. We rank the alignment hits of a given template according to their scores and select top-*k* ($k < 20$) alignments to generate the models.

Coarse-grain model generation

The main idea of MUFOLD is to construct models by optimally satisfying spatial restraints derived from the alignments, which is similar to the existing system such as Modeller.[5] However, there are two main differences: (1) instead of constructing full-atom models directly, we model the structure hierarchically, that is, retrieving the spatial restraints for Cα or backbone atoms to generate coarse-grain model at first and then constructing full-atom models based on the coarse-grain model. In this way, we speed up the modeling dramatically without significantly sacrificing accuracy. The fast modeling allows us to explore different folds and alignments for the same target in reasonable computing time automatically. (2) In Modeller, the spatial restraints derived from the aligned homologous proteins are expressed as probability density functions (pdfs). Optimizing pdf and physical energy function simultaneously in Modeller often works well when using a single template with good alignment or homogeneous templates with consistent restraints. When the alignment has significant gaps or templates are heterogeneous with inconsistent restraints, models generated by Modeller are often distorted, as it cannot satisfy the pdf restraints or energy function well. In contrast, MUFOLD can better accommodate diverse spatial restraints retrieved from heterogeneous alignments including global or local alignments tailored for individual target. This is achieved through an iterative, graph-based model generation method, which is free of any energy function. In the following sections, we describe our methods of coarse-grain model generation in detail.

Graph-based model generation formulation

We formulate the structure prediction problem as a graph realization problem[29] and then apply an efficient optimization approach to solve it. The basic idea is to estimate the distances between Cα (or backbone) atoms for each pair of amino acids in the target sequence (i.e., spatial distance restraints) and then calculate the corresponding coordinates for all amino acids by applying MDS to the distance restraints. Assume that there are *n* points (each representing the Cα atom of a residue) $X_k \in R^3$, $k = 1,\ldots,n$ in a 3D space. If we know the distances between some pairs of points specified in the edge set, for example, d_{ij} between residue *i* at X_i and residue *j* at X_j, then the graph realization problem is to determine the coordinates of the points from the partial distance restraints, such that the distance between each pair of points matches the given distance restraint, $||X_i - X_j|| = d_{ij}$ for all d_{ij}. If the distance restraints are inaccurate

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

(e.g., estimated based on the aligned positions of the structure template), usually there is no exact or unique solution to the overdetermined system of equations (in this article, we use "restraint" to indicate "soft restraint," which does not have to be satisfied as a hard condition). Instead, the problem is formulated as an optimization problem that minimizes the sum of squared errors. The basic realization problem can be formulated as minimizing the squared error for a smoother function:

$$\min_{X_1 \ldots X_n \in R^3} \sum_{i,j=1,\ldots,n} \left(\left\| X_i - X_j \right\| - d_{ij} \right)^2 \quad (1)$$

The optimization problem of Eq. (1) is generally non-convex with many local minima. Traditional local optimization techniques, such as the Levenberg-Marquardt method,[30] require good initial points to produce good solutions. Global search methods such as simulated annealing or genetic algorithms are very slow in the large continuous search space. However, MDS is suitable for this optimization problem.

Multidimensional scaling

MDS method is efficient for solving the graph realization problem, which has been applied in many fields such as machine learning and computational chemistry. MDS starts with one or more distance matrices (or dissimilarity matrices) that are derived from points in a multidimensional space, and it finds a placement of the points in a low-dimensional space, where the distances between points resemble the original dissimilarities. Here, we estimate the distances between Cα (or backbone) atoms for each pair of amino acids in the target sequence as distance matrix and then calculate the coordinates of the Cα atom for each amino acid. We obtain various distance restraints through the alignments to structural templates and then build one model for each set of distance restraint through applying MDS. In MUFOLD, we generate models using different techniques of MDS: classical metric MDS (noted as CMDS),[31] weighted MDS (noted as WMDS),[32] and split-and-combine MDS (noted as SC-MDS).[33]

CMDS is the first and simplest MDS algorithm. If the distance constraints are applied without error in the Euclidean space, CMDS will exactly recreate the configuration of points (or its mirror configuration) with a computational complexity of $O(n^3)$, where n is the number of points. When there are errors in the distance constraints, CMDS minimizes the sum of least squared errors between the estimated distances and the actual distances in the output model for all pairs of points. In practice, the technique can gracefully tolerate errors due to the overdetermination of the solution. This is important in the protein structure prediction, as our distance restraints can be incomplete, inaccurate, and inconsistent.

Besides CMDS, we also apply WMDS in MUFOLD. At first, we roughly assess the confidence of a distance constraint using the number of alignments that cover the residue pair and the quality of each alignment. For a residue pair covered by high-quality alignments (e.g., PSI-BLAST alignments with low *E*-values), we set higher weight for the corresponding restraint to force WMDS to satisfy these restraints with high priority. The experiments show that WMDS can generate more accurate results than CMDS.

We also use the SC-MDS technique to accelerate computation of large proteins. The main idea of SC-MDS is divide-and-conquer.[33] First, the whole set of points is divided into overlapping subsets with relatively small intersections. Then, an MDS algorithm, such as CMDS or WMDS, is applied to compute the coordinates of the points in each subset. Finally, the overlapping subsets are combined through affine transformation between the coordinates of their overlapping points. Let the size of intersections be p and the total number of points be n, the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

complexity of SC-MDS is $O(p^2n)$. When $p \ll n$, its complexity is $O(n)$, much faster than CMDS. Note that SC-MDS speeds up the computation but does not improve the accuracy of the models.

Spatial distance restraints

As mentioned in the earlier section, predicted distance restraints are often noisy. Hence, our strategy is to keep refining the initial models by sampling and improving the distance restraints (or contact maps) iteratively.

The initial contact maps of a target protein are retrieved from alignments between the target sequence and various template proteins in PDB obtained in the above step of "template selection and alignment." More specifically, we classify the alignments into two categories: one is global (longer) alignment between the target protein and its homologous or analogous template; the other is local (shorter) alignment between the target protein and some protein fragment. The number of fragments depends on two factors: (1) the gaps in the template hits and (2) the structural consistency or compatibility between the local alignments and the global alignments. We first select the fragments that can cover the gaps in global alignments. We then further select those fragments that have more consistent structures in the overlapping regions between the local alignment and the global alignment. The number of fragments is typically 50–200.

For a given alignment between the target and a template, we first estimate the pair-wise distance of the aligned residues in the target by the distance of the corresponding residues in the template. For those residues that have not been covered by the alignment (i.e., gaps in the alignment), we search for more fragments and then calculate the distance restraints from the corresponding alignments. Nevertheless, there may still be residues of the target that are aligned to gaps or two residues that are not covered by any single hit simultaneously so that related pair-wise distances cannot be derived. For these missing pair-wise distances, we estimate them by the shortest path distance. We know that the distance between the Cα atoms of adjacent amino acids is about 3.8 Å, which means that any two Cα atoms can be connected at least by adjacent Cα atoms. In fact, there are many different paths to connect two Cα atoms with adjacent Cα atoms or Cα pairs that have estimated distances from alignments. Each path has a length that equals to the sum of the distances of these connecting Cα pairs. The shortest path distance is used to estimate the unknown pair-wise distance. Although the shortest path often overestimates the distance, it provides an initial complete contact map for calculating a model by MDS.

Although using multiple templates and fragments can generate models that are closer to the native structure than any template alone, inconsistent restraints from different alignments and distances estimated by the shortest path method may compromise the quality of the models. Our strategy is to refine and improve the restraints iteratively by combining the original restraints derived from the alignments ($D_{\text{alignment}}$) and the measured distances from the generated models (D_{model}) as: $D_{\text{refine}} = \lambda * D_{\text{alignment}} + (1 - \lambda) * D_{\text{model}}$, $0 \leq \lambda \leq 1$. There are different ways to set the value of λ. For example, a simple way is to set $\lambda = 0.5$ if $D_{\text{alignment}}$ is available, otherwise $\lambda = 0$. Another way is to set λ according to the confidence level of $D_{\text{alignment}}$. By performing this iterative generation, the quality of models often gets better and better, while many deficiencies in the models are fixed over iterations.

An example of iteratively improved coarse-grain model generation is shown in Figure 2, where we show the original and improved contact maps in (a) and (c) and the corresponding models in (b) and (d), respectively. In the image of contact map, we use colors to illustrate the distances between pair-wise residues, where the lighter the color is, the larger the distance is. By comparing the data between Figure 2(a) and Figure 2(c), we can observe some changes in colors. For example, the colors around the column (and row) of 20, 40, 80, and 180 are heavier

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

in Figure 2(c) than those in Figure 2(a), which means that the distances of the corresponding residues to the other residues have been shortened.

It should be mentioned that MDS generates two mirror models for any given contact map. As we derive the model from template alignment, we can easily distinguish between correct and mirror configurations. Technically, we superimpose the model configuration to the template and calculate the reflection factor of the superimposition. If the reflection factor equals to 1, it indicates that the configuration is correct; otherwise, it is the incorrect mirror.

Coarse-grain model evaluation

Model generation using MDS leads to a large number of candidate structures. In MUFOLD, a process is used to select the near-native ones from the candidate set. Currently, structure quality assessment and model selection generally use the scoring functions in two categories[34]: physics-based energy functions and knowledge-based statistical potentials. The knowledge-based statistical potentials are typically fast, easy to construct, and hence are most widely used in the structure quality assessment. In MUFOLD, we apply a clustering-based and knowledge-based scoring method to evaluate and select the models for the next iteration of model improvement.

The main idea of MUFOLD model evaluation is an integrative approach by applying machine-learning methods based on the values of various scoring functions for the candidate structures. These scoring functions are normalized to *z*-scores, and the selected models are combined from the representative of clustered structures and ranked by *z*-score. Specifically, the method consists of four steps: filtering, clustering, cluster reduction, and centroid construction.

In the filtering step, poor structures are removed based on the thresholds and values of existing scoring functions. We have investigated some state-of-the-art scoring functions, such as OPUS, [35] Dope,[36] Model Evaluator,[37] Rapdf,[38] Dfire energy,[39] Hopp score,[40] and a geometric potential (Li and Liang, Submitted). The correlations between RMSDs of the models to native structures and the values of these functions vary significantly for different model sets: some work well on high-accuracy CM models, whereas others work well on hard cases. Through extensive experiments, we chose to use a subset of the scoring functions to filter and empirically set the corresponding threshold values.

Next, the remaining structures are grouped into clusters based on pair-wise similarity measured by RMSD. Most clustering methods including SPICKER[41] use pairwise root-mean-squared distance (pRMSD) between structures to indicate how these structures are similar to each other. Here, we use an efficient way to approximate pRMSD, referred to as reference root-mean-squared distance (rRMSD) developed by Li and Zhou.[42] Our basic clustering process consists of computing rRMSD and finding a proper cutoff of rRMSD to form clusters.

For cluster reduction, we try to find a subset of good models in a cluster, which often leads to a better cluster centroid. We first apply the CMDS to map protein structures onto two-dimensional space using their rRMSD values as similarity measures. Empirically, we found that in the 2D map, points corresponding to good structures are often densely located together, whereas points corresponding to poor structures tend to scatter throughout the space. Thus, we develop several cluster reduction techniques, including cluster border shrinking, half-plane pruning, and quadrant-based pruning to iteratively remove sparsely distributed points.[43] In cluster border shrinking, a bounding rectangle along the *x*- and *y*-axis is created for all the points in the two-dimensional space. The distances from the medoid of all points to the four sides of the bounding box are computed. In each round, the four sides will move toward the medoid for a certain ratio, for example, 10%. Points outside the reduced bounding box are removed. A new medoid is computed based on remaining points, and so on. In half-plane

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

pruning, the distances of the medoid to the four borders (up, down, left, and right) are compared, and then the longest distance side is pruned with all points on the opposite side of the medoid are removed in each round. In quadrant-based pruning, a quadrant centering at the medoid is pruned in each round. The two-dimensional plane is divided into four quadrants by drawing two lines parallel to the two axes crossing at the cluster medoid. The quadrant with the largest area is removed. This technique is more conservative than the half-plane pruning. These cluster reduction techniques can be applied alone or combined. In our experiments, we found that a combination of these techniques usually generated better results.

Finally, one model is selected or generated from each cluster. One method is to select the medoid of the cluster, that is, to select a structural model whose average RMSD to all the models in the cluster is minimal. Another way is to use the centroid of a cluster, that is, a model that averages the coordinates of all the models in the cluster. A centroid model is usually better than the medoid in the RMSD measure, but is often more likely to produce clashes among atoms. Therefore, we allow users to set either medoid or centroids as the final models. A user can also set the number of final models. For example, in CASP8 predictions, five final models were generated for each target.

Full-atom model evaluation

The coarse-grain model generation and evaluation method described in the earlier section provides various structures with significantly different conformations. How to identify the best of them, that is, the one with the smallest RMSD compared with the unknown native structure is highly challenging. Existing methods generally use static scoring functions[35–40] to rank models. However, the dynamics properties of a model may reveal its structural quality better than static scoring functions. We assume that near native models are more stable than poor-quality models during simulated heating, that is, the latter unfold at lower temperatures than the former. Thus, the quantitative assessment of relative stabilities of structural models against gradual heating provides an alternative way of ranking the structures' quality. In this article, we propose an MD-Ranking (MDR) method based on full-atom MD simulations[44] to evaluate and rank protein models according to their stabilities against external perturbations, for example, change in temperature or externally applied forces. The basic idea is to build all-atom models from the coarse-grain (Cα or backbone level) models, optimize models by energy minimization, gradually heat the models, and then rank the models based on their structural changes during heating.

More specifically, first, an all-atom model is built for each of the top 5–20 selected structures by the above coarse-grain model generation and evaluation process. The coordinates of the missing backbone and side-chain heavy atoms are predicted by using the program Pul-chra, [45] and the hydrogen atoms are added by using psfgen, which is part of the VMD package.[46] Next, the obtained structures are optimized by removing the bad contacts through energy minimization. Finally, the stability of a structure is tested by monitoring the change of its Cα RMSD (cRMSD) with respect to its initial structure during the MD simulation of a scheduled heating at a rate of 1 K/ps. The MD simulations are carried out in vacuum by coupling the system to a Langevin heat bath whose temperature can be varied (i.e., the dynamics of protein atoms is described by a Langevin equation). All energy minimization and MD simulations were performed by using the CHARMM27 force field and the parallel NAMD2.6 MD simulation program.[44] It should be noted that besides our own generated models, the MDR method can evaluate models from other software tools, such as Rosetta.[21]

To demonstrate how the MDR method works, Figure 3 shows three typical plots of the changes in cRMSD during the heating MD simulations. For the first case [Fig. 3(a)], it contains a structure with RMSD < 3 Å to the native one, and the selection of the best structure from the dataset often works well. When the best structure in the set has RMSD > 3 Å, the RMSD of

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

the top ranked structure is within 0.5 Å from the best structure in most cases [Fig. 3(b)]. In a few cases, however, the MDR method yielded only mediocre results, as shown in Figure 3(c), where the curves of different cRMSD changes mostly overlap with the lack of discerning power. In summary, the performance of MDR varies for different cases, while it is most efficient when the pool of decoys contains high-quality models (RMSD < 3 Å) besides poor ones.

RESULTS

Over the past 2 years, MUFOLD has been consistently improved with the same 200 target proteins and the same template database of 2248 proteins as the benchmark. In addition, MUFOLD was applied to the category of "tertiary structure predictions" in CASP8. Here, we evaluate MUFOLD models on the 200 target proteins and CASP8 targets.

Structure predictions for 200 benchmark sequences

The 2248 database proteins are selected from the December 2006 release of PDBSelect database[47] with length varying from 50 to 1500. The PDBSelect database consists of proteins such that the sequence identity between any two proteins in the database does not exceed 25%. The 200 target proteins were selected from the July 2005 release of Astral database,[48] and they satisfy the following criteria: (1) length ≤150 and (2) the sequence identity to any protein in the template database does not exceed 25%.

We compared the performance of MUFOLD by implementing five different methods, (1) AA +CMDS, (2)(AA+SS)+CMDS, (3) (AA+SS)+WMDS, (4) and (5) iterative (AA+SS)+CMDS, over time as shown in Table I. Here "AA" represents sequence alignment information, "SS" means secondary structure alignment information, and "CMDS" and "WMDS" indicate generating model by classical MDS and weighted MDS, respectively. As the SC-MDS method just speeds up the calculation but does not improve the accuracy of the models, we did not include it in this table. In the first four methods, (1), (2), (3), and (4), we only used our in-house tool MUPRED to predict the secondary structure and PSI-BLAST to obtain alignments, whereas in method (5), we applied PSIPRED and SSPro to do the secondary structure prediction, and HHSearch to obtain more alignment information. In future, we will apply our in-house threading tool Prospect[10] to obtain more alignments information.

For the model generation of each target protein, about 10,000 models are produced, and the best model against the native structure is recorded to assess the model generation capacity. From the data of methods (1) and (2), we can see that there is a significant improvement after using secondary structure alignments. This is because secondary structure alignments can help us find some hits without significant sequence similarity but still having similar structures. The data of method (3) show that WMDS can obtain more accurate models than CMDS. However, WMDS needs much more computational resources (an order of magnitude more CPU time). From the data in method (4), we can see that the iterative method with CMDS is very powerful to improve the prediction results. The data in method (5) indicate that by using different predicted secondary structure information and applying different alignment tools, we can obtain more alignment information, which can further help improve the prediction accuracy. This is probably because we can sample more diverse folds and alignments to include more accurate distance restraints. In addition, the GDT_TS scores of various methods have similar improvements as RMSD. Note that GDT_TS (Global Distance Test Total Score)[49] measurements are used as a major assessment criterion in the structure prediction. Here GDT_TS = 100*(GDT_P1 + GDT_P2 + GDT_P4 + GDT_P8)/4, where GDT_Pn denotes the percentage of residues in the model structure falling within a defined distance cutoff nÅ of their position in the experimental structure. For example, the average gain of GDT_TS score of method (5) over method (4) on the 200 target proteins is 2.79.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Structure predictions for CASP8 targets

A total of 81 automated servers participated in the "tertiary structure predictions" of CASP8. The CASP8 organizers used two broad categories to classify modeling difficulties of the targets: the "template-based modeling" category, including domains where a suitable template can be identified that covers all or nearly all of the target, and the "template-free modeling" (FM) category including 13 proteins for which no suitable template can be identified. As in previous CASPs, independent assessors evaluated five submitted models for each target sequence.

In the category of "tertiary structure predictions," we registered two servers in the CASP8 competition: MUFOLD-Server (with team code of 462) and MUFOLD-MD server (with team code of 404). MUFOLD-Server included the template selection and alignment, and coarse-grain model generation and evaluation described earlier. MUFOLD-MD applied Rosetta (version 2.2.0) for *ab initio* generation of models. When Rosetta failed to provide structures for large proteins, the models were generated by MUFOLD-Server. MUFOLD-MD applied Rosetta to generate 10,000 *ab initio* models, selected the top 64 models using the Rosetta energy function (when Rosetta failed, the top 64 models were provided by the MUFOLD-server), ran MD simulation with scheduled heating from 40 to 140 K, at a rate of 1 K/ps, and ranked models using their relative change in cRMSD during heating.

As a completely new framework for protein structure prediction, MUFOLD is at its early stage of development and its performance as an automated server had not been fine-tuned before CASP8 (that is why we used Rosetta models for MUFOLD-MD). Nevertheless, our servers still worked very well for some targets, including both template-based targets and template-free targets. For example, for some of template-based targets, such as T0398, T0443, T0444, and T0499, MUFOLD-Server models are ranked top 3 among 81 competing servers based on their GDT-TS scores. For some FM targets, such as for T0510_D3, a MUFOLD-MD model was ranked the best model by the three independent assessors. For target T0405_D2, which was a particularly difficult target, a MUFOLD-MD prediction was rated in the top 3 by two of the assessors. For target T0416-D2, two of the assessors rated MUFOLD-MD model as No.1.

For target T0444 as shown in Figure 4, the first MUFOLD-Server model (T0444_TS462_1) has the closest structural alignment to the native structure with an RMSD of 1.02 Å and the GDT_TS score of 94.6. A comparison of top 20 models generated by various servers for T0444 in terms of the percentage of correctly aligned residues (by structure alignment) shows that the MUFOLD-Server model is capable of achieving more accurate alignment at the C-terminal region than other server models. One advantage of MUFOLD over some other servers is that it applied fragments alignment information, which helps cover regions that do not have significant structural information in the full-length template.

T0510 is a template-free target and it has three domains: D1: 1–163, D2: 168–235, and D3: 236–279. We did remarkably well on target T0510_D3, as shown in Figure 5. The best model of MUFOLD-MD is the fourth model (T0510_4_D3), which was rated "the best" by the three independent assessors with RMSD of 6.8 Å and GDT_TS of 55.81. The C-terminal region (residues 255–279) has better prediction quality than the N-terminal one. This example demonstrates the proof of principles that dynamic properties can better rank models than static scoring functions.

For target T0416_D2, two of the assessors rated the second model T0416_2_D2 as the best model, as shown in Figure 6. Furthermore, this model is in a different "structural" cluster from models generated by other groups for this target. This is because most of existing methods use static scoring functions to rank models, and they tended to select models from similar "structural" clusters while we used the dynamics properties, which not only give a better

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

ranking to the best candidate in the pool but also select a distinct structure from the candidate pool. This example illustrates the uniqueness and novelty of our method.

Based on the official CASP8 results, in the FM category MUFOLD-MD is ranked No. 1 among all server predictions, as shown in Table II. The No. 1 ranking has been confirmed by the official CASP8 evaluation.[51] The ranking of the servers was done according to the total GDT-TS score (last column) obtained by summing up the GDT-TS scores of the best of the submitted five models for each of the 13 FM targets. It should be noted that this evaluation does not include manual predictions and various evaluation methods may give different rankings of servers. The success of MUFOLD-MD in the FM category is attributed not only to Rosetta but also to our MDR method that overall outperformed (though only marginally) the structure ranking based on the Rosetta scoring function, referred to as the RSR method. This point is illustrated in Figure 7, where we compare GDT-TS scores for structures of the 13 FM targets obtained as follows: (i) the best of the 10^4 structures generated by Rosetta (dark blue); (ii) the best of the top 64 structures obtained using the RSR method (light blue); (iii) the best of the top 5 structures selected by RSR (red); (iv) the best of the top 5 structures selected with our MDR method (green); and (v) the best of the five models submitted by the BAKER-ROBETTA server. In most of the cases, the MDR method outperformed both RSR and ROBETTA. The MDR method might have performed even better if the quality of the top 64 structures determined through RSR would have been better.

To further demonstrate the advantages of MDR with respect to RSR, we have compared the GDT-TS scores of the best of top 5 structures, corresponding to 41 CASP8 targets from the Human and Server category, generated by MUFOLD-MD and selected by using the MDR and RSR methods, respectively. The results are shown in Table III and Figure 8. Although in most of the cases the difference between the GDT-TS scores was smaller than 5% (blue bars in Fig. 8), the MDR result (green bars) was substantially better than the RSR one (red bars) in almost twice as many cases.

Improvement of MUFOLD-server since CASP8

The MUFOLD-Server that we applied in CASP8 was newly developed and did not go through sufficient tests and validations. Hence, its performance was not satisfactory. In particular, it failed completely in a number of easy targets where most servers predicted well. We have kept improving the algorithms and the software of MUFOLD since CASP8. Major changes include (1) database upgrading, (2) template and fragment selection, and (3) fine-tuning of the program. For example, instead of the 30% nonredundant template database used in CASP8, the new version uses a complete template database consisting of all PDB entries. Therefore, for each target the new version often obtains closer templates and better alignments. In addition, the new version uses more stringent criteria to select templates and fragment alignments, and fewer models with poor quality are included in the structural model pool. Consequently, the clustering and ranking methods are able to work better on these new models with narrower structural diversities, especially for “easy” and “medium” cases. Furthermore, the new version better handles many technical details, such as nonstandard amino acids in templates and optimization of various parameters used in MUFOLD.

We tested the latest MUFOLD on the 50 prediction targets in the CASP8 Human and Server section (http://predictioncenter.org/casp8/targetlist.cgi) that have available PDB files. Here, we used a template database as of May 13, 2008 (which was earlier than submission deadline of any CASP8 target). For each target, around 6000 models were generated, and finally, only the top-1 ranked model by MUFOLD was selected. We compared the performance between the latest MUFOLD-Server (named as MUFOLD-Server-New) and the MUFOLD-Server in CASP8 (named as MUFOLD-Server-Old), as shown in Figure 9(a). We also evaluated the capacity of clustering and ranking of models generated by MUFOLD by comparing the best

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

model that MUFOLD-Server-New generated and the top-1 model that MUFOLD-Server-New ranked, as shown in Figure 9(b). Figure 9(a) shows a dramatic improvement of MUFOLD since CASP8. Especially for eight targets, MUFOLD-Server models in CASP8 have less than 40 GDT-TS scores and new models have more than 60 GDT-TS scores. Figure 9(b) shows that the ranking method works particularly well for easier cases (with the GDT-TS score larger than 50), although it still fails for most of the hard cases.

We also compared the MUFOLD-Server-New with two top servers in the field, that is, Zhang-server and BAKER-ROBETTA, both of which participated in CASP8. As shown in Figure 10, Zhang-Sever achieved an average GDT-TS score of 51.42 on the 50 targets used in Figure 9 and BAKER-ROBETTA obtained 48.12, whereas MUFOLD-Server-New got 46.44. It is clear that MUFOLD has not reached the performance of these two servers. Nevertheless, MUFOLD-Server-New outperform them in some cases; in particular, 10 of 50 are better than Zhang-Server and 23 of 50 are better than BAKER-ROBETTA, which shows some proof of principles of our new approach.

An important advantage of MUFOLD is its short computing time. We measured the CPU times per models that MUFOLD-Server-New took on these 50 targets, as shown in Figure 11. In addition, we fit the computing time y with the length of target x using a polynomial curve of $y = 5.60 * 10^{-8} * x^3 + 5.04 * 10^{-5} * x^2 - 0.01 * x + 0.8$. The nice fitting demonstrates that the computational complexity is up to the cube of target length, as expected for CMDS. For targets with less than 100 amino acids, MUFOLD-Server just needs a few minutes on a single CPU to generate thousands of models, whereas for targets with 200–300 amino acids, it takes a few hours to finish the prediction. From our experience of running Rosetta, MUFOLD-Server is typically more than 10 times faster than Rosetta given the same number of models generated. In particular, MUFOLD works better than Rosetta for large proteins (>200 amino acids), where the latter often fails to generate models. This is probably because Rosetta uses a Monte Carlo scheme for searching conformations, which does not work well for large proteins as the number of possible conformations grows exponentially with the protein size.

As protein structure prediction is a highly complex problem, MUFOLD needs a number of functionalities to be developed. For example, lack of important modules such as domain parsing and disorder region recognition in the MUFOLD system can lead to some poor predictions. For target T0405_D1 (residue 2–73 of the 282 residues), our model was poor in terms of GDT_TS score (29.30), GDT_TS ranking (#50), and RMSD = 2.82 Å. A comparison between the target and our best-predicted model is shown in Figure 12. In contrast, the best model from group (489) has GDT_TS score = 39.14 and GDT_TS rated = #1, with RMSD = 2.01 Å. Our poor performance was partially due to the prediction without any domain parsing.

DISCUSSION

We have developed MUFOLD, a hybrid methodology of using whole and partial templates information and along with new computational techniques for protein tertiary structure prediction. MUFOLD covers both template-based and *ab initio* predictions using the same framework and aims to achieve high accuracy and fast computing for automated prediction. Two major novel contributions of MUFOLD are graph-based model generation and MDR. By formulating the structure prediction problem as a graph realization problem, we can apply an efficient optimization approach of MDS to speed up the prediction dramatically. In addition, under this graph-based frame, we can iteratively improve the distance restraints by using the information from the models and thus improve the predictions consistently. MDS' computational efficiency allows MUFOLD to explore different folds and alignments, which can be evaluated in 3D models by energy functions instead of optimizing structures using these energy functions. This bypasses the computationally expensive process of optimization in

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

typical model constructions by other methods. MDR, in contrast to widely used static scoring functions, exploits dynamics properties of structures to evaluate their qualities, which can identify best structures from a handful of structures selected by static energy functions more effectively. This is a rare success in applications of MD simulation for general protein structure predictions.

We have kept improving MUFOLD and used the same 200 target proteins and the same template database of 2248 proteins to benchmark various methods. We have implemented the MUFOLD methodology into a releasable software package, which is available upon request. The executable programs will be available for general download by the end of 2009 after the system is further debugged and documented. MUFOLD demonstrated some success in CASP8, showing some unique advantages in comparison with other systems that have been developed for many more years. In particular, MUFOLD demonstrated some success in CASP8.

As a completely new framework for protein structure prediction, there are various limitations to address and new functionalities to implement for MUFOLD. MUFOLD currently can only handle protein monomers but not protein oligomers or complexes. Like other tools, it is important to combine predicted structural models and wet-lab experiments to take advantage of the power of protein structure prediction. We are improving the system in many aspects. For example, the current MDS method for model generation enforces partial template information and structural quality but does not use any energy function to do so. We are applying energy functions into the system to refine the predicted models. Another issue is that models generated by MUFOLD have unique features compared with general decoys, where general knowledge-based scoring functions were tested. We are exploring ways of training new scoring functions[52] specifically targeting our own structural models. Furthermore, the lack of solvent in the MD simulations may lead to further errors of ranking, especially for structures that show comparable change in cRMSD during heating. Therefore, the MDR ranking method can be improved by considering a longer heating interval, using the GDT-TS or TM score instead of cRMSD, and including implicit solvent in the simulation, although adding implicit solvent may not be feasible in general protein structure prediction due to long computational time. Finally, we will add more modules, such as multiple sequence alignment, disorder region recognition and removal, domain parsing, and local quality assessment/refinement into the automated predictions to achieve higher prediction accuracy. Our vision is to deliver MUFOLD to the desktop machines of experimental biologists. An experimentalist using the Windows or Mac operating system should easily make predictions and analyze the prediction results through a graphical user interface to be developed. The computing time would be much faster than the widely used tools like Modeller and Rosetta. This will bring substantial new impact of protein structure prediction on biomedical studies.

Acknowledgments

Major computer time was provided by the University of Missouri Bioinformatics Consortium. The authors thank Jianlin Cheng, Yang Zhang, and Joel L. Sussman for helpful discussions.

Grant sponsor: National Institutes of Health; Grant number: R21/R33-GM078601

References

1. The UniProt Consortium. The Universal Protein Resource (Uni-Prot). Nucleic Acids Res 2008;36:D190–D195. [PubMed: 18045787]
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242. [PubMed: 10592235]
3. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170. [PubMed: 1853201]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

4. Ring CS, Cohen FE. Modeling protein structures: construction and their applications. FASEB J 1993;7:783–790. [PubMed: 8330685]
5. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815. [PubMed: 8254673]
6. Fiser A, Do RK, Sali A. Modeling of loops in protein structures. Protein Sci 2000;9:1753–1773. [PubMed: 11045621]
7. Soding J, Biegert A, Lupas A. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 2005;33:W244–W248. [PubMed: 15980461]
8. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225. [PubMed: 9149153]
9. Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. J Mol Biol 1999;281:565–577. [PubMed: 9698570]
10. Xu Y, Xu D. Protein threading using PROSPECT: design and evaluation. Proteins: Struct Funct Bioinformatics 2000;40:343–354.
11. Inbar Y, Benyamini H, Nussinov R, Wolfson HJ. Protein structure prediction via combinatorial assembly of sub-structural units. Bioinformatics 2003;19:158–168.
12. Chikenji G, Fujitsuka Y, Takada S. A reversible fragment assembly method for *de novo* protein structure prediction. J Chem Phys 2003;119:6895–6903.
13. Skolnick J, Kihara D, Zhang Y. Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. Proteins: Struct Funct Bioinformatics 2004;56:502–518.
14. Lee J, Kim SY, Joo K, Kim I, Lee J. Prediction of protein tertiary structure using PROFESY, a novel method based on fragment assembly and conformational space annealing. Proteins: Struct Funct Bioinformatics 2004;56:704–714.
15. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci USA 2004;101:7594–7599. [PubMed: 15126668]
16. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 2008;9:40. [PubMed: 18215316]
17. Li Z, Scheraga HA. Monte carlo-minimization approach to the multiple-minima problem in protein folding. Proc Natl Acad Sci USA 1987;84:6611–6615. [PubMed: 3477791]
18. Liwo A, Lee J, Ripoll DR, Pillardy J, Scheraga HA. Protein structure prediction by global optimization of a potential energy function. Proc Natl Acad Sci USA 1999;96:5482–5485. [PubMed: 10318909]
19. Simons KT, Strauss C, Baker D. Prospects for ab initio protein structural genomics. J Mol Biol 2001;306:1191–1199. [PubMed: 11237627]
20. Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. Biophys J 2003;85:1145–1164. [PubMed: 12885659]
21. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res 2004;32:526–531.
22. Borg, I.; Groenen, P. Modern multidimensional scaling—theory and applications. New York: Springer-Verlag; 1997.
23. Moult J, Fidelis K, Kryshtafovych A, Rost B, Tramontano A. Critical assessment of methods of protein structure prediction—Round VIII. Proteins: Struct Funct Bioinformatics 2009;77(S9):1–4.
24. Altschul S, Madden T, Schäffer A, Zhang J, Zhang Z, Miller W, Lipman D. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402. [PubMed: 9254694]
25. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics 2005;21:951–960. [PubMed: 15531603]
26. Bondugula R, Xu D. MUPRED: a tool for bridging the gap between template based methods and sequence profile based methods for protein secondary structure prediction. Proteins: Struct Funct Bioinformatics 2007;66:664–670.
27. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202. [PubMed: 10493868]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

28. Cheng J, Randall A, Sweredoski M, Baldi P. SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 2005;33:72–76.
29. Shang, Y.; Bondugula, R.; Xu, D.; Wang, Q. A new method for protein tertiary structure prediction. Proceedings of the IASTED International Conference on Computational Intelligence; Banff, Canada. July 2007;
30. Levenberg K. A method for the solution of certain non-linear problems in least squares. Q Appl Math 1944;2:164–168.
31. Torgerson WS. Multidimensional scaling. I. Theory and method. Psychometrika 1952;17:401–419.
32. Torgerson WS. Multidimensional scaling of similarity. Psychometrika 1965;30:379–393. [PubMed: 5217606]
33. Tzeng J, Lu H, Li W. Multidimensional scaling for large genomic data sets. BMC Bioinformatics 2008;9:179. [PubMed: 18394154]
34. Xu, Y.; Xu, D.; Liang, J. Computational methods for protein structure prediction and modeling. Vol. 1, 2. New York: Springer-Verlag; 2006.
35. Wu Y, Lu M, Chen M, Li J, Ma J. OPUS-Ca: a knowledge-based potential function requiring only Ca positions. Protein Sci 2007;16:1449–1463. [PubMed: 17586777]
36. Shen M, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci 2006;15:2507–2524. [PubMed: 17075131]
37. Wang Z, Tegge A, Cheng J. Evaluating the absolute quality of a single protein model using structural features and support vector machines. Proteins: Struct Funct Bioinformatics 2009;75:638–647.
38. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1998;275:895–916. [PubMed: 9480776]
39. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11:2714–2726. [PubMed: 12381853]
40. Sims GE, Kim S-H. A method for evaluating the structural quality of protein models by using HOPP scoring. Proc Natl Acad Sci USA 2006;103:4428–4432. [PubMed: 16537409]
41. Zhang Y, Skolnick J. SPICKER: a clustering approach to identify near-native protein folds. J Comput Chem 2004;25:865–871. [PubMed: 15011258]
42. Li H, Zhou Y. SCUD: fast structure clustering of decoys using reference state to remove overall rotation. J Comput Chem 2005;26:1189–1192. [PubMed: 15954080]
43. Wang, Q.; Shang, Y.; Xu, D. A new clustering-based method for protein structure selection. Proceedings of the IEEE International Joint Conference on Neural Networks; Hong Kong. June 2008;
44. Phillips J, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel R, Kale L, Schulten K. Scalable molecular dynamics with NAMD. J Comput Chem 2005;26:1781–1802. [PubMed: 16222654]
45. Feig M, Rotkiewicz P, Kolinski A, Skolnick J, Brooks CL III. Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. Proteins: Struct Funct Bioinformatics 2000;41:86–97.
46. Humphrey W, Dalke A, Schulten K. VMD—Visual molecular dynamics. J Molec Graphics 1996;14:33–38.
47. Hobohm U, Scharf M, Schneider R, Sander C. Selection of a representative set of structures from the Brookhaven Protein Data Bank. Protein Sci 1992;1:409–417. [PubMed: 1304348]
48. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res 2000;28:254–256. [PubMed: 10592239]
49. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–374. [PubMed: 12824330]
50. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins: Struct Funct Bioinformatics 2004;57:702–710.
51. Moshe BD, Orly NB, Aviv P, Jaime P, Joel LS, Yaakov L. Assessment of CASP8 structure predictions for template free targets. Proteins: Struct Funct Bioinformatics 2009;77(S9):50–65.
52. Hu C, Li X, Liang J. Developing optimal nonlinear scoring function for protein design. Bioinformatics 2004;20:3080–3098. [PubMed: 15217818]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

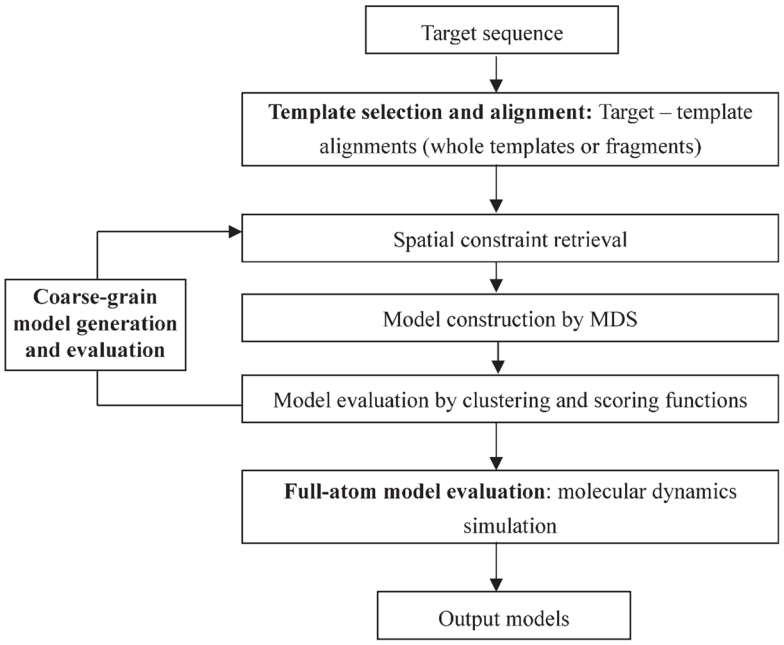

Figure 1.
Flowchart of the MUFOLD structure prediction method.

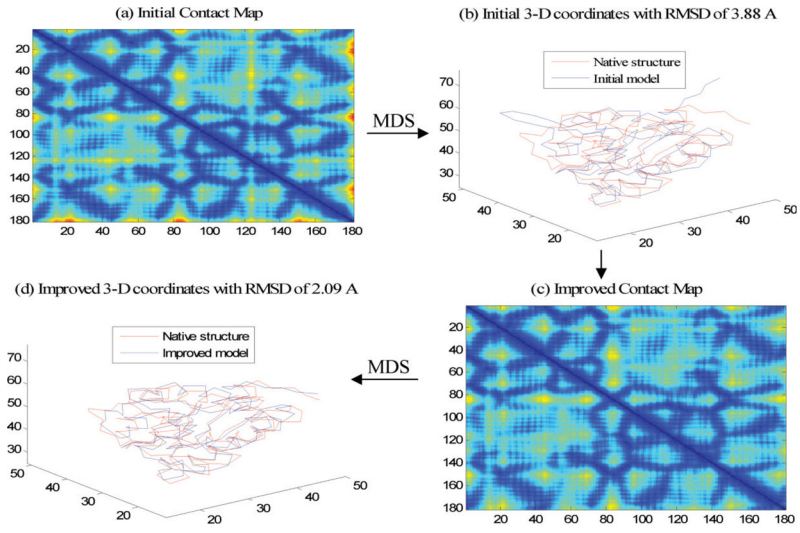

Figure 2.
An example of iterative coarse-grain model generation.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

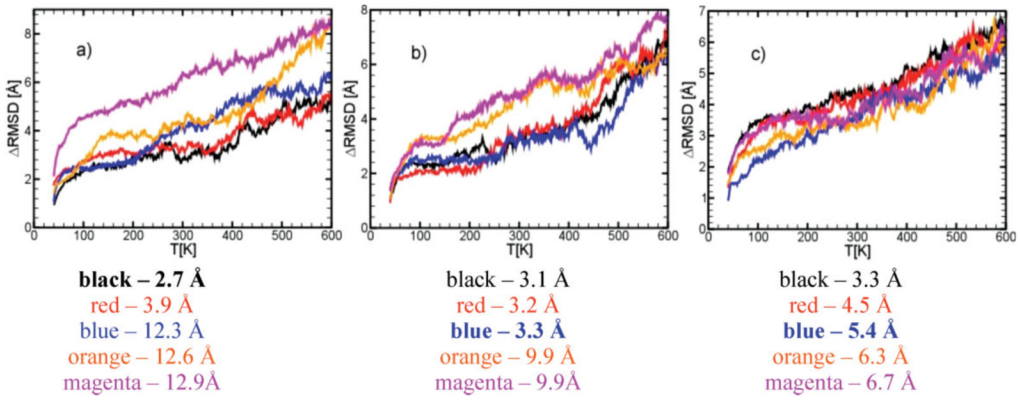

Figure 3.

Changes in cRMSD during the heating in MD simulations for three proteins. The colored curves correspond to five models with various RMSD values to the native structures as labeled. The bold fonts indicate top models ranked by MDR.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

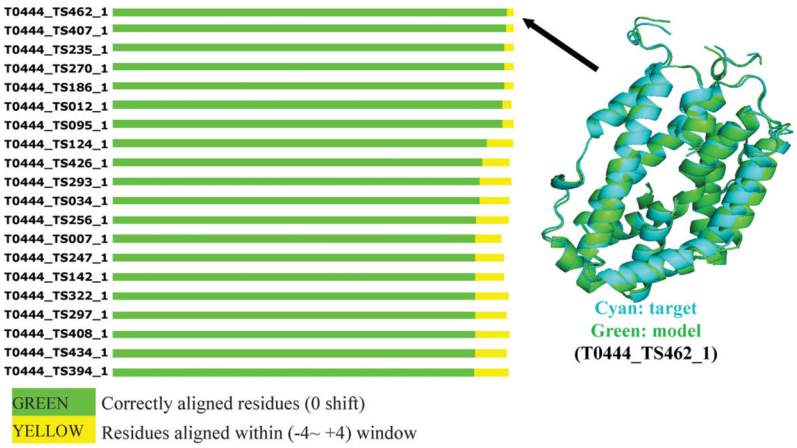

Figure 4.

Performance of MUFOLD-Server models on target T0444. A comparison of top 20 models by various CASP8 prediction groups for T0444 in terms of the percentage of residues in agreement with the structure alignment is shown on the left (picture generated at the CASP8 assessment website http://www.predictioncenter.org/casp8/results.cgi). The right side shows the structural superimposition between the model of T0444_TS462_1 and the native structure of T0444.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Models	RMSD(Å)	GDT-TS
T0510_ TS404_1_D3	10.141	44.77
T0510_ TS404_2_D3	9.392	40.70
T0510_ TS404_3_D3	8.942	45.93
T0510_ TS404_4_D3	6.845	55.81
T0510_ TS404_5_D3	9.230	43.60

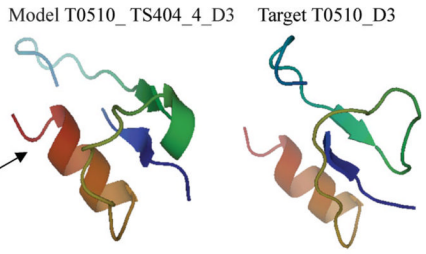

Figure 5.

Performance of five selected MUFOLD-MD models for target T0510_D3 (residues 236–279) in CASP8. The RMSD and GDT_TS scores were calculated by the TM-score package.[50] [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Models	RMSD(Å)	GDT-TS
T0416_ TS404_1_D2	5.554	55.70
T0416_ TS404_2_D2	4.213	70.18
T0416_ TS404_3_D2	4.414	61.84
T0416_ TS404_4_D2	5.133	57.46
T0416_ TS404_5_D2	5.145	56.14

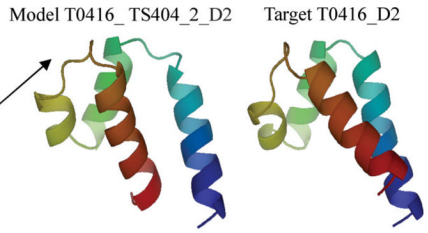

Figure 6.
Performance of five selected MUFOLD-MD models for target T0416_D2 in CASP8. The RMSD and GDT_TS scores were calculated by TM-score package.[50] [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

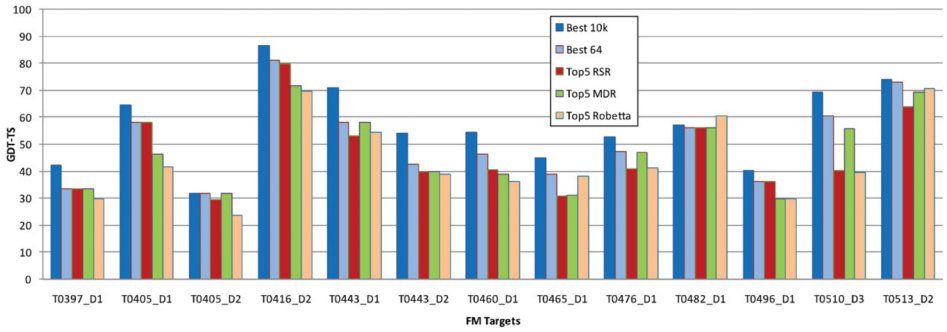

Figure 7.

Comparison of GDT-TS scores for structures, corresponding to 13 target proteins from the Free Modeling section in CASP8, obtained as follows: Best 10k = best of the 10,000 structures (i.e., the one closest to the native structure) generated by Rosetta used in the MUFOLD-MD server; Best 64 = best of the top 64 structures selected by the RSR method; Top5 RSR = best of the top 5 structures selected with RSR; Top5 MDR = best of the top 5 structures selected with MDR among “Best 64”; and Top5 Robetta = best of the five structures produced by the BAKER-ROBETTA server.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

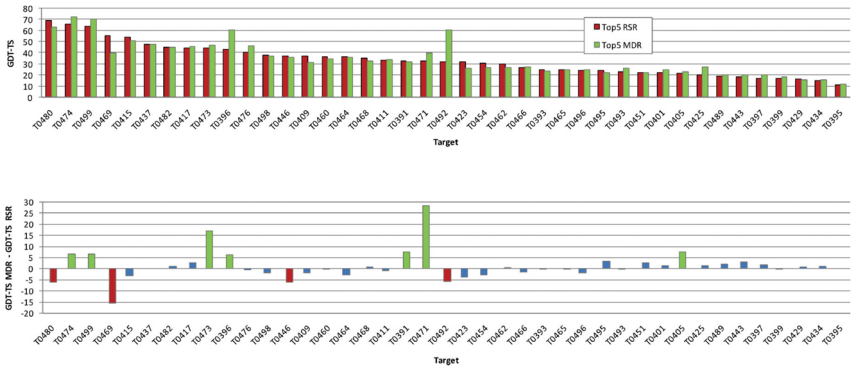

Figure 8.

Top: Comparison of the GDT-TS scores of the best of top 5 structures, corresponding to 41 target proteins from the Human and Server section of CASP8, selected by the MUFOLD-MD server using the MDR (“Top5 MDR”) and RSR (“Top5 RSR”) methods, respectively. Bottom: Difference between the GDT-TS scores obtained with the MDR and RSR methods for the same structures as in the top panel. The green (or red) bars indicate that the GDT-TS score of the structure picked by the MDR (or RSR) method was at least five points bigger than the other one. For the blue bars, the difference between the GDT-TS scores (in absolute value) was less than five points.

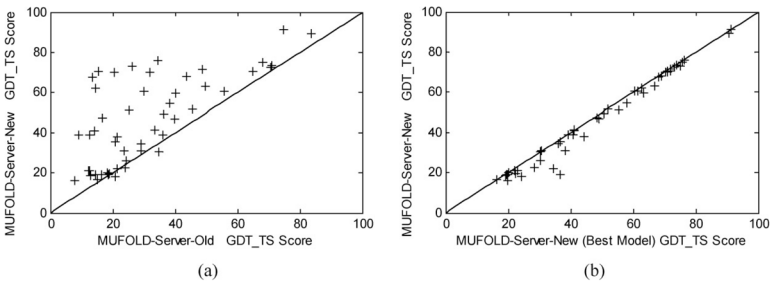

Figure 9.

(**a**) Comparison between the MUFOLD-Server-New and the MUFOLD-Server-Old that was used in CASP8; (**b**) comparison between the best-generated model (against the native structure) and the top-1 model picked by MUFOLD-Server-New. The GDT_TS scores were calculated by the TM-score package.[50]

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

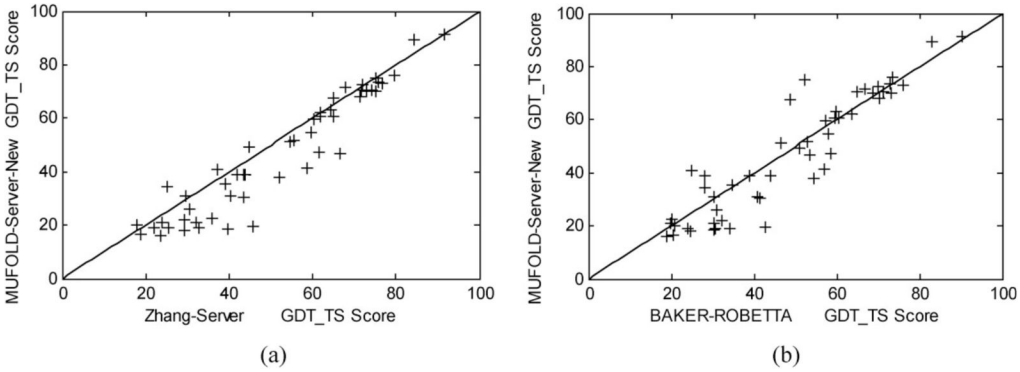

Figure 10.
Comparison between the MUFOLD-Server-New and Zhang-Server (**a**) or BAKER-ROBETTA (**b**) in CASP8. The GDT_TS scores were calculated by the TM-score package.[50]

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

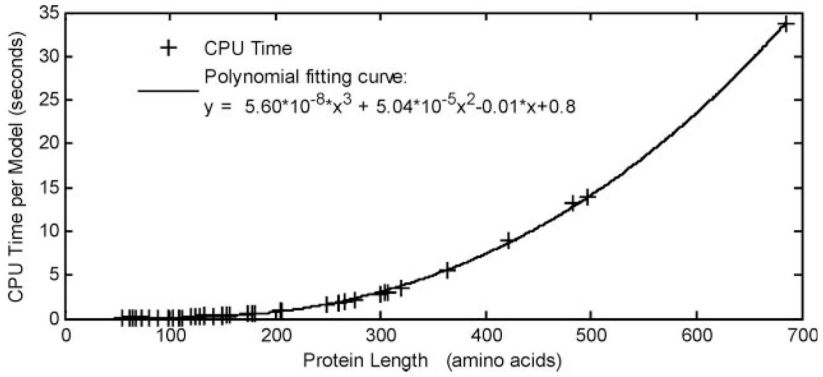

Figure 11.
Single CPU time (Xeon E5440 Processor, 2.83GHz) per model that MUFOLD-Server-New took for model generation on the 50 prediction targets in the CASP8 Human and Server section.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

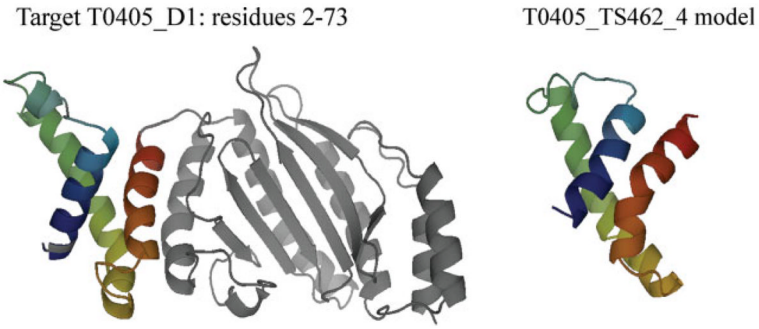

Figure 12.
MUFOLD prediction for target T0405_D1.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table I

Comparison of Five Methods in Prediction Quality

	Quality of the best model						
Method of model generation	**<2 Å**	**<3 Å**	**<4 Å**	**<6 Å**	**<10 Å**	**<12 Å**	**Mean (Å)**
(1) (AA)[a] + CMDS	1	5		40	132	180	8.58
(2) (AA + SS[b])[a] +CMDS	4	14	18	81	178	197	6.86
(3) (AA + SS[b])[a] + WMDS	6	20	43	111	192	200	5.91
(4) Iterative (AA + SS[b])[a] + CMDS	11	34	79 (99 < 4.5 Å)	147	200	200	4.74
(5) Iterative (AA + SS[c])[d] + CMDS	19	53	104 (118 < 4.5 Å)	164	200	200	4.28

AA represents sequence alignment information; SS means secondary structure alignment information; CMDS and WMDS indicate generating model by classical MDS and weighted MDS, respectively.

[a] Sequence-profile alignments by PSI-BLAST.

[b] Predicted secondary structures by MUPRED.

[c] Predicted secondary structures by MUPRED, SSPro, and PSIPRED.

[d] Sequence-profile alignments by PSI-BLAST and profile–profile alignments by HHSearch.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table II

Top 10 Servers in the CASP8 Free Modeling (FM) Category

	Target													
Server	**T0397_D1**	**T0405_D1**	**T0405_D2**	**T0416_D2**	**T0443_D1**	**T0443_D2**	**T0460_D1**	**T0465_D1**	**T0476_D1**	**T0482_D1**	**T0496_D1**	**T0510_D3**	**T0513_D2**	**GDT-TS Sum**
MUFOLD-MD	33.54	46.18	31.85	71.49	57.95	40.00	40.00	31.51	47.13	41.79	29.79	55.81	69.20	596.24
BAKER-ROBETTA	29.57	41.67	23.68	69.74	54.55	38.75	38.75	38.28	41.09	60.45	29.79	39.53	70.65	576.50
Zhang-Server	29.57	43.40	20.67	66.23	53.41	41.67	41.67	36.72	41.38	49.25	28.54	35.47	61.23	549.21
RBO-Proteus	35.06	37.50	24.76	51.32	43.94	38.33	38.33	35.68	25.57	65.30	29.38	43.02	43.48	511.67
pro-sp3-TASSER	27.44	58.68	16.95	39.47	49.24	43.33	43.33	28.39	36.49	57.84	23.33	45.35	41.30	511.14
PSI	25.61	57.99	21.39	38.60	45.83	41.67	41.67	28.91	35.34	55.60	29.79	50.58	36.23	509.21
FALCON	32.32	55.56	21.39	38.60	41.67	41.67	41.67	37.50	31.32	49.63	25.42	43.02	47.83	507.60
FALCON_CONSENSUS	32.32	55.56	21.39	38.60	38.64	41.67	41.67	35.94	31.32	48.51	25.83	43.02	47.83	502.30
METATASSER	33.54	47.92	17.79	51.75	40.53	41.67	41.67	26.82	32.76	55.22	26.25	39.53	31.52	486.97
fais-server	25.00	48.96	20.79	52.19	51.14	40.83	40.83	30.21	30.46	41.42	23.54	41.86	38.41	485.64

The ranking is based on the total GDT-TS (last column) obtained by summing up the GDT-TS scores of the best of the submitted five models for each of the 13 FM models.

Proteins. Author manuscript; available in PMC 2010 June 15.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table III

GDT-TS Scores of the Best of Top 5 Structures Ranked by the Rosetta Scoring Function (RSR) and the MDR Method

	GDT-TS			GDT-TS	
Target	**Top5 RSR**	**Top5 MDR**	**Target**	**Top5 RSR**	**Top5 MDR**
T0480	69.12	63.24	T0423	31.71	26.01
T0474	65.76	72.28	T0454	30.86	26.95
T0499	63.84	70.54	T0462	29.62	26.71
T0469	55.00	39.62	T0466	26.72	27.30
T0415	53.67	50.69	T0393	24.90	23.54
T0437	47.50	47.50	T0465	24.79	24.59
T0482	44.74	44.74	T0496	24.29	24.57
T0417	44.18	45.28	T0495	23.97	22.26
T0473	44.12	46.69	T0493	22.83	26.24
T0396	43.14	60.29	T0451	22.18	21.99
T0476	40.06	46.31	T0401	21.97	24.81
T0498	37.50	36.98	T0405	21.53	23.04
T0446	37.39	35.68	T0425	19.89	27.62
T0409	37.14	31.07	T0489	18.69	20.00
T0460	36.52	34.55	T0443	18.00	20.13
T0464	36.30	35.96	T0397	16.83	20.00
T0468	35.34	32.47	T0399	16.67	18.36
T0411	33.33	33.96	T0429	16.07	15.89
T0391	32.72	31.99	T0434	15.08	15.92
T0471	32.25	39.69	T0395	11.06	12.02
T0492	32.19	60.62			
Average GDT-TS				33.15	34.34

The ranking was performed using the top 64 Rosetta-generated structures determined through RSR. In this analysis, we used only 41 target proteins out of the listed 57 proteins from the Human and Server section of CASP8. We dropped 16 targets for various reasons, that is, 3 were canceled by the organizers (T0387, T0403, T0467), 2 were not included in the official CASP8 results (T0484, T0500), 7 for which Rosetta generated five or less structures (T0389, T0413, T0440, T0407, T0421, T0449, T0457), and 4 for which the quality of the generated structures is extremely low, with GDT-TS score less than 10 (T0419, T0427, T0431, T0487).