Protein Structure Refinement by Iterative Fragment Exchange

Debswapna Bhattacharya
Department of Computer Science
University of Missouri
Columbia, MO 65211, USA
+1-(801)-819-5828
db279@mail.missouri.edu

Jianlin Cheng
Department of Computer Science,
Informatics Institute, C. Bond Life
Science Center
University of Missouri
Columbia, MO 65211, USA
+1-(573)-882-7306
chengji@missouri.edu

ABSTRACT

Despite significant advancement of computational methods in protein structure prediction during the last decade, these techniques often cannot achieve allowable prediction accuracy to be applied in solving biological problems. Bringing these low-resolution predicted models to high-resolution structures close to their native state, called the protein structure refinement problem, however, has proven to be extremely challenging and a largely unsolved problem in the field of protein structure prediction. Here, we propose a new approach to protein structure refinement by iterative fragment exchange, called REFINEpro. The protocol first identifies the less conserved local regions in the initial model by consensus approach using ensemble of models produced for the same protein target. We call these regions problematic regions (PRs). The qualities of the PRs are then iteratively improved by exchanging better-modeled fragments corresponding to these PRs from structures in the ensemble. This method has been tested on benchmark datasets comprising of decoys generated through both template-based and ab-initio protein structure prediction methods and exhibits promising improvement over the initial models in both global and local model quality measures, indicating a new avenue to solve the protein structure refinement problem. REFINEpro web server is freely available at http://sysbio.rnet.missouri.edu/REFINEpro/.

Categories and Subject Descriptors

I.6.5 [**Modeling methodologies**]: Computational modeling of protein structure, improving quality of modeling.

General Terms

Design

Keywords

protein structure refinement, protein structure prediction, iterative improvement, unreliable local regions, fragment exchange, consensus quality assessment.

1. INTRODUCTION

The advancement of template based modeling (TBM) techniques and the expansion of protein sequence and structure spaces have certainly improved the average qualities of protein models over the previous decades. However, the contemporary computational protein structure prediction methods still lacks the consistent accuracy needed to be successfully applied to address biological problems. Refinement of these predicted models in order to enhance the qualities thereby bringing them closer to the native state is, therefore, an integral part of protein structure prediction pipelines.

Efforts to solve the protein structure refinement problem have usually been rooted in two schools of thoughts. One is physics based methods which is governed by the thermodynamic hypothesis proposed by Anfinsen that the native structure of a protein corresponds to the global minimum of its free energy [1]. Consequently, a force-field is first developed to calculate the potential energy of the initial protein model. Then the potential energy is minimized through conformation changes with the goal to find the free-energy minimum in the protein energy landscape using traditional molecular mechanics (MM) potentials like AMBER99 [2, 3], OPLS-AA [4], etc. A number of noteworthy studies have been performed in this direction [5, 6]. However, there are two major bottlenecks of these methods: (1) limited accuracy of physics based empirical force fields and (2) "multiple-minima problem" arising from the presence of many local minima in protein's multidimensional energy landscape [7]. The other school of thoughts is "knowledge-based" methods that utilize the statistical potentials derived from the analysis of recurrent patterns in experimentally derived protein structures and sequences [8]. Molecular Dynamics (MD) simulation is widely used in this kind of protocols (Fan and Mark, 2004; Lee, et al., 2001) to move every atom of a protein. Apart for some isolated cases, however, no systematic structural improvement has been attained [9].

Some promising progress has been made in the recent past by combining the two school of thoughts, that is, by using composite physics and knowledge-based potentials [10-12] to solve protein structure refinement problem. Although encouraging, these techniques highlight a key issue in protein structure refinement – that is, majority of these methods follow a conservative local sampling strategy around the starting structures producing improvement only in local qualities of the models rather than substantially improving the backbone positioning. Development of a method capable of performing global refinement aiming to resolve differences in the overall fold of the protein model is, therefore, a crucial step forward for solving protein structure refinement problem and more generally, towards the improvement of computational protein structure prediction.

We previously developed a refinement procedure, called 3Drefine [13] by optimizing the hydrogen bonding network and atomic level energy minimization using a composite physics and knowledge-based force field. We now extend the approach by partitioning the structure refinement process into two stages: (1) global refinement with the goal of improving the overall fold of the starting model and (2) local refinement which aims at correction of local errors like irregular hydrogen bonding, steric clashes, unphysical bond length, unrealistic bond angles, torsion angles and side-chain χ angles. These two stages demands different approaches and in accordance with the previous studies [14, 15], we decided to perform them sequentially. The first stage is accomplished by iterative fragment exchange based on an ensemble and we use our previous method 3Drefine to perform the refinement of the general physicality of the models. The new method, named REFINEpro, has been tested on diverse and independent benchmark datasets and has demonstrated significant potential in simultaneous improvement of the overall fold and resolving the local errors, thereby improving both the global and local qualities of the starting models.

2. MATERIALS AND METHODS

The REFINEpro protocol is shown in Figure 1. Given a target protein structure and a model ensemble consisting of numerous structures generated for the same target, the method first identifies the problematic regions (PRs) in the initial model using a consensus method. An iterative refinement is then applied for each PR via generation of hybrid models by combining the initial model and the PR exchanged by structures from the ensemble followed by quality assessment to select the best hybrid model. Finally, atomic level energy minimization is performed on the best hybrid model using 3Drefine to optimize the hydrogen-bonding network and to improve the local qualities in order to produce the refined structure. The procedure is fully automated and the average running time for each refinement target is less than 4 hours at a 2.4 GHz CPU. The REFINEpro web server is freely available at http://sysbio.rnet.missouri.edu/REFINEpro/.

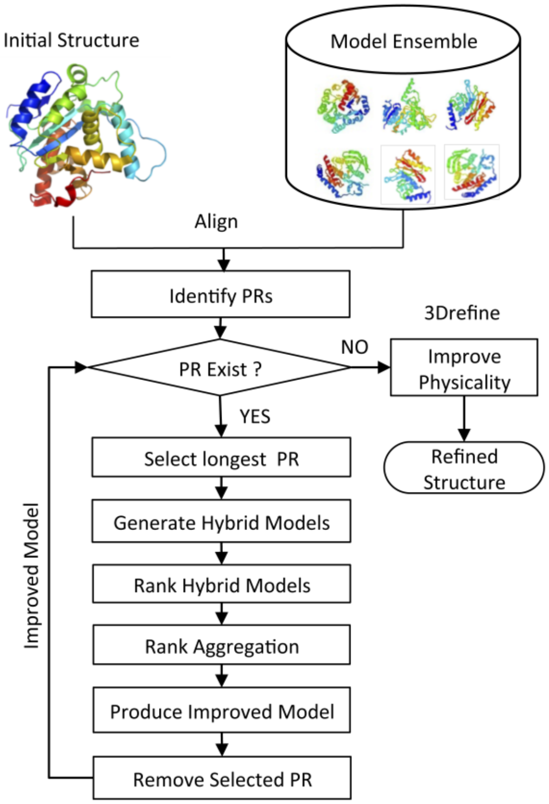

Figure 1. Flow Chart of the REFINEpro protocol.

The protocol indicates stages of the iterative refinement including identification of PRs in the initial model, generation of hybrid models, quality assessment to produce the improved model followed by correcting local errors using 3Drefine to produce refined model.

2.1 Predicting PRs

It is vital to identify appropriate regions of refinement in the initial model. In template based modeling, the regions build from reliable template information tend to be largely correct and any attempt to refine them may pose the risk of degrading the model quality. The key is, therefore, to detect the problematic regions (PRs) in the starting structure and then try to refine them while keeping the conformation of the reliable regions unaltered. In REFINEpro, consensus local quality assessment technique was adapted using a model ensemble approach. For identification of PRs, we used a quality measure, similar to one originally developed by Levitt and Gerstein [16] and later widely used in developing global structural similarity measures such as TM-score [17] or MaxSub [18] and for local quality assessment of protein models [19-21]. This is called S-score.

In the first step, the models in the ensemble with missing residues were discarded and the valid structures were trimmed to exactly match the residues in the starting model. Then, each model in the ensemble were superposed into the initial model using TM-score program [17]. Once the superposed model pool was organized, we calculated the distance between C_α atoms after optimal structural superposition between the initial structure and each model in the ensemble. The distance was converted into the S-score using the equation:

$$S_i^j = \frac{1}{1+\left(\frac{d_i^j}{d_0}\right)^2} \qquad (1)$$

where S_i^j is the S-score of ith residue of the starting model with respect to the jth model in the ensemble, d_i^j is the calculated C_α distance between residue i of the starting model and the corresponding residue of the jth model in the ensemble and d_0 is the distance threshold. The purpose of converting the distance into S-score is that there is no upper limit to the distance and it is, therefore, often dominated by outliers. We used $d_0 = 5Å$ as in LGscore [16]. For smoothing purpose and to avoid local fluctuations, we adopted a sliding window approach. The central residue and their sequence neighbors were selected in a window of fixed size of three residues and then the average S-score was calculated in the window to determine the correctness of the central residue. Therefore, the μ_i^j, the average S-score of ith residue with respect to the jth model in the ensemble is defined as:

$$\mu_i^j = \frac{1}{3}\sum_{k=i-1}^{i+1} S_k^j \qquad (2)$$

We then use the decision function p_i^j to determine whether residue i in the starting model have more than 5Å deviation compared to model j in ensemble as below:

$$p_i^j = \begin{cases} 1 : \mu_i^j < 0.5 \\ 0 : otherwise \end{cases} \quad (3)$$

where p_i^j assumes the value 1 if the ith residue in the initial model deviates more than 5Å (i.e. $S_i^{win} < 0.5$) with respect to jth model in the ensemble and 0 otherwise. The empirical likelihood of residue conservation in a model ensemble is defined using the equation:

$$RC_i = \frac{1}{n}\sum_{k=1}^{n} p_i^k \quad (4)$$

where RC_i is the probability of the ith residue in the initial model to have average fluctuation more than 5Å for all models in the ensemble and n is the number of valid models in ensemble. Higher RC_i values indicates more fluctuation and therefore, more likely to be problematic. The rationale behind this step is: reliable residues tend to be conserved across the model ensemble, but the problematic residues prone to have higher diversity in terms of backbone C_α positioning. The threshold of RC_i (hereafter called the residue conservation index) was chosen to be 0.5. This means if majority of structures in the ensemble have average fluctuation of more than 5Å compared to initial model (i.e. $RC_i > 0.5$), we consider the residue as problematic. Any region in the initial model having more than five consecutive problematic residues was considered as PR.

2.2 Generating hybrid models and quality assessment

Once the PRs in the initial model are identified, we attempt to refine each PR by exchanging better modeled fragments from the ensemble for the corresponding PR while keeping the reliable regions fixed. Our hypothesis is: if a better modeled fragment for a PR exists in the ensemble, then exchanging the conformation for that PR from the better quality fragment without any changes to the rest of the structure should improve the overall fold for the PR. We adopted a hybrid model generation approach to assemble the better modeled fragments with the starting structure. First, the reliable regions were kept fixed in the initial model and the coordinates of the PR were replaced from each model in ensemble. Then, we used this model as a template for Modeller to generate hybrid model. The automodel protocol of Modeller 9v8 [22] was used to do template-based modeling with default parameter settings. We performed this operation for all the models in the ensemble to generate same number of hybrid models as the number of valid models in the ensemble with the PR replaced.

The next step is to perform quality assessment of the generated hybrid models to select structures having enhanced overall qualities compared to the initial model. To this end, six complementary single-model quality evaluation methods were applied to rank each hybrid model in the ensemble: (1) RWplus [23] which is a side-chain orientation dependent atomic statistical potential, (2) Distance-scaled, Finite Ideal-gas REference (DFIRE) [24, 25] that is based on the orientation angles involved in dipole-dipole interactions, (3) Discrete Optimized Protein Energy (DOPE) [26] which is an atomic distance-dependent statistical potential derived from a pool of known protein structures by applying probability theory, (4) FRST [27] that is based on weighted combination of four complementary knowledge-based potentials: (i) the RAPDF potential [28], (ii) solvation potential, (iii) hydrogen bond potential and (iv) torsion angle potential, (5) TAP [29] which measures the local sequence to structure fitness of the protein model depending on the torsion angle propensities, and (6) ModelEvaluator [30] that is a machine learning approach based on features derived from secondary structure, relative solvent accessibility, contact map, and beta sheet structure.

It is worth mentioning here that along with the hybrid models, we also used the starting structure during ranking. This was performed with the vision that in case there is no fragment present in the ensemble which can improve the quality of the PR, then the initial model should remain unaltered.

2.3 Consensus ranking of hybrid models

After ranking each of the hybrid models using six above mentioned model quality assessment methods, the obvious subsequent phase is to select the top ranked model and designate that as the improved model for the PR. But, because of the complementary nature of the evaluation methods, the ranking is often inconsistent across different protocols. One way to overcome this obstacle is to apply cumulative or average ranking in order to identify the top ranked model. However, if average rank (or cumulative rank) is applied to derive the consensus ranking, we see many ties between different hybrid models which make it difficult to identify the best model. Thus, arriving at the optimal consensus ranking for all the hybrid models in order to select the best model is a non-trivial problem. Our goal here is to find an optimal ranking which would be as close as possible to all individual ranking schemes simultaneously. This problem, therefore, can be viewed as an optimization problem.

The objective function takes the following form in its abstract representation:

$$\varphi(\delta) = \sum_{i=1}^{m} w_i d(\delta, L_i) \quad (5)$$

where δ is a proposed ordered list of length $k = |L_i|$ with m number of individual ordering, w_i is the importance weight associated with list L_i, d is a distance function described below and L_i is the i^{th} ordered list. The aim would be to find δ^* which would minimize the total distance between δ^* and L_i's. This is denoted as:

$$\delta^* = \arg\min \sum_{i=1}^{m} w_i d(\delta, L_i) \quad (6)$$

We adopted a weighted rank aggregation method to derive the optimal solution using a R implementation [31]. Cross-Entropy Monte Carlo algorithm (CE) [32, 33] was applied using the Kendall's Tau distance for partially ordered lists [34]. The Kendall's Tau distance counts the number of pair-wise disagreements between different lists, and normalizes by the maximum possible disagreements. When the Kendall's Tau distance is 0, the two lists are exactly the same, and when it is 1, they are in reverse order. Two random lists have, on average, a distance of 0.5. The CE algorithm involves 200,000 maximum iterations or until convergence with a convergence indicator of 20 meaning that if the smallest value of the objective function does not change during 20 consecutive iterations during the optimization process, the algorithm is assumed to be converged to its optimal solution. We used the same weight (w_i =1) for all the six different ranking schemes. The optimal solution derived by CE algorithm is finally chosen as the consensus ranking of the hybrid models. We term this "optimal ranking". To the best of our knowledge, weighted rank aggregation has not been used before in consensus model quality assessment. Moreover, the application of rank aggregation in the field of protein structure refinement is new.

2.4 Improving overall fold in the initial model through iterative fragment exchange

With the optimal ranking at hand, the models ranked above the starting structure were extracted to construct what we call "superior model set". Then, a maximum of top three models were chosen from the superior model set as templates to feed into automodel class of Modeller 9v8 [22] to derive the "improved model" for the PR using multiple template alignment. It has to be noted here that selecting the number of models to be fed into Modeller depend on the structures present in the superior model set. For instance, if only two hybrid models are ranked above the starting structure after optimal ranking, then the superior model set consists of two models and we use only them as templates. In case the superior model set is empty, suggesting no hybrid model is ranked above the starting structure, the improved model is same as the initial structure.

When the starting structure consists of multiple PRs, we employed iterative refinement strategy to gradually improve the initial model with each PR getting improved in a single iteration. The PRs were sorted based on their length and longer PRs getting higher priority than shorter PRs. After each iteration, the improved model corresponding to a PR becomes the starting model for the next round of iteration aiming to improve the next PR. This process continues until all the PRs in the initial model are consumed.

2.5 Improving general physicality to produce the refined structure

When all the PRs in the initial model are iteratively refined and the final improved model is generated, the overall fold of the starting structure is supposedly improved. We now aim to improve the local errors and general physical reasonableness of the final improved model like reducing any unfavorable steric clashes or staggered χ angles. This was achieved by applying atomic level energy minimization using our previously developed refinement procedure, 3Drefine. Our previous study shows that 3Drefine has been reliable in consistent improvement of the local qualities of protein models [13]. Also, because of the fast running time of 3Drefine, it does not pose any computational overhead to the REFINEpro pipeline. The energy minimized model is the refined model.

2.6 Metrics used for evaluation

From the flowchart of REFINEpro (Figure 1), it is clear that we need two-fold evaluation method for our refinement pipeline. First, we are interested to see how accurately REFINEpro predicts the PRs in the starting models and second, how consistently our method can improve the global and local qualities of the initial structures to bring it closer to native state.

2.6.1 Assessment criteria for PR prediction

In order to identify the true problematic residues in the initial model, we superposed the initial model on the native structure using TM-score and calculated the distance between C_α atoms after optimal structural superposition. The distance is converted to the S-score using Eq. (1). Once again, we used a sliding window of 3 residues around the central residue to avoid local fluctuation and then calculated the average S-score in that window to determine the correctness of the central residue using Eq. (2).

We use receiver operating characteristic (ROC) curve to evaluate the overall prediction accuracy of problematic residues. ROC curve is a plot of the sensitivity versus (1 - specificity) for a binary classifier as its decision boundary is moved. Sensitivity measures the capability of predicting positive samples (problematic residues in our case) correctly and specificity determines if any non-problematic residues are incorrectly predicted as problematic residues.

Since problematic residues prediction becomes a binary classification problem when the residue conservation index is set at 0.5, we measure its performance by using the following widely used criteria functions:

$$Precision = \frac{TP}{TP + FP} \qquad Specificity = \frac{TN}{TN + FN}$$

$$Recall = \frac{TP}{TP + FN} \qquad Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where true positive (TP) is the number of true problematic residues that are predicted correctly, true negative (TN) is the number of true non-problematic residues that are predicted correctly, false positive (FP) is the number of true non-problematic residues that are predicted to be problematic and false negative (FN) is the number of true problematic residues that are predicted to be non-problematic.

2.6.2 Model quality evaluation measures

We assess the quality of the models from two perspectives: (1) similarity to the native structures and (2) physical reasonableness of the models. C_α Root Mean Square Deviation (RMSD) [35] is used to evaluate global positioning of C_α atoms purpose where a lower RMSD value indicates that the protein model is close to its native state. However, RMSD is very sensitive to small structural errors. Even if the coordinates of only a few atoms undergo large atomic changes, RMSD becomes high making it difficult to assess the overall correctness of the structure. Global quality measures like GDT-HA [36] or TM-score [17] overcomes this difficulty to a large extent. TM-score counts all the residues and tends to be more sensitive to the global topology, whereas GDT-HA count the residue pairs with distances in (0.5Å, 1Å, 2Å and 4Å), and tend to be more sensitive in capturing the errors in local fragments. Both GDT-HA and TM-score lie in [0, 1] with a higher value indicating better similarity to the native structures. GDT-HA score has been a widely used scoring function to measure the global positing of C_α atoms in CASP experiments [37-39]. In order to evaluate the physical reasonableness and the local errors, we use MolProbity [40] – a single and composite score to measure local model quality. MolProbity score is a log-weighted combination of the rotamer outliers, torsion-angle outliers, and steric clashes that have values outside the region of experimentally derived standard protein structures. The MolProbity score denotes the expected resolution of the protein model with respect to standard experimental structures and therefore, lower MolProbity score indicates more physically realistic model.

2.7 Data Sets used for assessment

To benchmark the performance of REFINEpro, we collected a test set containing 163 targets: (1) 107 targets from 9th edition of

Critical Assessment of Techniques for Protein Structure Prediction (CASP9) structure prediction category and (2) 56 targets from I-TASSER decoy set.

2.7.1 107 CASP9 targets

This dataset consists of 107 CASP9 TS targets taken from http://predictioncenter.org/download_area/CASP9/. We used the first models submitted to CASP9 TS category by our structure prediction method, MULTICOM-CONSTRUCT [41] as the initial model for each of these targets. The complete archive of submitted models by all servers had been used as the ensemble. The CASP9 dataset in most interesting in terms of practical applications: (1) it contains the best models submitted by the best research groups around the world and therefore represents the state-of-the-art in the field of protein structure prediction; (2) the high diversity of targets both in terms of length and complexity reduces bias in testing our protocol and (3) because of the popularity of TBM methods amongst the CASP predictors, REFINEpro can be evaluated in its ability to refine models produced by TBM techniques.

2.7.2 I-TASSER decoy set

Models in this dataset contain 56 non-homologous small proteins with lengths from 47 to 118 residues. I-TASSER ab-initio modeling [42] were used to generate the backbone structure and 12,500 - 32,000 conformations were selected from the trajectories of 3 lowest-temperature replicas of the simulations. Then, iterative structure clustering [43] followed by energy minimization was performed on the selected decoys using GROMACS 4.0 simulation package [44] with OPLS-AA force field [45] for improving local qualities while keeping the topologies unchanged. The decoy set is available at http://zhanglab.ccmb.med.umich.edu/decoys. For each target, the best structure (having lowest RMSD to native) was used as the initial model while the rest of the decoy set serves as model ensemble during REFINEpro run.

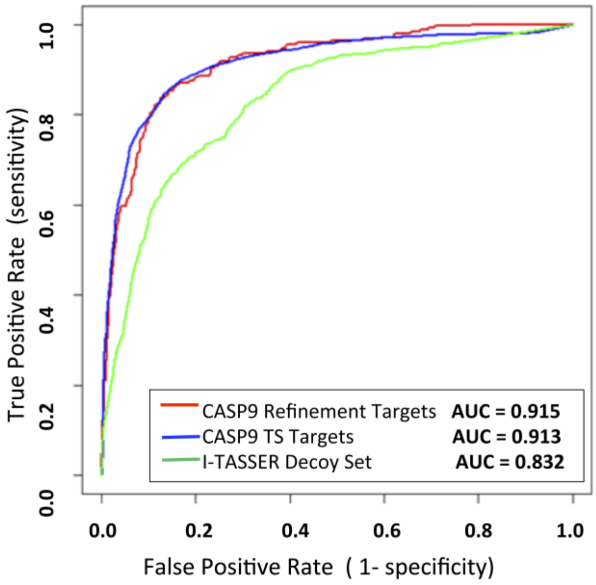

Figure 2. ROC curves for prediction of PRs.

107 CASP9 Targets (red) and I-TASSER Decoy Set (blue). The numbers beside each legend represent the values of Area Under the Curve (AUC).

3. RESULTS AND DISCUSSION

We begin by evaluating the performance of REFINEpro to detect the PRs in the initial model for all our datasets. Then, the overall improvement in global and local qualities of the initial models generated by MULTICOM-CONSTRUCT [41] for 107 CASP9 targets are presented. Finally, we judge how significant the refinement is when the initial models and the ensembles are generated I-TASSER ab-initio simulation.

3.1 Accuracy of PR prediction

In Figure 2, the ROC curves are shown for each of the two datasets we used with the corresponding area under the curves (AUC). The values of the criteria functions are presented in Table 1 at residue conservation index threshold of 0.5. It can be observed that, the dataset of 107 CASP9 targets yields better performance while prediction accuracy is less significant in I-TASSER decoy set. This is mainly because in I-TASSER decoy set, the initial models and the ensembles are generated by the same prediction pipeline, and thus, the conformations of the decoys are largely same, making it difficult to successfully apply consensus prediction. These results, therefore, demonstrate that our hypothesis of residue conservation works best when complementary structures are present in the ensemble.

Table 1. Accuracy measures for the prediction of PRs at residue conservation index threshold of 0.5

Dataset	Precision	Specificity	Recall	Accuracy
107 CASP9 Targets	85.51	94.14	71.2	86.63
I-TASSER Decoy Set	41.02	86.22	64.7	83.43

3.2 Performance on 107 CASP9 targets

Due to the presence of potentially disordered regions in the native structures, often there are mismatches in the CASP9 sequence with that of the experimental structures. After executing REFINEpro in blind mode, we identified the residues in the target sequences that did not have coordinates in the experimental structures by performing alignment between CASP9 sequence and the corresponding sequences for the native structures using ClustalW [46]. These residues were removed from both the initial and refined models during refinement assessment.

Figures 3A and 3B show the scatter plot of GDT-HA, TM-score RMSD and MolProbity score difference before and after refinement against the initial TM-scores. Out of total 107 targets, REFINEpro refinement resulted in improving the model qualities for 95, 91, 78 and 80 cases with respect to RMSD, TM-score, GDT-HA and MolProbity scores respectively. Overall, 5.1% and 4.9% improvement in cumulative TM-score and cumulative GDT-HA score respectively has been observed while the average RMSD and average MolProbity improvement is 12.7% and 5.8% respectively. These results clearly demonstrate the promising ability of REFINEpro to improve the overall fold of the starting models together with enhancement in the general physicality of the models in a large benchmark set comprised of different lengths and target complexities.

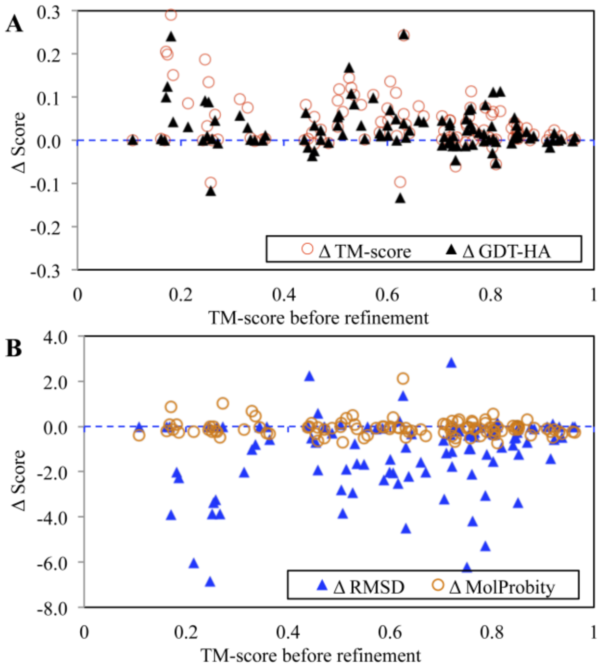

Figure 3. Changes in structural qualities using REFINEpro on 107 CASP9 targets.

(A) Scatter plot of changes in and TM-score and GDT-HA. A positive change indicates the quality of the model has been improved by refinement.

(B) Scatter plot of changes in RMSD and MolProbity score. A negative change indicates the quality of the model been improved by refinement.

A representative example of refinement is presented in Figure 4 for target T0559. The initial model has an RMSD of 6.188 Å with a large deviation in the N-terminal helix region compared to the native structure. After refinement, the RMSD is drastically improved to 1.694 Å with 51.1%, 38.5% and 3.6% improvement in GDT-HA, TM-score and MolProbity score respectively. The definite improvement in the N-terminal region is obvious even by simple visual inspection.

3.3 Performance on I-TASSER decoy set

Similar to the CASP9 dataset, we performed the refinement on the I-TASSER decoy set, where initial models are generated by I-TASSER ab-initio simulation in a strict blind mode, that is, without the knowledge of the native structure.

A consistent improvement is observed in qualities of the starting structures as measured by the GDT-HA, TM-score, and RMSD scores. There were 35, 37 and 31 cases when REFINEpro brings the starting models closer to the native ones. In Figure 5A and 5B, we present the scatter plot of GDT-HA, TM-score and RMSD score difference before and after refinement against the initial TM-score for all the 56 targets. Although encouraging, the refinement for the I-TASSER decoy set is not as pronounced as the CASP9 dataset. This is primarily because the starting models in I-TASSER decoy set are already the best models selected from the ensemble with majority of the initial models have RMSD less than 3Å compared to the native structures, resulting in less PRs being identified by REFINEpro, thereby hindering the ability of REFINEpro for drastic improvement in the backbone positioning. These results, therefore, suggest that most prominent improvement in model qualities are observed in REFINEpro when the starting structure is further away from native state.

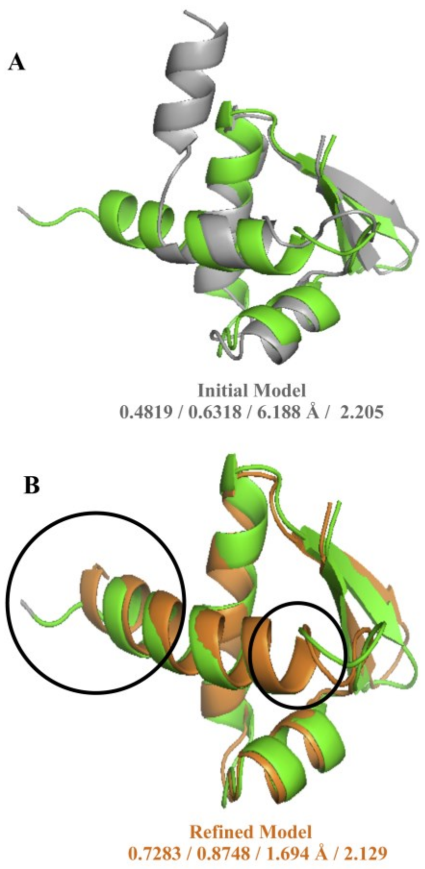
A
Initial Model
0.4819 / 0.6318 / 6.188 Å / 2.205
B
Refined Model
0.7283 / 0.8748 / 1.694 Å / 2.129

Figure 4. Example of REFINEpro refinement for CASP9 target T0559.

(A) Structural superposition of initial model (grey) on native structure (green). The values under the model indicate GDT-HA, TM-score, RMSD and MolProbity score respectively before refinement.

(B) Structural superposition of refined model using REFINEpro (orange) on native structure (green). The values under the model indicate GDT-HA, TM-score, RMSD and MolProbity score respectively after refinement. The black circles highlight the regions with prominent structural improvements.

A typical example of refinement from the I-TASSER decoy set has been shown in Figure 6A and 6B for the target 1cqkA. The starting structure is quite accurate with initial RMSD of 1.448Å. The refinement is distributed across the whole sequence with reorientation of several loops to beta-strands, thereby bringing the model closer to the native state. The RMSD of the refined model is improved to 1.299Å with a 5.72% increase in GDT-HA, 2.38% increase in TM-score and 47.2% improvement in MolProbity score.

4. CONCLUSION

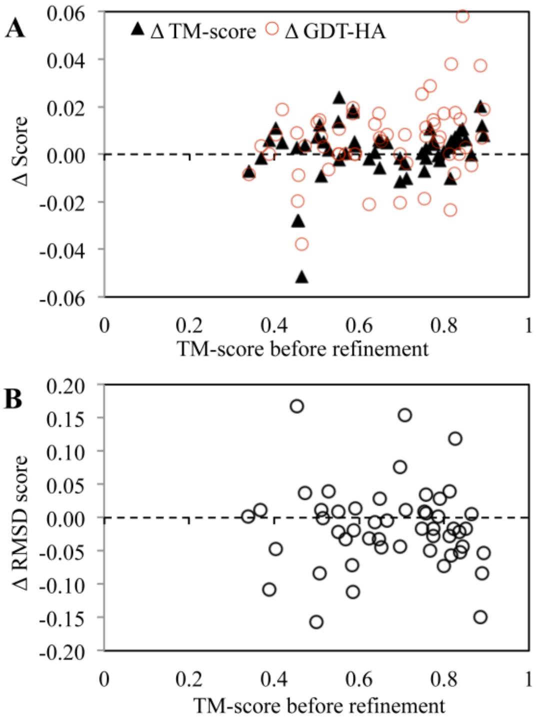

Figure 5. Changes in structural qualities using REFINEpro on I-TASSER decoy set.

(A) Scatter plot of changes in GDT-HA and TM-score. A positive change indicates the quality of the model has been improved by refinement.

(B) Scatter plot of changes in RMSD score. A negative change indicates the quality of the model been improved by refinement.

Development of a method capable of improving the overall fold in predicted protein models has been a major challenge in the protein structure refinement field. The existing state-of-the-art refinement algorithms often rely on "conservative" strategies to sample locally around the starting structures producing improvement only in the physicality of the models as opposed to improvement of the global positioning of the backbone atoms [47]. In this article, we presented a new conformation ensemble-based iterative refinement method aimed at resolving this bottleneck. Coupled with our previous study on protein structure refinement [13] the method can often drastically improve the overall fold of the initial models through refinement of loop and terminal regions or rearrangements of disoriented secondary structure segments, accompanied by correction of local errors. By performing a large-scale benchmark study on 163 targets, we demonstrated that the protocol is capable of simultaneous improvement in global and local qualities of protein models generated by both TBM and ab-initio methods. More prominent results were achieved when the model ensemble for the target structure contains diverse and complementary alternative models. To our knowledge, a fully automated ensemble based approach has not been used before in refinement problem. We hope the promising aspects of our refinement protocol provide useful insights for advancement in the field of protein structure refinement, thereby enhancing the accuracy of contemporary computational protein structure prediction methods.

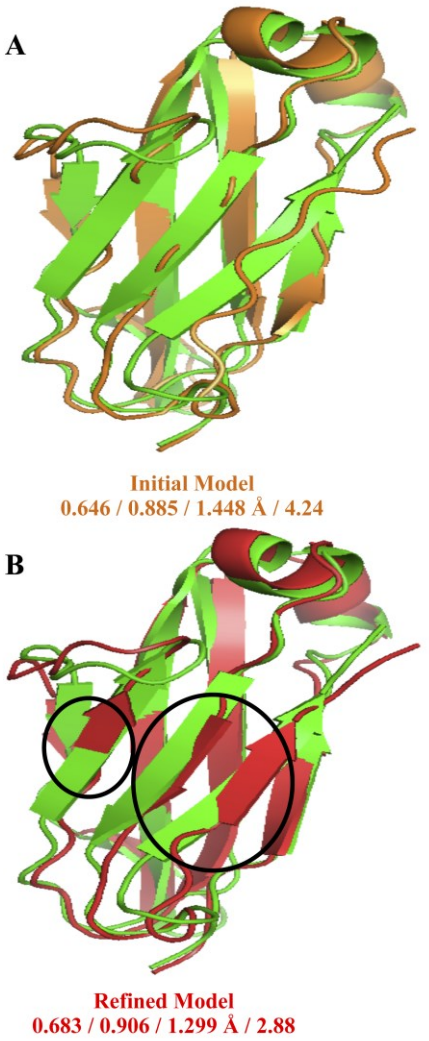
A
Initial Model
0.646 / 0.885 / 1.448 Å / 4.24
B
Refined Model
0.683 / 0.906 / 1.299 Å / 2.88

Figure 6. Example of REFINEpro refinement for I-TASSER target 1cqkA.

(A) Structural superposition of initial model (orange) on native structure (green). The values under the model indicate GDT-HA, TM-score, RMSD and MolProbity score respectively before refinement.

(B) Structural superposition of refined model using REFINEpro (red) on native structure (green). The values under the model indicate GDT-HA, TM-score, RMSD and MolProbity score respectively after refinement. The black circles highlight the regions with prominent structural improvements.

Even though encouraging success has been obtained in the present study, there is still large room for improvement. The major challenges encountered were: (1) accurate prediction of PRs in the starting structures, (2) availability of diverse and complementary models in the ensemble and (3) quality assessment method aimed at selecting the "best" hybrid model. Future directions would be to extend our consensus-based method of PR prediction to more a robust and precise approach, preferably by applying machine learning techniques. Also, significant success of REFINEpro in CASP9 dataset compared to the I-TASSER decoy set demonstrates that it is essential to have structures in the ensemble with independent and various folds. We need to investigate in future on how to automatically generate a large pool of models with various folds in a practical and efficient manner. Finally, a better model quality assessment technique is desirable which can select the "best" alternative structure from the hybrid model pool, and consequently, the accuracy of our refinement method can be improved further.

5. ACKNOWLEDGMENTS

The work was supported by a NIGMS grant (R01GM093123) to JC.

6. REFERENCES

[1] C. B. Anfinsen, "Principles that Govern the Folding of Protein Chains," 1973.

[2] J. Wang, P. Cieplak, and P. A. Kollman, "How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?," *Journal of Computational Chemistry,* vol. 21, no. 12, pp. 1049-1074, 2000.

[3] E. J. Sorin, and V. S. Pande, "Exploring the helix-coil transition via all-atom equilibrium ensemble simulations," *Biophys J,* vol. 88, no. 4, pp. 2472-2493, 2005.

[4] G. A. Kaminski, R. A. Friesner, J. Tirado-Rives *et al.*, "Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides," *The Journal of Physical Chemistry B,* vol. 105, no. 28, pp. 6474-6487, 2001.

[5] A. Jagielska, L. Wroblewska, and J. Skolnick, "Protein model refinement using an optimized physics-based all-atom force field," *Proceedings of the National Academy of Sciences,* vol. 105, no. 24, pp. 8268-8273, 2008.

[6] C. M. Summa, and M. Levitt, "Near-native structure refinement using in vacuo energy minimization," *Proceedings of the National Academy of Sciences,* vol. 104, no. 9, pp. 3177-3182, 2007.

[7] H. A. Scheraga, "Recent developments in the theory of protein folding: searching for the global energy minimum," *Biophysical chemistry,* vol. 59, no. 3, pp. 329-339, 1996.

[8] A. Kolinski, "Protein modeling and structure prediction with a reduced representation," *ACTA BIOCHIMICA POLONICA-ENGLISH EDITION-,* vol. 51, pp. 349-372, 2004.

[9] M. R. Lee, J. Tsai, D. Baker *et al.*, "Molecular dynamics in the endgame of protein structure prediction1," *Journal of molecular biology,* vol. 313, no. 2, pp. 417-430, 2001.

[10] G. Chopra, N. Kalisman, and M. Levitt, "Consistent refinement of submitted models at CASP using a knowledge-based potential," *Proteins: Structure, Function, and Bioinformatics,* vol. 78, no. 12, pp. 2668-2678, 2010.

[11] J. Zhang, Y. Liang, and Y. Zhang, "Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling," *Structure,* vol. 19, no. 12, pp. 1784-1795, 2011.

[12] D. Xu, and Y. Zhang, "Improving the Physical Realism and Structural Accuracy of Protein Models by a Two-Step Atomic-Level Energy Minimization," *Biophys J,* vol. 101, no. 10, pp. 2525, 2011.

[13] D. Bhattacharya, and J. Cheng, "3Drefine: Consistent protein structure refinement by optimizing hydrogen bonding network and atomic-level energy minimization," *Proteins: Structure, Function, and Bioinformatics*, 2012.

[14] D. Baker, and A. Sali, "Protein structure prediction and structural genomics," *Science's STKE,* vol. 294, no. 5540, pp. 93, 2001.

[15] J. Zhu, H. Fan, X. Periole *et al.*, "Refining homology models by combining replica-exchange molecular dynamics and statistical potentials," *Proteins: Structure, Function, and Bioinformatics,* vol. 72, no. 4, pp. 1171-1188, 2008.

[16] M. Levitt, and M. Gerstein, "A unified statistical framework for sequence comparison and structure comparison," *Proceedings of the National Academy of Sciences,* vol. 95, no. 11, pp. 5913, 1998.

[17] Y. Zhang, and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," *Proteins: Structure, Function, and Bioinformatics,* vol. 57, no. 4, pp. 702-710, 2004.

[18] N. Siew, A. Elofsson, L. Rychlewski *et al.*, "MaxSub: an automated measure for the assessment of protein structure prediction quality," *Bioinformatics,* vol. 16, no. 9, pp. 776-785, 2000.

[19] M. Fasnacht, J. Zhu, and B. Honig, "Local quality assessment in homology models using statistical potentials and support vector machines," *Protein Science,* vol. 16, no. 8, pp. 1557-1568, 2007.

[20] P. Larsson, M. J. Skwark, B. Wallner *et al.*, "Assessment of global and local model quality in CASP8 using Pcons and ProQ," *Proteins: Structure, Function, and Bioinformatics,* vol. 77, no. S9, pp. 167-172, 2009.

[21] B. Wallner, and A. Elofsson, "Identification of correct regions in protein models using structural, alignment, and consensus information," *Protein Science,* vol. 15, no. 4, pp. 900-913, 2006.

[22] A. Fiser, and A. Šali, "Modeller: generation and refinement of homology-based protein structure models," *Methods in enzymology,* vol. 374, pp. 461-491, 2003.

[23] J. Zhang, and Y. Zhang, "A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction," *PloS one,* vol. 5, no. 10, pp. e15386, 2010.

[24] Y. Yang, and Y. Zhou, "Specific interactions for ab initio folding of protein terminal regions with secondary structures," *Proteins: Structure, Function, and Bioinformatics,* vol. 72, no. 2, pp. 793-803, 2008.

[25] Y. Yang, and Y. Zhou, "Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions," *Protein Science,* vol. 17, no. 7, pp. 1212-1219, 2008.

[26] M. Shen, and A. Sali, "Statistical potential for assessment and prediction of protein structures," *Protein Science,* vol. 15, no. 11, pp. 2507-2524, 2006.

[27] S. C. E. Tosatto, "The victor/FRST function for model quality estimation," *Journal of Computational Biology,* vol. 12, no. 10, pp. 1316-1327, 2005.

[28] R. Samudrala, and J. Moult, "An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction1," *Journal of molecular biology,* vol. 275, no. 5, pp. 895-916, 1998.

[29] S. C. E. Tosatto, and R. Battistutta, "TAP score: torsion angle propensity normalization applied to local protein structure evaluation," *BMC bioinformatics,* vol. 8, no. 1, pp. 155, 2007.

[30] Z. Wang, A. N. Tegge, and J. Cheng, "Evaluating the absolute quality of a single protein model using structural features and support vector machines," *Proteins: Structure, Function, and Bioinformatics,* vol. 75, no. 3, pp. 638-647, 2009.

[31] V. Pihur, and S. Datta, "RankAggreg, an R package for weighted rank aggregation," *BMC bioinformatics,* vol. 10, no. 1, pp. 62, 2009.

[32] R. Rubinstein, "The cross-entropy method for combinatorial and continuous optimization," *Methodology and computing in applied probability,* vol. 1, no. 2, pp. 127-190, 1999.

[33] P. T. De Boer, D. P. Kroese, S. Mannor *et al.*, "A tutorial on the cross-entropy method," *Annals of operations research,* vol. 134, no. 1, pp. 19-67, 2005.

[34] L. M. K. Adler, "A modification of Kendall's tau for the case of arbitrary ties in both rankings," *Journal of the American Statistical Association,* vol. 52, no. 277, pp. 33-35, 1957.

[35] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography,* vol. 32, no. 5, pp. 922-923, 1976.

[36] A. Zemla, "LGA: a method for finding 3D similarities in protein structures," *Nucleic acids research,* vol. 31, no. 13, pp. 3370-3374, 2003.

[37] J. Kopp, L. Bordoli, J. N. D. Battey *et al.*, "Assessment of CASP7 predictions for template-based modeling targets," *Proteins: Structure, Function, and Bioinformatics,* vol. 69, no. S8, pp. 38-56, 2007.

[38] D. Cozzetto, A. Kryshtafovych, K. Fidelis *et al.*, "Evaluation of template-based models in CASP8 with standard measures," *Proteins: Structure, Function, and Bioinformatics,* vol. 77, no. S9, pp. 18-28, 2009.

[39] D. A. Keedy, C. J. Williams, J. J. Headd *et al.*, "The other 90% of the protein: Assessment beyond the Cαs for CASP8 template-based and high-accuracy models," *Proteins: Structure, Function, and Bioinformatics,* vol. 77, no. S9, pp. 29-49, 2009.

[40] V. B. Chen, W. B. Arendall, J. J. Headd *et al.*, "MolProbity: all-atom structure validation for macromolecular crystallography," *Acta Crystallographica Section D: Biological Crystallography,* vol. 66, no. 1, pp. 12-21, 2009.

[41] Z. Wang, J. Eickholt, and J. Cheng, "MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8," *Bioinformatics,* vol. 26, no. 7, pp. 882-888, 2010.

[42] S. Wu, J. Skolnick, and Y. Zhang, "Ab initio modeling of small proteins by iterative TASSER simulations," *BMC biology,* vol. 5, no. 1, pp. 17, 2007.

[43] Y. Zhang, and J. Skolnick, "SPICKER: A clustering approach to identify near-native protein folds," *Journal of Computational Chemistry,* vol. 25, no. 6, pp. 865-871, 2004.

[44] B. Hess, C. Kutzner, D. Van Der Spoel *et al.*, "GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation," *Journal of chemical theory and computation,* vol. 4, no. 3, pp. 435-447, 2008.

[45] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, "Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids," *Journal of the American Chemical Society,* vol. 118, no. 45, pp. 11225-11236, 1996.

[46] M. Larkin, G. Blackshields, N. Brown *et al.*, "Clustal W and Clustal X version 2.0," *Bioinformatics,* vol. 23, no. 21, pp. 2947-2948, 2007.

[47] J. L. MacCallum, A. Pérez, M. J. Schnieders *et al.*, "Assessment of protein structure refinement in CASP9," *Proteins: Structure, Function, and Bioinformatics*, 2011.