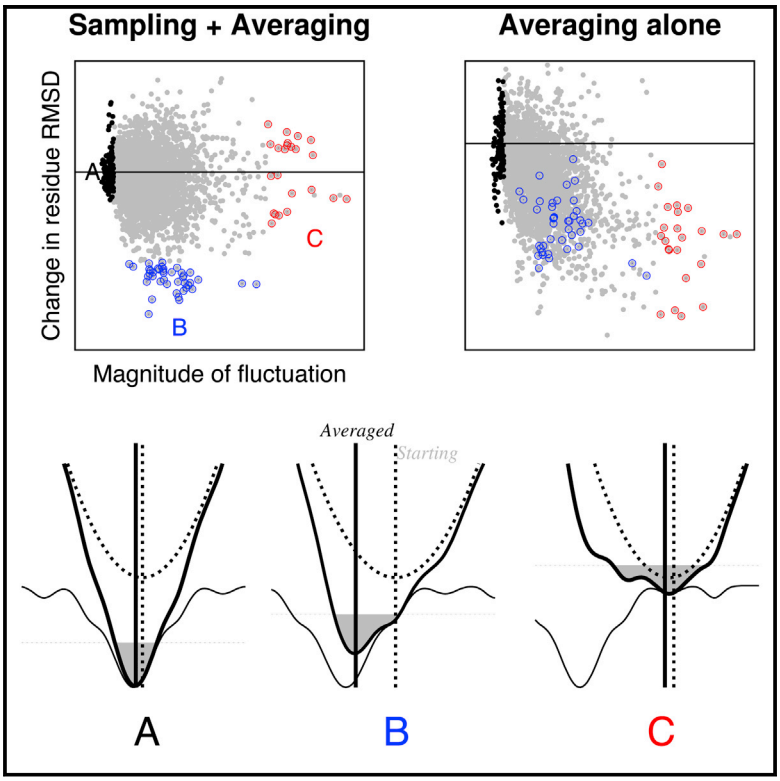
Short Article

Structure

The Origin of Consistent Protein Structure Refinement from Structural Averaging

Graphical Abstract



Authors

Hahnbeom Park, Frank DiMaio, David Baker

Correspondence

dabaker@u.washington.edu

In Brief

Park et al. investigate the origin of protein structure refinement from structural averaging at residue level. Improvement upon averaging was found to be correlated with residue fluctuations and is the superposition of two limiting effects: amplifying improvements and dampening divergences.

Highlights

- Consistent refinement upon averaging is found in implicit water MD/MCM simulations
- The origin of refinement upon averaging is addressed at residue level
- Improvements upon averaging are related to the extent of residue fluctuations
- Improvements result from amplifying improvements and dampening divergences

Park et al., 2015, Structure *23*, 1123–1128
June 2, 2015
http://dx.doi.org/10.1016/j.str.2015.03.022

Structure
Short Article

The Origin of Consistent Protein Structure Refinement from Structural Averaging

Hahnbeom Park,[1,2] Frank DiMaio,[1,2] and David Baker[1,2,3,*]
[1]Department of Biochemistry, University of Washington, Seattle, WA 98195, USA
[2]Institute for Protein Design, University of Washington, Seattle, WA 98195, USA
[3]Howard Hughes Medical Institute, University of Washington, Box 357370, Seattle, WA 98195, USA
*Correspondence: dabaker@u.washington.edu
http://dx.doi.org/10.1016/j.str.2015.03.022

SUMMARY

Recent studies have shown that explicit solvent molecular dynamics (MD) simulation followed by structural averaging can consistently improve protein structure models. We find that improvement upon averaging is not limited to explicit water MD simulation, as consistent improvements are also observed for more efficient implicit solvent MD or Monte Carlo minimization simulations. To determine the origin of these improvements, we examine the changes in model accuracy brought about by averaging at the individual residue level. We find that the improvement in model quality from averaging results from the superposition of two effects: a dampening of deviations from the correct structure in the least well modeled regions, and a reinforcement of consistent movements towards the correct structure in better modeled regions. These observations are consistent with an energy landscape model in which the magnitude of the energy gradient toward the native structure decreases with increasing distance from the native state.

INTRODUCTION

In the current protein structure rich era, an important challenge in the protein structure prediction field is the structure refinement problem (Nugent et al., 2014). The ultimate aim of protein structure refinement is to improve homology models to the level of experimentally determined structures. Feig and coworkers recently made a breakthrough in this area (Mirjalili and Feig, 2012; Mirjalili et al., 2014), obtaining consistent blind improvements to homology models in the recent Critical Assessments of techniques for protein Structure Prediction (CASP10) experiment (Nugent et al., 2014; Moult et al., 2014).

Although the results of Mirjalili et al. are very encouraging, the origins of these improvements are not completely clear. Their approach employed explicit water molecular dynamics (MD) simulations with a molecular mechanics force field (Best et al., 2012) and restraints to the starting coordinates, followed by filtering the sampled ensemble using a knowledge-based potential (Yang and Zhou, 2008) and, finally, generating a single representative model using structural averaging. A first question is practical: can this expensive calculation be made more efficient so as to be more broadly applicable? A second question is more fundamental: what aspect of the Mirjalili et al.'s protocol contributes to the consistency of refinement, which has been a long-standing challenge in the field (Moult et al., 2014)?

Here we investigate these questions by adapting Mirjalili et al.'s approach to less computationally intensive sampling methods. We show that structural averaging has a clearly beneficial effect independent of simulation type (MD versus Monte Carlo minimization [MCM]) and force field (explicit versus implicit water model). We dissect the improvements in model quality at the individual residue level, and find that in an ensemble of trajectories, the improvements in the close to correct regions are generally similar to one another and hence are reinforced by averaging, while the divergences in the incorrect regions are generally different from one another and hence dampened by averaging.

RESULTS

Robust Improvements in Homology Models Using Short Implicit Solvent Simulations

We show that consistent refinement can be achieved using short simulations with an implicit solvation model. Implicit solvent simulation using Rosetta (Leaver-Fay et al., 2014) *CartesianRefiner* (see the Experimental Procedures section) followed by filtering and averaging consistently improves homology models (Figure 1A). Changes in radius of gyration are subtle, indicating that the improvements are not an artifact of uniformly compressing or expanding structures, and the stereochemistry also improves (Figure 1 and Table S1). The improvements in RMSD are smaller than those with high accuracy global distance test (GDT-HA) (Kopp et al., 2007) and Cα local distance difference test (LDDT) (Mariani et al., 2013), as has been found for explicit water MD simulations (Mirjalili and Feig, 2012), suggesting that refinement approaches based on trajectory averaging are rather conservative in refining incorrect parts of the structures (GDT-HA and Cα LDDT are more tolerant of large local errors).

Structural averaging can generate improved models even when less than half of the structures sampled in a trajectory are closer to the native structure than the starting model (Figure 1B). The change in GDT-HA (ΔGDT-HA) (Kopp et al., 2007) from the starting models to the ensemble-averaged models is correlated with the fraction of improved structures in each ensemble

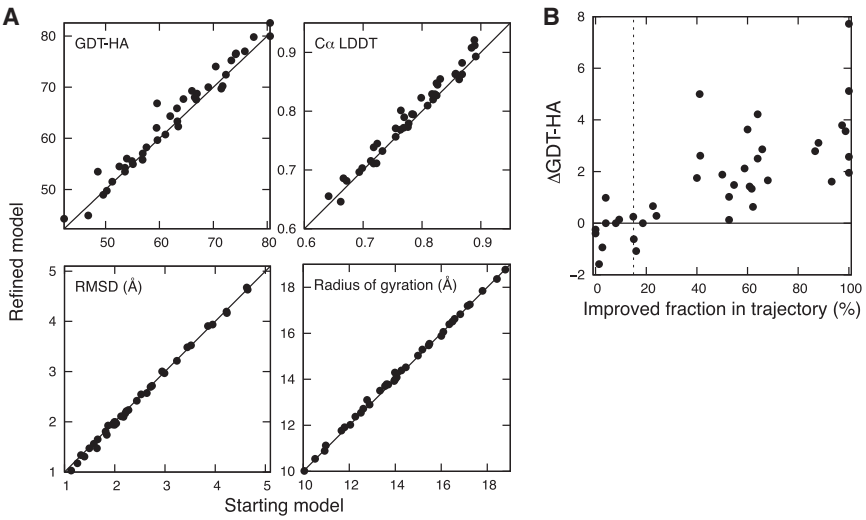


Figure 1. Consistent Refinement Using Implicit Solvent Simulations

Rosetta implicit solvent simulations were carried out starting from CASP8-10 refinement targets (see the Experimental Procedures section for details).

(A) Comparison of starting (X axis) and refined models (Y axis) by four measures: GDT-HA (Kopp et al., 2007), Cα LDDT (Mariani et al., 2013), RMSD (Å), and radius of gyration (Å). GDT-HA measures the fraction of residues to within 0.5, 1.0, 2.0, and 4.0 Å of the native position after structural superimposition. Cα LDDT measures the similarity of the Cα-Cα pairwise distance map within 5-Å cutoff without structural superimposition. Overall, the fraction of targets improved by refinement is 77.5%, 72.5%, 77.5% with average improvements of 1.56%, 1.13%, and 0.13Å in GDT-HA, Cα LDDT, and RMSD, respectively.

(B) Correlation between the fraction of structures improved in the sampled trajectory (by GDT-HA, X axis) and GDT-HA change in the final model brought by refinement (Y axis). Each dot represents a single target among 40 test cases. Averaging results in improvements over the starting model when more than 15% of the structures (dotted line) sampled in the trajectory are closer to the native structure than the starting model.

(compared with the starting models). When more than 15% of the sampled structures are improved over the starting structure, the average structure is generally better than the starting structure. Averaging outperforms selecting a structure: (1) with the lowest Rosetta energy, (2) with the lowest statistical potential (Zhou and Skolnick, 2011), and (3) nearest to the cluster center. The first two approaches, which purely rely on energy functions (the first one used in sampling and the second orthogonal to that used in sampling), do not provide consistent improvements (with an average ΔGDT-HA of 0.0). Clustering only produces marginal improvements (with an average ΔGDT-HA of 0.5), while averaging yields an average ΔGDT-HA of 1.56.

The consistency of improvement upon averaging is similar using MD and MCM simulations (Table 1). Using ensembles from MD trajectories, 80.0% of targets improve or stay on par with an average GDT-HA increase of 1.2. Similarly, using MCM-based methods, structure quality improved or stayed on par for 75% of targets with an average GDT-HA increase of 0.6. Combining models generated using different methods yields improved results (Table 1). These results suggest that the robustness of the improvement is not dependent on the sampling method as long as the trajectory samples reasonable structural diversity. Restraints to the starting structure are important: unrestrained simulations yield only marginal improvements in structure accuracy even after averaging (Figure S1; mean GDT-HA increase is 0.38).

Analysis of Refinement at the Individual Residue Level

To determine the origins of the improvement in model accuracy, we investigate the effects of refinement on the accuracy of placement of individual residues. We first examined the fluctuations in Cα positions of individual residues over samples of ~100 structures generated from MCM and MD trajectories for each of 40 different targets. We found that the magnitude of these fluctuations are correlated with the deviation of the starting position of the residues from their positions in the native structure: residues close to the native structure (<1.0 Å) to start out fluctuate relatively little (0.32 Å with SD 0.12 Å), while those that are far from the native structure (>4.0 Å) undergo considerable fluctuations (0.45 Å with SD 0.26 Å) (Figure 2A).

The Role of Structural Averaging

The impact of structural averaging at the residue level can be measured by the difference between the average accuracy in the individual members of the ensemble (Figure 2B) and the accuracy in the averaged structure (Figure 2C). The first of these quantities reports on sampling in the individual trajectories. As shown in Figure 2B (blue), residues undergoing intermediate levels of fluctuations (0.3–0.6 Å) often improve considerably in the trajectories. These improvements are offset by a general deterioration in accuracy of residues undergoing larger levels of fluctuations (Figure 2B, red); as described above, these are the residues that in the starting structure are furthest from the native structure. Averaging further increases the structural improvements in the intermediate fluctuation range (Figure 2C, blue) and considerably reduces the deterioration in accuracy of

Table 1. Effect of Structural Averaging Using Different Sampling Methods

Sampling Methods	ΔGDT-HA[a]	Fraction Equal or Improved (%)
MCM only[b]	0.56	75.0
MD only[c]	1.18	80.0
MCM+MD[d]	1.56	82.5
MCM+MD, without restraint	0.38	47.5

[a]Average GDT-HA change from the starting models. Final models are generated by averaging the selected trajectories in the first column.
[b]MCM by Rosetta *FastRelax* (Tyka et al., 2011) protocol.
[c]Implicit solvent MD simulation in Rosetta.
[d]Trajectories are selected among multiple MD and MCM simulations based on their median statistical potential score (Zhou and Skolnick, 2011) (see the Experimental Procedures section for details).

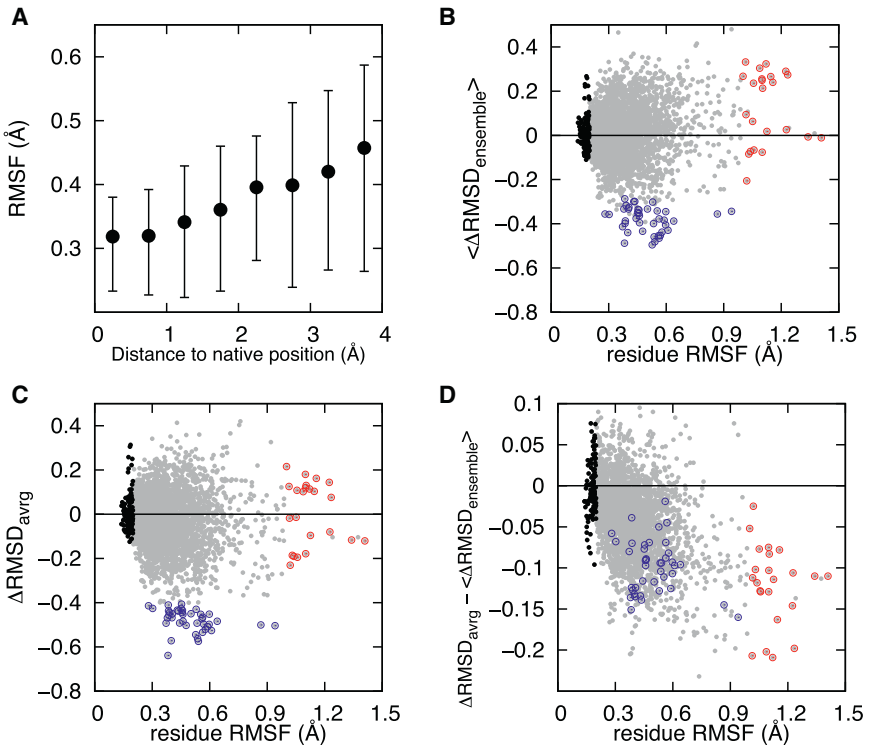


Figure 2. Analysis of the Effect of Averaging at the Individual Residue Level

(A) The magnitude of the fluctuations in the coordinates of a residue during the trajectories increases with increasing distance of the starting coordinates from the native structure. The first and third quartiles are shown as error bars. From (B) to (D), residue model quality changes are plotted (Y axis) as a function of the fluctuations of the residue during the simulation (X axis, RMSF in Å) for the 40 targets in the study.

(B) The mean value of the per residue change in RMSD (ΔRMSD) for each member of the ensemble. ΔRMSD values are computed for each residue in each member of the ensemble and the resulting ΔRMSD values are then averaged.

(C) The per residue ΔRMSD of the ensemble-averaged structure. Members of the ensemble are first structurally averaged and then the per residue ΔRMSD is computed for this averaged structure.

(D) The difference between the pre (B) and post (C) structural averaging per residue ΔRMSDs. The Y axis indicates the contribution of structural averaging: negative values represent improvements upon averaging. Residues with low fluctuation (<0.2 Å), high fluctuation (>1.0 Å), or with large improvements (<−0.4 Å) in the averaged structures are indicated with black, red, and blue circles, respectively. Residue-level model quality is measured by residue RMSD on a nine-residue window with the target residue at the center. For comparison, the same analyses on unrestrained simulations are provided in Figure S1.

the residues undergoing large fluctuations (Figure 2C, red). The effects of averaging are isolated in Figure 2D by subtracting the individual trajectory results in Figure 2B from the post averaging results in Figure 2C. Compared with the ensemble structures, averaging improves residue accuracy across the fluctuation range, with the magnitude of the improvement increasing with the fluctuation magnitude.

The net improvements to the starting structures during refinement are consistent with these observations. In Table 2, residues are binned based on the magnitude of their fluctuations, and the average change in RMSD is computed for residues in each bin. The biggest contribution to improvement comes from the medium-size fluctuation bin (root-mean square fluctuation [RMSF] range from 0.42 to 0.53 Å). The net ΔRMSD during refinement in this bin changes from −13.3 (average value of ensemble structures) to −58.0 Å (averaged structure). In regions undergoing large fluctuations (RMSF range over 0.64 Å, red circles in Figure 2), significant errors in the ensemble of structures (+24.0 Å) are reduced considerably by averaging (−5.0 Å). Because the magnitude of residue-level fluctuations is correlated with distance from the native structure, similar trends are observed when residues are binned based on their RMSD to the native structure in the starting model. Except for residues essentially already in the correct positions, which have no further room for improvement, residues closer to their native positions tend to show lower fluctuations (Figure 2A), bigger decreases in net RMSD (Figure 3A) and higher frequencies of improvement (Figure 3B). Residues far from their native positions tend to move further away in the individual trajectories and these divergences are to some extent canceled out by averaging.

The effects of averaging are similar but even more pronounced for the trajectories carried out in the absence of restraints. The overall shape of the dependence of the effect of averaging on fluctuation magnitude (Figure S1) is similar although the absolute magnitude of the fluctuations is larger. The mean RMSD per residue in the absence of restraints decreases from 1.1 Å to −0.02 Å upon averaging, and in the presence of restraints, from 0.0 Å to −0.04 Å (negative means improvement). The distributions of the RMSD changes in the presence and absence of restraints are shown in the lower panels of Figure 3. Much of the improvement by averaging in unrestrained simulation results from dampening of the largest fluctuations: ~30% of residues have mean deviations of greater than 0.5 Å before averaging and only ~5% after. Hence, the contributions of averaging and the trajectory restraints to the overall success of refinement are independent: averaging leads to considerable improvements in either case and the restraints considerably improve the ensemble structures being averaged.

The Role of Explicit Water and Loop Modeling in the Refinement Problem

Despite its successful reproduction of robust refinement, our implicit solvent approach shows a smaller extent of improvement on average compared with explicit water simulations. For the same 40 targets, the GDT-HA improvement is 1.56 in this study compared with 2.8 by long explicit solvent simulations (Mirjalili and Feig, 2012). While some of this reduction is due to the reduced sampling of our approach (approximately

Table 2. Dependence of Net Change in RMSD during Refinement on Magnitude of Fluctuations

RMSF Range (Å)	% Res[a]	ΔRMSD (Å)[b], Ensemble Average		ΔRMSD (Å)[b], Averaged Structure		ΔPer Residue[d] (Å)
		Net[c]	Per Residue	Net[c]	Per Residue	
~0.24	10%	+4.9	+0.010	+0.2	0.000	−0.010
0.24–0.28	10%	−4.8	−0.010	−17.7	−0.037	−0.027
0.28–0.34	20%	−0.6	−0.001	−29.1	−0.030	−0.029
0.34–0.42	20%	−5.7	−0.006	−45.5	−0.047	−0.041
0.42–0.53	20%	−13.3	−0.014	−58.8	−0.061	−0.047
0.53–0.64	10%	+2.3	+0.005	−20.3	−0.042	−0.047
0.64–1.41	10%	+24.0	+0.050	−5.0	−0.010	−0.060
Sum	100%	+6.9	+0.001	−176.1	−0.036	−0.037

[a]Percentage of residues in the indicated RMSF range.
[b]RMSD change from starting structures. Negative values are improvements.
[c]Sum over all residues in the RMSF bin.
[d]Per residue ΔRMSD change from ensemble average to the averaged structure; the improvement resulting purely from structure averaging. Negative values are improvements.

100-fold less), there is also a clear limitation in implicit solvent simulations, as pointed out in other studies (Fennel et al., 2010). When the 40 target proteins are categorized based on the importance of explicit waters, there are eight cases with more than five buried water molecules hydrogen bonding to the protein, 22 cases where there are few explicit water-protein hydrogen bonds, and ten cases where the water-protein interactions are uncertain (due to nuclear magnetic resonance [NMR] or low-resolution crystal structures). The ΔGDT-HA for the 22 targets that do not have buried water-protein interactions are comparable between the two methods, 2.1 to 2.4; for the remaining 18 targets, the differences are dramatic, 0.8 to 3.1, for implicit and explicit solvent simulations, respectively.

A common limitation of trajectory averaging, with methods based on both implicit and explicit solvent, is in refinement of the most incorrect regions, where dynamics or minimization alone is insufficient to move the backbone into the native energy attractor. Complementary to this approach are loop and terminus backbone modeling methods (Park and Seok, 2012; Stein and Kortemme, 2013). To highlight this complementarity, we show an example combining both methods on a homology model of CASP target TR723, with starting GDT-HA = 66.0 and RMSD = 2.2 Å. Applying the approach from the previous section, there is a GDT-HA improvement of 3.3 but no change in RMSD, with improvements entirely in the core region. If we apply *RosettaCM* (Song et al., 2013) to reconstruct the N terminus on top of the model with refined core, we further improve GDT-HA and RMSD by 4.6 and 0.4 Å, respectively. Achieving consistent improvements in model quality through loop modeling will likely require additional method development.

DISCUSSION

Our analysis of trajectory averaging at the individual residue level suggests that the increase in success of refinement upon averaging results from the superposition of two limiting effects. The trajectories may be viewed as diffusive processes in very high dimensional spaces; in one limit, the free energy landscape is flat, and in the other, harmonic. In the first limit, which dominates for residues that start out far from the native structure and free energy minimum, averaging dampens the random (and hence nonreinforcing) changes to the starting structure. In the second limit, which holds for residues closer to the native structure and free energy minimum, averaging better locates the position of the harmonic minimum than any individual structure since it is unlikely for the many structural degrees of freedom to all move in the right direction in a single trajectory.

An alternative explanation of the improvement due to ensemble averaging is that it better describes the ensemble of structures present in a crystal during X-ray data collection. While it is possible some of this improvement stems from this effect, it is unlikely this is responsible for the majority of the improvement. First, averaging yields improved results independent of the quality of the starting model, even when the starting model is quite non-native and unlike the structures sampled in the crystal environment. Second, successful refinement of targets whose native structures are determined by NMR suggests that the result is not crystal specific (and NMR measurements are a different type of ensemble average).

The Anna Karenina principle is a generalization of the novel's opening sentence: ''Happy families are all alike; every unhappy family is unhappy in its own way.'' In the context of the protein refinement problem, ''happy'' residues near the native structure experience forces in the direction of the native minimum and undergo consistent motions, while ''unhappy'' residues far from the native minimum experience diverse forces and undergo diverse motions. This picture helps rationalize why iteratively reapplying refinement approaches based on trajectory averaging does not result in continued improvements. The large improvements in structure quality come from the residues at intermediate distances from the native structure; once these become close to native, further improvements in the structure require improvements in the more divergent regions where the large fluctuations are mostly canceled out by averaging. Improvements in model quality beyond the first iteration will likely require improvements in energy functions so that a larger fraction of residues feel a strong force toward the native conformation.

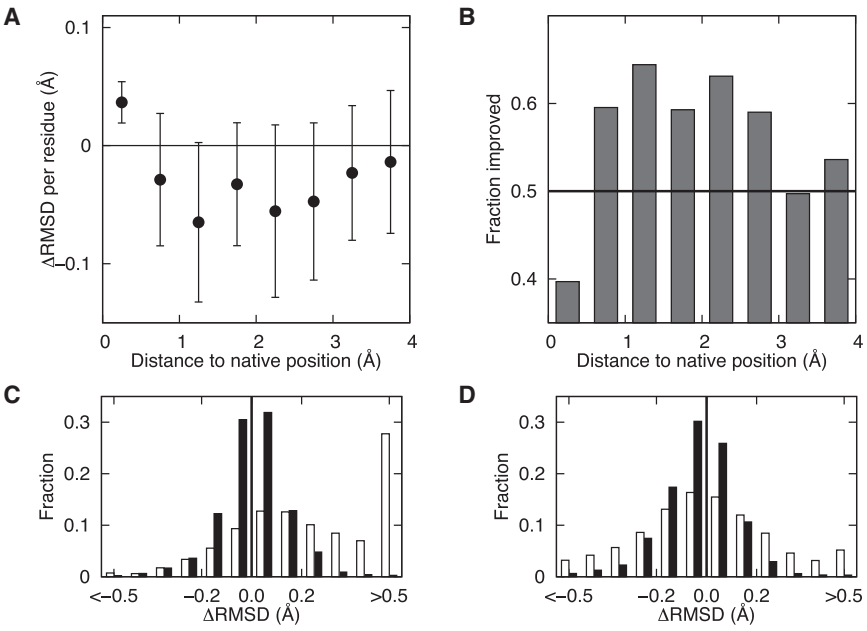


Figure 3. Dependence of Per Residue Changes During Refinement on Starting RMSD to the Native Structure

(A) Dependence of per residue ΔRMSD (see Figure 2 legend) on starting RMSD to the native structure. The first and third quartiles are shown as error bars.
(B) The fraction of residues improved as a function of the starting RMSD to the native structure.
(C and D) The distribution of per residue improvements (C) before and (D) after structural averaging are shown as histograms. Black and white columns correspond to restrained and unrestrained simulations, respectively.

The physical basis for the Anna Karenina effect in structure refinement is that native interactions must generally be stronger and consistent (less frustrated) (Taketomi et al., 1975; Bryngelson et al., 1995) than non-native interactions for the folded state to be a sufficiently deep energy minimum to overcome the large entropic cost to folding. Fragments of structure close to this energy minimum experience consistent forces, while fragments of structure far from the minimum experience less consistent and more rapidly varying forces; hence in different trajectories there is much more variation in the motions undergone in the latter than in the former.

EXPERIMENTAL PROCEDURES

Dataset

For validation of the method, targets consisted of all refinement category targets from CASP8 to CASP10. Targets with a starting model with GDT-HA below 40.0 were removed, as conservative refinement is likely limited in those cases. In total, 40 targets were used. Parameter optimization was done on a separate set composed of homology models from server predictions on other CASP targets.

Rosetta *CartesianRefiner*: Improved Sampling Efficiency by a Multimethod Approach

Here we describe the *CartesianRefiner* protocol developed for protein homology model refinement implemented within Rosetta (Leaver-Fay et al., 2014). The protocol begins with a homology model and returns a single refined model; no additional information, e.g. template structure, known contacts, and so on, is assumed. First, a given starting structure is distributed into multiple trajectories on each of seven different methods utilizing either MD or MCM simulations. Individual methods vary in their energy functions, initial structure preparation, and simulation parameters. In all methods, Cα atoms of all residues are restrained through harmonic force at their starting positions with the restraint strength of 1.0 REU (Rosetta energy unit)/mol. The four MD methods are combinations of two energy functions and two variants on the side-chain initialization protocol. The energy functions are the standard Rosetta energy as well as FACTS energy: standard Rosetta energy employs an effective solvation term (Lazaridis and Karplus, 1999) while FACTS energy describes the solvation effect by a Generalized Born/Surface Area (GB/SA) approach using the FACTS model (Haberthur and Caflisch, 2008). Side chains were initially optimized using the Rosetta packer (Leaver-Fay et al., 2014) with either standard energy weights, or softened energy weights inspired from other refinement methods (Heo et al., 2013) where van der Waals interactions are dampened to reduce the sensitivity to inaccurate initial backbone placement. For each MD method, 12 replicas of 20-ps simulations are performed from which structures are collected every 1 ps. The temperature is set uniformly at 150 K, which roughly corresponds to room temperature with Rosetta energy (Liu et al., 2012). The three MCM methods employed are: Rosetta *FastRelax* (Tyka et al., 2011) with standard Rosetta energy in Cartesian space and torsion space (Conway et al., 2014), and *FastRelax* in Cartesian space with FACTS energy as described above. The *FastRelax* protocol consists of several rounds of Monte Carlo side-chain modeling and energy minimization while slowly ramping the weight of the repulsive part of the van der Waals potential up and down to anneal the structure. Once sampling is done, trajectories from three methods among seven are selected based on their median statistical potential score (Zhou and Skolnick, 2011), followed by sampling enrichment to double the ensemble structures. Finally, structural averaging is carried out on the ensemble, combining the structures with lowest 50% statistical potential from each method.

There are two reasons for using different methods simultaneously. First, it allows for diverse sampling within a short simulation time. As pointed out above, diverse sampling is crucial for deriving sufficient statistics on residue fluctuations. Second, it increases the probability of any sampled structure to overcome energetic barriers. Energetic barriers may differ depending on the starting model, thus, they may be more easily overcome by employing different sampling techniques and energy functions. This is especially important, as short simulations may not guarantee enough sampling.

The overall simulation time for a 200-residue target is about 30 CPU hours on a 2.0 GHz Intel Xeon CPU. The simulation time scales linearly to the number of protein residues. Our approach is in the order of hundreds of times faster compared with explicit water simulations carried out by Mirjalili et al. using the NAMD package (Phillips et al., 2005). Rosetta *CartesianRefiner* is freely available to academic users as part of the Rosetta software suite. Details of usage are provided in the Supplemental Information.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Method Availability, Running Scripts, one figure, and one table and can be found with this article online at http://dx.doi.org/10.1016/j.str.2015.03.022.

AUTHOR CONTRIBUTIONS

H.P., F.D., and D.B. designed the research; H.P. performed the research; H.P. and F.D. developed the methods; H.P. and D.B. analyzed the data; and H.P., F.D, and D.B. wrote the paper.

ACKNOWLEDGMENTS

This work was supported by NIH under award numbers R01GM092802 (H.P. and D.B.). An award of computer time was provided by the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program. This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.

Received: December 19, 2014
Revised: March 3, 2015
Accepted: March 26, 2015
Published: May 7, 2015

REFERENCES

Best, R.B., Zhu, X., Shim, J., Lopes, P.E., Mittal, J., Feig, M., and Mackerell, A.D., Jr. (2012). Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi(1)$ and $\chi(2)$ dihedral angles. J. Chem. Theory Comput. *8*, 3257–3273.

Bryngelson, J.D., Onuchic, J.N., Socci, N.D., and Wolynes, P.G. (1995). Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins *21*, 167–195.

Conway, P., Tyka, M.D., DiMaio, F., Konerding, D.E., and Baker, D. (2014). Relaxation of backbone bond geometry improves protein energy landscape modeling. Protein Sci. *23*, 47–55.

Fennel, C.J., Kehoe, C.W., and Dill, K.A. (2010). Modeling aqueous solvation with semi-explicit assembly. Proc. Natl. Acad. Sci. USA *108*, 3234–3239.

Haberthur, U., and Caflisch, A. (2008). FACTS: fast analytical continuum treatment of solvation. J. Comput. Chem. *29*, 701–715.

Heo, L., Park, H., and Seok, C. (2013). GalaxyRefine: protein structure refinement driven by side-chain repacking. Nucleic Acids Res. *41*, W384–W388.

Kopp, J., Bordoli, L., Battey, J.N., Kiefer, F., and Schwede, T. (2007). Assessment of CASP7 predictions for template-based modeling targets. Proteins *69*, 36–56.

Lazaridis, T., and Karplus, M. (1999). Effective energy function for proteins in solution. Proteins *35*, 133–152.

Leaver-Fay, A., Tyka, M.D., Lewis, S.M., Lange, O.F., Thopson, J., Jacak, R., Kaufman, K., Renfrew, P.D., Smith, C.A., Sheffler, W., et al. (2014). Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. *487*, 545–574.

Liu, Y., Kellog, E., and Liang, H. (2012). Canonical and micro-canonical analysis of folding of trpzip2: an all-atom replica exchange Monte Carlo simulation study. J. Chem. Phys. *137*, 045103.

Mariani, V., Biasini, M., Barbato, A., and Schwede, T. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics *29*, 2722–2728.

Mirjalili, V., and Feig, M. (2012). Protein structure refinement through structure selection and averaging from molecular dynamics ensembles. J. Comput. Chem. Theor. *9*, 1294–1303.

Mirjalili, V., Noyes, K., and Feig, M. (2014). Physics-based protein structure refinement through multiple molecular dynamics trajectories and structure averaging. Proteins *82*, 196–207.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. (2014). Critical assessment of methods of protein structure prediction (CASP) - round X. Proteins *82*, 1–6.

Nugent, T., Cozzetto, D., and Jones, D.T. (2014). Evaluation of predictions in the CASP10 model refinement category. Proteins *82*, 98–111.

Park, H., and Seok, C. (2012). Refinement of unreliable local regions in template-based protein models. Proteins *80*, 1974–1986.

Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhoshid, E., Villa, E., Chipot, C., Skeel, R.D., Kalé, L., and Schulten, K. (2005). Scalable molecular dynamics with NAMD. J. Comput. Chem. *26*, 1781–1802.

Song, Y., DiMaio, F., Wang, R.Y.-R., Kim, D.E., Miles, C., Brunette, T., Thompson, J., and Baker, D. (2013). High resolution comparative modeling with RosettaCM. Structure *21*, 1735–1742.

Stein, A., and Kortemme, T. (2013). Improvements to robotics-inspired conformational sampling in rosetta. PLoS One *8*, e63090.

Taketomi, H., Ueda, Y., and Go, N. (1975). Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions. Int. J. Pept. Protein Res. *7*, 445–459.

Tyka, M.D., Keedy, D.A., Andre, I., DiMaio, F., Song, Y., Richardson, D.C., Richardson, J.S., and Baker, D. (2011). Alternate states of proteins revealed by detailed energy landscape mapping. J. Mol. Biol. *405*, 607–618.

Yang, Y., and Zhou, Y. (2008). Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. Protein Sci. *17*, 1212–1219.

Zhou, H., and Skolnick, J. (2011). GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophys. J. *101*, 2043–2052.

