Refining protein structures using enhanced sampling techniques with restraints derived from an ensemble-based model

Tianqi Ma,¹ Tianwu Zang,¹ Qinghua Wang,² and Jianpeng Ma^{1,2*}

¹Applied Physics Program and Department of Bioengineering, Rice University, Houston, Texas, 77005

²Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, Texas, 77030

Received 30 May 2018; Accepted 18 July 2018

DOI: 10.1002/pro.3486

Published online 25 September 2018 proteinscience.org

Abstract: This paper reports a method for high-accuracy protein structural refinement, which is a direct extension of the method in our recent publication (Zang, *J Chem Phys* 2018; 149:072319). It combines a parallel continuous simulated tempering (PCST) method with a temperature-dependent restraint and a blind model selection scheme. In this work, a single-reference-based restraint in previous work was changed to an ensemble-based model (EBM), in which the non-bonded Lennard–Jones term for each contacting atomic pair in previous restraining potential was replaced by a multi-Gaussian function whose parameters are derived from an ensemble of structures such as the ones from various CASP participating groups. The purpose of EBM is to take advantage of partial “correctness” distributed among members of the structural ensemble. Totally 18 targets were refined from the refinement category of CASP10, CASP11 and CASP12. In Top-1 group, 11 out of 18 targets had better models (greater GDT_TS scores) than the CASPR participants. In Top-5 group, nine out of 18 were better. Our results show that PCST-EBM method can considerably improve the low-accuracy structures.

Keywords: structure prediction; high-accuracy refinement; enhanced sampling; empirical potential

Introduction

A very important goal in modern computational biology is to determine the three-dimensional structure of a protein solely from its primary sequence. Currently, the state-of-the-art methods^{1–7} can generate models that are 3–5 Å main chain root-mean-square deviation (RMSD) from their native structures.^{8–10} Improving these low-accuracy models to high-accuracy ones (with 1–2 Å

RMSD) is a monumentally difficult task. This can be seen from biennial event of Critical Assessment of Structure Prediction (CASP).^{11,12} Particularly, CASPR, the refinement category of CASP, progresses much slower than other CASP categories in past few decades.^{13–16}

In our previous paper,¹⁷ in order to improve the low-accuracy protein structural models, the parallel continuous simulated tempering (PCST)¹⁸ was combined with a temperature-dependent restraint using structure-based model (SBM).^{17,19,20} In fully solvated molecular dynamics simulation for refinement, it was shown that PCST-SBM method was able to achieve a more thorough sampling in the configurational space. Furthermore, a novel blind selection method was also introduced to select final models from very long simulation trajectories.

Grant sponsor: NIH AI067839GM067801GM116280; Grant sponsor: Welch FoundationQ-1512Q-1826Q-1826Q-1512; Grant sponsor: University of Texas at Austin; Grant sponsor: National Institutes of HealthR01-GM116280R01-GM127628.

*Correspondence to: Jianpeng Ma, Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, One Baylor Plaza, BCM-125, Houston, Texas 77030.

E-mail: jpma@bcm.tmc.edu

A distinct weakness of SBM restraint,¹⁷ which is derived from a single reference structure, is that the reference structure may itself carry large errors either partially or entirely. These errors can hinder the refinement accuracy. However, in the case that an ensemble of predicted structures is available, such as in CASP competition in which multiple groups generate an ensemble of predicted structures, “correct” structural information may be distributed among various members in the ensemble. For example, this member may have a particularly “correct” structural region while the other may carry a different “correct” region. Therefore, we want to design a restraining scheme such that the distributed structural “correctness” can be taken into account simultaneously. This is the purpose of developing the ensemble-based model (EBM) as a restraint in this paper.

In EBM restraint, the non-bonded Lennard–Jones term for each contacting atomic pair in SBM restraint¹⁷ is replaced by a multi-Gaussian function. The parameters for the multi-Gaussian function are determined by fitting the radial distribution function of the contact distance in an ensemble of structures. In this work, we used the predicted structures provided by various CASP participating groups as the structural ensemble for computing the radial distribution function. In actual implementation, for each target, the starting model of our refinement simulation and the high-temperature reference structure for restraint is the structure provided by CASP organizer, which is in general more accurate than models provided by individual participating groups. We thus weighted the distance information from this reference model in the multi-Gaussian function more than that taken from various participating groups. But the latter may still carry useful information or local “correctness” that is why they are included but weighted differently.

The effectiveness of PCST-EBM scheme is demonstrated on 18 targets selected from the refinement category of CASP10, CASP11, and CASP12 (four from CASP10, six from CASP11, and eight from CASP12). Judged by the global distance test total score, or GDT_TS,^{21–23} the final models generated by the simulation and blind selection scheme are significantly better than the initial models. In Top-1 group, 11 out of 18 cases, and in Top-5 group, nine out of 18 cases, are better than those of the CASPR participants. The results show that the low-accuracy models can be significantly improved by PCST-EBM method despite the large errors in current force fields.

Methods

PCST simulation method and simulation parameters

We apply PCST simulation method in this work.¹⁸ It uses generalized ensemble,²⁴ which changes the traditional Boltzmann distribution to other mathematical

forms $W(X, \beta)$. For example, multicanonical ensembles^{25–31} generates flat potential energy histograms $W(X) \propto \Omega^{-1}(E(X)) = e^{-\beta TS(E(X))}$, while tempering methods^{32–36} can create a flat temperature histogram with $W(T) = \text{const.}$ There are discussions about the efficiencies of these methods.³⁷ The PCST method combines the idea of simulated tempering (ST)^{32,38,39} and parallel tempering (PT).^{33–36} We use multiple copies with different temperature distributions and perform random walk in temperature space. In PCST method, we update the temperature via the following equation:

$$d\left(\frac{1}{\beta}\right) = \left(E - \bar{E}(\beta) - \frac{\gamma}{\beta} + \frac{\beta - \beta_i^0}{\sigma_i^2}\right)dt + \frac{\sqrt{2}}{\beta}dW_t,$$

where the reciprocal temperature β is defined in a large range ($\beta_{\min}, \beta_{\max}$) and dW_t is the Wiener process. The average potential energy $\bar{E}(\beta)$ is the current average potential energy calculated during the simulation, that is $\bar{E}(\beta) \equiv Z^{-1}(\beta) \int E(X) e^{-\beta E(X)} dX$ where $Z(\beta)$ is the canonical partition function at reciprocal temperature β . A probability distribution with three parameters $\{\gamma, \beta_i^0, \sigma_i\}$ in the form of $P(X, \beta) \propto Z^{-1}(\beta) e^{-\beta E(X)} \beta^{-\gamma} e^{-(\beta - \beta_i^0)^2 / 2\sigma_i^2}$ will be generated by the random walk process. We can treat the probability distribution as the sum of Boltzmann distributions at different temperatures with a polynomial-like weight function $\beta^{-\gamma}$ as well as a Gaussian-like weight function with the mean value of β_i^0 and width of σ_i . More details can be found in PCST paper.¹⁸

Besides, we introduce a specific parameter exchange protocol between copies to help the simulation process cross the energy barrier. We define the acceptance ratio for copy exchange between the i th copy and the j th copy as $\min\left\{1, e^{-(\Delta\beta\Delta q_0 + \Delta\beta^2\Delta q_1)}\right\}$ to satisfy the detailed balance condition, where $\Delta\beta \equiv \beta_i - \beta_j$, $\Delta\beta^2 \equiv \beta_i^2 - \beta_j^2$, $\Delta q_0 \equiv \beta_i^0/\sigma_i^2 - \beta_j^0/\sigma_j^2$ and $\Delta q_1 \equiv -1/(2\sigma_i^2) + 1/(2\sigma_j^2)$. The acceptance ratio is only related to the parameters in the temperature space, and it dependent on the extensive quantities implicitly, that is the potential energy E . Therefore, when the system size increases, the exchange rate will not decrease in PCST, this is different from parallel tempering, which usually employs more than 20 copies.^{40–45} In practice, only 2–3 copies are enough to maintain a high rate of exchange between neighboring copies.

In PCST, there are three important time parameters: the simulation time step dt_{CE} , the Langevin equation integration time step dt_{walk} and exchange time interval dt_{ex} . In our simulation, we set the “initial model” provided by the CASP organizers as the starting point of each simulation. The temperature range for our simulation was 293–500 K. Two copies

were employed in PCST, with the parameters $\{\beta^0, \sigma\}$ set of $\{0.38, 0.05\}$ for the low temperature copy and $\{0.27, 0.13\}$ for the high temperature copy (the peaks of Gaussian temperature distribution are at 316 K and 445 K, respectively). The time step for MD integration dt_{CE} was 0.002 ps. The time step for integrating the Langevin equation dt_{walk} was 0.04 ps, which was the same as the neighbor-list refreshing interval. For our PCST method, the time interval of exchange attempt dt_{ex} was 10 ns (5×10^6 steps). We conducted the simulation on Stampede2.⁴⁶

Structure-based model (SBM) restraints

The basis of our restraint is the SBM.^{19,20} It can increase the sampling efficiency around an important conformation. The mathematical form of the SBM restraining potential is

$$V = \sum_{bonds} \epsilon_r (r - r_0)^2 + \sum_{angles} \epsilon_\theta (\theta - \theta_0)^2 + \sum_{improper\ dihedrals} \epsilon_\chi (\chi - \chi_0)^2 + \sum_{backbone} \epsilon_{BB} F_D(\varphi) + \sum_{sidechain} \epsilon_{SC} F_D(\varphi) + \sum_{contacts} \epsilon_C \left[\left(\frac{\sigma_{ij}}{r} \right)^{12} - 2 \left(\frac{\sigma_{ij}}{r} \right)^6 \right] + \sum_{non-contacts} \epsilon_{NC} \left(\frac{\sigma_{ij}}{r} \right)^{12},$$

where $F_D(\varphi) \equiv [1 - \cos(\varphi - \varphi_0)] + [1 - \cos(3(\varphi - \varphi_0))]/2$, σ_{ij} and all the ϵ s are pre-set SBM parameters; r , θ , χ , and φ are derived from the current protein coordinates, r_0 , θ_0 , χ_0 , and φ_0 are parameters derived from the reference state or initial model in our refinement. The “shadow map” scheme is used to select the contacts, which only counts the atom pairs that are in each other’s “first coordination shell.” To be more specific, when the distance between two atoms is in a certain range (e.g. 4–8 Å) and there are no other atoms between them, we consider this atom pairs are considered as contacts. The details can be found in our previous PCST-SBM paper.¹⁷

Ensemble-based model (EBM) restraints

In EBM restraint, we replace the Lennard–Jones term in SBM potential^{19,20} with a multi-Gaussian function. For a particular contact pair, based on its contact distance in different member of the ensemble, we calculate its radial distribution function (RDF) and fit this distribution function with a multi-Gaussian function. We then construct the non-bounded EBM restraining potential term as

$$V = - \sum_{contacts} \sum_i h_i e^{-\frac{(r-r_i)^2}{2\sigma_i^2}}.$$

Three parameters are determined for the i th Gaussian peak in the RDF: the distance r_i , the width σ_i and the height h_i .

In this work, we focus on CASP systems, so we used the predicted structures provided by the CASP organizer from various participants as the structural ensemble for training the RDFs. We show the schematic procedure in Figure 1.

In our actual implementation on CASP systems, we weighted the initial model (provided by CASPR) more than the other models from various groups. Specifically, the restraint function can be rewritten as

$$V = - \sum_{contacts} \sum_i h_i e^{-\frac{(r-r_i)^2}{2\sigma_i^2}} w_i$$

where w_i is the weight of the i th Gaussian peak. We only considered Gaussian peaks whose r_i is no more than three peaks away from the peak of r_0 in the RDF. The rational of our doing so is that the initial model we used in refinement provided by CASPR organizer is in general more accurate than models by individual participating groups. However, the models taken from various participating groups may still carry useful information or local “correctness” in their structures. That is why they are included but weighted differently.

Temperature-dependent restraining protocol

In SBM¹⁷ and EBM, the amplitude of restraint is temperature dependent.⁴⁷ The restraining strength is set to zero at room temperature for recovering original energy landscape. The restraining strength reaches its maximum at the highest temperature for holding the system near the reference state during the simulation. In practice, a biased potential function $V(X)$ is used, depending on molecular coordinates X , as a restraint and it is added to the generalized ensemble without disturbing room-temperature properties. In other word, the Hamiltonian is changed to $H(X) = H_0(X) + \lambda(T)V(X)$ where $H_0(X)$ is the original Hamiltonian and $\lambda(T)$ is a linear function of temperature T . Two parameters $\{\lambda_{min}, \lambda_{max}\}$ define the shape of $\lambda(T)$ and the value of λ at T_{min} (0 in our case) and the value at T_{max} , respectively. The function $\lambda(T)$ is a linear function, and the equation is now

$$d\left(\frac{1}{\beta}\right) = \left(E' - \overline{E'}(\beta) - \frac{\gamma}{\beta} + \frac{\beta - \beta_i^0}{\sigma_i^2}\right)dt + \frac{\sqrt{2}}{\beta}dW_t,$$

where $E' \equiv E(X) - V(X)$ denotes the difference between the original potential energy and the biased

Figure 1. The schematic illustration of the construction of EBM restraint. (A) For each non-bonded atomic contacting pair, contact distance is measured in an ensemble of structures. (B) A histogram is first computed for population distribution as a function of contact distance. Then a continuous radial distribution function was generated by fitting the histogram. Example is shown for (9, 205) pair. The continuous line represents the fitted radial distribution function from the histogram.

energy. Therefore, the desired probability distribution is

$$P(X, \beta) \propto Z^{-1}(\beta) e^{-\beta[E(X) + \lambda(T)V(X)]} \beta^{-\gamma} e^{-(\beta - \beta_i^0)^2 / 2\sigma_i^2}.$$

This temperature-dependent restraining scheme helps the simulation specifically in high temperature, because it stops the system from drifting too far away in high temperature.

Model selection methods

The model selection scheme in this paper is the same as that in PCST-SBM.¹⁷ The structures are usually not stable at high temperature so, in Step I, a temperature threshold of 300 K is used to eliminate candidates of models. In this step, all models with temperature higher than the threshold are eliminated, and the number of candidate models can be reduced to $10^3 - 10^4$, which is 5–10% of the total number of states in the beginning. In Step II, GOAP potential⁴⁸ is used to score the models. The clusters are ranked by the average GOAP scores among the members inside the clusters. In the last step (Step III), an averaged structure for each cluster is generated to represent the cluster.

Refinement of CASPR targets using PCST-EBM method

PCST-EBM method was used to refine 18 targets selected from the refinement category of CASP10, CASP11, and CASP12 (four from CASP10, six from CASP11, and eight from CASP12). The targets have a great variety of sizes, ranging from 62 to 251 residues, and contain different secondary structure components. The C α -RMSD and the GDT_TS of the initial models, compared to the native structures, are in the range of 1.97–12.36 Å and 38.94–90.24, respectively. For each target, a 1000 ns all-atom MD simulation in explicit solvent was performed using the molecular simulation package GROMACS 4.5.^{49–52} TIP3P⁵³ model was used for explicit solvent, AMBER99SB-ILDN⁵⁴ was used as the force field, and explicit ions were used for eliminating net charges.

For each target, the initial model provided by the CASP organizers was used both as the “reference state” of EBM restraint and the “starting point” of the refinement simulation. These structures are also the starting points of CASPR. The SMOG server at Rice University⁵⁵ was used to extract the SBM parameters from the reference state. Then, using the predicted structures provided by CASP participating groups as the ensemble of structures, the parameters for multi-Gaussian function for non-bonded term was calculated so as to construct the EBM restraint.

Results

Best models in the refinement of CASPR targets using PCST-EBM method

In Table I, the best model in the refinement of each target, identified by the native structure, is compared with the initial model. The average GDT_TS enhancement is 11.23 over the initial models while the greatest GDT_TS enhancement reaches 30.16 for TR705. The best models from PCST-EBM also have much smaller C α -RMSD from the native structures. The average improvement in C α -RMSD is 1.33 Å, with the highest being 3.35 Å for TR705. Clearly, the results suggest that PCST-EBM can “sense” the “correct fragments” from an ensemble of predicted models and improve the refinement simulation. It is worth noting that the results in Table I are the best models from trajectories identified by the corresponding native structures, while the results after blind selection will be discussed in next section.

Blind selection of high-quality models from trajectories

In CASPR competition, the results are selected without any knowledge of native structure, and only five candidates are allowed. Thus the final results are presented in terms of five candidates and one candidate group. In Table II, we present the results after blind selection by the same rule. The blind model selection procedure is the same as the procedure in PCST-SBM.¹⁷ Overall, in terms of GDT_TS improvement, nine out of 18 targets from PCST-EBM

Table I. Improvements in C α -RMSD and GDT_TS of the Best PCST-EBM Models (Best) Compared with the Initial Models Provided by CASP (Initial). TR663 to TR705 are CASP10 Targets, TR780 to TR857 are CASP11 Targets, and TR862 to TR922 are CASP12 Targets

Target	Size	Initial		Best	
		RMSD (Å)	GDT_TS	Δ RMSD (Å)	Δ GDT_TS
TR663	152	3.37	69.24	-1.56	28.13
TR674	132	3.92	85.80	-1.78	7.38
TR704	235	3.52	70.21	-1.75	17.52
TR705	96	5.21	64.84	-3.35	30.16
TR780	95	3.70	74.21	-1.77	13.42
TR786	217	4.31	69.36	-1.47	13.59
TR803	134	6.58	52.98	1.18	2.98
TR811	251	2.10	90.24	-0.30	2.49
TR837	121	3.64	65.70	-1.11	7.23
TR857	96	4.97	57.29	-1.54	7.95
TR862	93	5.55	58.60	-1.17	9.41
TR866	104	3.27	79.57	-1.68	12.5
TR868	105	1.97	80.95	-0.69	11.29
TR869	104	12.36	38.94	-1.21	3.61
TR870	109	7.64	42.43	0.22	4.59
TR872	88	5.76	73.86	-4.23	20.17
TR893	169	2.89	87.28	-0.88	1.78
TR922	62	2.10	89.52	-0.93	8.06
Average	131	4.60	69.50	-1.33	11.23

refinement are better than those from CASPR in Top-5 group and 11 out 18 targets are better in Top-1 group. Moreover, in Top-1 group, the largest GDT_TS

improvement is in TR705 with a value of 24.35 (the CASPR is 5.21). The average improvement of GDT_TS is 7.19 for Top-5 group and 5.91 for Top-1

Table II. GDT_TS Scores in the Refinement (TR663 to TR705 are CASP10 Targets, TR780 to TR857 are CASP11 Targets, and TR862 to TR922 are CASP12 Targets). Numbers in Initial are the GDT_TS Scores for Initial Models Provided by CASP Organizers, that is, the Starting Points of CASPR. Numbers in Δ CSPR5 are the GDT_TS Score Improvements of the Best Models Among all CASPR Participants Evaluated by the Organizers using the Native Structures (Each Participating Group Contributed five Models). Numbers in Top5 are the GDT_TS Score Improvements of the Best Models Evaluated by the Native Structures Among the Top5 choices provided by Our Blind Selection Scheme. Numbers in Rank5 are the Ranking of the Models in Top5 Among the Five Blindly Selected Models, therefore, 1 in Rank5 Means the Best Model Indicated by the Blind Selection Scheme is the Same as the One Evaluated by the Native Structures (there are totally six targets with 1 in Rank5). Numbers in Δ CSPR1 are the GDT_TS Score Improvements for the Best Models Declared by the CASP Participants Among all Participating Groups. Numbers in Top1 are the GDT_TS Score Improvements for the Best Models Provided by Our Blind Selection Scheme (14 of them are Better than the Initial Models and 11 of them are Better than those in Δ CSPR1). The Systems that Appear in Figure 2 are Highlighted in Bold

Target	Initial	Δ CSPR5	Top5	Rank5	Δ CSPR1	Top1
TR663	69.24	8.06	18.92	2	5.60	15.30
TR674	85.80	1.7	5.49	4	1.7	3.59
TR704	70.13	5.85	11.63	1	3.09	11.63
TR705	64.84	6.51	25.52	5	5.21	24.35
TR780	74.21	5.00	1.87	1	4.74	1.84
TR786	69.36	4.14	7.60	4	4.14	5.28
TR803	52.98	3.36	0.56	1	3.36	0.56
TR811	90.24	1.89	0.20	4	1.39	-1.00
TR837	65.70	4.75	-2.91	2	3.52	-5.58
TR857	57.29	4.95	4.69	1	3.91	4.69
TR862	58.60	4.57	7.80	1	3.77	7.80
TR866	79.57	9.61	12.50	2	9.61	11.78
TR868	80.95	8.34	6.43	3	3.57	4.76
TR869	38.94	8.18	2.41	3	8.18	-0.24
TR870	42.43	9.41	1.61	5	5.73	-0.23
TR872	73.86	3.41	16.77	2	3.41	15.06
TR893	87.28	1.48	1.06	1	1.03	1.06
TR922	89.52	5.64	7.25	4	5.64	5.64
Average	69.50	5.38	7.19		4.31	5.91

group. These results are consistently better than the CASPR results, which are 5.34 and 4.31, respectively. We show the superpositions of some targets in Figure 2.

For the targets that were refined by both PCST-SBM¹⁷ method and PCST-EBM method (totally nine of them in Table III), we can see direct comparison of the two methods. In both Top-5 and Top-1 groups, the performance of EBM is only slightly worse than that of SBM for TR704 and TR870.

A very important fact is that the result of PCST-EBM listed in Table III are the results by our single group, but the results of CASPR (Δ CSPR5 and Δ CSPR1) are the best results among ALL CASPR participants. This indicates that our results are more significant than what the numbers indicate alone.

Concluding Discussion

The PCST-EBM method described in this paper is a direct extension of our previous PCST-SBM method.¹⁷ It combined an enhanced sampling method, parallel continuous simulated tempering (PCST),¹⁸ with a temperature-dependent coordinate-based restraint, the ensemble-based model (EBM), to conduct thorough search of configurational space in the process of protein structural refinement. The EBM restraint was constructed in such a way that the non-bonded Lennard–Jones term for each

contacting atomic pair, in single-reference-based SBM potential^{19,20} used in previous paper,¹⁷ was replaced by a multi-Gaussian function for including the contributions of partial structural “correctness” distributed among an ensemble of predicted structures. In this work, we used the predicted structures provided by various CASP participating groups as the structural ensemble. Finally, a blind selection scheme, also developed in our previous paper,¹⁷ was used to pick out models from the long trajectory.

We applied PCST-EBM refinement protocol to 18 targets from the refinement category of CASP10, CASP11, and CASP12. In Top-5 group, nine out of 18 targets had better final models (greater GDT_TS scores) than the CASPR participants. In Top-1 group, 11 out of 18 were better. The results seem to suggest that PCST-EBM method can considerably improve the low-accuracy structures.

Although the only difference between PCST-EBM method in this paper and PCST-SBM method in the previous paper¹⁷ is that the non-bonded term in the restraining potential in EBM takes a form that contains contributions from an ensemble of structures, rather than those from a single structure in SBM. EBM restraint, however, seems to deliver much better performance than that of SBM. It seems to allow the system to explore more basins in the free energy landscape than SBM. We found that such a

Figure 2. The superposition of refined structures in Top-1 group for four targets, (A) TR663, (B) TR704, (C) TR705, and (D) TR872 are shown. In each case, the native structure is shown in green, the initial model in purple and the final model in Top-1 group in Table II is shown in red. Thus, the closer the green and red models, the better our refinement performance is.

Table III. Comparison of Performance on the Common Targets Refined by PCST-SBM and PCST-EBM Methods. Numbers in Initial, Δ CSPR5, and Δ CSPR1 the Same as those in Table II. Terms with Suffix “_SBM” are the Results from PCST-SBM Refinement, and Terms with Suffix “_EBM” are the Results from PCST-EBM Refinement. Numbers in Top5 are the GDT_TS Score Improvements of the Best Models Evaluated by the Native Structures Among the Top5 Choices Provided by Our Blind Selection Scheme. Numbers in Top1 are the GDT_TS Score Improvements for the Best Models Provided by Our Blind Selection Scheme

Target	Initial	Δ CSPR5	Top5_SBM	Top5_EBM	Δ CSPR1	Top1_SBM	Top1_EBM
TR663	69.24	8.06	15.63	18.92	5.60	13.16	15.30
TR674	85.80	1.7	2.46	5.49	1.70	-0.95	3.59
TR704	70.13	5.85	12.77	11.63	3.09	12.77	11.63
TR705	64.84	6.51	7.82	25.52	5.21	7.82	24.35
TR862	58.60	4.57	1.08	7.80	4.57	1.08	7.80
TR866	79.57	9.61	10.81	12.50	9.61	10.81	11.78
TR868	80.95	8.34	-2.57	6.43	3.57	-7.09	4.76
TR869	38.94	8.18	-2.58	2.41	8.18	-2.88	-0.24
TR870	42.43	9.41	5.96	1.61	5.73	5.96	-0.23
Average	65.61	6.91	5.71	10.26	5.25	4.52	8.75

scheme is very effective in letting system explore distributed “correctness” among difference sources.

We also wish to point out that, from Figure 2, even for the systems with the largest improvement, for example, a small helix seems to be flipped in its orientation in Figure 2(D) and large loop movement can be seen in Figure 2(C), the shifts of structural elements are in general not dramatic in terms of the refolding of local features such as either secondary or tertiary structural elements. The reason for that could be that the restraints used in refinement simulation limited the sampling near the reference structure even though the EBM restraints offered greater flexibility in sampling distributed “correctness” from a structural ensemble. Another fact is that larger GDT_TS improvements usually happen in the systems with relatively large GDT_TS scores to begin with, that is, the initial model quality dependence. That seems to suggest that the choice of restraints play a very important role in achieving good refinement outcome. However, the results also suggest that if the overall topology of a polypeptide is reasonable, then the errors in local regions can be refined by PCST-EBM method. This makes the method very attractive to the refinement of structural models that are built for example against cryo-EM density maps at intermediate resolutions.

Acknowledgments

J. M. acknowledges support by grants from the National Institutes of Health (R01-GM127628, R01-GM116280) and the Welch Foundation (Q-1512). Q. W. acknowledges support by grants from the National Institutes of Health (R01-GM127628, R01-GM116280) and The Welch Foundation (Q-1826). The authors also acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin. The project in this paper used the Extreme Science and Engineering Discovery

Environment (XSEDE) Stampede2 at the TACC through allocation MCB150013.

References

1. Zhang Y (2008) Progress and challenges in protein structure prediction. *Curr Opin Struct Biol* 18:342–348.
2. Faraggi E, Yang Y, Zhang S, Zhou Y (2009) Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. *Structure* 17:1515–1527.
3. Zhang Y (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinform* 9:40.
4. Raman S, Vernon R, Thompson J, Tyka M, Sadreyev R, Pei J, Kim D, Kellogg E, DiMaio F, Lange O (2009) Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* 77:89–99.
5. Marti-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Šali A (2000) Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* 29:291–325.
6. Peng J, Xu J (2011) RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins* 79:161–171.
7. Wu S, Skolnick J, Zhang Y (2007) Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol* 5:17.
8. Huang YJ, Mao B, Aramini JM, Montelione GT (2014) Assessment of template-based protein structure predictions in CASP10. *Proteins* 82(Suppl 2):43–56.
9. Mariani V, Kiefer F, Schmidt T, Haas J, Schwede T (2011) Assessment of template based protein structure predictions in CASP9. *Proteins* 79:37–58.
10. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A (2014) Critical assessment of methods of protein structure prediction (CASP)–round x. *Proteins* 82(Suppl 2):1–6.
11. Moult J, Pedersen JT, Judson R, Fidelis K (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins* 23:ii–v.
12. Moult J (2005) A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol* 15:285–289.
13. Nugent T, Cozzetto D, Jones DT (2014) Evaluation of predictions in the CASP10 model refinement category. *Proteins* 82(Suppl 2):98–111.

14. MacCallum JL, Perez A, Schnieders MJ, Hua L, Jacobson MP, Dill KA (2011) Assessment of protein structure refinement in CASP9. *Proteins* 79(Suppl 10): 74–90.
15. Modi V, Dunbrack RL (2016) Assessment of refinement of template-based models in CASP11. *Proteins* 84: 260–281.
16. Hovan L, Oleinikovas V, Yalinca H, Kryshchukovych A, Saladino G, Gervasio FL (2018) Assessment of the model refinement category in CASP12. *Proteins* 86:152–167.
17. Zang T, Ma T, Wang Q, Ma J (2018) Improving low-accuracy protein structures using enhanced sampling techniques. *J Chem Phys* 149:072319.
18. Zang T, Yu L, Zhang C, Ma J (2014) Parallel continuous simulated tempering and its applications in large-scale molecular simulations. *J Chem Phys* 141:044113.
19. Whitford PC, Noel JK, Gosavi S, Schug A, Sanbonmatsu KY, Onuchic JN (2009) An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields. *Proteins* 75: 430–441.
20. Clementi C, Nymeyer H, Onuchic JN (2000) Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J Mol Biol* 298:937–953.
21. Zemla A, Venclovas C, Moulton J, Fidelis K (1999) Processing and analysis of CASP3 protein structure predictions. *Proteins Suppl* 3:22–29.
22. Zemla A, Venclovas C, Moulton J, Fidelis K (2001) Processing and evaluation of predictions in CASP4. *Proteins* 45:13–21.
23. Zemla A (2003) LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res* 31: 3370–3374.
24. Mitsutake A, Sugita Y, Okamoto Y (2001) Generalized-ensemble algorithms for molecular simulations of biopolymers. *Peptide Sci* 60:96–123.
25. Berg BA, Neuhaus T (1992) Multicanonical ensemble: a new approach to simulate first-order phase transitions. *Phys Rev Lett* 68:9–12.
26. Baumann B (1987) Noncanonical path and surface simulation. *Nucl Phys B* 285:391–409.
27. Wang F, Landau DP (2001) Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys Rev Lett* 86:2050–2053.
28. Dayal P, Trebst S, Wessel S, Wurtz D, Troyer M, Sabhapandit S, Coppersmith SN (2004) Performance limitations of flat-histogram methods. *Phys Rev Lett* 92:097201.
29. Yan Q, Faller R, de Pablo JJ (2002) Density-of-states Monte Carlo method for simulation of fluids. *J Chem Phys* 116:8745–8749.
30. Yan Q, de Pablo JJ (2003) Fast calculation of the density of states of a fluid by Monte Carlo simulations. *Phys Rev Lett* 90:035701.
31. Shell MS, Debenetti PG, Panagiotopoulos AZ (2002) Generalization of the Wang-Landau method for off-lattice simulations. *Phys Rev E Stat Nonlin Soft Matter Phys* 66:056703.
32. Marinari E, Parisi G (1992) Simulated tempering: a new Monte Carlo scheme. *Europhys Lett* 19:451–458.
33. Swendsen RH, Wang J-S (1986) Replica Monte Carlo simulation of spin-glasses. *Phys Rev Lett* 57:2607–2609.
34. Falcioni M, Deem MW (1999) A biased Monte Carlo scheme for zeolite structure solution. *J Chem Phys* 110: 1754–1766.
35. Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett* 314:141–151.
36. Earl DJ, Deem MW (2005) Parallel tempering: theory, applications, and new perspectives. *PCCP* 7:3910–3916.
37. Jiang P, Yaşar F, Hansmann UH (2013) Sampling of protein folding transitions: multicanonical versus replica exchange molecular dynamics. *J Chem Theory Comput* 9:3816–3825.
38. Zhang C, Ma JP (2009) Enhanced sampling in generalized ensemble with large gap of sampling parameter: case study in temperature space random walk. *J Chem Phys* 130: 194112.
39. (2010) Enhanced sampling and applications in protein folding in explicit solvent. *J Chem Phys* 132:244101.
40. Bussi G, Gervasio FL, Laio A, Parrinello M (2006) Free-energy landscape for beta hairpin folding from combined parallel tempering and metadynamics. *J Am Chem Soc* 128:13435–13441.
41. Camilloni C, Provati D, Tiana G, Broglia RA (2008) Exploring the protein G helix free-energy surface by solute tempering metadynamics. *Proteins* 71:1647–1654.
42. Barducci A, Bonomi M, Prakash MK, Parrinello M (2013) Free-energy landscape of protein oligomerization from atomistic simulations. *Proc Natl Acad Sci USA* 110:E4708–E4713.
43. Zhou R, Berne BJ, Germain R (2001) The free energy landscape for beta hairpin folding in explicit water. *Proc Natl Acad Sci USA* 98:14931–14936.
44. Plazinska A, Plazinski W, Jozwiak K (2014) Fast, metadynamics-based method for prediction of the stereochemistry-dependent relative free energies of ligand-receptor interactions. *J Comput Chem* 35: 876–882.
45. Garcia AE, Onuchic JN (2003) Folding a protein in a computer: an atomic description of the folding/unfolding of protein A. *Proc Natl Acad Sci USA* 100:13898–13903.
46. Towns J, Cockerill T, Dahan M, Foster I, Gaither K, Grimshaw A, Hazlewood V, Lathrop S, Lifka D, Peterson GD (2014) XSEDE: accelerating scientific discovery. *Comput Sci Engin* 16:62–74.
47. Zhang C, Ma J (2012) Folding helical proteins in explicit solvent using dihedral-biased tempering. *Proc Natl Acad Sci USA* 109:8139–8144.
48. Zhou H, Skolnick J (2011) GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophys J* 101:2043–2052.
49. Berendsen HJC, van der Spoel D, van Drunen R (1995) GROMACS: a message-passing parallel molecular dynamics implementation. *Comput Phys Commun* 91:43–56.
50. Lindahl E, Hess B, Van Der Spoel D (2001) GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J Mol Model* 7:306–317.
51. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ (2005) GROMACS: fast, flexible, and free. *J Comput Chem* 26:1701–1718.
52. Hess B, Kutzner C, Van Der Spoel D, Lindahl E (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput* 4:435–447.
53. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79:926–935.
54. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, Shaw DE (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* 78:1950–1958.
55. Noel JK, Levi M, Raghunathan M, Lammert H, Hayes RL, Onuchic JN, Whitford PC (2016) SMOG 2: A versatile software package for generating structure-based models. *PLoS Comput Biol* 12:e1004794.