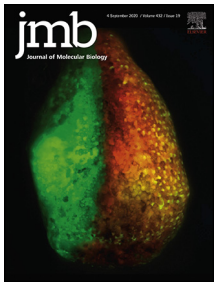

A New Protocol for Atomic-Level Protein Structure Modeling and Refinement Using Low-to-Medium Resolution Cryo-EM Density Maps

Biao Zhang [1,2], Xi Zhang [2], Robin Pearce [2], Hong-Bin Shen [1] and Yang Zhang [2,3,]

*1 - **Institute of Image Processing and Pattern Recognition,** Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, China*
*2 - **Department of Computational Medicine and Bioinformatics,** University of Michigan, Ann Arbor, MI 48109, USA*
*3 - **Department of Biological Chemistry,** University of Michigan, Ann Arbor, MI 48109, USA*

***Correspondence to Hong-Bin Shen and Yang Zhang:** Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, China. Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI, 48109-2218, USA. hbshen@sjtu.edu.cn, zhng@umich.edu*
https://doi.org/10.1016/j.jmb.2020.07.027
Edited by Michael Sternberg

Abstract

The rapid progress of cryo-electron microscopy (cryo-EM) in structural biology has raised an urgent need for robust methods to create and refine atomic-level structural models using low-resolution EM density maps. We propose a new protocol to create initial models using I-TASSER protein structure prediction, followed by EM density map-based rigid-body structure fitting, flexible fragment adjustment and atomic-level structure refinement simulations. The protocol was tested on a large set of 285 non-homologous proteins and generated structural models with correct folds for 260 proteins, where 28% had RMSDs below 2 Å. Compared to other state-of-the-art methods, the major advantage of the proposed pipeline lies in the uniform structure prediction and refinement protocol, as well as the extensive structural re-assembly simulations, which allow for low-to-medium resolution EM density map-guided structure modeling starting from amino acid sequences. Interestingly, the quality of both the image fitting and subsequent structure refinement was found to be strongly correlated with the correctness of the initial I-TASSER models; this is mainly due to the different correlation patterns observed between force field and structural quality for the models with template modeling score (or TM-score, a metric quantifying the similarity of models to the native) above and below a threshold of 0.5. Overall, the results demonstrate a new avenue that is ready to use for large-scale cryo-EM-based structure modeling and atomic-level density map-guided structure refinement.

Introduction

The determination of high-resolution protein structures is critical to biological function annotation and rational drug discovery. Recent progress in single-particle cryogenic electron microscopy (cryo-EM) has brought about exciting opportunities for direct determination of protein structures without the need for crystallization [1–3]. A rapid accumulation of low-to-medium resolution protein structures derived from EM data has been recently witnessed in the Electron Microscopy Data Bank (EMDB) [4]. While the surge in the number of cryo-EM-derived structures is mainly due to advancements in electron detector technology and image-processing techniques [5], the field still lacks advanced computational methods to create and refine atomic-level structure models. This is particularly challenging when the resolutions of the EM density maps are in the intermediate range (5–10 Å) [6].

Considerable effort has been made to address this challenge [7], where the approaches can be largely categorized into two groups. The first constructs protein (mainly complex) models by fitting known component structures into EM density maps, which can be done either by normal mode analysis [8,9] or through fast Fourier transform (FFT) based rigid-body docking [10]. The second group tries to refine existing structure models by maximizing the correlation

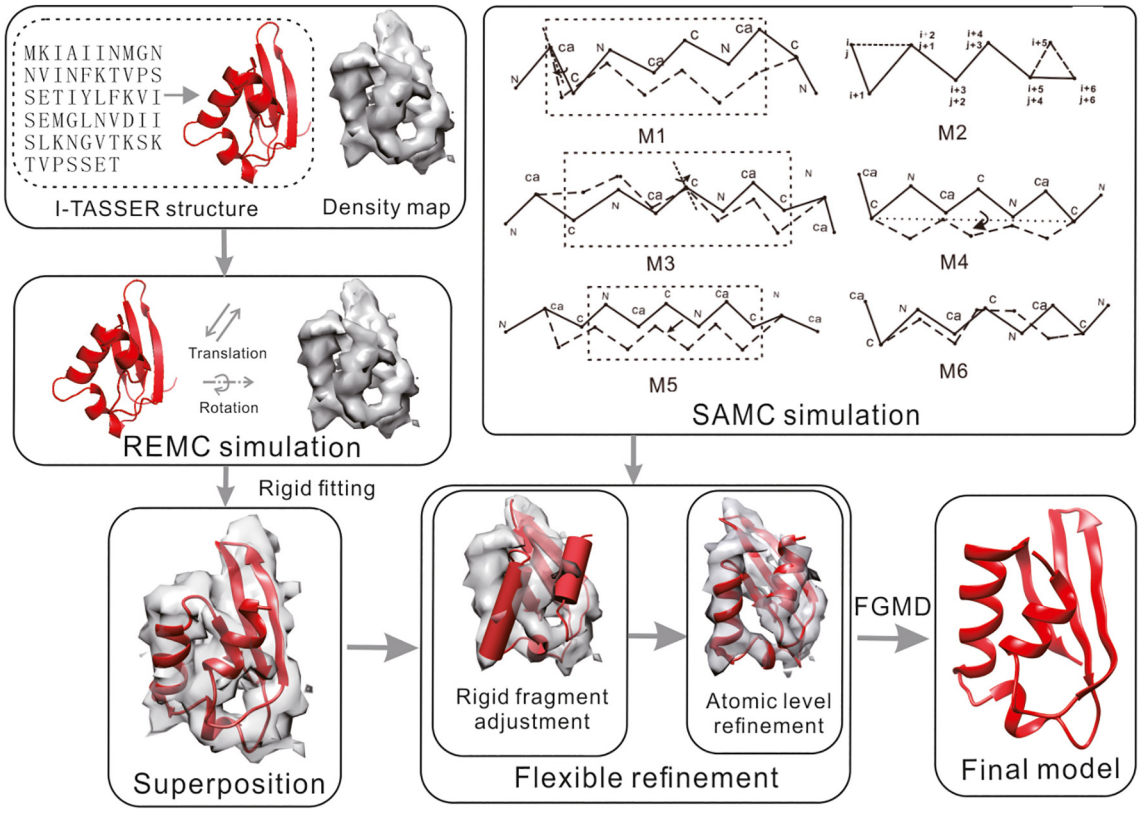

Figure 1. Flowchart of EM-Refiner, which fits and refines low-resolution predicted protein structures generated by I-TASSER using cryo-EM density maps. Following I-TASSER structure prediction, the EM density map-guided modeling procedure consists of three steps of structure-density map superposition, fragment adjustment and atom-level refinement simulations.

between the refined structures and the experimental EM density maps through molecular dynamics simulations (e.g., MDFF [11], Flex-EM [12]) or Monte Carlo simulations (e.g., Rosetta [13]). Despite their successes, many of the approaches perform rigid-body fitting and flexible refinement separately. For example, "fit-in-map" in Chimera [14], ADPEM [15] and EMFIT [16] are designed for rigid-body structure and density map superposition; Rosetta [13] and Flex-EM [11] require users to input the initial structure superposed with the density map; and CCP-EM [17], Scipion [18] and Situs [19] perform rigid-body fitting and flexible refinement separately. These requirements and constraints limit the usefulness and efficiency of the approaches, since fitting and refinement are two intertwined processes and most proteins do not have known crystal structures.

In this work, we propose a new uniform pipeline, EM-Refiner, which first generates low-resolution structure models using the cutting-edge I-TASSER protein structure prediction method. Next, the models are superposed with the EM density map, and atomic-level, density map-guided flexible refinement simulations are carried out (Figure 1). To examine the effectiveness of the pipeline, we performed a large-scale benchmark test on 280 non-redundant proteins with density maps created from both noise-free simulations and cryo-EM experiment data. The results demonstrate the significant advantages of the EM-Refiner pipeline for cryo-EM-based protein structure refinement and modeling over current state-of-the-art methods [10,12,13]. The source code and online server for EM-Refiner are freely available at https://zhanglab.ccmb.med.umich.edu/EM-Refiner/.

Results

Benchmark results using simulated density maps

Benchmark dataset

The benchmark dataset contained 278 single-domain proteins with sequence lengths ranging from 76 to 620 residues and pair-wise sequence identities <30%. It included 48 α-, 34 β-, and 196 $\alpha\beta$-proteins, following their SCOP categorization [20]. Simulated noise-free density maps were generated from the target structures using EMAN2 (pdb2mrc) [21], where the resolution for each domain was randomly selected

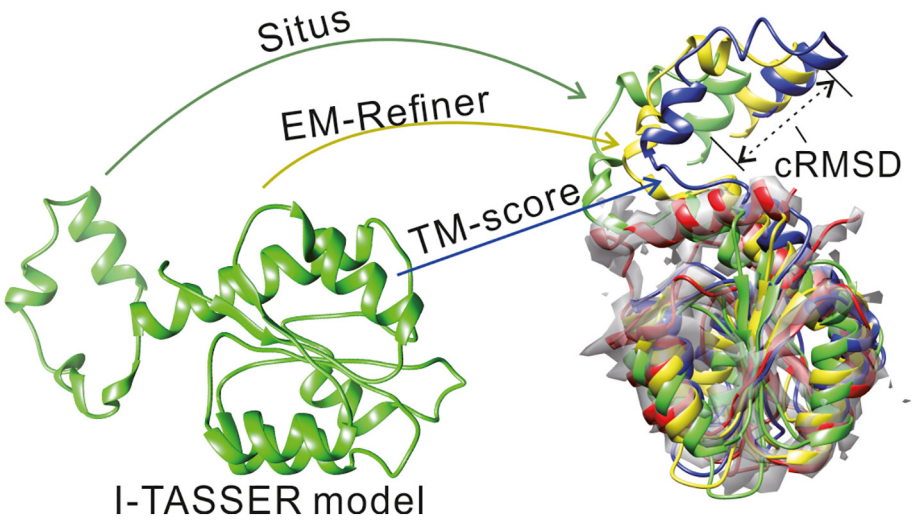

Figure 2. Illustration of the rigid-body superposition of the predicted model onto the EM density map for the circadian clock protein kaiA (PDB ID: 4g86A). The initial I-TASSER model (lime) was superposed onto the EM density-map by EM-Refiner (yellow) and Situs (green), which had cRMSDs of 3.84 and 5.17 Å, respectively, compared to the ideal position (blue) obtained by TM-score superposition onto the target structure (red).

in the integer range from 5 to 10 Å. For a given resolution value, each atom in the simulated density map is represented as a 3D Gaussian distribution with the deviation width proportional to the resolution and the amplitude proportional to the atom mass. The other parameters were set to their default values. In Table S3 in the Supporting Information (SI), we list the selected resolution parameters for all test domains, where a histogram is given in Figure S1. Here, one reason for us to choose the range of 5–10 Å is that it represents a typical low-to-medium resolution range for which I-TASSER and other structural modeling tools are most needed, as high-resolution structure determination is difficult to achieve from the density maps alone, although the average resolution of the cryo-EM data in the community has kept improving in the last years [4].

For each target, the initial structure models were predicted by I-TASSER [22], for which all homologous templates with sequence identities $\geq$30% to the query were excluded. The average TM-score of the resulting I-TASSER models was 0.713, where 254 out of the 278 targets had TM-scores $\geq$0.5 and 24 had TM-scores $<$0.5. A more detailed TM-score distribution is displayed in Figure S2, which is largely consistent with the typical accuracy range of I-TASSER in blind tests [23]. Here, TM-score is a metric used to measure the similarity between two protein structures [24]. The value of TM-score falls in the range (0,1], where a TM-score of 1 indicates a perfect match between two structures and a value $\geq$0.5 indicates that two structures share the same fold [25]. Our assessments of the initial and final model quality will be mainly based on TM-score, since it is more sensitive than RMSD to the topological similarity of protein structures, although the latter is also listed to provide additional reference information. The initial model set, target structures and the simulated density map data can be download at https://zhanglab.ccmb.med.umich.edu/EM-Refiner/.

Initial model and density map superposition

Many of the commonly used cryo-EM structure refinement programs require users to manually specify the initial protein structure conformation superposed with the density map. Because correct superposition between the input structure and density map can reduce the conformational search time and impact the final modeling results, EM-Refiner performs a quick replica-exchange Monte Carlo (REMC) simulation search to superpose the initial model and density map, which is guided purely by the correlation coefficient between the model and EM map data (see Methods).

To assess the effectiveness of this procedure, we calculated the coordinate RMSDs, $\text{cRMSD} = \sqrt{\sum_{i=1}^{L}|\overrightarrow{x_i}-\overrightarrow{y_i}|^2/L}$, of the superposed models, where L is the protein length, $\overrightarrow{x}$ represents the coordinates of the I-TASSER model superposed onto the target structure using the TM-score program [24], and $\overrightarrow{y}$ represents the coordinates of the model generated by EM-Refiner-based superposition with the EM density map. On average, the cRMSD by EM-Refiner was 2.94 Å, which was slightly lower than that obtained by Situs (3.45 Å), which uses an FFT-based search technique [10]. Among the 278 test cases, EM-Refiner created lower cRMSD superpositions for 152 cases, while Situs did so for 126

Table 1. Summary of modeling results by I-TASSER structure prediction and the follow-up EM density map refinement methods on 278 test proteins

Methods	Starting from predicted superposition		Starting from ideal superposition	
	TM-score (*p*-value)	RMSD (Å)	TM-score (*p*-value)	RMSD (Å)
I-TASSER	0.713 (7.5×10^{-49})	5.71	0.713 (4.4×10^{-82})	5.71
Flex-EM	0.714 (6.3×10^{-50})	5.41	0.718 (1.8×10^{-41})	5.37
Rosetta	0.761 (6.1×10^{-4})	4.76	0.780 (3.7×10^{-5})	4.62
EM-Refiner	0.779	4.74	0.796	4.37

P-values were calculated using two-tailed Student's t-tests between the TM-scores produced by EM-Refiner and the other programs.

cases. Part of the reason for the better performance by EM-Refiner is that it uses different poses and structural conformations as initial structures for its REMC simulations, which increases the diversity of the superposition search results, while the FFT-based Situs search only corresponds to one energy basin. In Figure 2, we show a representative example from the circadian clock protein kaiA (PDB ID: 4g86A), for which the superposed model by EM-Refiner had a 3.84 Å cRMSD, while the cRMSD of the Situs model was 5.17 Å.

Performance of EM-Refiner on structure refinement

Table 1 (columns 2–3) lists a summary of the structural refinement results for EM-Refiner starting from the predicted initial model superpositions. On average, EM-Refiner achieved a TM-score of 0.779 for the 278 test proteins, which was significantly higher than that of the initial I-TASSER models (0.713), the difference of which corresponded to a *p*-value of 7.5E-49 as determined by a two-tailed Student's *t*-test. The RMSD of the I-TASSER models was also significantly reduced from 5.71 to 4.74 Å by EM-Refiner refinement, corresponding to a *p*-value of 1.26E-32. These results demonstrate the effectiveness of EM-Refiner at refining low-to-medium resolution models.

In Figure 3 (column 1), we present a head-to-head comparison of the TM-score and RMSD values between the models produced by I-TASSER and EM-Refiner. Interestingly, the majority of the significantly

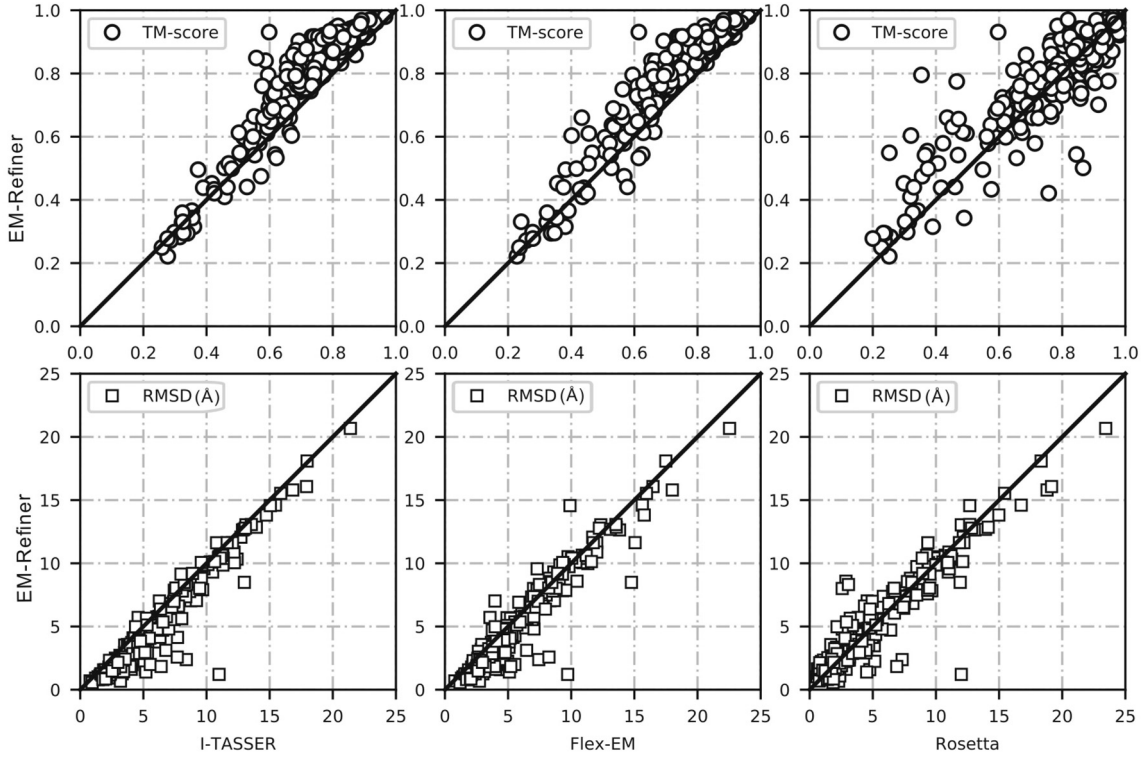

Figure 3. Comparison of structure models obtained by different methods on 278 test proteins. Rows 1 and 2 are the TM-score and RMSD to the native structure, respectively.

refined cases are located in the region where the initial I-TASSER models had TM-scores above 0.5, which largely corresponds to models with correct folds [25]. More specifically, for initial models with correct fold (TM-score≥ 0.5), the improvement was significant with a *p*-value = 1.7E-46 between the EM-Refiner and I-TASSER results, where the average TM-score increased 10.5% from 0.746 (I-TASSER) to 0.824 (EM-Refiner). For initial models with incorrect fold (TM-score < 0.5), however, the improvement was much less significant (the average TM-scores were 0.368 and 0.365 for EM-Refiner and I-TASSER, respectively, corresponding to a *p*-value of 0.75). This happened mainly due to the special energy landscape of EM-Refiner, which is dominated by the structure and density map correlation coefficient (*CC*, see Methods), where *CC* and the TM-score have a pronounced correlation only when the TM-score is high (Figure 4). Apparently, with such golf-course like energy landscape, it is likely the initial models could be driven in a correct direction toward the native, if the initial predicted model is of correct fold and with the starting point near the native basin of the landscape [26–28]. On the other hand, if the initial model is further away from the native, the current energy landscape lacks a long-range energy funnel to guide the conformational search toward the native energy basin and therefore fails on structural refinements.

However, we did notice a few cases where EM-Refiner pushed the initial models with incorrect folds toward the correct folds. Figure S3 shows an illustrative example of this (T0892_dom1) from the UDO-Glucose Glycoprotein Glucosyltransferase of *Chaetomium thermophilum* double mutant D611C:

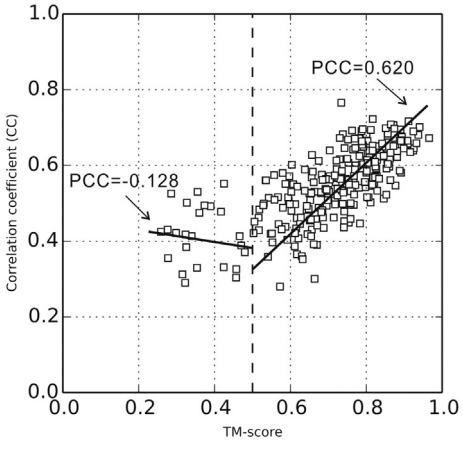

Figure 4. The correlation coefficients (CC) to the EM density map *versus* TM-score of the initial models for the 278 test proteins. The dashed-dotted line divides the models into two regions with TM-score below and above 0.5, where the solid lines represent the fitting results by linear regression for the samples in the two regions. The Pearson correlation coefficients between CC and TM-score are −0.128 and 0.620, respectively, in the two regions.

G1050C (PDBID: 5NV4). This is a small domain with 69 residues and the resolution of the simulated density map was 5.0 Å. For this target, I-TASSER built an initial structure with an incorrect fold (TM-score = 0.37), and EM-Refiner pushed the model toward the correct fold with a TM-score of 0.53; this suggests that the inherent knowledge-based force field of EM-Refiner is probably capable of refining the structural model for some small proteins, when coupled with cryo-EM data, although the procedure failed in most other cases.

In Figure 4, we present the *CC versus* TM-score for the initial models of the 278 test proteins following superposition. The figure shows that *CC* has a largely golf course-like landscape, where a strong Pearson correlation coefficient (PCC) of 0.620 between CC and TM-score occurs when the decoys have a model close to the target structure (typically with a TM-score above 0.5). When the decoys are further away from the native structure, the correlation becomes negligible, corresponding to a PCC of −0.128 in the region where the TM-score is <0.5. A similar tendency can be seen in Figure S4 when we list all the decoy conformations from the EM-Refiner Monte Carlo simulations, in which PCC vanishes in the regions with TM-scores below 0.5. Therefore, when the initial models have incorrect folds, as shown in Figure 3, the EM-Refiner simulation cannot be efficiently guided by EM density map data to identify correct conformations. This also explain the poor performance of the initial model-density map superposition procedure for the incorrectly predicted I-TASSER models, since the procedure is purely driven by the *CC* score, which cannot correctly guide the search toward the correct superposition. Development of efficient approaches, such as those that use long-range, deep-learning-based contact and distance maps [29,30], may help draw the structures into the correlation zone.

Control results compared to Flex-EM and Rosetta

As a control, we present in Table 1 and Figure 3 the results from two other commonly used density map-based refinement methods, Rosetta [13] and Flex-EM [12]. For Rosetta, the protein structure PDB file, cryo-EM density map file, density map resolution option, refinement protocol file and other parameters were set up following the tutorial instructions, and the default values were used for all other parameters. For Flex-EM, the voxel size, origin of the density map and rigid body file were also provided, along with the above input files, with all the parameters set to their default values following the Flex-EM tutorial. The data of voxel size and origin of the density maps were obtained from the head file of each density map, with the rigid body files identified by RIBFIND [31]. The Hierarchical flexible fitting was implemented for Flex-EM to

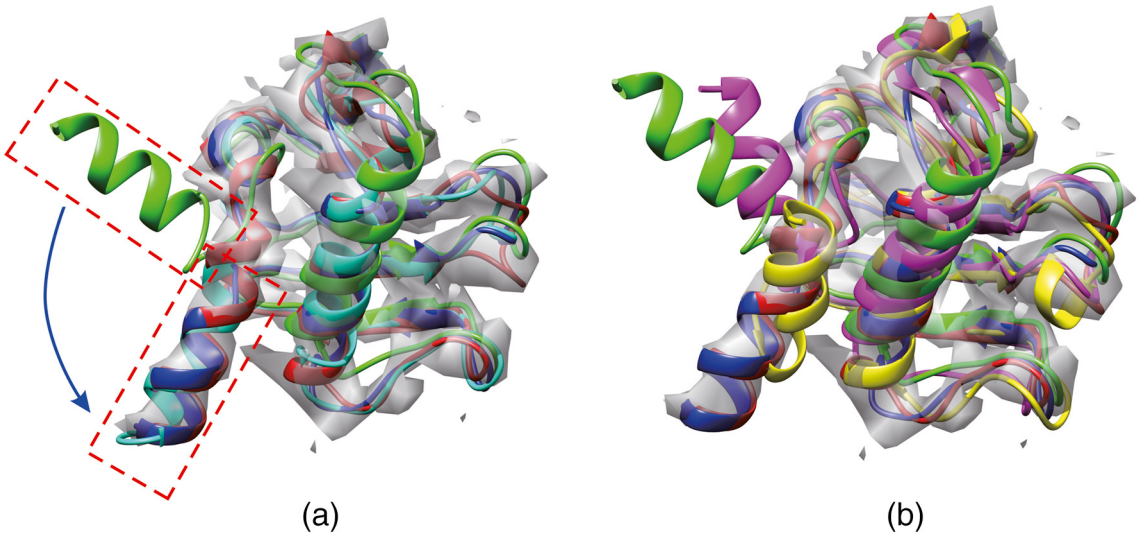

Figure 5. Illustrative example of refinement using simulated experimental data at a resolution of 5 Å for a single domain of 2gzaC. (a) Cartoons show the initial model from I-TASSER (green), the EM-Refiner models after Step-2 (cyan) and Step-3 (blue), and the native structure (red) overlaid on the density map (gray). (b) Cartoons show the superposed model from I-TASSER (green), Flex-EM refined model (magenta), Rosetta refined model (yellow), the EM-Refiner model (blue), and the native model (red) overlaid on the density map (gray). The EM-Refiner model had a TM-score of 0.921, an RMSD of 1.35 Å to the native structure and a CC of 0.811 with the density map, compared to a TM-score of 0.743, an RMSD of 5.25 Å and a CC of 0.641 for the Flex-EM model and a TM-score of 0.800, an RMSD of 4.52 Å and a CC of 0.733 for the Rosetta refined model.

reduce the overfitting. Since both Flex-EM and ROSETTA do not automatically generate initial model and model superpositions, we used the same I-TASSER models as the starting point and applied the same superposed models from the EM-Refiner Step-1 program (see Methods), to ensure a fair comparison. Similarly, the density maps also came from the same simulated noise-free maps produced by the EMAN2 program, with resolutions ranging from 5 to 10 Å as listed in Table S3 of the SI.

On average, although both of the control methods were able to achieve some level of success at refining the initial I-TASSER models, Rosetta significantly outperformed Flex-EM, while the magnitude of the improvement by EM-Refiner was larger than both programs. If we directly compare EM-Refiner and the control methods, the average TM-scores were 0.779, 0.761 and 0.714 for EM-Refiner, ROSETTA and Flex-EM, respectively, where the *p*-values between EM-Refiner and the control methods were all below 0.05, showing the differences were statistically significant (Table 1).

To examine the robustness of EM-Refiner, we also tested it on simulated, noisy density maps, where Gaussian noises were introduced, with a standard deviation of 0.01 and an average of 1, to the simulated density maps of the 278 test proteins using Xmipp [32,33]. Compared to the noise free density maps, the model quality of EM-Refiner with the noisy density maps is slightly reduced, with the average TM-score and RMSD being 0.776 and 4.76 Å, respectively. Nevertheless, the EM-Refiner models compare favorably with the control programs that have TM-score and RMSD of 0.760 and 4.77 Å for Rosetta, and 0.712 and 5.43 Å for Flex-EM, respectively. The differences were statistically significant with *p*-values of 3.1×10^{-3} and 8.3×10^{-46} for EM-Refiner *versus* Rosetta and Flex-EM, respectively, as calculated by two-tailed Student's *t*-tests between the TM-scores. These results suggest that EM-Refiner could outperform Rosetta and Flex-EM with noisy density maps.

Figure 5 shows an illustrative example of a single domain of the VirB11 ATPase from the *Brucella suis* type IV secretion system in complex with sulfate (PDBID: 2gzaC), where the resolution of its simulated density map was 5 Å. For this target, I-TASSER built an initial structure with a correct fold (TM-score = 0.734), and EM-Refiner put the model in a marginally correct position by rigid-body density map superposition with a *CC* of 0.567 between the superposed model and density map. After the Step-2 fragment adjustment simulations, EM-Refiner created a significantly refined model with a TM-score of 0.900 and an RMSD of 1.43 Å, where the CC increased to 0.794. The Step-3 atomic-level refinement simulations further increased the TM-score to 0.921 and reduced the RMSD to 1.35 Å, where the CC increased to 0.811 accordingly. As shown in Figure 5(a), the major improvements by EM-Refiner following Step-2 were in the regions with poor local CC scores, where the fragment adjustment procedure identified and correctly adjusted

these regions into the density map. After Step-3, the conformation was further improved in terms of TM-score, RMSD and CC, demonstrating the usefulness of the atomic-level simulations for detailed local structure refinement. For this same target, the Flex-EM model had a TM-score of 0.743, a *CC* of 0.641 and an RMSD of 5.25 Å, while the Rosetta model had a TM-score of 0.800, a *CC* of 0.733 and an RMSD of 4.52 Å (Figure 5(b)). This example highlights the importance and effectiveness of the hierarchical three-step refinement procedure utilized by EM-Refiner, which helps achieve a more significant model improvement based on the EM-density data over the control methods.

Impact of initial model superposition on final model quality

Current programs cannot always identify the best structural superposition onto the EM density maps. To examine the impact of initial model superposition on the final model quality, we list in Table 1 (right column) the structural refinement results when each program started from the optimal superstition obtained by the TM-score program, which directly superposed the I-TASSER models onto the target structures. The results showed that better initial model superposition can indeed result in improved refinement quality. Quantitatively, the average EM-Refiner TM-score increased from 0.779 to 0.796, which corresponds to a *p*-value of 3.2E-10. Furthermore, the quality of the models produced by Rosetta and Flex-EM also improved with TM-score increases from 0.761 and 0.714 to 0.780 and 0.718, respectively. Nevertheless, the TM-score of EM-Refiner was still significantly higher than Rosetta and Flex-EM with *p*-values of 3.7E-5 and 1.8E-41, respectively.

In Figure S5, we provide a detailed comparison between EM-Refiner models and the initial I-TASSER models, where EM-Refiner started from the optimal superpositions by the TM-score program. Interestingly, nearly all of the EM-Refiner models (except for 3 out of the 278 cases) had better TM-scores than the corresponding initial I-TASSER models, including those cases where the initial model had an incorrect fold (TM-score < 0.5). It is worth noting that for the initial models with TM-scores <0.5, the TM-score-based superposition did not have a perfect overlap with the density map. Given the approximately correct orientation of the initial conformations, however, EM-Refiner could still draw the conformations closer to native with the aid of the physics-based force field, although the *CC* term did not have a strong correlation with the model quality in this region as shown in Figure 4. Accordingly, there were a higher number of cases for which EM-Refiner outperformed the two control methods, compared to the models starting from predicted superpositions. This was particularly true for the cases with initial I-TASSER models with TM-scores <0.5, showing the efficiency of EM-Refiner at refining low-to-medium resolution models after correct model superposition.

Case studies on refining atomic structure using experimental cryo-EM density maps

While the above tests were based on simulated map data for the target structures, here, we tested EM-Refiner on two examples using experimental density maps. One example is from GroEL (EMD-2221) [34] in the EMDB, for which both its fitted and target structures have been released; the second is from the Ash2L domain of a recently solved human MLL1 core complex [35], for which EM-Refiner was applied for cryo-EM map-guided structure refinement. To obtain the initial structures that were used as input to EM-Refiner, we first extracted the sequences from the native structures. Next, we predicted the atomic structures starting from the sequences using I-TASSER, excluding homologous templates with ≥30% sequence identity to the query. For the experimental density map of Ash2L, which does not have a released structure in the PDB database, we estimated the accuracy using the correlation coefficient between the predicted structure and the density map.

Refining a single domain protein from GroEL

GroEL is a protein of high interest in the field of biosensors for gut microbiota sensing and has been utilized in sublingual vaccination against various disorders including atherosclerosis, diarrhea and colitis [36]. The density map of GroEL with an 8.4 Å resolution was taken from the EMDB using code EMD-2221 [34]. Next, a single domain density map of GroEL was segmented from EMD-2221 using UCSF Chimera [14]. The corresponding native structure was obtained from the PDB (the residue numbers were 190 to 338 from chain D of the PDB ID: 2ynj).

From the query sequence, I-TASSER created a model with a TM-score of 0.709 to the native structure. The EM-Refiner Step-1 program superposed the model to the GroEL density map and achieved a cRMSD of 4.04 Å and a *CC* of 0.701 (Figure 6). After EM-Refiner refinement, the TM-score of the final model increased to 0.771, where the RMSD and *CC* improved to 3.19 Å and 0.835, respectively. We also ran Rosetta starting with the same initial superposition, which resulted in a refined model with a lower TM-score of 0.729. Additionally, compared to EM-Refiner, both the RMSD (3.34 Å) and *CC* (0.824) of the model produced by Rosetta were slightly worse. Given that the overall correlation between the model and density map for EM-Refiner and Rosetta was largely equivalent at the resolution

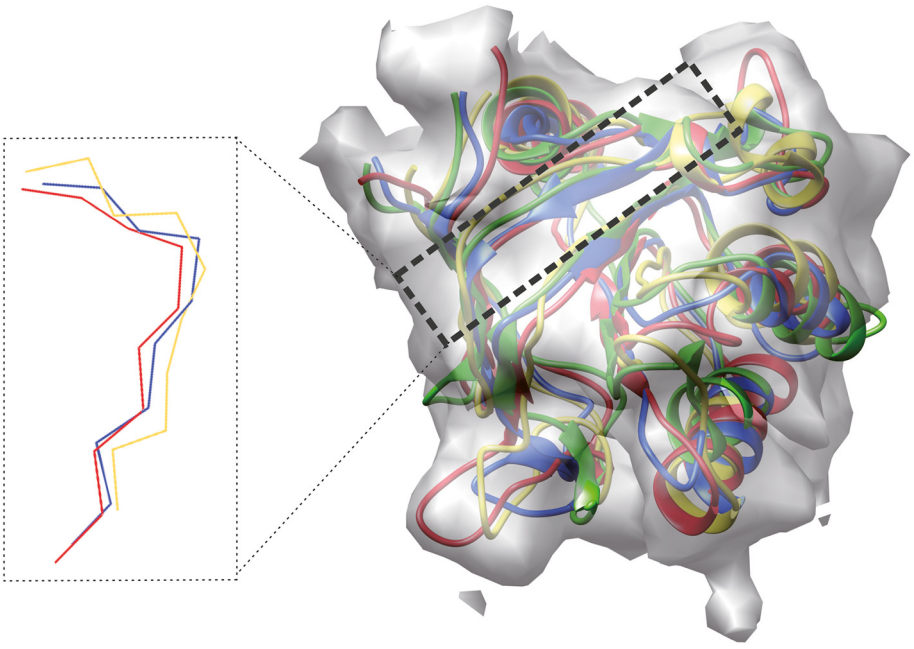

Figure 6. Refinement of a single domain of the GroEL protein based on an 8.4 Å experimental EM density map. The refined model by EM-Refiner (blue) is overlaid onto the density map (gray) with a correlation coefficient (CC) of 0.835 and a TM-score of 0.771 to the native structure (red). As a comparison, the initial I-TASSER model (green) and the Rosetta refined model (yellow) had *CC*s of 0.701 and 0.824 and TM-scores of 0.709 and 0.729, respectively.

of the density map, the difference in model quality was most likely not a result of EM-Refiner fitting of the model with the density map. Rather, the increased model quality by EM-Refiner can be probably attributed to the effectiveness of the inherent knowledge-based force field combined with the CC score and the REMC simulation strategy, which work harmoniously to perform the atomic-level structural refinement. To further examine the local structure of the models, we plot in Figure S7 the residue-level distance of the model to the native. After TM-score superposition, the EM-Refiner model has 97 residues whose Cα atom is closer to the native than Rosetta while the Rosetta model does so in 52 residues. In addition, the EM-Refiner and Rosetta models have 82 and 47 residues, respectively, with a distance below 2 Å to the native. These data show that EM-Refiner has also a better local structure than Rosetta in this example.

In Table S2, we provide a summary of the modeling results by I-TASSER and the follow-up EM density map refinement methods on an additional set of 5 proteins randomly selected from the EMDB that have experimentally determined density maps. For these five proteins, all the methods (except for Flex-EM on EMD6708_dom2) achieved some level of refinements over the I-TASSER models. The average TM-score and RMSD of EM-refiner was 0.815 and 7.78 Å, respectively, compared to 0.800 and 8.55 Å for Rosetta and 0.711 and 8.63 Å for Flex-EM, respectively. The differences were statistically significant with *p*-values of 0.05 and 0.01 for EM-Refiner *versus* Rosetta and Flex-EM, respectively, as calculated by two-tailed Student's *t*-tests between the TM-scores. These results show that EM-Refiner outperforms Rosetta and Flex-EM at structure refinement using experimental cryo-EM density maps as well.

Refining Ash2L

Ash2L is a component of the Set1/Ash2 histone methyltransferase complex, where human mixed lineage leukemia (MLL) histone methyltransferases utilize the Ash2L domain to anchor themselves to the nucleosome core particle (NCP) [37]. The cryo-EM structure of the human MLL1 and NCP complex with a 6.2 Å resolution was most recently determined with the aid of a former version of EM-Refiner by Park *et al.* [35]. Although the interactions between the interface residues (residues 205–207) in Ash2L and the nucleosomal DNA were validated by Park *et al.*, the atomistic Ash2L structure remains to be solved without a high-resolution density map. In the study, we applied I-TASSER to build the initial model for the human Ash2L domain, and then segmented the density map from MLL1 using Chimera. Next, the I-TASSER model of human Ash2L was superposed and refined by EM-Refiner using the segmented density map. Finally, the Ash2L model was fit, together with other components of the human

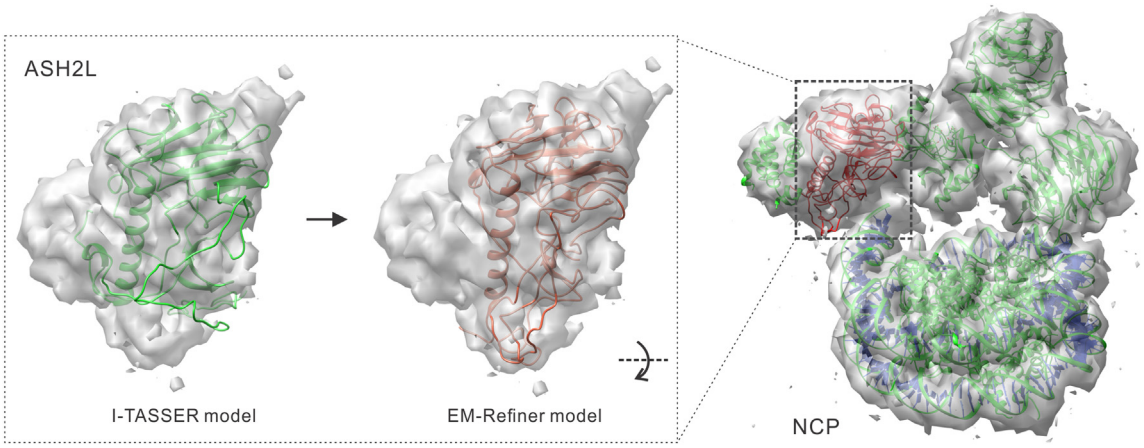

Figure 7. The fitting and refinement results for the Ash2L model into the cryo-EM density map of the human MLL1 and nucleosome core particle (NCP) with 6.2 Å resolution. The final model of Ash2L by EM-Refiner (red color) has a correlation coefficient (CC) of 0.696 with the density map, compared to the initial models from I-TASSER (green) with a *CC* of 0.517 (left panel).

COMPASS complex taken from Park *et al.*, into the density maps to obtain the final atomic models after removing clashes using UCSF Chimera.

Figure 7 shows the results of refinement and the fitted model of the MLL1-NCP complex. The final EM-Refiner model of Ash2L had a *CC* score of 0.696 with the experimental density map, which is slightly higher than that of the I-TASSER predicted model (0.517) and the refined model produced by Rosetta (0.674). Apparently, in the absence of other experimental structure information, it is difficult to objectively assess if the EM-Refiner model is closer to the native than the I-TASSER and Rosetta models, despite the slightly higher CC score. However, this case helps to illustrate how EM-Refiner can be used to predict 3D atomic protein structures in order to address the important, real-world biological problems, using medium- or lower-resolution density maps.

Conclusion

Atomic-level protein structure refinement using low-to-medium resolution cryo-EM density maps has become an increasingly important means of obtaining high-resolution structures. Many of the current approaches address this challenge either by performing rigid-body structure fitting, or through flexible structure refinement simulations that require user-specified initial model creation and density-map superposition. Despite the success of some of these methods, their usefulness and efficiency are limited by the separation of the fitting and refinement processes. In this work, we developed a new protocol to integrate structural fitting and structural refinement into a uniform pipeline that allows for automatic prediction and refinement of protein structures starting from amino acid sequences. The test results on a large set of 285 (=278 + 7) protein domains showed the significant advantages of the protocol for atomic-level structure refinement compared to many of the state-of-the-art methods. The EM-Refiner program is freely available for download and can be applied to large-scale, automated EM-based structure prediction and refinement.

Despite the encouraging results, the EM-Refiner pipeline is still more effective at refining structural models with correct folds (typically with TM-scores $\geq$0.5) than those with incorrect folds. This constraint stems essentially from the absence of correlation between the density map/predicted model correlation score and TM-score of the predicted models when their TM-scores to native are below 0.5 (Figures 4 and S3). This absence of long-range correlation significantly limits the accuracy of initial model superposition, since the procedure is purely driven by the *CC* between the model and density map. Second, although our benchmark tests showed that EM-Refiner is capable of refining low TM-score models when starting from optimal model superpositions, the absence of correlation also limits the flexible structure refinement simulations because the force field is strongly weighted by the *CC* score in addition to physics- and knowledge-based energy terms. In this regard, the recent efforts utilizing deep-learning-based contact and distance maps in the I-TASSER pipeline [30,38] will be important to provide improved initial models for non-homologous protein targets. Meanwhile, developments of advanced methods to improve the initial model and density map superposition should also be key to help improve the success rate of low-to-medium resolution EM-based structure prediction and refinement.

Currently, the EM-Refiner was trained and tested on single domain proteins. For multi-domain protein modeling, the user needs to utilize a segmentation program, such as "Segment Map" in UCSF chimera [14], to obtain density maps for each domain. EM-Refiner can then be used for automated domain-level structure fitting and refinements. Following the structural modeling of individual domains, the user can use DEMO [39], which was designed to assemble multi-domain protein structures from individual domains under the guidance of the inter-domain distance profiles automatically searched from homologous quaternary templates in the PDB. Nevertheless, an optimal solution to cryo-EM-guided multi-domain structural construction might be through the integration of original density map data with flexible domain refinement and inter-domain assembly simulations; the work along this line is under progress.

Methods

EM-Refiner consists of three steps of structure-to-density map superposition, rigid-body fragment adjustments, and atomic-level structure refinement, where the pipeline is depicted in Figure 1.

Step 1: structure-density map superposition

Starting from an initial predicted structure produced by I-TASSER and a cryo-EM density map (in MRC format), EM-Refiner first superposes the model onto the EM density map by maximizing their correlation coefficient. This procedure is performed through a rapid REMC simulation [40], which is guided by the energy $E = 1 - CC$, where CC is the correlation coefficient between the model and density map as defined in Eq. (3) below. The REMC simulation contains $N=20$ replicas with the temperature of the ith replica equal to

$$T_i = T_{\min}\left(\frac{T_{\max}}{T_{\min}}\right)^{\frac{i-1}{N-1}} \quad (1)$$

where $T_{\min} = 0.01$ and $T_{\max} = 1.0$ are the minimum and maximum temperatures respectively. To ensure diversity, each replica starts from a separate initial superposition pose, with movements composed of rigid-body rotation and translation of the initial model.

Step 2: rigid-body fragment adjustment

The conformation with the highest CC to the density map is selected from the REMC simulations for the next step of flexible structure refinement. Due to the errors in the I-TASSER models, the superposed conformations often have parts of fragments sticking outside their density maps. Since residue-level movements can break down the local secondary structures and reduce the efficiency of refinement, we employ a coarse-grained, fragment-level adjustment to rapidly shift the fragments into the density map. This process is performed by a simulated annealing Monte Carlo (SAMC) simulation [41]. Here, a fragment is defined as a region with regular secondary structure (α-helix or β-strand) and the movements, including rigid-body fragment rotation and translation, use the two end loops as the hinge points. For the fragments outside the density map, the length is expanded to include all residues that have a CC below 0.05, in order to enhance the simulation and movement efficiency. To calculate the CC for individual residues, the summation in Eq. (3) only includes the grids that have a distance below 8 Å. The SAMC simulation uses $30 \times 20 \times 500$ steps (i.e., a target has 30 annealing repeats each with 20 temperatures that range from 2 to 0.01 following Eq. (1), and 500 movements are attempted at each temperature). The full-atomic energy as defined in Eq. (2) below is employed to guide the SAMC simulation.

Step 3: atomic-level refinement

Five low-energy decoys are selected from Step-2 for atomic-level structure refinement, which is also performed using SAMC simulations. Here, 50 annealing repeats are performed, and each annealing repeat involves 30 temperatures ranging from 1 to 0.001 as calculated by Eq. (1) with 500 MC movements attempted at each temperature. Five runs are conducted, each starting from a different conformation from Step-2. The model is finally selected from the conformation with the lowest energy, which is submitted to FG-MD [28] for steric clash removal and atomic structure construction. We also tried model selection based on clustering [42], but it produced slightly worse results. The atomic-level refinement includes six movements (M1–M6) as illustrated by the upper-right panel of Figure 1, i.e., M1 rotates a randomly selected fragment around a random axis centered at the N or C terminus of the fragment; M2 shifts the conformation of four residues ($i + 1$, $i + 2$, $i + 3$, and $i + 4$) along the sequence by one residue; M3 is similar to M1 but with the rotation center located at the middle of the selected fragment; M4 rotates the conformation vertically with the rotation axis oriented from the N- to C-terminal region of the fragment; M5 translates a selected fragment a short distance along a random direction; finally, M6 randomly perturbs the coordinate of one residue of a fragment followed by local structure recovery. In case a movement results in residue discontinuity (normally at the ends of the moved fragments), a random walk is performed to connect the residues.

EM-Refiner force field

The energy function used to guide steps 2 and 3 of the EM-Refiner program is a linear combination of six energy terms:

$$E_{\text{main}} = w_1 \mathrm{E}_{\text{cc}} + w_2 \mathrm{E}_{\text{cl}} + w_3 \mathrm{E}_{\text{hb}} + w_4 \mathrm{E}_{\text{bl}} + w_5 \mathrm{E}_{\text{tor}} + w_6 \mathrm{E}_{\text{ref}} \quad (2)$$

where $E_{\text{cc}} = 1 - \text{CC}$ and CC is the correlation coefficient between the calculated and experimental density maps, i.e.

$$\text{CC} = \frac{\sum_g [\rho_c(g) - \overline{\rho_c}][\rho_e(g) - \overline{\rho_e}]}{\sqrt{\sum_g [\rho_c(g) - \overline{\rho_c}]^2} \sqrt{\sum_g [\rho_e(g) - \overline{\rho_e}]^2}} \quad (3)$$

Here, $\rho_e(g)$ is the experimental density map at the g-th grid point; $\rho_c(g)$ is the density map calculated from the decoy conformation by

$$\rho_c(g) = \sum_i^N \alpha \exp\left(-\beta \left| x_i - x_g \right|^2\right) \quad (4)$$

where x_i and x_g are the coordinates of the i-th atom and g-th grid point, respectively. $\beta = [\pi/(2.4 + 0.8R_0)]^2$ and $\alpha = m(\beta/\pi)^{1.5}$, where m is the atom mass and R_0 is the resolution of the density map, following Topf *et al.* [12].

The second term in Eq. (2) is the Lennard–Jones potential, which accounts for the van der Waals interactions:

$$\mathrm{E}_{\text{cl}} = \sum_{i>j} \varepsilon_{ij} \left[\left(\frac{r_{ij}}{d_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}}{d_{ij}} \right)^{6} \right] \quad (5)$$

where d_{ij} is the distance between atoms i and j; ε_{ij} and r_{ij} are the well depth and the sum of the two atoms' radii, respectively, which are taken from CHARMM19 [43].

E_{hb} is the main-chain hydrogen bonding energy adopted from QUARK [44]. Here, four geometric features are considered for a residue pair (i and j, Figure S6), including the distance between O_i and H_j (d_{OH}), the inner angle of $\mathrm{C}_i - \mathrm{O}_i - \mathrm{H}_j$ (A_{COH}), the inner angle of $\mathrm{O}_i - \mathrm{H}_j - \mathrm{N}_j$ (A_{OHN}), and the torsion angle of $\mathrm{C}_i - \mathrm{O}_i - \mathrm{H}_j - \mathrm{N}_j$ (T_{COHN}) (the torsion angle term is only used for helical residues). The H-bond energy is calculated by as follows

$$E_{\text{hb}} = \sum_{i>j} \sum_{k=1}^{4} \frac{\left[F_k(i,j) - \overline{F_k} \right]^2}{2\sigma_k^2} \quad (6)$$

where $F_k(i,j)$ denotes the kth feature for the candidate residue pair (i, j). $\overline{F_k}$ and σ_k are the mean and standard deviation of the kth feature, respectively, which were calculated from 3881 high-resolution X-ray PDB structures with their values listed in Table S1.

E_{bl} is a bond-length potential calculated by

$$E_{\text{bl}} = \sum_{i>j} \kappa_{i,i+1} \left(d_{i,i+1} - d_0 \right)^2 \quad (7)$$

where $d_{i,\,i+1}$ is the distance between neighboring heavy atoms connected by a covalent bond. The parameters for statistical bond-length d_0 and the force constant κ were taken from CHARMM [43].

E_{tr} is a main-chain dihedral torsion potential calculated by

$$E_{\text{tr}} = -\sum_{i=2}^{L-1} \log(P(\varphi_i, \psi_i | A(i), S(i))) \quad (8)$$

where φ_i/ψ_i are the Ramachandran torsion-angle pair for residue i, and $P(\varphi, \psi \,|\, A, S)$ is the conditional probability of φ and ψ given the residue type A and the secondary structure type S, which was taken from the QUARK force field [44].

Finally, E_{res} is designed to constrain the structural differences between the decoys and the initial model by

$$E_{\text{res}} = \sum_{i,j \in G, j-i>1} \left| d_{ij}^d - d_{ij}^i \right| \quad (9)$$

where G represents all residue pairs that have Cα–Cα distances below 8.0 Å in the initial model; d_{ij}^d and d_{ij}^i are the Cα–Cα distances in the decoy and initial model, respectively.

The weight parameters are set to $w_1 = 520$, $w_2 = 1.0$, $w_3 = 0.8$, $w_4 = 1.0$, $w_5 = 0.6$, $w_6 = 10$, which were tuned using a training set to balance the contributions of the different energy terms in Eq. (1).

Acknowledgments

We thank Chengxin Zhang, Wenyi Zhang and Wei Zheng for stimulating discussions. This work was supported in part by the National Institute of General Medical Sciences (GM083107, GM136422, to Y.Z.), the National Institute of Allergy and Infectious Diseases (AI134678 to Y.Z.), the National Science Foundation (IIS1901191, DBI2030790 to Y. Z.), the National Natural Science Foundation of China (Nos. 61725302, 61671288 to H.S.), and the Science and Technology Commission of Shanghai Municipality (No. 17JC1403500 to H.S.). This work was performed when B.Z. visited the University of Michigan.

Author's Contributions

H.S. and Y.Z. designed the research. B.Z. performed the research. X.Z. participated in discussions and analyzed the data. R.P. participated in

discussions and proofread the manuscript. B.Z. and Y.Z. wrote the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2020.07.027.

Received 12 April 2020;
Received in revised form 14 July 2020;
Accepted 31 July 2020
Available online 6 August 2020

Keywords
Cryo-EM;
protein structure prediction;
structure refinement;
structure and density-map fitting;
replica-exchange Monte Carlo simulation

Abbreviations used:
cryo-EM, cryogenic electron microscopy; EMDB, Electron Microscopy Data Bank; REMC, replica-exchange Monte Carlo; PCC, Pearson correlation coefficient; NCP, nucleosome core particle; SAMC, simulated annealing Monte Carlo.

References

1. Henderson, R., Baldwin, J.M., Ceska, T.A., Zemlin, F., Beckmann, E., Downing, K.H., (1990). Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J. Mol. Biol.*, **213**, 899–929, https://doi.org/10.1016/S0022-2836(05)80271-2.
2. Frank, J., (2006). Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State. 2nd edn Oxford University Press, 2006.
3. Li, X., Mooney, P., Zheng, S., Booth, C.R., Braunfeld, M.B., Gubbens, S., et al., (2013). Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat. Methods*, **10**, 584–590, https://doi.org/10.1038/nmeth.2472.
4. Patwardhan, A., (2017). Trends in the Electron Microscopy Data Bank (EMDB). *Acta Crystallogr. D Struct. Biol*, **73**, 503–508, https://doi.org/10.1107/S2059798317004181.
5. Vinothkumar, K.R., Henderson, R., (2016). Single particle electron cryomicroscopy: trends, issues and future perspective. *Q. Rev. Biophys.*, **49**, e13https://doi.org/10.1017/S0033583516000068.
6. Cheng, Y., (2015). Single-particle cryo-EM at crystallographic resolution. *Cell*, **161**, 450–457, https://doi.org/10.1016/j.cell.2015.03.049.
7. Malhotra, S., Trager, S., Dal Peraro, M., Topf, M., (2019). Modelling structures in cryo-EM maps. *Curr. Opin. Struct. Biol.*, **58**, 105–114, https://doi.org/10.1016/j.sbi.2019.05.024.
8. Bahar, I., Rader, A., (2005). Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struct. Biol.*, **15**, 586–592.
9. Tama, F., Miyashita, O., Brooks III III, C.L., (2004). Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM. *J. Struct. Biol.*, **147**, 315–326.
10. Chacón, P., Wriggers, W., (2002). Multi-resolution contour-based fitting of macromolecular structures. *J. Mol. Biol.*, **317**, 375–384.
11. Trabuco, L.G., Villa, E., Mitra, K., Frank, J., Schulten, K.J.S., (2008). Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure*, **16**, 673–683.
12. Topf, M., Lasker, K., Webb, B., Wolfson, H., Chiu, W., Sali, A., (2008). Protein structure fitting and refinement guided by cryo-EM density. *Structure*, **16**, 295–307.
13. DiMaio, F., Tyka, M.D., Baker, M.L., Chiu, W., Baker, D.J., (2009). Refinement of protein structures into low-resolution density maps using rosetta. *J. Mol. Biol.*, **392**, 181–190.
14. Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., et al., (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.
15. Garzón, J.I., Kovacs, J., Abagyan, R., Chacón, P., (2007). ADP_EM: fast exhaustive multi-resolution docking for high-throughput coverage. *Bioinformatics*, **23**, 427–433.
16. Rossmann, M.G., Bernal, R., Pletnev, S.V., (2001). Combining electron microscopic with X-ray crystallographic structures. *J. Struct. Biol.*, **136**, 190–200.
17. Nicholls, R.A., Tykac, M., Kovalevskiy, O., Murshudov, G.N., (2018). Current approaches for the fitting and refinement of atomic models into cryo-EM maps using CCP-EM. *Acta Crystallogr. D Struct. Biol*, **74**, 492–505.
18. De la Rosa-Trevín, J., Quintana, A., Del Cano, L., Zaldívar, A., Foche, I., Gutiérrez, J., et al., (2016). Scipion: a software framework toward integration, reproducibility and validation in 3D electron microscopy. *J. Struct. Biol.*, **195**, 93–99.
19. Kovacs, J.A., Galkin, V.E., Wriggers, W., (2018). Accurate flexible refinement of atomic models against medium-resolution cryo-EM maps using damped dynamics. *BMC Struct. Biol.*, **18**, 12.
20. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
21. Tang, G., Peng, L., Baldwin, P.R., Mann, D.S., Jiang, W., Rees, I., et al., (2007). EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.*, **157**, 38–46, https://doi.org/10.1016/j.jsb.2006.05.009.
22. Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., Zhang, Y., (2015). The I-TASSER Suite: protein structure and function prediction. *Nat. Methods*, **12**, 7–8.
23. Zhang, C., Mortuza, S.M., He, B., Wang, Y., Zhang, Y., (2018). Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12. *Proteins*, **86**, (Suppl. 1) 136–151, https://doi.org/10.1002/prot.25414.
24. Zhang, Y., Skolnick, J., (2004). Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.
25. Xu, J., Zhang, Y., (2010). How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, **26**, 889–895.
26. Baker, M.L., Jiang, W., Wedemeyer, W.J., Rixon, F.J., Baker, D., Chiu, W., (2006). Ab initio modeling of the herpesvirus VP26 core domain assessed by CryoEM density. *PLoS Comput. Biol*, **2**,.
27. Topf, M., Baker, M.L., John, B., Chiu, W., Sali, A., (2005). Structural characterization of components of protein assemblies by comparative modeling and electron cryo-microscopy. *J. Struct. Biol.*, **149**, 191–203.

28. Zhang, J., Liang, Y., Zhang, Y., (2011). Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. *Structure*, **19**, 1784–1795, https://doi.org/10.1016/j.str.2011.09.022.
29. Li, Y., Hu, J., Zhang, C., Yu, D.J., Zhang, Y., (2019). ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. *Bioinformatics*, https://doi.org/10.1093/bioinformatics/btz291.
30. Zheng, W., Li, Y., Zhang, C., Pearce, R., Mortuza, S.M., Zhang, Y. , (2019). Deep-learning contact-map guided protein structure prediction in CASP13. *Protein s*, https://doi.org/10.1002/prot.25792.
31. Pandurangan, A.P., Topf, M., (2012). Finding rigid bodies in protein structures: application to flexible fitting into cryoEM maps. *J. Struct. Biol.*, **177**, 520–531.
32. Sorzano, C., Marabini, R., Velázquez-Muriel, J., Bilbao-Castro, J.R., Scheres, S.H., Carazo, J.M., et al., (2004). XMIPP: a new generation of an open-source image processing package for electron microscopy. *J. Struct. Biol.*, **148**, 194–204.
33. Lindert, S., Alexander, N., Wötzel, N., Karakaş, M., Stewart, P.L., Meiler, J., (2012). EM-fold: de novo atomic-detail protein structure determination from medium-resolution density maps. *Structure*, **20**, 464–478.
34. Bartesaghi, A., Lecumberry, F., Sapiro, G., Subramaniam, S., (2012). Protein secondary structure determination by constrained single-particle cryo-electron tomography. *Structure*, **20**, 2003–2013, https://doi.org/10.1016/j.str.2012.10.016.
35. Park, S.H., Ayoub, A., Lee, Y.T., Xu, J., Kim, H., Zheng, W., et al., (2019). Cryo-EM structure of the human MLL1 core complex bound to the nucleosome. *Nat. Commun.*, **10**, 5540, https://doi.org/10.1038/s41467-019-13550-2.
36. Pechkova, E., Bragazzi, N.L., Nicolini, C., (2014). *Advances in Protein Chemistry and Structural Biology* Vol. 95. Elsevier, 2014 163–191.
37. Yokoyama, A., Wang, Z., Wysocka, J., Sanyal, M., Aufiero, D.J., Kitabayashi, I., et al., (2004). Leukemia proto-oncoprotein MLL forms a SET1-like histone methyltransferase complex with menin to regulate Hox gene expression. *Mol. Cell. Biol.*, **24**, 5639–5649, https://doi.org/10.1128/MCB.24.13.5639-5649.2004.
38. Li, Y., Zhang, C., Bell, E.W., Yu, D.J., Zhang, Y., (2019). Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13. *Proteins*, https://doi.org/10.1002/prot.25798.
39. Zhou, X., Hu, J., Zhang, C., Zhang, G., Zhang, Y., (2019). Assembling multidomain protein structures through analogous global structural alignments. *Proc. Natl. Acad. Sci.*, **116**, 15930–15938.
40. Swendsen, R.H., Wang, J.S., (1986). Replica Monte Carlo simulation of spin glasses. *Phys. Rev. Lett.*, **57**, 2607–2609.
41. Kirkpatrick, S., Gelatt, J., Vecchi, M.P., (1983). Optimization by simulated annealing. *Science*, **220**, 671–680.
42. Zhang, Y., Skolnick, J., (2004). SPICKER: a clustering approach to identify near-native protein folds. *J. Comput. Chem.*, **25**, 865–871.
43. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., Karplus, M., (1983). CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, **4**, 187–217.
44. Xu, D., Zhang, Y., (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins*, **80**, 1715–1735, https://doi.org/10.1002/prot.24065.