JCTC Journal of Chemical Theory and Computation

Letter

pubs.acs.org/JCTC

Coupling an Ensemble of Homologues Improves Refinement of Protein Homology Models

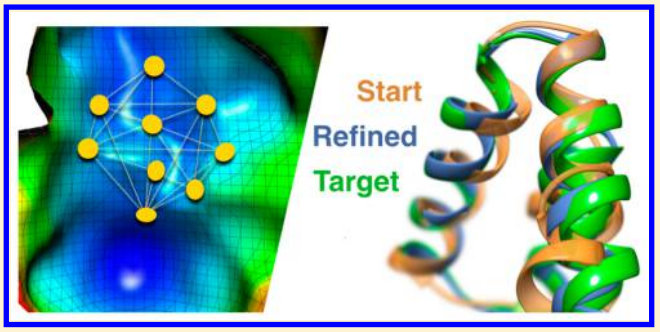
André Wildberg,[†,§] Dennis Della Corte,[†] and Gunnar F. Schröder*[,†,‡]

[†]ICS-6: Structural Biochemistry, Institute of Complex Systems, Forschungszentrum Jülich, 52425 Jülich, Germany

[‡]Physics Department, Heinrich-Heine Universität Düsseldorf, 40225 Düsseldorf, Germany

Supporting Information

ABSTRACT: Atomic models of proteins built by homology modeling or from low-resolution experimental data may contain considerable local errors. The refinement success of molecular dynamics simulations is usually limited by both force field accuracy and by the substantial width of the conformational distribution at physiological temperatures. We propose a method to overcome both these problems by coupling homologous replicas during a molecular dynamics simulation, which narrows the conformational distribution, and smoothens and even improves the energy landscape by adding evolutionary information.



The interpretation of genomics data in terms of protein structure is an important postgenomic challenge. Building atomic models for individual amino acid sequences becomes increasingly important in understanding the molecular effects of genetic variation. Homology modeling is a useful tool to build atomic models if the structure of a homologous protein is known. However, due to the limitation of current methodology such models of protein structures may contain considerable errors. Similarly, atomic models built with low-resolution (e.g., from X-ray diffraction or cryo-EM) or sparse experimental data might contain comparable errors.

Refinement approaches have the goal of correcting these errors in atomic protein models. The types of errors we consider as being amenable to refinement include disrupted hydrogen bond networks, small shifts of secondary structure elements, incorrect side-chain packing and rotamers, and wrong loop conformations. Correcting such errors is typically challenging since the energy differences between alternative, slightly different conformations are rather small. The critical assessment of structure prediction (CASP) experiment has a refinement category to test the performance of refinement methods.[1,2]

Regular molecular dynamics (MD) simulations are generally unable to refine homology models and do not consistently yield a structure that is moved closer to the native structure (as usually determined by high-resolution X-ray crystallography).[3] Even though regular unrestrained MD simulations can sample closer-to-native structures, reliably selecting these structures is not possible.[4]

The main causes for this limitation of MD simulations are (1) force field inaccuracies,[5] (2) high energy barriers that need to be crossed, and (3) the fact that the Boltzmann distribution, which is approximated by the MD simulation, is broad at physiological temperatures. Simulation at physiological temperatures is however necessary for proper contribution of the entropy; only then is the free energy of conformational states correctly described.

Position restraints have been used successfully to prevent the simulation from exploring the broad Boltzmann distribution; these restraints force the structure to sample a region around the starting model, which also leads to sampling closer-to-native structures with higher probability.[5−7] Position restraints however also set an upper limit to the extent of the conformational change, which might hinder sufficient sampling and refinement.

The goal of structure refinement is to determine the most probable conformation, which corresponds to the free energy minimum, rather than the conformational distribution. It has been shown that the most probable conformation can be approximated by averaging the structures from an MD ensemble more robustly than by selecting a single structure with a scoring function.[6,7] However, for the averaging to yield a good structure requires the simulation to predominantly sample near native structures.

We here present in two steps a modified MD protocol that addresses all three problems of regular MD simulations mentioned previously (force field inaccuracies, high energy barriers, and broad Boltzmann distribution).

In the first modification step, we simulate simultaneously eight identical replicas of the starting structure. These replicas are subjected to the same harmonic position restraints (on Cα-atoms), which forces them to remain similar to each other. The positions of the restraints are constantly updated during the simulation and slowly follow the motion of the center of mass of all replicas. These adaptive restraints were inspired by

Received: August 24, 2015
Published: October 28, 2015

DOI: 10.1021/acs.jctc.5b00942
J. Chem. Theory Comput. 2015, 11, 5578−5582

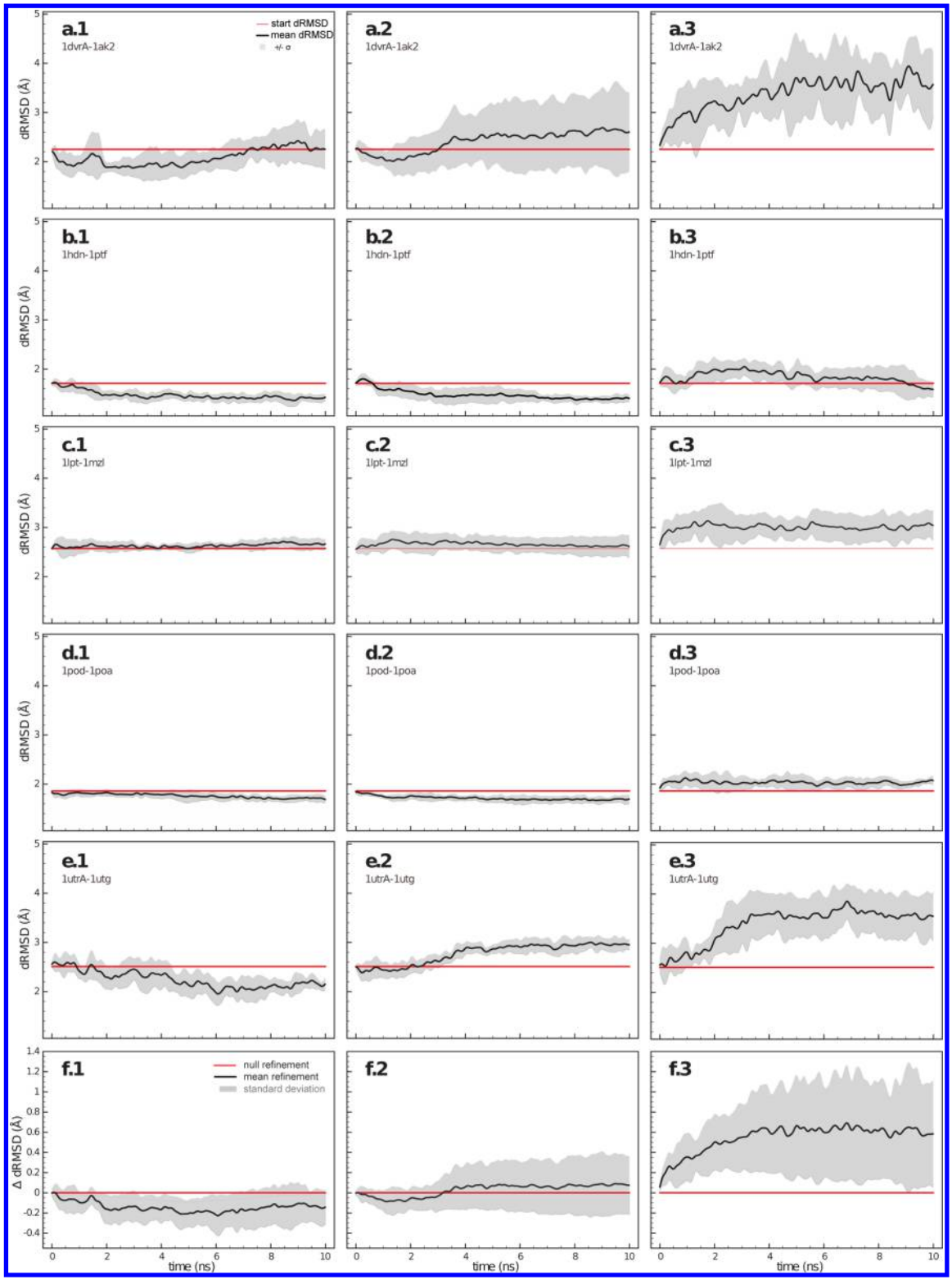


Figure 1. Comparison of different refinement protocols (1−3). Five test models (a−e) have been simulated 5 times with three MD protocols: replica restraints with coupled homologous sequences (1, left column), replica restraints with coupled identical sequences (2, middle column), and free MD simulations of a single protein (3, right column). dRMSD values were averaged over the five independent trajectories (black lines). The standard deviation is shown in gray. dRMSD values below the null refinement line (red line) indicate improved frames. The bottom row (f) shows the average and standard deviations over the five test cases for each refinement protocol.

deformable elastic network restraints (DEN), which have been shown to guide structure refinement against X-ray diffraction and cryo-EM data.[8−10] Since the restraints are adaptive, the coupled replicas are allowed to undergo any conformational motion as long as they stay close together. The number of replicas has been chosen here simply to limit the computational expense, and it is not yet clear how the results depend on the number of replicas.

DOI: 10.1021/acs.jctc.5b00942
J. Chem. Theory Comput. 2015, 11, 5578−5582

The harmonic restraints restrict larger motions more than smaller motions, which leads to a time-scale-dependent diffusion coefficient (cf. Supporting Information Figure S1). For small time scales the size of the diffusion coefficient is comparable to that of free MD simulations, which enables individual replicas to cross local energy barriers. In addition, entropic contributions of solvent and side chains (which are not restrained) are not strongly affected, which means that in particular the solvation free energy is mostly unperturbed. For longer time scales the diffusion coefficient decreases significantly, which reduces large conformational fluctuations. Smaller fluctuations mean that the system of coupled replicas is less likely to drift in random directions and will sample low free energy states more frequently than a free MD simulation. Furthermore, the coupling of replicas has an effect of smoothening the energy landscape, similar to particle swarm optimization, which has been applied to MD simulations before.[11] The motion of the center of mass is the result of an effective force averaged over all replicas. Because the replicas are in different positions on the energy landscape the center of mass moves on a locally averaged, i.e., smoothened, energy landscape (see Methods).

In the second modification step, the target sequence in seven of the replicas is replaced with homologous sequences. This is motivated by the observation that structure is much more conserved than sequence which causes homologous proteins to fold into similar structures.[12] This fact can be exploited by coupling homologous proteins (with pairwise sequence identity of at least 50%) instead of identical replicas during a MD simulation. Keasar et al.[13−15] have proposed that such a coupling of homologous proteins with slightly different energy landscapes results in an energy landscape that is smoothened not only in structure space but also in sequence space. The methodology was implemented in the GROMACS 4.5.3[16] software (see Methods).

Our method has been tested successfully in the recent CASP11 experiment and was among the best performing methods.[17] Of all 37 refinement targets 65% could be improved. For the improved models the average GDT-HA score was increased by 6.6. For those models that could not be improved the average GDT-HA score decreased by 3.9. To understand why and how our approach works, we studied here the refinement of five representative homology models in more detail. To avoid any bias, we selected five models from a publicly available set of decoys that was not generated by us (the Badretdinov decoy set;[18] see Supporting Information Table S1).

Each model was simulated 5 times for 10 ns each. We compared three different simulation protocols: (1) a free MD simulation of the homology model, (2) coupled identical replicas, and (3) coupled homologous replicas.

Figure 1 shows for each test case average and standard deviations of distance root-mean-square (dRMSD) values from five independent simulations. The dRMSD is used to measure the deviation of two atomic models. It is calculated as the root-mean-square deviation of corresponding pairs of Cα-atom distances in two structures. All possible pairs of Cα-atoms are considered.

Simulations with coupled replicas (with both homologous (Figure 1, first column) and identical replicas (Figure 1, second column)) show clear improvement over regular free MD simulations (Figure 1, third column). In free MD simulations the structure drifted away from the correct structure in all cases except for 1hdn-1ptf, where a small number of frames was improved. In simulations with identical replicas 51% of the frames were improved on average, but the improvements are small fluctuations around the null refinement (horizontal red line). In contrast, simulations with homologous replicas consistently improve the structure and sample conformations closer to the native structure most of the time.

The improvement of the structures from the different refinement protocols is quantified by the change in dRMSD of the average structure from each trajectory, compared to that of the starting structure (see Figure 2). Free MD simulations led in all five cases to the lowest structural quality. Comparison of the average improvements (Figure 2, dashed lines) shows an offset between free MD simulation and coupling of identical replicas, which can be attributed to the particle swarm optimization effect. More importantly another offset can be seen between coupling with identical and with homologous replicas, which represents the improvement that is due to the added evolutionary information and which we interpret as an improvement of the force field.

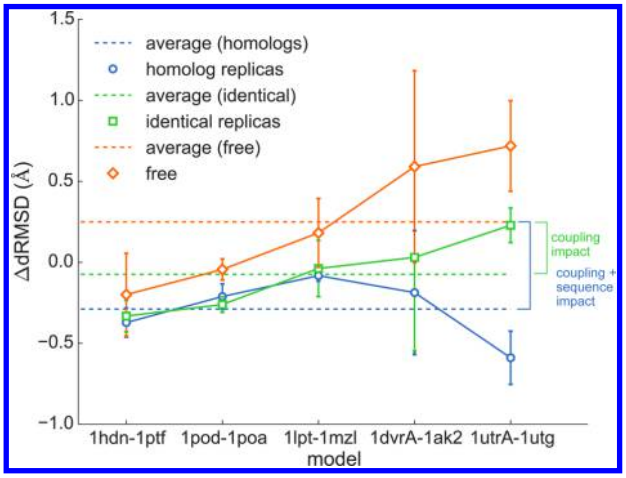


Figure 2. Average dRMSD for each test case and each refinement protocol is plotted as well as the average over the five test cases (dotted lines). Coupling identical sequences leads to a clear improvement over free MD simulations. Coupling of homologous replicas leads to a consistent refinement of all 5 test cases. Interestingly, the main improvement of using homologous versus identical replicas is observed for the two test cases (1dvrA-1ak2 and1utrA-1utg), for which coupling of identical replicas was not successful.

Only for 1lpt-1mzl, which has the lowest starting quality of 3.8 Å RMSD, the average dRMSD was not improved (see Figure 1c1); however, the dRMSD of the average structure was slightly improved (Figure 2), clearly showing the benefit of structural averaging.[6]

Figure 3 compares the result of the simulation with coupled homologous replicas (green) with the starting model (purple) and the native target structure (blue). Several secondary structure elements and loop regions are shifted toward the correct structure. In contrast to free MD simulations, the coupled-replica simulations yielded very consistent results: the average structures from five independent simulations are very similar, as visualized in Supporting Information Figure S2 by multidimensional scaling.[19]

We found that refinement with coupled homologous replicas outperforms regular MD simulations in all test cases. The additional evolutionary information and the reduction in global fluctuations through coupling of homologous replicas leads to consistently sampling structures closer to their native state

DOI: 10.1021/acs.jctc.5b00942
J. Chem. Theory Comput. 2015, 11, 5578−5582

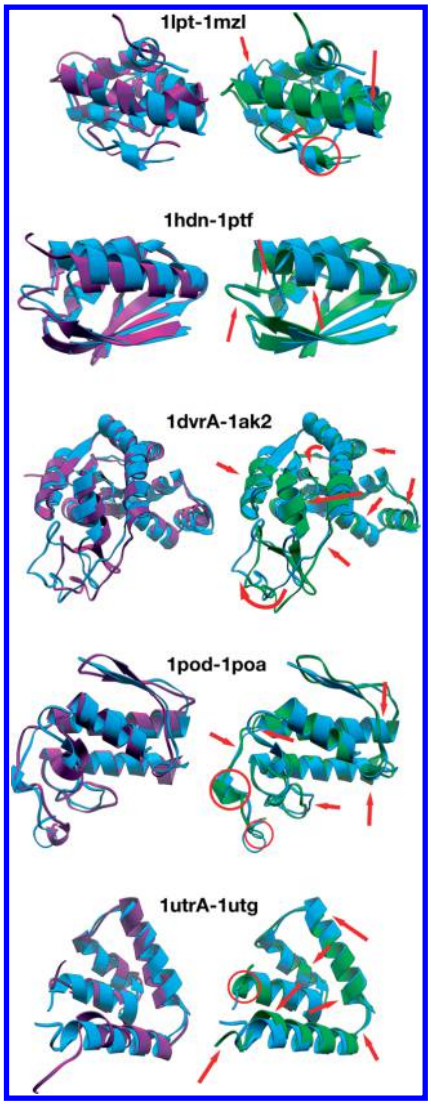


Figure 3. For each test case, the starting structure (purple) is shown together with the best model from the simulations with coupled homologous replicas (green) and the native structure (blue). Refined regions are highlighted by red arrows. Figures were made with Chimera.[23]

compared with free MD simulations. This insight will help to develop even more powerful refinement methods based on MD.

METHODS

Implementation of Replica Coupling in GROMACS 4.5.3. For the replica-coupled simulations, the simulation box was composed of eight replicas, which are positioned at the edges of a cube. The distance between the replicas needs to be large enough to avoid electrostatic interactions between the replicas.

The replicas are coupled through adaptive position restraints on all Cα-atoms. GROMACS does not by default support dynamic updates of position restraints during a simulation. We implemented the necessary changes into the source code of Gromacs 4.5.3[16] to enable updates without reducing the speed of GROMACS. We implemented the changes only for domain decomposition runs (in source code file *domdec.c*). For each Cα-atom i in each replica j a position restraint $\mathbf{p}_{i,j}$ is defined on its initial position. The total energy is then the sum of the energy of the eight uncoupled (solvated) replicas, and the time-dependent energy term for the position restraints, E_{posre}, is given by

$$E_{\text{posre}}(t) = w \sum_{i=1}^{N} \sum_{j=1}^{M} (\mathbf{x}_{i,j}(t) - \mathbf{p}_{i,j}(t))^2 \quad (1)$$

with the coordinates $\mathbf{x}_{i,j}$ of Cα-atom i in replica j, the number of atoms N, and the number of replicas M which we chose to be 8. The force constant w was set to 100 kJ/(mol nm^2). After a period, n, of 500 steps the position restraints are updated according to

$$\mathbf{p}_{i,j}(t + n\Delta t) = \mathbf{p}_{i,j}(t) + \kappa\langle \mathbf{x}_{i,j}(t) - \mathbf{p}_{i,j}(t)_j \rangle \quad (2)$$

with the integration time step Δt of 2 fs. The relaxation rate, κ, at which the position restraints follow the average coordinate displacement was set to 0.5. The same displacement vector $\langle \mathbf{x}_{i,j}(t) - \boldsymbol{p}_{i,j}(t) \rangle_j$, which is an average over the corresponding displacements in all replicas j, is added to all replicas, which leads to a coupling of the replicas. These adaptive restraints were inspired by deformable elastic network (DEN) restraints, which yield a similar effect for a γ-value of 1. The original DEN method employs a network of (also long) distance restraints, which cannot efficiently be parallelized with domain decomposition. We therefore decided to use adaptive position restraints.

For identical replicas the assignment of corresponding atoms in different replicas is trivial. However, in the case of homologous replicas, a multisequence alignment is performed to assign each Cα-atom from the starting sequence to the corresponding Cα-atoms in the homologues. If there are no gaps or insertions, the assignment is again trivial. If the alignment shows a gap for k sequences at a certain amino acid position, then position restraints are applied and averaged only for the remaining $(8 - k)$ residues that are present at this position, which means that the displacement vector will be averaged over $(8 - k)$ replicas. Insertions will not generate extra position restraints; those residues instead are kept unrestrained and are free to move. The total number of position restraints is therefore always identical to the number of Cα-atoms in the target sequence.

Method Availability. The modified GROMACS version with adaptive position restraints to couple multiple replicas is available from the SimTK Web site: http://simtk.org/home/adpt-gromacs.

MD Protocols. All simulations used the AMBER99SB-ILDN force field with TIP3P explicit water with an integration time step of 2 fs. Temperature was kept constant at 300 K by the Nosé–Hoover algorithm. Electrostatic long-range interactions were calculated with PME, and bond lengths were constrained by the P-LINCS approach. Na^+/Cl^- ions were added at physiological concentration. Before and after adding the solvent molecules, the structure was energy minimized to remove any sterical clashes that may be the result of homology model building.

Each simulation was repeated 5 times with duration of 10 ns. For comparison we performed three different simulation protocols: (1) free MD simulation of a single protein structure, (2) replica-coupled simulation with identical sequences, and (3) replica-coupled simulations with homologous sequences.

The computational expense for the replica-coupled simulations is much larger than for the single MD simulations, since the simulation system is eight times larger. The total amount of simulation time equals a single protein simulation in solvent for 4.25 μs.

Sequence Selection Strategy. To build the homologous replicas, seven homologous sequences were searched via BLAST[20] on the RefSeq[21] database. Sequences were manually

DOI: 10.1021/acs.jctc.5b00942
J. Chem. Theory Comput. 2015, 11, 5578−5582

selected that fulfilled two criteria: (1) their sequence identities with the target sequence needs to be between 50 and 80%, and (2) the sequence identities between all pairs of the eight sequences should ideally also be in the range 50−80%. However, for some test cases the second criterion could not be strictly fulfilled. The sequences chosen have an average sequence identity of 61.8% to the target structure and are shown in Supporting Information Table S2. The homology models used as the replicas were generated with MODELLERv9.[22]

Test Case Selection Strategy. Homology models from the Badretdinov decoy set[18] were chosen as test cases. We aimed to cover a wide range of protein properties, such as size, secondary structure composition, and shape. The five homology models that were selected represent starting qualities between 2 and 4 Å RMSD to the solved crystal structure. The sequence lengths vary between 70 and 220 amino acids. The details of the selected models are shown in Supporting Information Table S1. The naming scheme of the models is xxxxX-yyyy, where xxxx and yyyy are the PDB IDs of the template and the target, respectively, and X is the chain ID of the target.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.5b00942.

Mean square displacement and multidimensional scaling plots for different simulation protocols, and properties of test targets and homologous replicas (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: gu.schroeder@fz-juelich.de.

Present Address

§Department of Chemistry and Biochemistry, University of California, 9500 Gilman Dr., La Jolla, CA 92093, USA.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We gratefully acknowledge the computing time granted on the supercomputer JUROPA at Jülich Supercomputing Centre (JSC).

■ REFERENCES

(1) MacCallum, J. L.; Pérez, A.; Schnieders, M. J.; Hua, L.; Jacobson, M. P.; Dill, K. A. Assessment of Protein Structure Refinement in CASP9. *Proteins: Struct., Funct., Genet.* **2011**, *79*, 74−90.

(2) Nugent, T.; Cozzetto, D.; Jones, D. T. Evaluation of Predictions in the CASP10 Model Refinement Category. *Proteins: Struct., Funct., Genet.* **2014**, *82*, 98−111.

(3) Fan, H.; Mark, A. E. Refinement of Homology-Based Protein Structures by Molecular Dynamics Simulation Techniques. *Protein Sci.* **2004**, *13*, 211−220.

(4) Zhu, J.; Fan, H.; Periole, X.; Honig, B.; Mark, A. E. Refining Homology Models by Combining Replica-exchange Molecular Dynamics and Statistical Potentials. *Proteins: Struct., Funct., Genet.* **2008**, *72*, 1171−1188.

(5) Raval, A.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Refinement of Protein Structure Homology Models via Long, All-Atom Molecular Dynamics Simulations. *Proteins: Struct., Funct., Genet.* **2012**, *80*, 2071−2079.

(6) Mirjalili, V.; Feig, M. Protein Structure Refinement through Structure Selection and Averaging from Molecular Dynamics Ensembles. *J. Chem. Theory Comput.* **2013**, *9*, 1294−1303.

(7) Mirjalili, V.; Noyes, K.; Feig, M. Physics-based Protein Structure Refinement through Multiple Molecular Dynamics Trajectories and Structure Averaging. *Proteins: Struct., Funct., Genet.* **2014**, *82*, 196−207.

(8) Schröder, G. F.; Brunger, A. T.; Levitt, M. Combining Efficient Conformational Sampling with a Deformable Elastic Network Model Facilitates Structure Refinement at Low Resolution. *Structure* **2007**, *15*, 1630−1641.

(9) Schröder, G. F.; Levitt, M.; Brunger, A. T. Super-resolution Biomolecular Crystallography with Low-resolution Data. *Nature* **2010**, *464*, 1218−1222.

(10) Schröder, G. F.; Levitt, M.; Brunger, A. T. Deformable Elastic Network Refinement for Low-Resolution Macromolecular Crystallography. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2014**, *70*, 2241−2255.

(11) Huber, T.; van Gunsteren, W. F. SWARM-MD: Searching Conformational Space by Cooperative Molecular Dynamics. *J. Phys. Chem. A* **1998**, *102*, 5937−5943.

(12) Chothia, C.; Lesk, A. M. The Relation between the Divergence of Sequence and Structure in Proteins. *EMBO J.* **1986**, *5*, 823−826.

(13) Keasar, C.; Elber, R.; Skolnick, J. Simultaneous and Coupled Energy Optimization of Homologous Proteins: A New Tool for Structure Prediction. *Folding Des.* **1997**, *2*, 247−259.

(14) Keasar, C.; Tobi, D.; Elber, R.; Skolnick, J. Coupling the Folding of Homologous Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 5880−5883.

(15) Keasar, C.; Elber, R. Homology as a Tool in Optimization Problems: Structure Determination of 2D Heteropolymers. *J. Phys. Chem.* **1995**, *99*, 11550−11556.

(16) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: A High-throughput and Highly Parallel Open Source Molecular Simulation Toolkit. *Bioinformatics* **2013**, *29* (7), 845−854.

(17) *CASP11 Refinement Results*, 2014; http://predictioncenter.org/casp11/zscores_final_refine.cgi (accessed May 16, 2015).

(18) Badretdinov, A. (1997). *Sali Lab Decoy Sets for Model Assessment*, ftp://salilab.org/pub/azat/models-a/ (accessed Jan. 21, 2006).

(19) *MDSJ: Java Library for Multidimensional Scaling*, Version 0.2; Algorithmics Group, University of Konstanz, Germany, 2009. http://www.inf.uni-konstanz.de/algo/software/mdsj/ (accessed Jun. 12, 2015).

(20) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25* (17), 3389−3402.

(21) Pruitt, K. D.; Tatusova, T.; Maglott, D. R. NCBI Reference Sequences (RefSeq): A Curated Non-redundant Sequence Database of Genomes, Transcripts and Proteins. *Nucleic Acids Res.* **2007**, *35*, D61−D65.

(22) Fiser, A.; Sali, A. Modeller: Generation and Refinement of Homology-based Protein Structure Models. *Methods Enzymol.* **2003**, *374*, 461−491.

(23) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera — A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.

DOI: 10.1021/acs.jctc.5b00942
J. Chem. Theory Comput. 2015, 11, 5578−5582