JCTC Journal of Chemical Theory and Computation

Article

Cite This: J. Chem. Theory Comput. 2018, 14, 991−997

pubs.acs.org/JCTC

Fast Calculation of Protein−Protein Binding Free Energies Using Umbrella Sampling with a Coarse-Grained Model

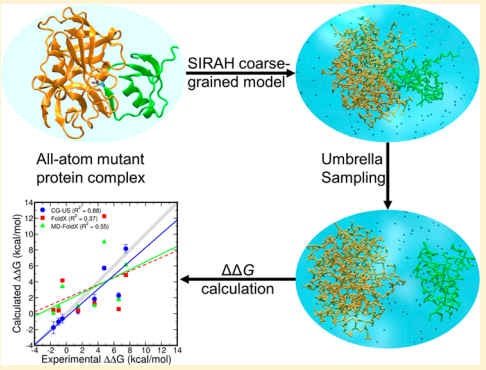
Jagdish Suresh Patel*,† and F. Marty Ytreberg*,‡

†Center for Modeling Complex Interactions, University of Idaho, Moscow, Idaho 83844, United States

‡Department of Physics, University of Idaho, Moscow, Idaho 83844, United States

Supporting Information

ABSTRACT: Determination of protein−protein binding affinity values is key to understanding various underlying biological phenomena, such as how missense variations change protein−protein binding. Most existing non-rigorous (fast) and rigorous (slow) methods that rely on all-atom representation of the proteins force the user to choose between speed and accuracy. In an attempt to achieve balance between speed and accuracy, we have combined rigorous umbrella sampling molecular dynamics simulation with a coarse-grained protein model. We predicted the effect of missense variations on binding affinity by selecting three protein−protein systems and comparing results to empirical relative binding affinity values and to non-rigorous modeling approaches. We obtained significant improvement both in our ability to discern stabilizing from destabilizing missense variations and in the correlation between predicted and experimental values compared to non-rigorous approaches. Overall our results suggest that using a rigorous affinity calculation method with coarse-grained protein models could offer fast and reliable predictions of protein−protein binding free energies.



INTRODUCTION

Protein−protein interactions are at the heart of regulation for all biological processes in a cell. Missense variations (or mutations) of the amino acids that make up these proteins play an essential role by introducing diversity into genomes. These missense variations can lead to an altered protein affinity and can result in dysfunction of the protein interaction network.[1] To understand living organisms, it is thus vital to have a comprehensive knowledge of how proteins interact under physiological conditions, that is, to determine their binding affinities and how these affinities can be modified.[2]

Many techniques have been successful in determining the Gibbs free energy change of protein−protein binding due to a missense variation (i.e., relative affinity, $\Delta\Delta G$). Experimental biophysical methods can quantitatively measure $\Delta\Delta G$ values for protein interactions, but these methods are typically costly, laborious, and time-consuming since all mutants must be expressed and purified.[3] Consequently, many researchers have developed and utilized computational methods to predict $\Delta\Delta G$ values. The most promising in terms of accuracy are rigorous methods based on statistical mechanics that use molecular dynamics (MD) simulations and are capable of addressing conformational flexibility and entropic effects; however, these approaches are computationally highly expensive.[4] By contrast, non-rigorous, computationally less expensive, methods have been developed using the static all-atom protein complex structure. Such methods typically involve the following: (i) empirical energy scoring function;[5] (ii) potentials derived using molecular mechanics principles that enumerate the interactions in physically meaningful terms;[6] (iii) statistical potentials based on the likelihood of similar interactions and local conformations occurring in the Protein Data Bank (PDB);[7] (iv) combination of the first three;[1c] and (v) protein−protein docking.[3b,8] Other approaches have also emerged relying on either coarse representation of the protein (use of Cα or Cβ backbone atoms) to derive a simple contact map potential[9] or machine learning on sequence conservation, solvent accessibility, and secondary structure information to predict $\Delta\Delta G$ values.[10] These approaches are fast and show some degree of success, but they do not account for the conformational changes that can be induced due to missense variation that can prevent clashes and allow residues to form more favorable interaction.[1c] Protein−protein docking has been successful in identifying the interface region but struggles to correctly predict $\Delta\Delta G$ values.[3b,8] Some efforts have also been put into addressing flexibility into non-rigorous binding affinity calculations.[6,11] However, such approaches could only model small deviations in the protein complex fearing the loss of computational speed.

A promising approach to increase MD simulation speed is to use coarse-grained (CG) force fields that rely on abstract descriptions of the biomolecular system including the solvent, yet retain essential physicochemical information. Several CG

Received: June 22, 2017
Published: December 29, 2017

DOI: 10.1021/acs.jctc.7b00660
J. Chem. Theory Comput. 2018, 14, 991−997

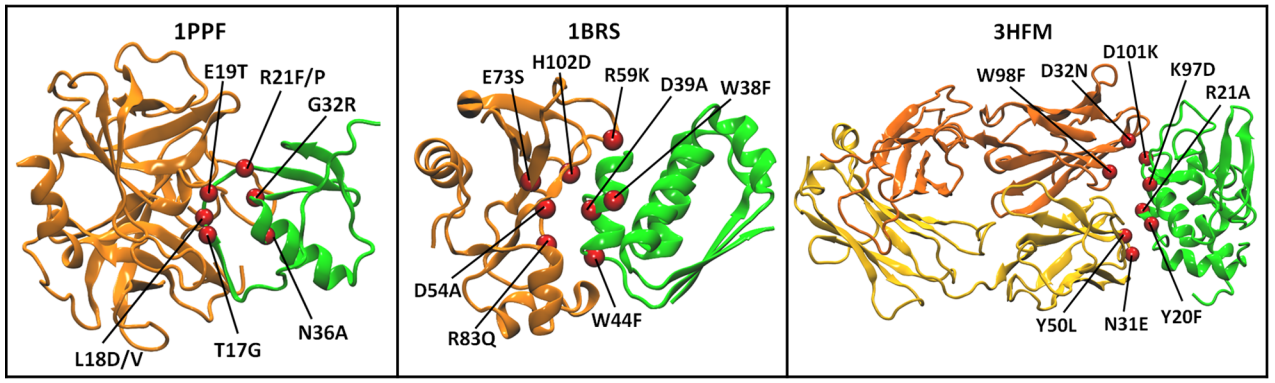


Figure 1. Three-dimensional structures of test protein−protein complexes. System names (PDB IDs) are given above each panel (1PPF, 1BRS, and 3HFM). Each protein pair is colored in orange and green. The red spheres along the interface of the protein complex indicate the sites of the single missense variations chosen for the present study.

models for water and proteins have emerged over the years, each with their strengths and limitations. These models have the potential to significantly increase the speed of a molecular simulation with a cost to biochemical accuracy as compared to atomistic force fields.[2,12] CG models came into existence mainly with the purpose of modeling the structure and the dynamics of the biomolecular systems.[12b] However, in recent years CG simulations have been employed in combination with enhanced sampling/biasing methods to obtain single-dimensional and multidimensional projections of the free energy landscape of association and dissociation processes of biomolecular assembly systems[13] (also see review by Baaden and Marrink[2] on this topic). CG modeling can enable researchers to extend the time scale of the simulation and increase the phase space exploration allowing the study of rare events and large-scale motions of the biomolecules at less computational expense compared to all-atom.[2,12b,14]

In this work, we investigate the performance of a strategy combining a SIRAH CG protein model[15] with rigorous umbrella sampling[16] molecular dynamics to predict the $\Delta\Delta G$ values of single amino acid missense variations. We are interested in predicting the effect of multiple missense variations and thus have developed a semi-automated strategy with default values for input simulation parameters that avoids fine-tuning each parameter to individual complex systems. To investigate whether this strategy has a good trade-off between speed and accuracy, we chose three protein−protein test systems with empirical $\Delta\Delta G$ values for observed missense variations. For each test system we selected eight different missense variations occurring at the different sites with varying empirical $\Delta\Delta G$ values. We calculated $\Delta\Delta G$ values for each missense variation using fast umbrella sampling simulations (i.e., short simulation time with similar input parameters) and compared the results with empirical $\Delta\Delta G$ values and with two non-rigorous approaches.[11a,17] We obtained significant improvement in the correlation between predicted and experimental $\Delta\Delta G$ values compared to faster approaches. Moreover, our strategy predicted the sign of $\Delta\Delta G$ values correctly at a much higher rate compared to the other tested methods. To our knowledge, there is only one study by May et al.[13a] that has previously applied a strategy of using Martini CG models[18] combined with restrained simulations to estimate effect of single missense variations on protein−protein binding affinities. In their study, absolute affinity values were calculated and found to be in reasonable agreement with those from atomistic simulation and correlated well with evolutionary likelihood. However, their predictions lacked experimental validation, and the study did not investigate the performance of this strategy in computing $\Delta\Delta G$ compared to other methods or in predicting the sign of $\Delta\Delta G$. Combining CG models with enhanced sampling techniques is slowly gaining traction for calculating the free energy of various physiological processes. However, previous studies have mainly employed the Martini CG model or a highly coarse Gō-like model and lacked a systematic evaluation of these models in predicting the effect of missense variations.

METHODS

Test Systems. To provide a test of the speed and accuracy of predicting relative binding free energy differences ($\Delta\Delta G$), we selected three different protein−protein complexes (see Figure 1) from the SKEMPI database:[19] (i) Complex between human leukocyte elastase (218 aa) and the third domain of the turkey ovomucoid inhibitor (56 aa) (PDB ID 1PPF);[20] (ii) Barnase (110 aa)−Barstar (89 aa) complex (PDB ID 1BRS);[21] (iii) an antigen−antibody complex of the lysozyme (129 aa)−HY/HEL-10 FAB (429 aa) (PDB ID 3HFM).[22] We chose eight missense variations for each of the three protein complex systems. The choice of these missense variations was driven by several factors: (i) the values for $\Delta\Delta G$ for reported experimental missense variations were varied in sign, important since negative, stabilizing values are often harder to predict than positive, destabilizing values; (ii) there were non-alanine-scanning point missense variations at differing sites; (iii) the structures in the PDB were not missing a large number of residues; (iv) there was a range in the size of the chosen protein complexes; and (v) missense variations were reported on one chain (1PPF), on both chains (1BRS), and on multiple chains (3HFM) (see Figure 1).

Preparation of the Wild-Type and Mutant Complexes. Each test complex was prepared in an identical manner using the following steps: (i) experimental structures were downloaded from the PDB Web site (http://www.rcsb.org/pdb/home/home.do); (ii) the structure files were edited to remove all but the two interacting chains listed in the SKEMPI database;[19] (iii) all missing residues or atoms in the PDB files were added using MODELER v9.15;[23] and (iv) mutant complexes were generated using Dunbrack rotamer library[24] in UCSF Chimera.[25]

DOI: 10.1021/acs.jctc.7b00660
J. Chem. Theory Comput. 2018, 14, 991−997

Coarse-Grained Molecular Dynamics Simulations. All MD simulations were carried out using GROMACS v5.1.2.[26] Biasing potentials necessary to carry out umbrella sampling[16] (US) with restraints were introduced via the PLUMED v2.2 plugin[27] integrated in the GROMACS code.

Coarse-grained simulations were performed using the SIRAH force field[15] (http://www.sirahff.com) for all three systems. SIRAH CG force field aims to address some common limitations of CG force fields such as the use of uniform dielectric constant, lack of long-range interactions, use of topological information to maintain the secondary structure, and implicit or no ionic strength effects, etc.[15] In contrast to the "four heavy atoms to one CG bead" rule used by the popular Martini force field,[18,28] SIRAH CG force field treats the peptide bonds with a relatively high degree of detail, preserving the positions of the nitrogen (N), α-carbon (Cα), and oxygen (O), while side chains are modeled more coarsely. WT4 water model[29] included in SIRAH CG force field is formed by four linked beads, each carrying a partial charge, thus allowing it to generate its own dielectric permittivity. Moreover, the CG electrolytes are capable of mimicking the ionic strength effects and osmotic pressure. This residue-based CG model provides all the interactions within a classical Hamiltonian, which is commonly found in most MD simulation packages.[15]

We followed the protocol reported in Darré et al.[15] for carrying out CG simulations. Coordinate mapping and analysis were performed with SIRAH tools.[30] Prior to mapping to the SIRAH CG model, protonation states were assigned based on the assumption of neutral pH using PDB2PQR server[31] and choosing the AMBER[32] naming scheme as an output. Following the all-atom to CG conversion, the protein complexes were placed in a dodecahedral box of SIRAH WT4 water and given neutral charge by adding Na^+ and Cl^- ions at a concentration of 0.15 mol/L. Thickness of the water layer was kept at 4 nm resulting in the following number of WT4 molecules added to each protein complex: 1PPF, 5472; 1BRS, 5687; 3HFM, 10943. The large box size was chosen to make sure that when the two proteins are at the maximum separation distance, they do not interact with each other via their periodic images. Each system was then minimized using the steepest decent for 10,000 steps. To allow for equilibration of the water around the protein complex, each system was then simulated for 1 ns with the positions of all CG atoms in the complex harmonically restrained. During the restrained simulations, the temperature of the systems was set to 300 K and the pressure to 1 atm using respectively the V-rescale thermostat[33] and Parrinello−Rahman barostat[34] with isotropic pressure coupling. Unrestrained simulations were then carried out for 2 ns. All the simulations used a time step of 20 fs and updated neighbor lists every 10 steps. Electrostatic interactions are calculated using particle mesh Ewald[35] with a direct cutoff of 1.2 nm and a grid spacing of 0.2 nm, and a 1.2 nm cutoff was used for van der Waals (vdW) interactions.

Coarse-Grained-Umbrella Sampling Simulations. To calculate the potential of mean force (PMF) for the wild-type and mutant protein complexes, we chose the widely used umbrella sampling (US) method.[16] Since we were interested in predicting the effects of missense variations on protein−protein binding affinity, we selected interprotein separation (i.e., distance) as the reaction coordinate (RC; i.e., pulling variable). To avoid any distortions of the protein as a consequence of application of external harmonic potential, this distance was defined between the center of mass of all the coarse-grained atoms of both proteins in the complex (see Supporting Information (SI) Figure S1). Suitable spring constants for each complex (1PPF, 500 kJ/mol/nm^2; 1BRS, 2000 kJ/mol/nm^2; 3HFM: 1500 kJ/mol/nm^2) were chosen by test simulations performed on wild-type complexes to be strong enough to separate the two proteins in a short amount of time without affecting the overall structure of the proteins. The maximum pulling length was chosen to be 1.7 nm, which ensured the complete solvation of both the proteins in the complex in their unbound states. (see SI Figure S2) The unbinding pathway was chosen to be a vector joining the center of masses of the two proteins. To prevent the drifting of the systems, a weak harmonic restraint with force constant of 20 kJ/mol/nm^2 was added to all the CG atoms of a largest protein in the case of 1PPF and 1BRS complexes and to the antibody in 3HFM antigen−antibody complex. The RC for the US simulation of each complex was discretized into 35 windows with a spacing of 0.05 nm adopted from May et al.[13a] for each complex, ensuring sufficient overlap of the probability distribution of each window. The simulation length for each window was 8 ns (coarse-grained time scale). The time scale for the US simulations was intentionally kept small to match the time scales of non-rigorous approaches used in this study and also to match the same order of magnitude of CPU hours (CPUh) time used in protein−protein simulations using the Martini coarse-grained model in May et al.[13a] To improve the convergence of the PMF, we used a cylindrical harmonic restraint to prevent interactions between the protein being pulled out of the pocket with the full surface of the other protein. (see SI Figure S1) This cylindrical restraint was implemented using the distance from the center of mass of all the CG atoms of the protein being pulled from an axis between the centers of mass of two groups of atoms of the other. One of these two groups was the same as that used to define the RC, and the atoms in the second group are denoted in the Supporting Information. A harmonic restraint of 500 (kJ/mol)/nm^2 on the center of mass of each unbinding protein was applied when the distance from the axis was 0.3 nm or larger. This cylindrical restraint was applied only to the US windows with pulling length greater than 1.5 nm, i.e., only in the unbound state. The effect of this cylindrical restraint was not factored into the calculation of relative binding free energy differences ($\Delta\Delta G$) since the same restraint was applied to all protein complexes in the same fashion.

Potential of Mean Force and $\Delta\Delta G$ Calculation. In the US method,[16] a biasing potential is used at a certain position along the RC (distance in our case) to enhance the sampling of the regions involved in high potential barriers. The RC was discretized into 35 windows, and a harmonic potential, eq 1, was added to the original potential (unbiased potential) in each window to drive the system from one thermodynamic state (bound) to another (unbound).

$$V_n(s) = \frac{k}{2}(s(q) - s_n)^2 \tag{1}$$

where $s(q)$ is the current RC (distance) and $V_n(s)$ is the biasing potential. Here, s_n is the reference distance for the nth window, k is the spring constant, and q are the microscopic coordinates. We used the weighted histogram analysis method (WHAM)[36] to eliminate the bias from the restrained US simulations and construct PMFs using 100 bins along the RC.

Binding free energy (ΔG) for a protein complex was calculated by taking a difference of free energy in the bound

DOI: 10.1021/acs.jctc.7b00660
J. Chem. Theory Comput. 2018, 14, 991−997

and unbound states. A cutoff distance of 1.5 nm was chosen to differentiate a bound from an unbound state as the protein being pulled had no residual contact beyond this distance (see SI Figure S2):

$$\Delta G = \left(-k_B T \ln \int^{\text{bound}} e^{-\Phi_i/k_B T}\right) - \left(-k_B T \ln \int^{\text{unbound}} e^{-\Phi_i/k_B T}\right) \quad (2)$$

where Φ_i is the PMF associated with the ith bin along the RC. The relative binding free energy difference ($\Delta\Delta G$) for a missense variation is then calculated using

$$\Delta\Delta G = \Delta G_{\text{mutant}} - \Delta G_{\text{wild-type}} \quad (3)$$

The average PMF profiles for each wild-type and mutant complex for all three chosen systems were calculated by averaging the outcomes of four independent trials per complex.

RESULTS AND DISCUSSION

The purpose of our study was to assess the ability of combining the SIRAH CG model and US (CG-US) using short simulation times with similar input parameters to calculate $\Delta\Delta G$ values for missense variations and to compare the results with both experimental $\Delta\Delta G$ values and two semiempirical modeling methods FoldX[5,17] and MD-FoldX.[11a] Our strategy was applied to three different protein complexes: 1PPF, 1BRS, and 3HFM (see Figure 1), for which experimental $\Delta\Delta G$ values for the reported missense variations are available in the literature. The total of eight missense variations for each protein complex were chosen as per our criteria listed in Methods (see Figure 1).

Figure 2 shows the PMF profiles resulting from CG-US simulations for each missense variation for all three protein complexes. Each PMF is an average from four independent simulation trials. The panels indicate our chosen 1.5 nm distance used to distinguish bound versus unbound as seen in eqs 2 and 3. In comparison to wild-type, a stabilizing ($\Delta\Delta G < 0$) missense variation will typically be indicated by a PMF profile with larger barrier, and destabilizing missense variations ($\Delta\Delta G > 0$) will typically have a lower barrier. As the interactions between the proteins diminish in the unbound state, the PMF profiles for most protein complexes approach a plateau. We acknowledge that these profiles have not yet reached full convergence due to the use of a short simulation time per US window (e.g., see 1BRS). Averaged PMF profiles were used to compute the corresponding $\Delta\Delta G$ value for each missense variation. The goal of this work is to compare our results with the experimental data and not with atomistic simulation results, but we note that the PMF profile obtained for wild-type 1BRS complex is similar to what was obtained using an atomistic model.[4c] This is an interesting observation considering the small simulation time per window and the granularity of the SIRAH CG model.

Figure 3 summarizes our ability to estimate $\Delta\Delta G$ values for single missense variations. In the case of 1PPF our strategy clearly outperforms FoldX and MD-FoldX; CG-US shows a high correlation (R^2 = 0.88) with the experimental data, and five out of eight missense variations were predicted with high accuracy of within ±1.0 kcal/mol (see SI Table S1). Although we obtained high correlation, the CG-US strategy led to a large error for the R21P mutant complex. We believe this is due to the fact that predicting missense variation to proline is difficult

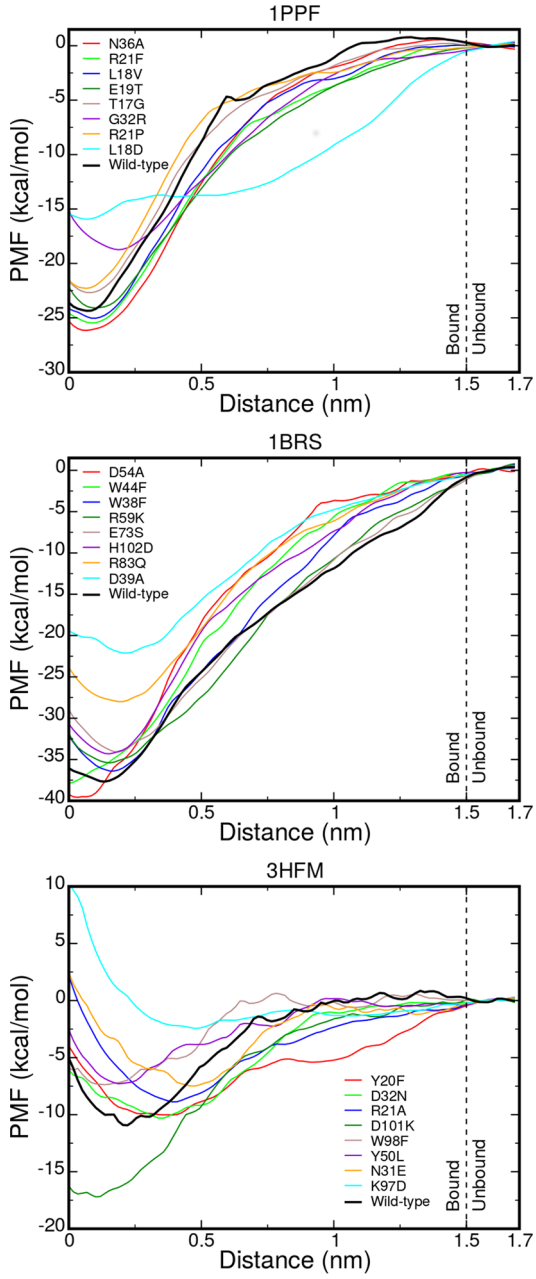


Figure 2. PMF profiles (kcal/mol) for the protein complexes as a function of separation distance (nm). System names (PDB IDs) are given above each panel (1PPF, 1BRS, and 3HFM). PMF profiles for wild-type protein complexes are shown using bold black lines, and mutant protein complex profiles are shown using other colors as denoted in the legends. The dashed black line at the 1.5 nm illustrates the bound from unbound state. Each PMF profile shown above is shifted so that the average PMF value in the unbound state is 0 kcal/mol.

due to their uniquely fused side chain, even for atomistic methods. CG-US performed equally well for 1BRS with R^2 = 0.92, but in this case FoldX and MD-FoldX also have high R^2 values. This is perhaps not surprising since the FoldX energy function was trained on a set of protein complexes that included 1BRS.[5] CG-US was able to estimate four out of eight missense variations with high accuracy (see SI Table S1). However, we observed significant overestimation of $\Delta\Delta G$ values in the cases of R83Q and D39A missense variations. We

DOI: 10.1021/acs.jctc.7b00660
J. Chem. Theory Comput. 2018, 14, 991−997

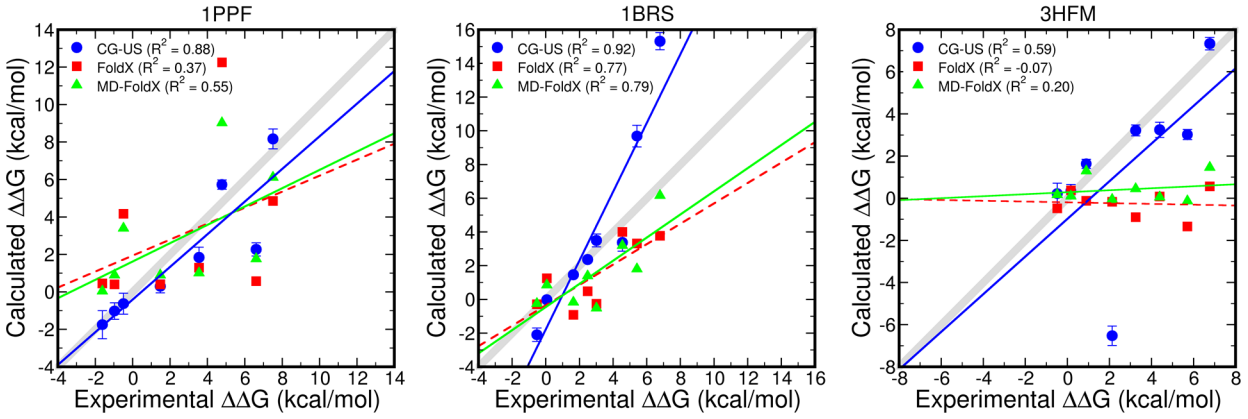


Figure 3. Experimentally observed $\Delta\Delta G$ compared to calculated $\Delta\Delta G$ (kcal/mol) for all three test protein complexes. System names (PDB IDs) are given above each panel (1PPF, 1BRS, and 3HFM). The three methods of $\Delta\Delta G$ are (i) using FoldX on the experimental structure (FoldX); (ii) using FoldX on each of 100 samples taken from a MD simulation and averaging (MD-FoldX); and (iii) using our strategy of combining SIRAH CG with US simulation (CG-US). A perfect fit to experimental data would fall along the gray diagonal line. The solid (green and blue) and dashed (red) lines show the linear relationship between calculated and experimental observation with method of prediction (FoldX, MD-FoldX, or CG-US), and corresponding R^2 values are given in the inset legends. Error bars shown on the CG-US data points represent the standard errors.

believe this is because the spring constant used for all missense variations in the case of 1BRS was tuned to wild-type protein complex and using the same constant for R83Q and D39A complexes led to probability distributions that were not sufficiently overlapped, causing larger errors. For the third and the larger protein complex 3HFM, CG-US significantly outperforms FoldX and MD-FoldX but yields a lower correlation to experiments compared to the other systems with $R^2 = 0.59$. CG-US still predicted five out of eight missense variations with high accuracy despite having a low R^2 value (see SI Table S1). It is important to note that the large error associated with the calculation of the D101 K mutant is significantly lowering the overall R^2 value (see Figure 3). Experimental data suggest this mutant has a positive $\Delta\Delta G$ value but interestingly all the approaches here predict it to have a negative $\Delta\Delta G$. We assume that the experimental data are correct, and this error is likely associated with modeling; however we also note that there are at least two serine residues in the neighborhood of the mutation site that can interact with the positively charged lysine.

It is worth noting that our CG-US strategy consistently outperforms FoldX and MD-FoldX in predicting the signs of the $\Delta\Delta G$ values even if we consider the associated standard errors (see Figure 4). We believe this is an important achievement of the CG-US strategy since correctly predicting the sign of $\Delta\Delta G$ allows discrimination between missense variations that enhance or disrupt binding, e.g., for predicting antibody escape missense variations.

FoldX, as expected, was the fastest among three approaches tested in this work, requiring ~0.42 CPUh to complete a single $\Delta\Delta G$ calculation for 3HFM, the largest among three test protein complexes. MD-FoldX and CG-US approaches for the same consumed ~4093 and ~425 CPUh, respectively (see the SI for more details). It should be emphasized that CG-US is trivially parallelizable in that each US window can run independently without relying on the completion of the previous simulation window; thus, the speed of the calculation depends on the availability of the computational resources.

Our current strategy assumes that conformations in the bound and unbound states do not significantly change due to missense variation. When the protein−protein interaction involves induced fit effects, it is unlikely that our strategy will be directly applicable because of our use of shorter simulation times and mild restraints used to prevent the drifting of stable protein. All-atom MD simulations would be equally unfeasible in this case due to the cost of achieving adequate conformational sampling.

Given that our interest is in calculating relative binding affinities, it should be noted that a more efficient implementation of our approach would be to use alchemical simulation, i.e., using the well-studied single- or dual-topology methods.[37] Such methods have the potential for shorter simulation times and smaller system sizes. However, these methods also require the generation of hybrid structures and topologies, significantly increasing the challenge associated with proper calculation of affinities, and thus will be investigated in future studies.

CONCLUSIONS

In this article, we have described a computational strategy combining the SIRAH coarse-grained (CG) force field with rigorous umbrella sampling (US) simulations using short simulation times with similar input parameters and tested it to predict the effects of single missense variations on protein−protein binding affinity. We have shown that our strategy is capable of delivering more accurate results than two non-rigorous, semiempirical methods. Moreover, it predicted the signs of relative binding free energy ($\Delta\Delta G$) values of the studied missense variations with high accuracy compared to those of non-rigorous approaches, which is remarkable given that the simulation times were intentionally kept short to match the speed of the non-rigorous approaches. With ever-increasing computational power, this strategy has the potential of becoming a routine tool to screen the effect of missense variations. In future work, we will test the generality of these findings by using a larger test set. In addition, we will test the ability of the CG-US strategy in predicting relative affinity changes due to missense variations far from the binding interface, and for multiple missense variations.

DOI: 10.1021/acs.jctc.7b00660
J. Chem. Theory Comput. 2018, 14, 991−997

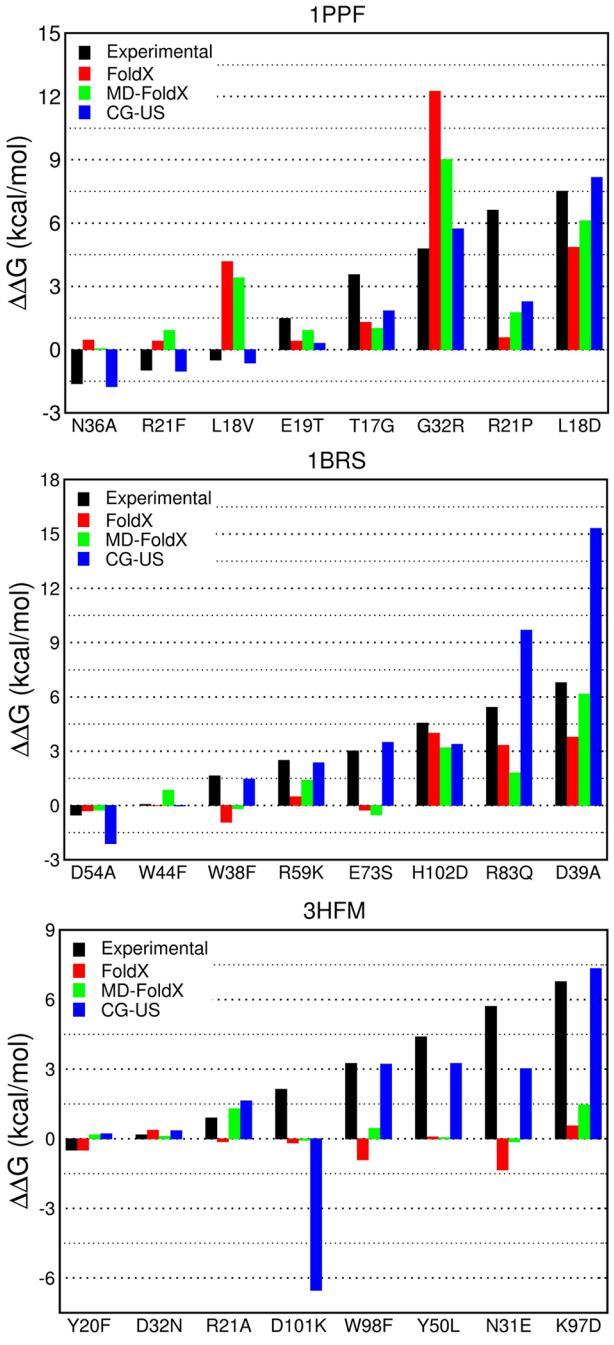


Figure 4. ΔΔ*G* (kcal/mol) for all three test protein complexes. System names (PDB IDs) are given above each panel (1PPF, 1BRS, and 3HFM). Black bars indicate experimentally observed ΔΔ*G*, whereas other colored bars indicated in the inset legends represent results from different methods of prediction (FoldX or MD-FoldX or CG-US).

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.7b00660.

Additional simulation details and tables containing ΔΔ*G* values of all the studied missense variations (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*(J.S.P.) E-mail: thejagdishpatel@gmail.com.

*(F.M.Y.) E-mail: ytreberg@uidaho.edu.

ORCID

Jagdish Suresh Patel: 0000-0003-4999-5347

Funding

This research was supported by the Center for Modeling Complex Interactions sponsored by the NIGMS under Award No. P20 GM104420. Computer resources were provided in part by the Institute for Bioinformatics and Evolutionary Studies Computational Resources Core sponsored by the National Institutes of Health (Grant No. P30 GM103324). This research also made use of the computational resources provided by the high-performance computing center at Idaho National Laboratory, which is supported by the Office of Nuclear Energy of the U.S. Department of Energy (DOE) and the Nuclear Science User Facilities under Contract No. DE-AC07-05ID14517. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) (a) Cargill, M.; Altshuler, D.; Ireland, J.; Sklar, P.; Ardlie, K.; Patil, N.; Lane, C. R.; Lim, E. P.; Kalyanaraman, N.; Nemesh, J.; Ziaugra, L.; Friedland, L.; Rolfe, A.; Warrington, J.; Lipshutz, R.; Daley, G. Q.; Lander, E. S. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **1999**, *22* (3), 231−238. (b) Sachidanandam, R.; Weissman, D.; Schmidt, S. C.; Kakol, J. M.; Stein, L. D.; Marth, G.; Sherry, S.; Mullikin, J. C.; Mortimore, B. J.; Willey, D. L.; Hunt, S. E.; Cole, C. G.; Coggill, P. C.; Rice, C. M.; Ning, Z.; Rogers, J.; Bentley, D. R.; Kwok, P. Y.; Mardis, E. R.; Yeh, R. T.; Schultz, B.; Cook, L.; Davenport, R.; Dante, M.; Fulton, L.; Hillier, L.; Waterston, R. H.; McPherson, J. D.; Gilman, B.; Schaffner, S.; Van Etten, W. J.; Reich, D.; Higgins, J.; Daly, M. J.; Blumenstiel, B.; Baldwin, J.; Stange-Thomann, N.; Zody, M. C.; Linton, L.; Lander, E. S.; Altshuler, D. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **2001**, *409* (6822), 928−933. (c) Brender, J. R.; Zhang, Y. Predicting the Effect of Mutations on Protein-Protein Binding Interactions through Structure-Based Interface Profiles. *PLoS Comput. Biol.* **2015**, *11* (10), No. e1004494.

(2) Baaden, M.; Marrink, S. J. Coarse-grain modelling of protein-protein interactions. *Curr. Opin. Struct. Biol.* **2013**, *23* (6), 878−886.

(3) (a) Ezkurdia, I.; Bartoli, L.; Fariselli, P.; Casadio, R.; Valencia, A.; Tress, M. L. Progress and challenges in predicting protein-protein interaction sites. *Briefings Bioinf.* **2009**, *10* (3), 233−246. (b) Kastritis, P. L.; Bonvin, A. M. Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. *J. Proteome Res.* **2010**, *9* (5), 2216−2225.

(4) (a) Bernardi, R. C.; Melo, M. C.; Schulten, K. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim. Biophys. Acta, Gen. Subj.* **2015**, *1850* (5), 872−877. (b) Spiwok, V.; Sucur, Z.; Hosek, P. Enhanced sampling techniques in biomolecular simulations. *Biotechnol. Adv.* **2015**, *33* (6), 1130−1140. (c) Gumbart, J. C.; Roux, B.; Chipot, C. Efficient Determination of Protein−Protein Standard Binding Free Energies from First Principles. *J. Chem. Theory Comput.* **2013**, *9* (8), 3789−3798.

(5) Guerois, R.; Nielsen, J. E.; Serrano, L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **2002**, *320* (2), 369−387.

(6) Li, M.; Petukh, M.; Alexov, E.; Panchenko, A. R. Predicting the Impact of Missense Mutations on Protein-Protein Binding Affinity. *J. Chem. Theory Comput.* **2014**, *10* (4), 1770−1780.

(7) Dehouck, Y.; Kwasigroch, J. M.; Rooman, M.; Gilis, D. BeAtMuSiC: Prediction of changes in protein-protein binding affinity on mutations. *Nucleic Acids Res.* **2013**, *41* (W1), W333−W339.

(8) (a) Gromiha, M. M.; Yugandhar, K.; Jemimah, S. Protein-protein interactions: scoring schemes and binding affinity. *Curr. Opin. Struct. Biol.* **2017**, *44*, 31−38. (b) Pons, C.; Grosdidier, S.; Solernou, A.; Perez-Cano, L.; Fernandez-Recio, J. Present and future challenges and limitations in protein-protein docking. *Proteins: Struct., Funct., Genet.* **2010**, *78* (1), 95−108.
(9) Clark, L. A.; van Vlijmen, H. W. A knowledge-based forcefield for protein-protein interface design. *Proteins: Struct., Funct., Genet.* **2008**, *70* (4), 1540−1550.
(10) Moal, I. H.; Fernandez-Recio, J. Comment on 'protein-protein binding affinity prediction from amino acid sequence'. *Bioinformatics* **2015**, *31* (4), 614−615.
(11) (a) Miller, C. R.; Johnson, E. L.; Burke, A. Z.; Martin, K. P.; Miura, T. A.; Wichman, H. A.; Brown, C. J.; Ytreberg, F. M. Initiating a watch list for Ebola virus antibody escape mutations. *PeerJ* **2016**, *4*, No. e1674. (b) Meroueh, S.; Liang, S.; Li, L. Computational Design of Protein Interfaces with Receptor Flexibility. *Biophys. J.* **2010**, *98* (3), 428a. (c) Humphris, E. L.; Kortemme, T. Prediction of protein-protein interface sequence diversity using flexible backbone computational protein design. *Structure* **2008**, *16* (12), 1777−1788. (d) Beard, H.; Cholleti, A.; Pearlman, D.; Sherman, W.; Loving, K. A. Applying physics-based scoring to calculate free energies of binding for single amino acid mutations in protein-protein complexes. *PLoS One* **2013**, *8* (12), No. e82849.
(12) (a) Saunders, M. G.; Voth, G. A. Coarse-graining methods for computational biology. *Annu. Rev. Biophys.* **2013**, *42*, 73−93. (b) Ingolfsson, H. I.; Lopez, C. A.; Uusitalo, J. J.; de Jong, D. H.; Gopal, S. M.; Periole, X.; Marrink, S. J. The power of coarse graining in biomolecular simulations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4* (3), 225−248.
(13) (a) May, A.; Pool, R.; van Dijk, E.; Bijlard, J.; Abeln, S.; Heringa, J.; Feenstra, K. A. Coarse-grained versus atomistic simulations: realistic interaction free energies for real proteins. *Bioinformatics* **2014**, *30* (3), 326−334. (b) Lelimousin, M.; Limongelli, V.; Sansom, M. S. Conformational Changes in the Epidermal Growth Factor Receptor: Role of the Transmembrane Domain Investigated by Coarse-Grained MetaDynamics Free Energy Calculations. *J. Am. Chem. Soc.* **2016**, *138* (33), 10611−10622. (c) Chavent, M.; Chetwynd, A. P.; Stansfeld, P. J.; Sansom, M. S. Dimerization of the EphA1 receptor tyrosine kinase transmembrane domain: Insights into the mechanism of receptor activation. *Biochemistry* **2014**, *53* (42), 6641−6652. (d) Janosi, L.; Prakash, A.; Doxastakis, M. Lipid-modulated sequence-specific association of glycophorin A in membranes. *Biophys. J.* **2010**, *99* (1), 284−292. (e) Prakash, A.; Janosi, L.; Doxastakis, M. Self-association of models of transmembrane domains of ErbB receptors in a lipid bilayer. *Biophys. J.* **2010**, *99* (11), 3657−3665. (f) Sengupta, D.; Marrink, S. J. Lipid-mediated interactions tune the association of glycophorin A helix and its disruptive mutants in membranes. *Phys. Chem. Chem. Phys.* **2010**, *12* (40), 12987−12996. (g) Dunton, T. A.; Goose, J. E.; Gavaghan, D. J.; Sansom, M. S.; Osborne, J. M. The free energy landscape of dimerization of a membrane protein, NanC. *PLoS Comput. Biol.* **2014**, *10* (1), No. e1003417. (h) Stark, A. C.; Andrews, C. T.; Elcock, A. H. Toward optimized potential functions for protein-protein interactions in aqueous solutions: osmotic second virial coefficient calculations using the MARTINI coarse-grained force field. *J. Chem. Theory Comput.* **2013**, *9* (9), 4176−4185. (i) Cao, H.; Huang, Y.; Liu, Z. Interplay between binding affinity and kinetics in protein-protein interactions. *Proteins: Struct., Funct., Genet.* **2016**, *84* (7), 920−933.
(14) Ganesan, A.; Coote, M. L.; Barakat, K. Molecular 'time-machines' to unravel key biological events for drug design. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2017**, *7* (4), No. e1306.
(15) Darre, L.; Machado, M. R.; Brandner, A. F.; Gonzalez, H. C.; Ferreira, S.; Pantano, S. SIRAH: a structurally unbiased coarse-grained force field for proteins with aqueous solvation and long-range electrostatics. *J. Chem. Theory Comput.* **2015**, *11* (2), 723−739.
(16) Kästner, J. Umbrella sampling. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1* (6), 932−942.
(17) Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L. The FoldX web server: an online force field. *Nucleic Acids Res.* **2005**, *33*, W382−W388.
(18) Periole, X.; Marrink, S. J. The Martini coarse-grained force field. *Methods Mol. Biol. (N. Y., NY, U. S.)* **2013**, *924*, 533−65.
(19) Moal, I. H.; Fernandez-Recio, J. SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. *Bioinformatics* **2012**, *28* (20), 2600−2607.
(20) Bode, W.; Wei, A.-Z.; Huber, R.; Meyer, E.; Travis, J.; Neumann, S. X-ray crystal structure of the complex of human leukocyte elastase (PMN elastase) and the third domain of the turkey ovomucoid inhibitor. *EMBO J.* **1986**, *5* (10), 2453−2458.
(21) Buckle, A. M.; Schreiber, G.; Fersht, A. R. Protein-protein recognition: crystal structural analysis of a barnase-barstar complex at 2.0-A resolution. *Biochemistry* **1994**, *33* (30), 8878−8889.
(22) Padlan, E. A.; Silverton, E. W.; Sheriff, S.; Cohen, G. H.; Smith-Gill, S. J.; Davies, D. R. Structure of an antibody-antigen complex: crystal structure of the HyHEL-10 Fab-lysozyme complex. *Proc. Natl. Acad. Sci. U. S. A.* **1989**, *86* (15), 5938−5942.
(23) Webb, B.; Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics*; 2016; Vol. *54*, pp 5.6.1−5.6.37, DOI: 10.1002/cpbi.3.
(24) Dunbrack, R. L., Jr Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.* **2002**, *12* (4), 431−440.
(25) 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* (13), 1605−1612.
(26) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. GROMACS: fast, flexible, and free. *J. Comput. Chem.* **2005**, *26* (16), 1701−1718.
(27) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **2014**, *185* (2), 604−613.
(28) Marrink, S. J.; Tieleman, D. P. Perspective on the Martini model. *Chem. Soc. Rev.* **2013**, *42* (16), 6801−6822.
(29) Darré, L.; Machado, M. R.; Dans, P. D.; Herrera, F. E.; Pantano, S. Another Coarse Grain Model for Aqueous Solvation: WAT FOUR? *J. Chem. Theory Comput.* **2010**, *6* (12), 3793−3807.
(30) Machado, M. R.; Pantano, S. SIRAH tools: mapping, backmapping and visualization of coarse-grained models. *Bioinformatics* **2016**, *32* (10), 1568−1570.
(31) Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* **2004**, *32*, W665−W667.
(32) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Struct., Funct., Genet.* **2006**, *65* (3), 712−725.
(33) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126* (1), 014101.
(34) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52* (12), 7182−7190.
(35) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98* (12), 10089−10092.
(36) Kumar, S.; Rosenberg, J.; Bouzida, D.; Swendsen, R.; Kollman, P. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **1992**, *13* (8), 1011−1021.
(37) Shirts, M. R.; Mobley, D. L. An Introduction to Best Practices in Free Energy Calculations. In *Biomolecular Simulations: Methods and Protocols*; Monticelli, L., Salonen, E., Eds.; Humana Press: Totowa, NJ, USA, 2013; pp 271−311, DOI: 10.1007/978-1-62703-017-5_11.

DOI: 10.1021/acs.jctc.7b00660
J. Chem. Theory Comput. 2018, 14, 991−997