JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
Article
pubs.acs.org/JACS

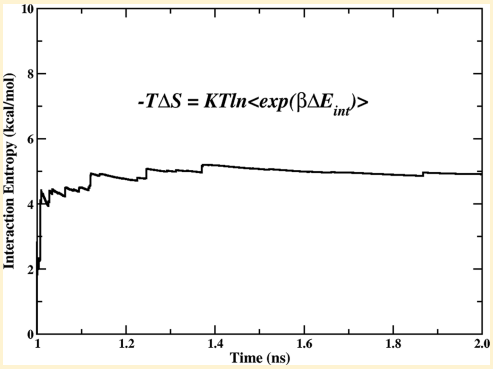
Interaction Entropy: A New Paradigm for Highly Efficient and Reliable Computation of Protein−Ligand Binding Free Energy

Lili Duan,[†,‡] Xiao Liu,[†] and John Z.H. Zhang*[,†,§,∥,⊥]

[†]Department of Physics, College of Chemistry and Molecular Engineering, East China Normal University, Shanghai 200062, China
[‡]School of Physics and Electronics, Shandong Normal University, Jinan 250014, China
[§]NYU-ECNU Center for Computational Chemistry at NYU Shanghai, Shanghai 200062, China
[∥]Department of Chemistry, New York University, New York, New York 10003, United States
[⊥]Collaborative Innovation Center of Extreme Optics, Shanxi University, Taiyuan, Shanxi 030006, PRC

Supporting Information

ABSTRACT: Efficient and reliable calculation of protein−ligand binding free energy is a grand challenge in computational biology and is of critical importance in drug design and many other molecular recognition problems. The main challenge lies in the calculation of entropic contribution to protein−ligand binding or interaction systems. In this report, we present a new interaction entropy method which is theoretically rigorous, computationally efficient, and numerically reliable for calculating entropic contribution to free energy in protein−ligand binding and other interaction processes. Drastically different from the widely employed but extremely expensive normal mode method for calculating entropy change in protein−ligand binding, the new method calculates the entropic component (interaction entropy or $-T\Delta S$) of the binding free energy directly from molecular dynamics simulation without any extra computational cost. Extensive study of over a dozen randomly selected protein−ligand binding systems demonstrated that this interaction entropy method is both computationally efficient and numerically reliable and is vastly superior to the standard normal mode approach. This interaction entropy paradigm introduces a novel and intuitive conceptual understanding of the entropic effect in protein−ligand binding and other general interaction systems as well as a practical method for highly efficient calculation of this effect.



1. INTRODUCTION

Molecular recognition plays a central role in biological systems. Many essential elements of life such as self-replication, metabolism, and information processing are controlled largely by specific interaction between biological molecules as those observed in between receptor−ligand, antigen−antibody, DNA−protein, sugar−lectin, RNA−ribosome, and so on. Thus, understanding how two molecules recognize each other is of fundamental importance in biology. One of the most important molecular recognition processes with direct medical importance and application is protein−ligand binding, which is at the heart of drug discovery and drug interaction and is an area of intensive experimental and computational studies in biomedical sciences.

Protein−ligand binding is essential to almost all biological processes. The underlying physical and chemical interactions determine the specific biological recognition at the molecular level. The essential element in drug discovery is to find a molecular ligand that either inhibits or activates a specific target protein through ligand binding. However, finding a ligand that binds a targeted protein with high affinity is a major challenge in early stage drug discovery. Modern technological advances in analytical methods and the availability of experimental tools such as X-ray crystallography and nuclear magnetic resonance (NMR)[1,2] have enabled researchers to obtain atomic resolution structures of protein−ligand complexes. The high-resolution structures of protein and protein−ligand complex provide a chemical basis for understanding protein−ligand interactions at atomic level,[3−8] and they can be effectively used as the basis for the design of small-molecule drugs for the treatment of diseases.

However, given the structure of a protein−ligand complex (such as from experiment or virtual molecular docking), it is not an easy task to calculate its binding affinity reliably, an extremely important but difficult undertaking in computational biology. The strength of binding of a ligand to a protein molecule is governed by the free energy change in the binding process. Besides the accuracy of force field and sufficient sampling of the phase space during molecular simulation, reliable calculation of entropy change is critical to the accuracy of the computed binding free energy. Currently, the most

Received: March 13, 2016
Published: April 8, 2016

DOI: 10.1021/jacs.6b02682
J. Am. Chem. Soc. 2016, 138, 5722−5728

rigorous approaches for accurate calculation of protein−ligand binding free energy are free energy perturbation (FEP)[9−14] and thermodynamic integration[15,16] methods. However, free energy calculations for protein−ligand binding using either FEP (free energy perturbation) or TI (thermodynamic integration) methods are extremely difficult; both can be prohibitively expensive and very difficult to converge numerically as one has to simulate many nonphysical intermediate states of the system. The linear interaction energy (LIE) approach is another class of methods in which the interaction energies are used with adjustable parameters to estimate protein−ligand binding free energies.[17,18] This class of methods often do well for systems with similar interaction characteristics. In contrast, the MM/PBSA approach,[19−25] which uses an implicit solvent model to compute solvation energy coupled with MD simulation in explicit water to obtain gas-phase component of the binding free energy, is more general for practical applications in computing binding free energies. However, a major problem in MM/PBSA method is the calculation of entropy change in protein−ligand binding. The current MM/PBSA approach calculates entropy change for protein−ligand binding by using the standard normal mode method, which is approximate in nature, extremely expensive in computation, and often unreliable for protein−ligand binding. As a result, many applications using MM/PBSA approach simply neglect the calculation of entropy change for protein−ligand binding and thus render the computed free energy even more uncertain.

In this report, we present a novel and conceptually more intuitive theoretical paradigm called "interaction entropy" or IE. This new paradigm introduces a novel but more intuitive conceptual understanding of the entropic effect in protein−ligand binding and other general interaction systems as well as a practical method for highly efficient calculation of its effect. This interaction entropy is theoretically rigorous and can be directly obtained from MD simulation of protein−ligand system without any extra computational cost. Thus, the new method is numerically superefficient compared to the normal mode calculation of entropy for protein−ligand binding. For free energy calculation of protein−ligand binding, we can simply employ the standard MM/PBSA method to calculate the solvation free energy component and then combine them with the calculated interaction entropy to obtain the binding free energy. Thus, the interaction entropy method is straightforward to implement and highly efficient to apply for practical computation of protein−ligand binding free energies. To fully demonstrate the efficiency and reliability of the present approach, we carried out computational studies for 15 randomly selected protein−ligand complexes with experimental binding affinities using both the interaction entropy method as well as the standard normal mode method for entropy calculations.

2. RESULTS

2.1. Details of the Numerical Studies. To demonstrate the computational superiority of the interaction entropy method against the standard normal mode approach in free energy calculation of protein−ligand binding, we randomly picked 15 protein−ligand systems with known experimental binding energies for comparison study. The native structures of these 15 protein−ligand complexes (Protein Databank ID: 1e66, 2brb, 2iwx, 2vw5, 2wbg, 2x00, 2xdl, 2yge, 2zjw, 3ao4, 3k5v, 3kgp, 3owj, 4des, and 4dew) from PDB are taken as the starting structures. These systems are randomly selected from a benchmark set called "core set" in the PDBbind database developed by Wang.[26,27] In our study, two separate MD simulations are carried out for each of these 15 systems: a 2 ns MD run with constraint (to be specified below) imposed on protein structures and a 6 ns run without any constraint on protein structures.

The ligands are optimized at HF/6-31G** level and single-point calculations at B3LYP/cc-PVTZ level are performed to generate electrostatic potentials (ESP) to fit their atomic charges using the restrained ESP (RESP) method.[28,29] All missing hydrogen atoms are added to their proper positions using the Leap module in AMBER12 package,[30] and the AMBER12SB force field is employed in all MD simulations. Each complex is solvated in a truncated periodic TIP3P water box, and the minimum distances from the surfaces of the box to the complex atoms are set to 12 Å. Counter ions are added to neutralize systems, and the complex systems are energy minimized by the steepest descent method followed by conjugate gradient minimization until convergence is reached. After that, the entire systems are heated from 0 to 300 K over 300 ps with 10 kcal mol^{-1} $Å^{-2}$ harmonic constraints on all solute atoms. Langevin dynamics[31] is used to regulate the temperature with a collision frequency of 1.0 ps^{-1}. All bonds involving hydrogen atoms are constrained by the SHAKE algorithm,[32] and a time step of 2 fs is used in the simulation.

Finally, two separate MD simulations are performed. In the 2 ns MD run, the proteins are constrained with 10 kcal mol^{-1} $Å^{-2}$ harmonic constraints on all atoms, and configurational sampling is taken every 10 fs from the last 1 ns trajectories. In the 6 ns MD run, no configurational constraints are imposed on proteins, and sampling is also taken every 10 fs from the last 1 ns trajectories. Thus, a total of 100 000 configurations or snapshots are extracted from the MD trajectories for the calculation of interaction entropies and MM/PBSA solvation energies. The rationale to run MD simulation with constraint on protein structure is as follows. In many protein MD simulations, many physical quantities are difficult or even impossible to converge often because of structural drift resulting from inaccurate force field. Thus, long simulation without constraint can often lead to incorrect protein structures and thus nonconvergent results.

For comparison study using normal mode[33] method to calculate entropy change in the standard MM/PBSA approach, only 10 configurations or snapshots, equally spaced in the last 1 ns trajectories, are used to compute averaged entropic contribution to the free energy. This is due to huge computational costs associated with normal mode calculation of entropy for protein systems.

2.2. Results of Comparisons. We first examine if the simulation time is reasonably converged for the systems we are studying. Figure 1 shows the RMSD of complex structures of a number of the 15 systems with respect to their native structures. As shown in Figure 1, all structures are stable with simulation time in 2 ns run, obviously due to the use of constraint. In the 6 ns run, the structures are generally stable within the time frame. The complete RMSD values for all the 15 systems are given in the Supporting Information. Thus, we believe that the two MD simulations with constrained protein structure for 2 ns and without constraint for 6 ns are reasonably converged for purpose of calculating binding free energies. Next, we need to establish that calculation of the interaction entropy using eq 8 is numerically convergent with respect to sampling configurations. Figure 2 shows the convergence of the

DOI: 10.1021/jacs.6b02682
J. Am. Chem. Soc. 2016, 138, 5722−5728

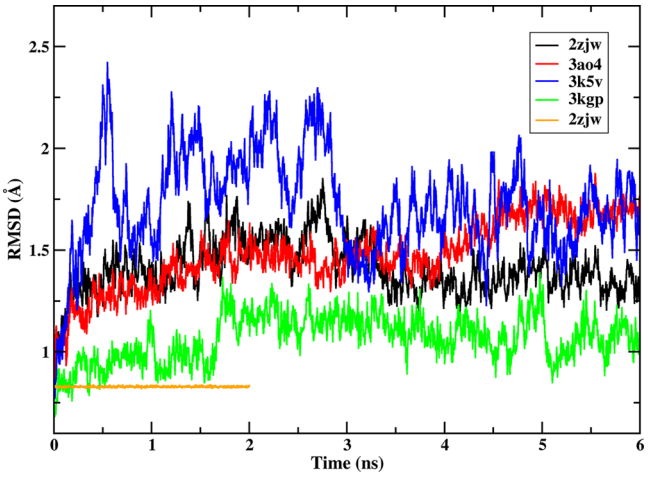


Figure 1. RMSD of the protein backbone as a function of MD simulation time for 4 of the 15 systems in the 6 ns MD run. The lower yellow line is the result from the 2 ns MD run. The structure of the initial time refers to the configuration after optimization of the crystal structure.

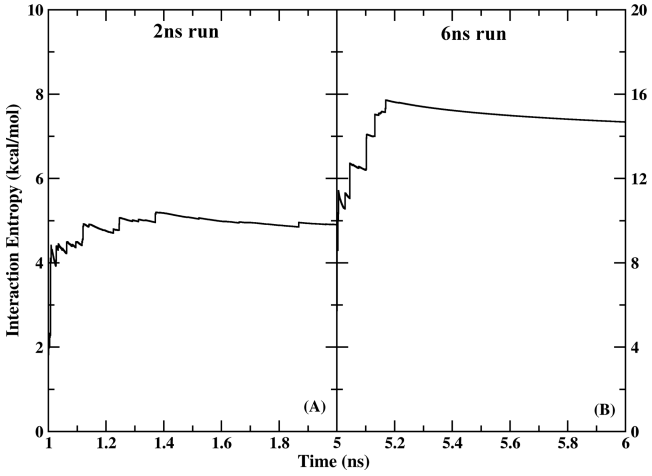


Figure 2. Calculated interaction entropy (in kcal/mol) for the 4dew system as a function of time used in configuration sampling. The left figure is the result from the 2 ns MD run (with constraint on protein structure) and the right one from the 6 ns MD run (without constraint on protein structure). In both calculations, configuration was sampled every 10 fs.

calculated interaction entropy with respect to sampling time or number of sampling configurations. It is shown that the values of interaction entropy are very well converged with respect to configuration sampling within the simulation time frame.

A major attractive feature of our method is the computational efficiency in calculating interaction entropy in comparison to the standard normal mode method for entropy calculation. Figure 3 shows the computer CPU time needed to calculate interaction entropy ($-T\Delta S$) for 15 protein−ligand systems by normal mode method. Because of computational expenses, only 10 configurations or snapshots that are evenly spaced in the last 1 ns trajectories are used for normal mode calculation of entropy in standard MM/PBSA free energy calculation. As shown in the figure, normal mode calculation of entropy takes hours or longer of CPU time just for a single configuration. In application for practical systems, one typically has to employ dozens or more configurations for entropy calculation in order

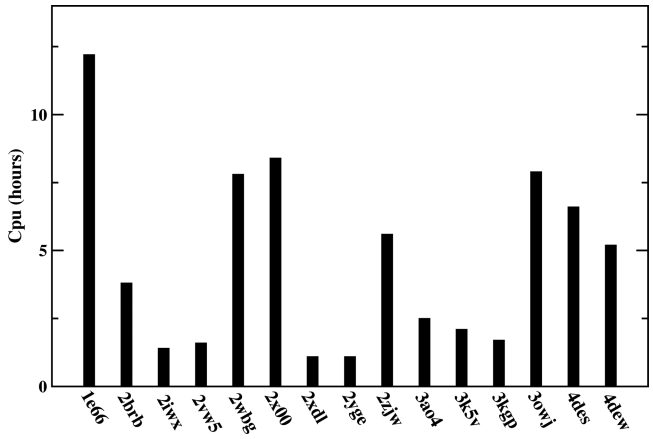


Figure 3. Computation costs for calculating the entropy changes for 15 protein−ligand systems using the normal mode method for a single configuration (snapshot). The computational cost using the interaction entropy method is fractional of minutes.

to be statistically meaningful. For example, if 100 snapshots are used, which is very common in MM/PBSA calculation of protein−ligand binding free energy, then the computer time will increase by 2 orders of magnitude in normal mode calculation of entropy. Although parallelization with many CPUs can shorten the wall clock time, the overall consumption of computer resources can still be extremely high. Furthermore, the entropy calculated by normal mode method is only approximate, and there is no simple way to measure the accuracy of such calculated entropy.

In contrast, the evaluation of interaction entropy in the present method takes just fractional minutes for sampling 100 000 configurations within the 1 ns time frame for a given protein−ligand system, so the calculation of interaction entropy does not require additional computer time beyond that used in MD simulation of the protein−ligand system and in MM/PBSA calculation of solvation energies.

Next, we compare the calculated components of the binding free energies including the interaction entropy using both normal mode and the present method. The calculated binding free energies are also directly compared to those of the experiment. Table 1 shows the results of free energy calculations from the 2 ns MD runs with constraint on protein backbones and various terms of the binding free energies calculated for the 15 protein−ligand complexes. Because we use the same MM/PBSA method to calculate the solvation component of the free energy in both the standard MM/BPSA method and the present interaction entropy method, the only differences in these calculations are the entropic component of the free energy. As we note, the entropic contributions to the free energies obtained from both calculation methods are quite comparable in magnitude, but the general trend appears to be that the result from the interaction entropy method is somewhat smaller than those calculated by the normal mode method. Of course, 15 systems may not be large enough to state that this is a general trend, but we should keep in mind that the interaction entropy computed using the new method is sampled over 100 000 configurations whereas that from the normal mode calculation samples only 10 configurations and the latter may not be well-converged because of insufficient sampling.

In the second study, we performed 6 ns MD runs for all 15 systems without constraints on proteins' structures. The

DOI: 10.1021/jacs.6b02682
J. Am. Chem. Soc. 2016, 138, 5722−5728

Table 1. Binding Free Energies of 15 Protein−Ligand Systems Computed from Nmode and IE Methods as Well as the Corresponding Experimental Values: 2 ns MD Run with Constraint on Protein Structure[a]

			$-T\Delta S$		ΔG_{bind}		
PDB code	$\langle E_{pl}^{int} \rangle$	ΔG_{sol}	N_{mode}	IE	N_{mode}	IE	ΔG_{exp}
1e66	−51.66	31.63	22.18	3.79	2.15	−16.24	−13.66
2brb	−61.35	38.20	13.62	7.08	−9.53	−16.07	−6.72
2iwx	−71.92	52.92	19.28	6.37	0.28	−12.63	−9.23
2vw5	−101.12	68.86	29.16	10.72	−3.10	−21.54	−11.77
2wbg	−118.61	96.09	19.13	10.60	−3.39	−11.92	−6.15
2x00	−71.90	44.99	25.47	11.18	−1.44	−15.73	−15.65
2xdl	−52.65	39.95	17.05	6.95	4.35	−5.75	−4.29
2yge	−85.30	61.74	27.57	12.81	4.01	−10.75	−7.00
2zjw	−54.06	46.33	4.76	10.68	−2.97	2.95	−10.64
3ao4	−64.09	41.76	20.39	7.27	−1.94	−15.06	−2.86
3k5v	−52.76	21.67	19.79	8.71	−11.30	−22.38	−8.71
3kgp	−107.74	84.89	10.80	13.68	−12.05	−9.17	−3.55
3owj	−25.58	9.27	10.50	5.65	−5.81	−10.66	−8.38
4des	−33.45	18.62	11.99	3.27	−2.84	−11.56	−8.09
4dew	−56.35	40.58	0.32	4.90	−15.45	−10.87	−9.67
MAE					7.11	5.88	

[a]MAE is the mean absolute error of the computed free energy from the experimental value.

Table 2. Binding Free Energies of 15 Protein−Ligand Systems Computed from Nmode and IE Methods as Well as the Corresponding Experimental Values: 6 ns MD Run without Constraint on Protein Structure[a]

			$-T\Delta S$		ΔG_{bind}		
PDB code	$\langle E_{pl}^{int} \rangle$	ΔG_{sol}	N_{mode}	IE	N_{mode}	IE	ΔG_{exp}
1e66	−50.25	30.60	14.69	5.16	−4.96	−14.49	−13.66
2brb	−63.57	36.90	17.77	9.66	−8.90	−17.01	−6.72
2iwx	−81.25	57.03	22.38	16.32	−1.84	−7.90	−9.23
2vw5	−105.80	76.36	27.78	26.69	−1.66	−2.75	−11.77
2wbg	−111.67	79.97	25.65	16.58	−6.05	−15.12	−6.15
2x00	−89.15	54.08	15.43	20.77	−19.64	−14.30	−15.65
2xdl	−55.38	37.29	18.42	11.19	0.33	−6.90	−4.29
2yge	−94.47	63.07	22.46	19.80	−8.94	−11.60	−7.00
2zjw	−49.63	32.79	5.16	11.43	−11.68	−5.41	−10.64
3ao4	−53.54	32.77	18.32	8.06	−2.45	−12.71	−2.86
3k5v	−57.75	29.82	15.11	18.53	−12.82	−9.40	−8.71
3kgp	−117.16	89.7	15.45	24.98	−12.01	−2.48	−3.55
3owj	−22.21	9.09	9.00	8.37	−4.12	−4.75	−8.38
4des	−53.78	36.85	17.87	10.44	0.94	−6.49	−8.09
4dew	−55.61	38.09	17.67	14.68	0.15	−2.84	−9.67
MAE					5.07	4.53	

[a]MAE is the mean absolute error of the computed free energies from the experimenxtal values.

computed results from this 6 ns MD run are also shown in Table 2. We notice that calculated free energies for most of these 15 systems are quite similar in value to those in the constrained 2 ns MD run. Also, the differences in the results between the IE method and the normal mode method are quite similar to that observed in the 2 ns MD run. The free energy deviations from the experimental values are also comparable to those in the 2 ns MD run. The results from these two independent and different MD runs are consistent with each other and demonstrate the reliability of the IE method.

3. DISCUSSION

In this report, we presented a theoretically rigorous, computationally efficient and numerically reliable method for calculation of interaction entropy for protein−ligand binding directly from the MD trajectory of the protein−ligand complex. The IE method has enormous advantages over the standard normal mode method for calculating entropic contribution to free energy change in protein−ligand binding or similar host−guest problems. First, the IE method is theoretically rigorous, whereas the normal method is approximate for entropy calculation. Second, in computational cost the IE method is highly efficient, whereas the normal method is extremely expensive. For example in protein−ligand binding, the "gas-phase" protein−ligand interaction energy is available at each time step of MD simulation, and there is no additional computer cost beyond the standard MD simulation of the protein−ligand complex system. Third, the result from IE method is numerically stable with respect to ample ensemble sampling, whereas that from the normal method is numerically less stable. In the normal mode approach, one computes the absolute entropies of the systems approximately and then calculates their differences. Because the absolute values of

DOI: 10.1021/jacs.6b02682
J. Am. Chem. Soc. 2016, 138, 5722−5728

entropies can be very large for biomolecules, their subtractions to produce a small number could cause numerical errors due to inherent errors in the computed absolute entropies. For example, the entropy term (TS) for the protein−ligand system 4dew is about 5027 kcal/mol, and the corresponding apo system is about 5002.5 kcal/mol. In contrast, the IE method directly computes the entropy change term without the need to calculate absolute entropies of the systems. Of course, it should also be noted that the normal mode method is used for calculation of absolute entropy and thus has more general applicability, whereas the IE method is only applicable to calculation of relative entropy or entropy change before and after protein−ligand binding or similar host−guest problems.

The difference in computed free energies between IE method and the normal method in this study is obviously due to a difference in the calculation of entropic terms in each MD run. There are two possible sources that could contribute to the difference in the result of two methods. First, the normal mode calculation of entropy involves errors, both from the inherently approximate nature of the method as well as from insufficient sampling of configurations (only 10 configurations are used in the present study because of computational cost). In the IE calculation, there is also a possible source of error, which is due to the choice of heat bath used in the MD simulation that could affect the ensemble average of IE. However, because the entire MD ensemble is based on the heat bath used, it is difficult to disintegrate it from other sources of errors in comparing the results of IE and normal mode methods.

To understand the difference between the present computational results with experimental binding free energies, there are quite many possible sources of error as listed below. (1) Experimental condition may not be exactly the same as in theoretical model, e.g., ionization states and complex environment of the system, among others. (2) Accuracy of the force field used in theoretical calculations. These include the accuracies of force fields for proteins, ligands, and solvent molecules; in particular, the lack of polarization of the force fields is a serious issue. (3) PBSA calculation of solvation energy is based on implicit solvent model with some molecular parameters. (4) The heat bath used in MD simulation is also artificial and could also be a source of error in computational results. (5) The present calculation is based on the assumption that the configurations of the protein and the ligand remain the same before and after binding. For many systems, this is not actually the case, and extra free energy changes due to conformational distortions of the protein and/or the ligand need to be included, e.g., by separate calculations.

It is worthwhile to mention the difference in computed free energies from the 2 ns MD run with constraint and the 6 ns MD run without constraint on protein structure. As we can see from Tables 1 and 2, the difference between the two results for a given system may come from interaction energy, solvation, or entropic term. This is an indication that the computed binding free energies can be sensitive to sampled complex structures in MD simulation. Because of uncertain error in the force field, it is difficult to fully converge the result with respect to configuration sampling because the system may never converge to the correct configuration distribution. It often is the case that the longer the MD simulation, the worse the result because longer MD simulation can actually cause the system to drift further away from the correct structure. Thus, we believe that by constraining the system near its natural structure should give more desirable as well as numerically more stable binding free energy from MD simulation. Contribution to free energy from structural distortion in protein−ligand binding could be obtained by some correction procedure to account for the energy cost from structural distortion.

Finally, it is important to mention that the present IE method not only provides a computational method for highly efficient calculation of binding free energies, it also brings about new conceptual understanding about entropic effect in protein−ligand binding and other general interactions. For this we note that there are a number of critical features for IE in the formulation of eq 6. First, the IE is always positive, meaning that the system entropy is always decreased upon interaction of the two partners. Although this result is generally expected, eq 6 rigorously proves this result. Second, the entropic loss of the system is closely correlated with the fluctuation of the interaction energy around its average value. That means that the more fluctuation of the interaction energy, the greater the entropic loss in the binding free energy. This may seem to be counterintuitive to the general perception of entropy about freedom of movement, but it is actually correct because we are dealing with the fluctuation of the interaction energy, which is related to the relative motion of the interacting partners, not individual partners. This means that the tighter the two partners interacting with each other, the less entropic the loss in the binding free energy.

4. METHODS

4.1. MM/PBSA Approach. The free energy for protein−ligand binding can be expressed as the sum of two components, the gas-phase binding free energy and the solvation free energy[19−25]

$$\Delta G = \Delta G_{gas} + \Delta G_{sol} \tag{1}$$

where

$$\Delta G_{gas} = \langle E_{pl}^{int} \rangle - T\Delta S \tag{2}$$

where $\langle E_{pl}^{int} \rangle$ is the ensemble-averaged protein−ligand interaction energy and the term $(-T\Delta S)$ is the entropic contribution. In the MM/PBSA approach, the averaged protein−ligand interaction energy $\langle E_{pl}^{int} \rangle$ is defined as the difference of gas-phase energy between that of the protein−ligand complex and those of the separate protein and ligand. The solvation free energy is obtained using an implicit solvent model by solving the Poisson−Boltzmann (PB) equation with an added empirical surface term to account for the cavitation free energy, and ΔG_{sol} is the difference between that of the protein−ligand complex and those of the separate protein and ligand systems. The protein−ligand solvation free energy is thus

$$\Delta G_{sol} = \Delta G_{pb} + \Delta G_{np} \tag{3}$$

where ΔG_{pb} is the electrostatic solvation free energy, which is obtained by solving PB equation using the PBSA program in AMBER suite. In the PB calculation, the interior and exterior dielectric constants are set to 1 and 80, respectively. ΔG_{np} is the nonpolar solvation free energy term, which is obtained by using an empirical solvent-accessible surface area (SASA) formula

$$\Delta G_{np} = \gamma \mathrm{SASA} + \beta \tag{4}$$

The values γ and β we used in the calculation are the standard values of 0.00542 kcal/(mol·Å^2) and 0.92 kcal/mol, respectively. The contribution of entropy $(-T\Delta S)$ to the binding free energy, which arises from the changes of the translational, rotational, and vibrational degrees of freedom, is calculated using classical statistical thermodynamics and normal mode approximation using the AMBER NMODE module. $-T\Delta S$ is the difference of entropy between the protein−ligand complex and those of the separate protein and ligand. In the PBSA approach, the free energy calculation is performed at multiple

DOI: 10.1021/jacs.6b02682
J. Am. Chem. Soc. 2016, 138, 5722−5728

configurations that are sampled by MD simulation of the protein−ligand system in explicit water. The difficulty is in the computation of the entropic term $T\Delta S$ in eq 2. In the MM/PBSA approach, the standard normal mode approximation is used to compute the entropic change. However, this normal mode approach is hugely expensive computationally for protein systems with thousands or even tens of thousands of degrees of freedom. Furthermore, such expensive calculations need to be performed at many configurations in order to obtain meaningful ensemble average. In addition, evaluation of entropy by normal mode method can contain uncertain errors especially for large biomolecules. As a result, such entropy calculation could be prohibitively expensive and thus is often neglected in practical applications.

For comparison in the present study, 10 configurations or snapshots from the last 1 ns of each MD trajectories with an interval of 100 ps are selected to calculate the entropic terms using the normal mode method, and each configuration is minimized using a maximum of 500 000 steps with the RMS gradient of 10^{-4} kcal mol^{-1} Å^{-2}. Sampling of more configurations is more desirable, but they can be extremely costly as is seen in the Results section of this study.

4.2. Interaction Entropy Method. In our new interaction entropy method, the gas-phase component of the binding free energy is derived by the simple steps as follows

$$\Delta G_{gas} = -KT \ln \frac{\int dq_w\, dq_p\, dq_l\, e^{-\beta(E_p+E_l+E_{pl}^{int}+E_w+E_{pw}^{int}+E_{lw}^{int})}}{\int dq_w\, dq_p\, dq_l\, e^{-\beta(E_p+E_l+E_w+E_{pw}^{int}+E_{lw}^{int})}}$$

$$= -KT \ln\left[\frac{1}{\langle e^{\beta E_{pl}^{int}}\rangle}\right] = KT \ln[\langle e^{\beta E_{pl}^{int}}\rangle]$$

$$= KT \ln[e^{\beta\langle E_{pl}^{int}\rangle}\langle e^{\beta(E_{pl}^{int}-\langle E_{pl}^{int}\rangle)}\rangle]$$

$$= \langle E_{pl}^{int}\rangle + KT \ln\langle e^{\beta\Delta E_{pl}^{int}}\rangle$$

$$= \langle E_{pl}^{int}\rangle - T\Delta S \quad (5)$$

Here E_p, E_l, and E_w are internal energies of the protein, ligand, and waters, respectively, E_{pl}^{int}, E_{pw}^{int}, and E_{lw}^{int} are interaction energies of protein−ligand, protein−water, and ligand−water, respectively, $\langle E_{pl}^{int}\rangle$ is the ensemble averaged protein−ligand interaction energy, and $\Delta E_{pl}^{int} = E_{pl}^{int} - \langle E_{pl}^{int}\rangle$ is the fluctuation of protein−ligand interaction energy around the average energy. Thus, we define the IE as

$$-T\Delta S = KT \ln\langle e^{\beta\Delta E_{pl}^{int}}\rangle \quad (6)$$

The relevant ensemble averages can be evaluated by averaging over MD simulation,

$$\langle E_{pl}^{int}\rangle = \frac{1}{T}\int_0^T E_{pl}^{int}(t)\, dt = \frac{1}{N}\sum_{i=1}^{N} E_{pl}^{int}(t_i) \quad (7)$$

and

$$\langle e^{\beta\Delta E_{pl}^{int}}\rangle = \frac{1}{N}\sum_{i=1}^{N} e^{\beta\Delta E_{pl}^{int}(t_i)} \quad (8)$$

In the above derivation, we use MD simulation of the protein−ligand complex in explicit water to generate ensemble average for interaction entropy as given in eq 6. The above derivation for interaction entropy is theoretically rigorous as compared to the inherently approximate nature of normal mode calculation of entropy.

It is not difficult to understand why eq 6 is computationally superior to that using normal mode approach. The calculation of the interaction entropy by eqs 6 or 8 simply involves the natural log of an ensemble average of $e^{\beta\Delta E_{pl}^{int}}$, which can be readily extracted along with MD simulation without extra computational cost. Here we note that the interaction energy E_{pl}^{int} includes both electrostatic and van der Waals interactions between the protein and the ligand.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/jacs.6b02682.

RMSDs of MD trajectories for the 15 protein−ligand systems in two different MD runs. (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*john.zhang@nyu.edu

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Jinfeng Liu for providing technical help for this project. This work was supported by the National Natural Science Foundation of China (Grant Nos. 21433004, 11574184, 31200545, and 11274206) and Shanghai Putuo District (Grant 2014-A-02). We thank the Supercomputer Center of East China Normal University for providing us computational time.

■ REFERENCES

(1) Bieri, M.; Kwan, A. H.; Mobli, M.; King, G. F.; Mackay, J. P.; Gooley, P. R. *FEBS J.* **2011**, *278*, 704−715.
(2) Kwan, A. H.; Mobli, M.; Gooley, P. R.; King, G. F.; Mackay, J. P. *FEBS J.* **2011**, *278*, 687−703.
(3) Kasai, K.; Ishii, S. *J. Biochem* **1975**, *77*, 261−264.
(4) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.
(5) Richarme, G.; Kepes, A. *Biochim. Biophys. Acta, Protein Struct. Mol. Enzymol.* **1983**, *742*, 16−24.
(6) Gohlke, H.; Klebe, G. *Angew. Chem., Int. Ed.* **2002**, *41*, 2644−2676.
(7) Muegge, I.; Martin, Y. C. *J. Med. Chem.* **1999**, *42*, 791−804.
(8) Olsson, T. S. G.; Williams, M. A.; Pitt, W. R.; Ladbury, J. E. *J. Mol. Biol.* **2008**, *384*, 1002−1017.
(9) Bash, P. A.; Field, M. J.; Karplus, M. *J. Am. Chem. Soc.* **1987**, *109*, 8092−8094.
(10) Rao, S. N.; Singh, U. C.; Bash, P. A.; Kollman, P. A. *Nature* **1987**, *328*, 551−554.
(11) Kita, Y.; Arakawa, T.; Lin, T. Y.; Timasheff, S. N. *Biochemistry* **1994**, *33*, 15178−15189.
(12) Rao, B. G.; Singh, U. C. *J. Am. Chem. Soc.* **1990**, *112*, 3803−3811.
(13) Kollman, P. A. *Chem. Rev.* **1993**, *93*, 2395−2417.
(14) Jorgensen, W. L.; Thomas, L. L. *J. Chem. Theory Comput.* **2008**, *4*, 869−876.
(15) Beveridge, D. L.; DiCapua, F. M. *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431−492.
(16) Zacharias, M.; Straatsma, T. P.; McCammon, J. A. *J. Chem. Phys.* **1994**, *100*, 9025−9031.
(17) Wang, J.; Dixon, R.; Kollman, P. A. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 69−81.
(18) Aqvist, J.; Luzhkov, V. B.; Brandsdal, B. O. *Acc. Chem. Res.* **2002**, *35*, 358−365.
(19) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. *Acc. Chem. Res.* **2000**, *33*, 889−897.
(20) Massova, I.; Kollman, P. A. *Perspect. Drug Discovery Des.* **2000**, *18*, 113−135.
(21) Kuhn, B.; Gerber, P.; Schulz-Gasch, T.; Stahl, M. *J. Med. Chem.* **2005**, *48*, 4040−4048.

DOI: 10.1021/jacs.6b02682
J. Am. Chem. Soc. 2016, 138, 5722−5728

(22) Luo, R.; David, L.; Gilson, M. K. *J. Comput. Chem.* **2002**, *23*, 1244−1253.
(23) Nicholls, A.; Honig, B. *J. Comput. Chem.* **1991**, *12*, 435−445.
(24) Rocchia, W.; Alexov, E.; Honig, B. *J. Phys. Chem. B* **2001**, *105*, 6507−6514.
(25) Chen, Z.; Baker, N. A.; Wei, G. W. *J. Comput. Phys.* **2010**, *229*, 8231−8258.
(26) Li, Y.; Han, L.; Liu, Z. H.; Wang, R. X. *J. Chem. Inf. Model.* **2014**, *54*, 1717−1736.
(27) Li, Y.; Liu, Z. H.; Li, J.; Han, L.; Liu, J.; Zhao, Z. X.; Wang, R. X. *J. Chem. Inf. Model.* **2014**, *54*, 1700−1716.
(28) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269−10280.
(29) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollmann, P. A. *J. Am. Chem. Soc.* **1993**, *115*, 9620−9631.
(30) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Swails, J.; Götz, A. W.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wolf, R. M.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Salomon-Ferrer, R.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *AMBER 12*; University of California, San Francisco, CA, 2012.
(31) Pastor, R. W.; Brooks, B. R.; Szabo, A. *Mol. Phys.* **1988**, *65*, 1409−1419.
(32) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327−341.
(33) Nguyen, D. T.; Case, D. A. *J. Phys. Chem.* **1985**, *89*, 4020−4026.

DOI: 10.1021/jacs.6b02682
J. Am. Chem. Soc. 2016, 138, 5722−5728