JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

MTS-MD of Biomolecules Steered with 3D-RISM-KH Mean Solvation Forces Accelerated with Generalized Solvation Force Extrapolation

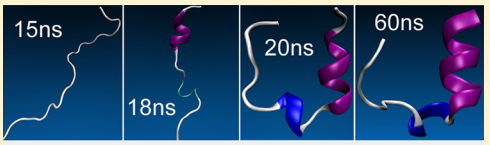
Igor Omelyan*[,†,‡,¶] and Andriy Kovalenko*[,‡,†]

[†]Department of Mechanical Engineering, University of Alberta, Mechanical Engineering Building 4-9, Edmonton, Alberta T6G 2G8, Canada

[‡]National Institute for Nanotechnology, 11421 Saskatchewan Drive, Edmonton, Alberta T6G 2M9, Canada

[¶]Institute for Condensed Matter Physics, National Academy of Sciences of Ukraine, 1 Svientsitskii Street, Lviv 79011, Ukraine

ABSTRACT: We developed a generalized solvation force extrapolation (GSFE) approach to speed up multiple time step molecular dynamics (MTS-MD) of biomolecules steered with mean solvation forces obtained from the 3D-RISM-KH molecular theory of solvation (three-dimensional reference interaction site model with the Kovalenko-Hirata closure). GSFE is based on a set of techniques including the non-Eckart-like transformation of coordinate space separately for each solute atom, extension of the force-coordinate pair basis set followed by selection of the best subset, balancing the normal equations by modified least-squares minimization of deviations, and incremental increase of outer time step in motion integration. Mean solvation forces acting on the biomolecule atoms in conformations at successive inner time steps are extrapolated using a relatively small number of best (closest) solute atomic coordinates and corresponding mean solvation forces obtained at previous outer time steps by converging the 3D-RISM-KH integral equations. The MTS-MD evolution steered with GSFE of 3D-RISM-KH mean solvation forces is efficiently stabilized with our optimized isokinetic Nosé−Hoover chain (OIN) thermostat. We validated the hybrid MTS-MD/OIN/GSFE/3D-RISM-KH integrator on solvated organic and biomolecules of different stiffness and complexity: asphaltene dimer in toluene solvent, hydrated alanine dipeptide, miniprotein 1L2Y, and protein G. The GSFE accuracy and the OIN efficiency allowed us to enlarge outer time steps up to huge values of 1−4 ps while accurately reproducing conformational properties. Quasidynamics steered with 3D-RISM-KH mean solvation forces achieves time scale compression of conformational changes coupled with solvent exchange, resulting in further significant acceleration of protein conformational sampling with respect to real time dynamics. Overall, this provided a 50- to 1000-fold effective speedup of conformational sampling for these systems, compared to conventional MD with explicit solvent. We have been able to fold the miniprotein from a fully denatured, extended state in about 60 ns of quasidynamics steered with 3D-RISM-KH mean solvation forces, compared to the average physical folding time of 4−9 μs observed in experiment.



Folding miniprotein by OIN quasidynamics steered with extrapolated 3D-RISM-KH mean solvation forces

1. INTRODUCTION

Prediction of the structure and functioning of proteins at the molecular level from the first-principles, i.e. solely from amino acids sequences and interaction potentials between the solute and solvent atoms, remains a challenging task.[1] The main problem is that the processes responsible for conformational and folding equilibria in these complex systems take place on time scales ranging from picoseconds to micro- and milliseconds. During the last decades, molecular dynamics (MD) simulations originally designed to study simple liquids[2−5] have been developed into a powerful tool[6−18] that enabled understanding the mechanisms of protein folding, one of the most fundamental biochemical operations. However, such simulations must be very long (at least several microseconds) to stand a good chance of observing a single folding event even for the simplest proteins.[1] In the course of development, the MD simulation length has continuously increased, first reaching one microsecond[8,9] and then ten microseconds.[14] Recently, a barrier of one millisecond has been broken.[15−17] However, the consideration was restricted to relatively simple systems. With the present capabilities of high performance computing, MD simulations of relatively large proteins are bounded, as a rule, to tens to hundreds of nanoseconds.[19] This is, of course, quite insufficient for usual MD to obtain a full pattern on the folding behavior.

A promising way to significantly accelerate molecular simulations has been to combine the MD method with the 3D-RISM integral equation theory of molecular liquids (three-dimensional reference interaction site model)[20−29] complemented with the Kovalenko-Hirata closure relation.[24,27,29] In the hybrid MD/3D-RISM-KH approach, individual trajectories and dynamics of solvent molecules are contracted to quasiequilibrium 3D density distribution functions of their interaction sites around the biomolecule in successive conformation snapshots. The evolution of the biomolecule thus becomes quasidynamics steered with mean solvation forces obtained for each conformation of the biomolecule from the 3D-RISM-KH molecular theory of solvation.[30−33] The latter produces mean solvation forces by converging the 3D-

Received: November 21, 2014
Published: March 3, 2015

DOI: 10.1021/ct5010438
J. Chem. Theory Comput. 2015, 11, 1875−1895

RISM-KH integral equations derived from the first-principles of statistical mechanics, beginning from an input of the interaction potentials and geometries of the biomolecule and solvent molecules (molecular force field). The 3D-RISM-KH mean solvation forces statistically mechanically averaged over the distributions of an infinite number of solvent molecules are thus added to the direct intramolecular interactions for integrating the equations of motion of atoms in the biomolecule. A chief advantage of this hybrid approach is that slow solvation processes in the confined geometry of the biomolecule, particularly under its conformational changes, such as solvent exchange and re-equilibration, localization of structural solvent, ions distribution and localization, and protein−ligand binding, which constitute a major challenge for conventional MD are readily accounted for by 3D-RISM-KH mean solvation forces and thus excluded from the biomolecule quasidynamics. This leads to drastic compression of time scale in protein quasidynamics compared to real dynamics and thus enables fast access to structural and folding properties of large biomolecular systems in solution.

Worth emphasizing is a profound advantage of the 3D-RISM-KH molecular theory of solvation as compared to continuum solvation methods which represent polar solvation forces with either the Poisson−Boltzmann (PB)[34] or the Generalized Born (GB)[35−37] models and empirically account for nonpolar solvation forces with the solvent accessible surface area (SASA, or SA) model supplemented with additional volume and dispersion integral terms.[38,39] These continuum solvation approaches are parametrized for hydration of biomolecules and are not transferable to other solvent or solvent system with cosolvent, ions at a finite concentration (physiological concentration in biomolecular systems), and other solvent species, in particular, in the recent methods[40−43] treating ligand fragments as part of a solvent mixture. Continuum solvation models based on the concept of a solvation cavity in dielectric structureless medium representing solvent entirely ignore effects of finite size of solvent molecules on mean solvation forces between solute molecules, for example, a desolvation barrier due to expelling solvent molecules from the gap between the surfaces of proteins (parts of protein) when bringing them together in contact. Furthermore, SASA is well-defined for an outer surface of a biomolecule but loses physical meaning and becomes inadequate inside small inner cavities of biomolecules like a narrow channel accommodating an ion and a few water molecules. As distinct, the 3D-RISM-KH molecular theory of solvation readily accounts for all such and other molecular solvation effects and yields both the structure as 3D maps of solvent density distributions and the solvation thermodynamics at the level of molecular simulations. (A comparison attainable if molecular simulations are feasible; for many biomolecular systems, affordable simulation times are too short to gain meaningful statistics of rare essential solvation events, while the molecular theory of solvation provides the solvation structure and thermodynamics in the equilibrium ensemble.) It is useful to clearly position these approaches in the context of solvation models nomenclature used in the literature: GB(PB)SA are *implicit* and *continuum* solvation models; whereas the 3D-RISM-KH molecular theory of solvation is *implicit* as it produces 3D density distributions which are an average of individual trajectories of solvent molecules, and *not continuum* as it uses all molecular interaction potentials in the system (molecular force field) at input to produce the full molecular picture of solvation in terms of the solvent distributions and solvation thermodynamics at output with full account of all chemical specificities and molecular geometries as specified in the force field.

Miyata and Hirata performed a pioneering MD/3D-RISM-KH simulation for hydrated acetylacetone.[30] They exploited the standard reference system propagator algorithm (RESPA)[44−46] in the microcanonical ensemble, with calculation of mean solvation forces by converging the 3D-RISM-KH integral equation at each outer step without resorting to extrapolation of solvation forces. With this propagator algorithm, it was impossible to apply outer time steps larger than 5 fs as a result of resonance instabilities[47−52] that appear in MD/3D-RISM-KH as well as in conventional MD simulations due to the multiple time step (MTS) interplay between strong intramolecular (solute−solute) and weak intermolecular (solute−solvent) forces. In conventional MD, the accuracy of MTS simulation can be increased by carrying out processed phase-space transformations.[53,54] Utilizing them within an energy-constrained scheme, it was shown[55] in MD simulations of water that outer time steps up to 16 fs are acceptable. However, such steps cannot exceed the theoretical limiting value of 20 fs inherent in the microcanonical description. Furthermore, in hybrid MD/3D-RISM-KH without solvation force extrapolation, the procedure of converging the 3D-RISM-KH integral equations has to be repeated too frequently (every 5 fs),[30] which drastically slows down the computations.

In order to damp MTS instabilities, the MD/3D-RISM-KH approach has been extended[31] to the canonical ensemble within the Langevin dynamics.[56,57] Introducing the method of solvation force extrapolation (SFE) for mean solvation forces acting on the solute, it has been demonstrated for hydrated alanine dipeptide that larger outer time steps of up to 20 fs are feasible.[31] They, however, are still smaller than those achievable in conventional MD simulations by the best previously known isokinetic Nosé−Hoover chain RESPA (INR) integrator, for which an outer time step of 100 fs is possible.[58,59] Recently, we introduced the optimized isokinetic Nosé−Hoover chain (OIN) canonical ensemble for more efficient elimination of MTS instabilities in MD simulations.[32] It improves over the INR method[58,59] and other canonical-isokinetic schemes[60−62] by coupling each set of Nosé−Hoover chain thermostats to some optimal number of degrees of freedom in the system. Slightly modifying the original SFE scheme,[31] the OIN integrator has been combined with the MD/3D-RISM-KH approach. It has been shown on an example of hydrated alanine dipeptide[32] that the OIN ensemble is superior to the Langevin and INR schemes. In particular, large outer time steps of order of several hundred femtoseconds can be employed, providing a speedup up to 20 times in comparison with conventional MD with explicit solvent.

Very recently, we introduced a method of advanced solvation force extrapolation (ASFE) in MD/3D-RISM-KH simulations.[33] The pivoting idea was to apply a global non-Eckart-like rotation of atomic coordinates to minimize the distances between the biomolecule sites in different conformations at successive time steps. We showed that carrying out the force extrapolation in the transformed space and extending the set of outer conformations provide evaluation of mean solvation forces with a much better accuracy than the previous extrapolation scheme. This allowed us, without affecting the equilibrium and conformational properties, to apply huge outer time steps up to tens of picoseconds in the hybrid MD/3D-

DOI: 10.1021/ct5010438
J. Chem. Theory Comput. 2015, 11, 1875−1895

RISM-KH simulations and thus to get a 100- to 500-fold acceleration compared to MD with explicit solvent. However, the validation was limited to hydrated alanine dipeptide, a relatively simple molecule of only 22 atoms, and it remained unknown whether this degree of acceleration would hold for larger biomolecules.

In the present paper, we significantly modify and generalize our recent ideas and techniques[33] so as to obtain the desired speedup for larger solute molecules, including proteins. First of all, rather than performing a full rotational transformation of the whole molecule, we introduce an individual non-Eckart-like transformation for each atom of the biomolecule in the presence of a smooth weighing function. For macromolecules, this appreciably enhances convergence of the extrapolated forces to their exact values with increase of the number of basis outer coordinates. Other techniques, such as the least-squares minimization, extension of the force-coordinate pair set to select the best subset, and balancing the normal equations have been modified, too. In addition, we propose the so-called frequency scheme which appreciably reduces computational overhead of the extrapolation without loss of precision. With all the above improvements, the good accuracy attainable with the original SFE scheme[31] for outer time steps of up to 20 fs can now be held with our new approach of generalized solvation force extrapolation (GSFE) for much longer steps of order of 1 to 4 ps, even for large biomolecules.

The paper is organized as follows. In Section 2, we introduce the GSFE approach. Section 3 describes the 3D-RISM-KH theory and the canonical OIN scheme for integration of the multiscale equations of motion in the presence of the extrapolated forces. In Section 4, we validate the resulting OIN/GSFE/3D-RISM-KH algorithm against MD simulations on an asphaltene molecule in toluene solvent and on miniprotein 1L2Y and protein G in water solvent. For comparison, we examine hydrated alanine dipeptide as well. In Section 5, we perform folding of the miniprotein. Final conclusions are made in Section 6.

2. GENERALIZED SOLVATION FORCE EXTRAPOLATION

2.1. 3D-RISM-KH Mean Solvation Forces. Let us consider a solute macromolecule consisting of M atoms solvated in liquid comprising a large number of solvent molecules with M' atomic sites. Conventional MD simulation deals with forces $-\partial U/\partial \mathbf{r}_i$ acting on all atoms $i = 1, ..., M+M'$ at positions $\mathbf{r}_i$ in the solute–solvent system with the total potential energy $U(\mathbf{r}_1, ..., \mathbf{r}_M, \mathbf{r}_{M+1}, ..., \mathbf{r}_{M+M'})$.

The latter can be split up as $U = U_1 + U_2$ into the solute–solute interaction potential $U_1(\mathbf{r}_1,...,\mathbf{r}_M)$ and the remaining term $U_2(\mathbf{r}_1,...,\mathbf{r}_{M+M'})$ comprising the solute–solvent and solvent–solvent interaction potentials. Accordingly, the forces acting on solute atoms split up into two parts coming respectively from solute–solute and solute–solvent atomic interactions, $-\partial U/\partial \mathbf{r}_i = -\partial U_1/\partial \mathbf{r}_i - \partial U_2/\partial \mathbf{r}_i$, where $i = 1, ..., M$. In MD simulations of biomolecules, the total number of atoms of solvent molecules has to be much larger than that of the solute biomolecule ($M' \gg M$) to have good statistics and neglect finite size effects. Even at infinite dilution, for a biomolecule of $M \sim$1,000–10,000 atoms the number of solvent atoms has to be $M \sim$ 10,000–100,000. This considerably complicates conventional MD because a large portion of the computational costs is spent on evaluation of solute–solvent and solvent–solvent atomic forces $-\partial U_2/\partial \mathbf{r}_i$, where $i = 1, ..., M+M'$. In common practice, the concentration of solute macromolecules (or aggregates of macromolecules) is small and interactions between solutes (or composite solutes) are neglected, thus reducing the consideration to infinite dilution.

Since we are interested exclusively in conformational and folding behavior of the solute biomolecule, one way to improve the efficiency of MD simulations consists of contracting the degrees of freedom of solvent (a huge number of molecules) and evaluating the dynamics of the biomolecule on the solvation free energy surface. Such quasidynamics of the biomolecule is then steered with mean solvation forces[63,64] that are defined as a statistical average of solute–solvent atomic forces $-\partial U_2/\partial \mathbf{r}_i$ acting on each solute atom $i = 1, ..., M$ over all arrangements of all M' solvent atoms around the biomolecule at a frozen conformation $\{\mathbf{r}_1, ..., \mathbf{r}_M\}$

$$\mathbf{f}_i(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_M) = -\frac{\int \frac{\partial U_2}{\partial \mathbf{r}_i} e^{-U/(k_B T)} d\mathbf{r}_{M+1} d\mathbf{r}_{M+2} ... d\mathbf{r}_{M+M'}}{\int e^{-U/(k_B T)} d\mathbf{r}_{M+1} d\mathbf{r}_{M+2} ... d\mathbf{r}_{M+M'}} \quad (1)$$

where $k_B T$ is the Boltzmann constant times system temperature. Without explicit solvent treatment, mean solvation forces can be obtained either from continuum solvation models or from molecular theory of solvation.

In continuum solvation models, mean solvation forces (1) are empirically constructed and parametrized. In the context of hydration of biomolecules, polar solvation forces in dilute systems are reproduced with either the Poisson–Boltzmann (PB)[34] or the Generalized Born (GB)[35−37] models, and nonpolar solvation forces at atomic length scale are accounted for with the solvent accessible surface area (SASA) model supplemented with additional volume and dispersion integral terms.[38,39] This approach works well for the values of the hydration free energy of biomolecules; unfortunately, it has all inherent disadvantages of continuum solvation theories: nontransferable to other solvents and solvent systems, in particular, electrolyte solutions, missing solvent size effects such as a desolvation barrier in protein aggregation, inadequacy for solvation of internal cavities such as narrow channels.

As distinct, the 3D-RISM-KH molecular theory of solvation[24−29] is transferable and yields both solvation structure and mean solvation forces from the first-principles of statistical mechanics with an accuracy (up to the closure approximation used) at the level of molecular simulation with a very large number of solvent molecules that has converged. (Explicit solvent simulation with viable statistical sampling is a huge challenge for systems with slow rate processes such as solvent exchange and localization in protein confined spaces, preferential adsorption of cosolvent, partitioning of ions, binding of ligands, and solvent and ions mediated protein–protein interactions). By converging the 3D-RISM-KH integral equations discretized on a 3D grid the solvation structure is obtained in terms of 3D maps of site correlation functions (including density distributions) (see Section 3.1). The solvation free energy and mean solvation forces $\mathbf{f}_i$ acting on each interaction site of the solute molecule in a solvation box of infinite size ($M' \to \infty$) are then readily calculated in a closed form analytically obtained by thermodynamic integration of the 3D-RISM-KH equations as a single 3D spatial integral of the 3D site correlation functions (see Section 3.2). The predictive capability of the 3D-RISM-KH theory has been validated for the solvation structure and thermodynamics of chemical species and biomolecules in various solvents and solution sys-

DOI: 10.1021/ct5010438
J. Chem. Theory Comput. 2015, 11, 1875–1895

tems,[27,29,65] in particular, in the context of hybrid MD/3D-RISM-KH for hydrated biomolecules.[30−33] 3D-RISM-KH mean solvation forces $\mathbf{f}_i$ can then be employed together with direct intramolecular solute−solute interactions $-\partial U_1/\partial \mathbf{r}_i$ to integrate the equations of motion just for solute atoms.

A further, much more important advantage of the hybrid MD/3D-RISM-KH approach is that slow rate processes of solvent exchange and re-equilibration in confined spaces due to conformational changes of the solute biomolecule are entirely eliminated from the MD/3D-RISM-KH quasidynamics. In fact, it performs essential dynamics of protein in solution with effective account for molecular steric forces and chemical specificities, such as desolvation barrier in hydrophobic interaction and hydrogen bonding. This drastically contracts the time scale of quasidynamics compared to real dynamics and so leads to dramatically shorter simulation times required to gain adequate statistics. This intrinsic acceleration of MD steered with 3D-RISM-KH mean solvation forces grows with complexity of the biomolecule (e.g., protein). Indeed, in conventional MD, solvent enters pockets and inner cavities of the biomolecule through its conformational changes. This is a very slow process with rare statistics which is as difficult to model explicitly as protein folding conformational changes themselves. (Note again that continuum solvation models miss steric effects such as desolvation barrier and hydrogen bonding and do not adequately reproduce solvation in inner cavities of biomolecules.) As distinct, the 3D-RISM-KH theory yields the solvent distribution in the inner cavity or pocket at once for the final conformation in chemical equilibrium with the bulk solvent outside the protein by construct of the theory, bypassing the intermediate conformational states. Calculation of the 3D-RISM-KH mean solvation forces requires significant numerical efforts if the integral equations are solved at each inner time step of the MD trajectory. However, 3D-RISM-KH mean solvation forces $\mathbf{f}_i$ vary with solute atomic coordinates and so with time much smoother than solute−solvent atomic interaction forces $-\partial U_2/\partial \mathbf{r}_i$ evaluated directly in conventional MD. The reason is that mean solvation forces at a given solute conformation are obtained by statistical averaging over all arrangements of equilibrated solvent molecules, and so all core repulsion forces and other strong short-range components typical to explicit solvent interactions are smoothed out in the averaging. Therefore, 3D-RISM-KH forces can be efficiently extrapolated, which allows the 3D-RISM-KH integral equations to be to converged much less frequently and thus drastically increases the efficiency of hybrid MD/3D-RISM-KH simulation.

2.2. Generalized Extrapolation of Mean Solvation Forces for Biomolecules. As mentioned in the Introduction, the original SFE scheme[31] is restricted to small outer time steps, whereas the advanced solvation force extrapolation (ASFE) method[33] still needs significant modification and generalization to remain efficient for large flexible macromolecules as well. We will now show how 3D-RISM-KH mean solvation forces can be extrapolated in the best way for a general case of solvated protein.

2.2.1. Local Non-Eckart-like Rotational Transformations. Let $\mathbf{f}_{i,k}$ be the solvation forces acting on solute sites $i = 1, 2, ..., M$ at N previous outer time steps $k = 1, 2, ..., N$ for which the 3D-RISM-KH integral equations are converged. The atomic positions at these steps will be denoted by $\mathbf{r}_{i,k}$. The forces $\mathbf{f}_{i,k}$ and positions $\mathbf{r}_{i,k}$ for a given i are ordered in such a way that larger values of k correspond to earlier moments t_k of time, i.e., $t_N < t_{N-1} < \cdots < t_2 < t_1$. The next outer moment is denoted by $t_0 > t_1$. Let $\mathbf{r}_i(t)$ be the current coordinate of atom i at some inner time point t belonging to the interval $]t_1, t_0[$. The total number of these points is equal to $P = (t_0 - t_1)/\Delta t \gg 1$, where Δt denotes the inner time step. Note also that the actual force $\mathbf{f}_i(t)$ exerted on atom i at time t will depend on the multidimensional vector $\{\mathbf{r}_{1,k}, \mathbf{r}_{2,k}, ..., \mathbf{r}_{M,k}\}$ via the relative positions $\mathbf{r}_{ij}(t) = \mathbf{r}_i(t) - \mathbf{r}_j(t)$ of $M-1$ neighbors j (with $j \neq i$). This follows from the translational invariance of solvation interactions when the total system (solute plus solvent) is arbitrarily shifted as a whole.

One of the main ideas of the new approach is to find such local rotational transformations $\mathbf{R}_{ij} = \mathbf{S}_i \mathbf{r}_{ij}$ of the relative positions $\mathbf{r}_{ij}$ for each atom $i = 1, 2, ..., M$ (where $j = 1, 2, ..., M$) that provide the most smooth behavior of $\mathbf{F}_i = \mathbf{f}_i(\{\mathbf{R}_{ij}\})$ in the new coordinates. This can be achieved by reducing the coordinate region of force extrapolation. For the discrete set $(k = 1, 2, ..., N)$ of the basis coordinate knots $\mathbf{r}_{i,k}$ the desired transformation $\mathbf{R}_{ij,k} = \mathbf{S}_{i,k}\mathbf{r}_{ij,k}$ with $\mathbf{r}_{ij,k} = \mathbf{r}_{i,k} - \mathbf{r}_{j,k}$ can be determined by minimizing the normalized distances between all the transformed outer coordinates $\mathbf{R}_{ij,k}$ and some origin (where $\mathbf{S} \equiv \mathbf{I}$) point $\mathbf{r}_{ij}^*$ lying in the extrapolating region as

$$\frac{1}{M_i}\sum_{j=1}^{M}{}' w(r_{ij}^*)(\mathbf{S}_{i,k}\mathbf{r}_{ij,k} - \mathbf{r}_{ij}^*)^2 = \min \tag{2}$$

Here w is a weighting function, $\sum'$ stands for $j \neq i$, and M_i is the total number of neighbors with $w(r_{ij}^*) \neq 0$. The current inner coordinate $\mathbf{r}_{ij}(t)$ should also be transformed analogously by $\mathbf{R}_{ij}(t) = \mathbf{S}_i(t)\mathbf{r}_{ij}(t)$ with

$$\frac{1}{M_i}\sum_{j=1}^{M}{}' w(r_{ij}^*)(\mathbf{S}_i(t)\mathbf{r}_{ij}(t) - \mathbf{r}_{ij}^*)^2 = \min$$

$$\text{for each } i = 1, 2, ..., M \tag{3}$$

Any choice for $\mathbf{r}_{ij}^* = \mathbf{r}_{ij}(t^*)$ with $t_1 \leq t^* \leq t$ can be in principle acceptable, where t is the current inner time and t_1 is the most recent point from the basis outer steps. However, the limiting values $t^* = t_1$ and $t^* = t$ are not recommended in the context of efficiency. Note that in eq 2 we should carry out the transformation for each $k = 1, 2, ..., N$ (and $i = 1, 2, ..., M$) whenever $\mathbf{r}_{ij}^*$ is changed, i.e., up NMP times if $t^* = t$, but only NM ones for $t^* = t_1$. In the latter case, however, the origin $\mathbf{r}_{ij}^*$ being equal to $\mathbf{r}_{ij,1}$ may appear to be too far from the current point $\mathbf{r}_{ij}(t)$ when the size of the outer time step $h = t_0 - t_1$ is large. Thus, an optimal choice is when the origin $\mathbf{r}_{ij}^*$ of the transformation is updated after every $1 \ll p \ll P$ inner time step during the outer interval $(t_0 - t_1)$. This constitutes the so-called frequency reuse scheme.

The necessity of introducing the weighting function $w(r_{ij})$ is dictated by the fact that the neighbors with smaller interatomic distances r_{ij} contribute to the mean solvation force $\mathbf{f}_i$ more significantly and thus are more important for the minimization. The most natural way to model such a situation is to put $w(r_{ij}) = 1/r_{ij}^2$, meaning that the relative (and not absolute) interatomic distances are minimized. Then the left-hand sides of eqs 2 and 3 become dimensionless. At long enough $r_{ij} > R$ the correlations between $\mathbf{r}_{ij}$ and $\mathbf{f}_i$ are diminished, and so we can put $w(r_{ij}) = 0$ in this range. It should be emphasized that the above truncation concerns only the coordinate transformations 2 and 3 but not the actual solvation forces $\mathbf{f}_{i,k}$ which are calculated for all atomic pairs, without any cutoff. Furthermore, for small solute molecules of radius less than R, no truncation is

DOI: 10.1021/ct5010438
J. Chem. Theory Comput. 2015, 11, 1875−1895

performed at all. For large macromolecules with $M \gg 1$, setting a finite cutoff radius R can considerably improve the quality of the mean force extrapolation using only a relatively small number $N \ll M$ of the basis outer points.

The simplest way to obtain explicit expressions for the rotational matrix $\mathbf{S}_{i,k}$ or $\mathbf{S}_i(t)$ is to represent them in terms of the four components quaternion $\mathbf{q} = \{\chi,\eta,\xi,\zeta\}$ as[66]

$$\mathbf{S} = \begin{pmatrix} \chi^2 + \eta^2 - \xi^2 - \zeta^2 & 2(\eta\xi - \chi\zeta) & 2(\chi\xi + \eta\zeta) \\ 2(\chi\zeta + \eta\xi) & \chi^2 + \xi^2 - \eta^2 - \zeta^2 & 2(\xi\zeta - \chi\eta) \\ 2(\eta\zeta - \chi\xi) & 2(\chi\eta + \xi\zeta) & \chi^2 + \zeta^2 - \eta^2 - \xi^2 \end{pmatrix} \quad (4)$$

with $\mathbf{q}^2 \equiv \mathbf{q}^+ \mathbf{q} = \chi^2 + \eta^2 + \xi^2 + \zeta^2 = 1$. Inserting eq 4 into the superposition eq 2 or 3 yields

$$\frac{1}{2}\mathbf{q}^+\Theta_i\mathbf{q} - \frac{1}{2}\vartheta(\mathbf{q}^+\mathbf{q} - 1) = \min \quad (5)$$

where $\Theta_i = 1/(M_i)\sum_{j=1}^{M}{}' w(r_{ij}^*)\Theta_{ij}$

$$\Theta_{ij} = \begin{pmatrix} (\mathbf{r}'_{ij} - \mathbf{r}_{ij}^*)^2 & 2(\mathbf{r}'_{ij} \times \mathbf{r}_{ij}^*)^+ \\ 2(\mathbf{r}'_{ij} \times \mathbf{r}_{ij}^*) & \mathbf{I}(\mathbf{r}'_{ij} + \mathbf{r}_{ij}^*)^2 - 2(\mathbf{r}'_{ij}\mathbf{r}_{ij}^{*+} + \mathbf{r}_{ij}^*\mathbf{r}_{ij}'^{+}) \end{pmatrix} \quad (6)$$

are the symmetric 4×4 matrices, $\mathbf{r}'_{ij}$ is equal either to $\mathbf{r}_{ij,k}$ or $\mathbf{r}_{ij}(t)$ for the cases $\mathbf{S}_{i,k}$ or $\mathbf{S}_i(t)$, respectively, ϑ is the Lagrange multiplier, $\mathbf{I}$ is the identity 3×3 matrix, and $\times$ denotes the vector product. Differentiating eq 5 with respect to all four components of $\mathbf{q}$ leads to the eigenvalue problem

$$\Theta_i\mathbf{q} = \vartheta\mathbf{q} \quad (7)$$

Because the right-hand sides of eqs 2 and 3 are always ≥ 0, the matrix Θ_i is positive semidefinite, having four eigenvectors $\mathbf{q}_{1,2,3,4}$ and the same number of nonnegative associated eigenvalues $\vartheta_{1,2,3,4} \geq 0$. The latter can be sorted in the ascending order, such that ϑ_1 is the smallest eigenvalue. It coincides with the global minimum in eqs 2, 3, and 5 since for any normalized eigenvectors the following equality takes place: $\mathbf{q}^+ \Theta_i\mathbf{q} = \vartheta$. The normalized eigenvector $\mathbf{q}_1$ corresponding to the smallest eigenvalue $\vartheta \equiv \vartheta_{1,(i,k)}$ or $\vartheta_{1,i}$ is thus the quaternion describing the optimal transformation by the rotational matrix $\mathbf{S}$ [eq 4].

2.2.2. Individual Minimization by Weighted Least-Squares. Having the transformed coordinates

$$\mathbf{R}_{ij,k} = \mathbf{S}_{i,k}\mathbf{r}_{ij,k}, \qquad \mathbf{R}_{ij}(t) = \mathbf{S}_i(t)\mathbf{r}_{ij}(t) \quad (8)$$

the solvation forces can be extrapolated as follows. First, for each atom i, the actual neighboring positions $\mathbf{R}_{ij}(t)$ are virtually approximated at a given inner point t of the next outer time interval $]t_1, t_0[$ by a linear combination of their previous outer values as

$$\tilde{\mathbf{R}}_{ij}(t) = \sum_{k=1}^{N} A_k^{(i)}(t)\mathbf{R}_{ij,k} \quad (9)$$

The expansion coefficients $A_k^{(i)}(t)$ in eq 9 can then be obtained as the best representation of the solute neighboring coordinates $\mathbf{R}_{ij}(t)$ at time t in terms of their projections onto the basis of N previous outer positions $\mathbf{R}_{ij,k}$ by minimizing a weighting norm of the difference between $\mathbf{R}_{ij}(t)$ and their approximated counterparts $\tilde{\mathbf{R}}_{ij}(t)$. Additionally imposing the normalizing conditions

$$\sum_{k=1}^{N} A_k^{(i)} = 1, \qquad \sum_{k=1}^{N} A_k^{(i)^2} = \min \quad (10)$$

the above minimization leads for each $i = 1, 2, ..., M$ to the following modified least-squares problem

$$\frac{1}{M_i}\sum_{j=1}^{M}{}' w\left(R_{ij}^*\right)\left(\mathbf{R}_{ij} - \sum_{k=1}^{N} A_k^{(i)}\mathbf{R}_{ij,k}\right)^2 + 2\Lambda_i\left(\sum_{k=1}^{N} A_k^{(i)} - 1\right) + \varepsilon_i^2 \sum_{k=1}^{N} A_k^{(i)^2} = \min \quad (11)$$

where Λ_i is the Lagrangian multiplier, and $\varepsilon_i^2 \geq 0$ is a balancing parameter. Note that for coordinate deviations, the weighting function $w(R_{ij}^*) \equiv w(r_{ij}^*)$ used in eq 11 is the same as in local rotations 2 and 3, meaning again that neighbors j lying more closely to the reference atom i should be mapped more accurately. Note also that $R_{ij}^* = r_{ij}^*$ because the rotation transformation is unitary ($\mathbf{S} \equiv \mathbf{I}$) in the origin point $\mathbf{r}_{ij}^* \equiv \mathbf{R}_{ij}^*$.

Now, the forces $\mathbf{F}_i(t)$ at any inner time $t \in]t_1, t_0[$ can be extrapolated on the basis of their outer values $\mathbf{F}_{i,k}$ employing a linear expansion procedure which is quite similar to that [eq 9] for coordinates $\mathbf{R}_{ij}(t)$. This yields

$$\tilde{\mathbf{F}}_i(t) = \sum_{k=1}^{N} A_k^{(i)}(t)\mathbf{F}_{i,k} \quad (12)$$

where $i = 1, 2,..., M$ and the expansion coefficients $A_k^{(i)}(t)$ are the same as those in eq 9. This is justified by the fact that $\mathbf{F}_i(t)$ is a function of only $\mathbf{R}_{ij}(t)$. Thus, a better representation of the coordinates by $\tilde{\mathbf{R}}_{ij}(t)$ should provide a more accurate extrapolation of the interactions, expecting a small difference between the exact forces $\mathbf{F}_i(t)$ and their approximated values $\tilde{\mathbf{F}}_i(t)$. Note that the coordinate mapping is virtual in the sense that $\mathbf{R}_{ij}(t)$ are never replaced by $\tilde{\mathbf{R}}_{ij}(t)$. It is necessary only to find the coefficients $A_k^{(i)}$ for the actual force approximation. Remember also that in eq 11 the weighting function $w(r)$ being equal to $1/r^2$ for $r \leq R$ is truncated by $w(r) = 0$ at longer $r > R$. Such a truncation in eqs 2, 3, and 11 may lead to some boundary effects. However, these effects can be neglected by choosing the radius R to be large enough.

An issue now arises how to obtain $\mathbf{F}_{i,k}$ from $\mathbf{f}_{i,k}$ without direct recalculations $\mathbf{F}_{i,k} = \mathbf{f}(\{\mathbf{R}_{ij,k}\})$, and in which way to return back from $\tilde{\mathbf{F}}_i$ to the desired extrapolated forces $\tilde{\mathbf{f}}_i$ in the usual coordinate space. This issue can be solved by taking into account that the original solvation forces $\mathbf{f}_i$ are not only translationally invariant but also satisfy the following orientational condition

$$\mathbf{f}(\{\mathbf{S}_i\mathbf{r}_{ij}\}) = \mathbf{S}_i\mathbf{f}(\{\mathbf{r}_{ij}\}) \quad (13)$$

where $\mathbf{S}$ is an arbitrary 3×3 rotational matrix. Equation 13 merely states that if the solute molecule is rotated as a whole, the total solvation forces acting on each atom of this molecule will be transformed according to the same rotation. Note that in our case of the detailed extrapolation, the atomic coordinates are defined relatively to the current reference site $i = 1, 2, ..., M$. This means that the virtual rotation of $\mathbf{r}_{ij} \equiv \mathbf{r}_{ij,k}$ or $\mathbf{r}_{ij}(t)$ by $\mathbf{S}_i \equiv \mathbf{S}_{i,k}$ or $\mathbf{S}_i(t)$ is performed in eq 13 at each given i around this site for all other atoms j. On the other hand, the rotations in eqs 2 and 3 are performed only for groups of atoms for which $r_{ij}^* \leq R$ with $j \neq i$ and the truncation of neighbors is possible.

From eq 13 it follows that

$$\mathbf{F}_{i,k} = \mathbf{f}(\{\mathbf{R}_{ij,k}\}) = \mathbf{f}(\{\mathbf{S}_{i,k}\mathbf{r}_{ij,k}\}) = \mathbf{S}_{i,k}\mathbf{f}_{i,k} \quad (14)$$

$$\mathbf{F}_i(t) = \mathbf{f}(\{\mathbf{R}_{ij}(t)\}) = \mathbf{f}(\{\mathbf{S}_i(t)\mathbf{r}_{ij}(t)\}) = \mathbf{S}_i(t)\mathbf{f}_i(t) \quad (15)$$

DOI: 10.1021/ct5010438
J. Chem. Theory Comput. 2015, 11, 1875−1895

In view of eqs 14 and 15, no additional direct recalculations are needed, and the desired approximated forces in the usual coordinate space at each inner point t can be readily reproduced from eq 12 using the inverse rotational transformation

$$\tilde{\mathbf{f}}_i(t) = \mathbf{S}_i^{-1}(t)\tilde{\mathbf{F}}_i(t) = \mathbf{S}_i^{-1}(t)\sum_{k=1}^{N} A_k^{(i)}(t)\mathbf{S}_{i,k}\mathbf{f}_{i,k} \tag{16}$$

Furthermore, the inverse matrix can easily be evaluated taking into account that the rotational transformation is orthonormal, i.e., $\mathbf{S}^{-1} = \mathbf{S}^{+}$, where $\mathbf{S}^{+}$ denotes the transposed matrix.

We see, therefore, that the proposed individual transformation efficiently excludes local rotations of the solute molecule, which can be large enough due to the interactions with the solvent and thermostats. This reduces the volume of the local coordinate space around each solute atom in view of eqs 2 and 3. Obviously, then a better accuracy of the force extrapolation is provided. In other words, the differences between the approximated values $\tilde{\mathbf{f}}_i(t)$ and their original counterparts $\mathbf{f}_i(t)$ will decrease. In particular, the new extrapolation scheme 16 is exact, i.e., $\tilde{\mathbf{f}}_i(t) = \mathbf{f}_i(t)$, already at $N = 1$ for the case of rotating rigid segments constituting the molecule, where the transformed forces $\mathbf{F}_i$ are constant. It should also be very precise for flexible segments, since the magnitudes of the atomic vibrational oscillations are small. The influence of torsion movements on $\mathbf{F}_i$ can also be minimized by extending the basis set and choosing the best subset (see Section 2.2.4).

2.2.3. Normal Equations with Balancing. The most gentle way to find the coefficients $A_k^{(i)}$ in the force extrapolation 14 is to reduce the least-squares minimization 11 to a normal representation.[67,68] Differentiating eq 11 with respect to these coefficients and Λ_i leads to the following set of N+1 linear equations

$$\begin{pmatrix} G_{11}^{\varepsilon(i)} & G_{12}^{(i)} & .. & G_{1N}^{(i)} & 1 \\ G_{21}^{(i)} & G_{22}^{\varepsilon(i)} & .. & G_{2N}^{(i)} & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ G_{N1}^{(i)} & G_{N2}^{(i)} & .. & G_{NN}^{\varepsilon(i)} & 1 \\ 1 & 1 & .. & 1 & 0 \end{pmatrix} \begin{pmatrix} A_1^{(i)} \\ A_2^{(i)} \\ \vdots \\ A_N^{(i)} \\ \Lambda_i \end{pmatrix} = \begin{pmatrix} G_1^{(i)} \\ G_2^{(i)} \\ \vdots \\ G_N^{(i)} \\ 1 \end{pmatrix} \tag{17}$$

which should be solved for the same number of unknowns $A_k^{(i)}$ at $k = 1, 2, ..., N$ and Λ_i at each $i = 1, 2, ..., M$, where $G_{kk}^{\varepsilon(i)} = G_{kk}^{(i)} + \varepsilon_i^2$ with

$$G_{kl}^{(i)} = \frac{1}{M_i}\sum_{j=1}^{M}{}' w(R_{ij}^*)\mathbf{R}_{ij,k}\cdot\mathbf{R}_{ij,l} \tag{18}$$

$$G_k^{(i)} = \frac{1}{M_i}\sum_{j=1}^{M}{}' w(R_{ij}^*)\mathbf{R}_{ij,k}\cdot\mathbf{R}_{ij} \tag{19}$$

and $l = 1, 2, ..., N$. Note that the $(N+1) \times (N+1)$ square matrix in eq 17 remains to be symmetrical, since the ε_i^2-addition concerns only diagonal elements.

The Lagrange multiplier Λ normalizes the linear equations with the imposed constraint $\sum_k A_k = 1$ [see eq 10]. It is necessary to make the extrapolation to be exact for the spatially homogeneous part of the interactions in the transformed space. Indeed, the solvation force $\mathbf{F}_i(t) = \mathbf{F}_i(\mathbf{R}(t))$ can be expanded in the power series of a deviation of the current coordinate vector $\mathbf{R}$ from the origin $\mathbf{R}^* = \mathbf{S}\mathbf{r}^* \equiv \mathbf{r}^*$ at each $i = 1, 2, ..., M$ as

$$\mathbf{F}_i(\mathbf{R}) = \mathbf{F}_i(\{\mathbf{R}_{ij}^*\}) + \sum_{j=1}^{M} \left.\frac{\partial \mathbf{F}_i}{\partial \mathbf{R}_{ij}}\right|_{\mathbf{R}_{ij}^*} (\mathbf{R}_{ij} - \mathbf{R}_{ij}^*) \tag{20}$$

where $\partial\mathbf{F}_i/\partial\mathbf{R}_{ij}$ is the Hessian $(3M \times 3M)$ matrix, and the second and higher order spatial inhomogeneities $O[(\mathbf{R}_{ij}-\mathbf{R}_{ij}^*)^2]$ are neglected. Thus, $\mathbf{F}_i(\mathbf{R})$ has the constant zeroth-order part representing by the first term in the right-hand side of eq 20. It immediately follows from eq 12 that this term can be reproduced exactly, provided $\sum_k A_k = 1$. The second term in the right-hand side of eq 20 is linear in coordinates. That is why it can be extrapolated using the duplex linear expansions 9 and 12. In fact, each of the $3M \times 3M \gg 1$ elements of the Hessian matrix is mapped in a very complicated way via the obtained solutions for the extrapolation coefficients A_k involving a finite number $(N \gg 1)$ of the basis knots $\mathbf{R}_{ij,k}$.

The balancing parameter $\varepsilon^2 > 0$ appears as a result of the required minimization $\sum_k A_k^2 = \min$ for the norm of the expansion coefficients [see eq 10]. Such an additional minimization is also needed for the following reason. The used dual (virtual coordinate and actual force) extrapolation tentatively assumes that lowering of the coordinate residuals should immediately lead to a decrease of the deviations between the approximated and original forces, but this is not so when N approaches the number of local internal degrees of freedom $3M_i$ of the neighboring atoms. Then the least-squares solver to eq 11 will try to reduce the coordinate residuals to the global (zeroth) minimum with no regarding to the values of expansion coefficients A_k which are exploited in both the coordinate and force extrapolations. As a result, a lot of these coefficients may accept large negative and positive values, despite the presence of the linear normalizing condition $\sum_k A_k = 1$. It is well-known from the general theory of extrapolative and quadrature formulas that the existence of weights large in magnitude decreases the stability range, leading to an appreciable increase of the uncertainties outside of this region.

The minimization $\sum_k A_k^2 = \min$ is introduced just to avoid the above singularity at $N \sim 3M_i$ when $\varepsilon = 0$. The nonzero values of $\varepsilon^2 > 0$ allow us to effectively balance between the two kinds of the extrapolations. Of course, ε^2 cannot be chosen too large because then the main effort is directed to minimizing the squared norm $\varepsilon^2\sum_k A_k^2 = \min$ rather than the coordinate residuals. This parameter should be treated as a small quantity aiming at improving the quality of solvation force extrapolation. Optimal values of ε^2 can be found in actual simulations to obtain the best accuracy.

2.2.4. Extending the Basis Set and Selecting the Best Subset. Evidently, the accuracy of the force extrapolation should increase with the number of basis points N. However, we cannot put N to be too large because then the number of linear equations increases, too. These eqs 17 need to be solved frequently (in total $P = h/\Delta t \gg 1$ times per h), namely, at each inner point inside the outer interval $h \gg \Delta t$ for the minimization of coordinate residuals 11. As a result, the computational overhead can be unacceptably high at large enough values of N, reducing the efficiency of MD/3D-RISM-KH simulation.

A way to remedy the above situation lies in the following. We can extend the basis set from a relatively small number of $N \lesssim 100$, say, to a larger value $N' \gg N$ by collecting the force-

DOI: 10.1021/ct5010438
J. Chem. Theory Comput. 2015, 11, 1875−1895

coordinate pairs during a broad previous time interval $\Delta H = N'h \gg Nh$. Then the weighting squared distances in the $3M$-dimensional space between the transformed basis outer coordinates $\mathbf{R}_{ij,k'}$ and the current origin point $\mathbf{r}_{ij}^*$

$$\mathcal{R}_{ik'}^2(t) = \frac{1}{M_i}\sum_{j=1}^{M} {}'w(r_{ij}^*)(\mathbf{R}_{ij,k'} - \mathbf{r}_{ij}^*)^2 \equiv \vartheta_{1,(i,k')} \tag{21}$$

can be readily expressed for each $i = 1, 2, ..., M$ and $k' = 1, 2, ..., N'$ in terms of the smallest eigenvalues $\vartheta_{1,(i,k')}$ [see the text after eq 7]. Now these distances can be sorted in the ascending order, and the first N most closest points can be selected among the extended set to satisfy the condition $\mathcal{R}_{i1} < \mathcal{R}_{i2} < ... < \mathcal{R}_{iN}$. The forces $\mathbf{F}_{i,k'}$ must be resorted synchronically with the coordinates $\mathbf{R}_{ij,k'}$ to form the best pair subset with N points. It should then be used when performing the advanced extrapolation 16.

The above procedure can further improve the quality of the extrapolation, especially at $N' \gg N$. The reason is that the choice of the nearest outer pairs in the transformed space additionally reduces the coordinate region in which the extrapolation is performed. This leads to a decrease of the coordinate residuals and, as a consequence, to an increase of the accuracy. In fact, such an additional reduction minimizes the change in the transformed solvation forces during torsion motion of the solute. Note that such motion (characterizing by large amplitudes) is responsible for transitions of the biomolecule from one conformational pool to another where the torsion potential has a local minima. Thus, an optimal value for the expanded interval $\Delta H = N'h$ should be of order of the mean lifetime in local conformational minima. Then, whenever the transition to other conformations occurs, we can quickly reselect the subset to fit the basis outer points to the current solute conformation. The accuracy of such fitting is especially high if the molecule has already been near this conformation at previous times.

Worth remarking is that the selecting procedure at $N \ll N'$ requires only little extra numerical efforts even for large enough N' of order of several thousands. This is because the computational cost grows with N' just slightly. Indeed, the selection operates on only the smallest eigenvalues $\vartheta_{1,(i,k')}$ (and not on eigenvectors) of small 4×4 matrices, and so the computational time scales linearly with N' (with a small proportionality factor). The eigenvectors $\mathbf{q}_{1,(i,k)}$ are necessary only for the best subset with $k = 1, 2, ..., N \ll N'$ to build the transformation matrix $\mathbf{S}_{i,k}$ for the extrapolation 16. On the other hand, the overhead increases much more rapidly with N, namely, proportionally to $(N+1)^3$, which is required to find solutions to $(N+1)$ linear eqs 17. In addition, the selection procedure is performed only once per many ($p \gg 1$) inner time steps, further lowering the computational costs.

2.2.5. The Whole Algorithm of GSFE. With the techniques laid out in the preceding subsections, the resulting generalized solvation force extrapolation (GSFE) algorithm can be briefly described as follows.

At the very beginning, the 3D-RISM-KH integral equations are converged after each Δt of the N first inner steps with no extrapolation to fill out the basis set. Then the extrapolation starts with N points, and the extended N'-set is accordingly completed step by step in the integration process. Since h can be much larger than Δt, we cannot put the outer step to be immediately equal to $h \gg \Delta t$. The reason is that then the extrapolation skews because of the significant nonuniformity of the time intervals between the points from the set. This issue can be remedied in such as a way that the outer time interval is smoothly increased every inner step from Δt to h with an increment of Δt.

Further, after each $p\Delta t$ step, we solve the eigenvalue problem 6 and 7 for the extended set with N' coordinates. The first $N < N'$ points are selected by sorting the corresponding smallest eigenvalues 21 in the ascending order. The coordinates and forces related to the subset obtained are then transformed by the local non-Eckart-like rotations 14 in terms of the $\mathbf{S}$-matrix 4 constructed on the N eigenvectors. Having the transformed coordinates, we build the system of $(N+1)$ linear eqs 17 and solve it for the expansion coefficients. Note that the inversion of the $(N+1) \times (N+1)$ matrix in eq 17 is performed only once per p inner steps because it remains unchanged during time $p\Delta t$ [see eq 18], while the right-hand side vector in eq 17 varies every Δt [eq 19].

Using the expansion coefficients, the solvent forces are extrapolated at each inner step Δt within the outer time interval $t \in]t_1, t_0[$ of length h as the weighted sum of their N previous outer transformed values, followed by the inverse transformation 16. The extrapolation procedure is applied $h/\Delta t$ times to achieve the next outer point. At that point, the solvent forces are calculated explicitly by solving the 3D-RISM-KH integral equations. The extended N'-set is then updated by the new outer force-coordinate pair, while the oldest one is discarded. All these actions are repeated H/h times for the next outer intervals until the desired simulation time length H is achieved.

This completes the derivation of the GSFE algorithm. It improves and generalizes our previous ASFE approach[33] to the case of arbitrary solute biomolecules. Formally replacing $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ with $\mathbf{r}_i - \mathbf{r}_c$ and $\mathbf{r}_{ij}^*$ with $\mathbf{r}_{i,1}$, where $\mathbf{r}_c$ is the center of mass of the solute molecule, as well as putting $w \equiv 1$ with $R \to \infty$ and $M_i \equiv M$, we come to the global non-Eckart transformation used in ASFE.[33] For large molecules, however, the global rotations appear to be small, thus having practically no effect on extrapolation improvement. Furthermore, in ASFE the extended set and the inversion of the $(N+1) \times (N+1)$ matrix were carried out at each inner time step ($p = 1$),[33] which significantly lowered the efficiency of the computations. In GSFE, applying the frequency regime with $1 \ll p \ll P = h/\Delta t$ appreciably reduces computational cost. This is achieved without loss of precision since the increased distances during time $p\Delta t$ are compensated by the inverse non-Eckart-like local transformations.

It is worth pointing out also that the rotational superpositions 2 and 3 introduced for GSFE look somewhat similar to those originally derived in the context of analyzing macromolecular structures obtained in conventional MD or experiment.[69] Note that there is no unique approach to separate translational, angular, and internal motions of these molecules. Within the well-known Eckart scheme[70−75] such a separation can be carried out unambiguously only for molecular structures with one equilibrium state. It does not work for more complicated molecules where two or more local equilibrium states can exist. This problem was solved[69] by exploiting Gauss' principle of least constrain by minimizing the coordinate deviation norm $(1/M)\sum_{i=1}^{M} m_i(\delta\mathbf{S}(t,\delta t)\mathbf{r}_i(t+\delta t) - \mathbf{r}_i(t))^2 = \min$, where m_i is the mass of the ith atom and Δt is the time step. For instance, in the limit $\delta t \to 0$, this allows us to uniquely define the angular velocity $\mathbf{\Omega}(t) = \mathbf{1}(t)$ of an arbitrary molecule at any time t, where $\delta\mathbf{S}(t,\delta t) \equiv \mathbf{S}(\mathbf{\Delta q}(t))$ with $\mathbf{\Delta q}(t)$

DOI: 10.1021/ct5010438
J. Chem. Theory Comput. 2015, 11, 1875−1895

$=\{\cos(\delta\phi/2), \mathbf{1}(t)\sin(\delta\phi/2)\}$ is the matrix of rotation of the whole molecule by angle ϕ around the unit vector $\mathbf{1}$ passing through the center of mass. The standard Eckart method appears as a particular case of the non-Eckart approach when the number of local equilibrium states is equal to one. In the absence of internal degrees of freedom the non-Eckart angular velocity completely coincides with the well-known definition for rigid bodies.

Our non-Eckart-like superposition scheme differs in several aspects from the original non-Eckart method.[69] It is modified by normalizing weights and applied individually for each reference atom of the solute molecule at discrete moments of time. This results in local reorientations of atomic groups instead of those in rotation of the molecule as a whole. Moreover, the newly introduced non-Eckart-like scheme is aimed at optimizing the performance of MD simulations rather than at only analyzing simulation or experimental data.[69,76−78]

3. COUPLING GSFE WITH MTS-MD/3D-RISM-KH IN THE OPTIMIZED ISOKINETIC NOSÉ−HOOVER ENSEMBLE

3.1. 3D-RISM-KH Molecular Theory of Solvation. The 3D-RISM-KH theory[24−29] yields the solvation structure in terms of the normalized 3D density distribution functions $g_\alpha(\mathbf{r})$ of interaction site α of solvent molecules at spatial position $\mathbf{r}$ around the whole solute macromolecule or supramolecule, starting from the input of intermolecular potentials for explicit solvent and solute molecules (molecular force field). The 3D-RISM integral equation for 3D solute−solvent site correlation functions[20−24,27] can be derived either within the site formalism of density functional theory of molecular liquids[20−22] or from the six-dimensional Ornstein−Zernike integral equation for molecular liquids[79] by orientations averaging centered at interaction sites of solvent molecules to contract orientational degrees of freedom of solvent.[23,24,27] It reads

$$h_\alpha(\mathbf{r}) = \sum_\gamma \int d\mathbf{r}' c_\gamma(\mathbf{r}-\mathbf{r}')\chi_{\gamma\alpha}(r') \tag{22}$$

where $h_\alpha(\mathbf{r})$ and $c_\alpha(\mathbf{r})$ are respectively the 3D total and direct correlation functions of solvent site α around the solute molecule, $\chi_{\gamma\alpha}(r)$ is the radially dependent site−site susceptibility function of solvent which is an input to 3D-RISM and is calculated beforehand, and the indices γ and α enumerate all interaction sites on all sorts of solvent species. Diagrammatic analysis of the total and direct correlation functions[79] relates the former to the density distribution function as $h_\alpha(\mathbf{r}) = g_\alpha(\mathbf{r})-1$, and so $h_\alpha(\mathbf{r})$ has the meaning of a normalized 3D distribution of spatial correlations, or normalized deviations of solvent site density around the solute molecule from the average value in the solution bulk. The long-range asymptotics of the 3D direct correlation function $c_\alpha(\mathbf{r})$ is given by the 3D interaction potential $u_\alpha(\mathbf{r})$ scaled by k_BT between the whole solute molecule and solvent interaction site α: $c_\alpha(\mathbf{r}) \sim -u_\alpha(\mathbf{r})/(k_BT)$ for $\mathbf{r}$ outside the short-range repulsive core region (typically comprising the repulsive slope down to the attractive well minimum); the values of $c_\alpha(\mathbf{r})$ inside the repulsive core are related to the solvation free energy. Usually but not necessarily, the 3D solute−solvent site interaction potential is given by a sum of pairwise potentials (typically Coulomb and Lennard-Jones) dependent on separations between solute and solvent interaction sites, $u_\alpha(\mathbf{r}) = \sum_i u_{i\alpha}(|\mathbf{r}-\mathbf{r}_i|)$, where $\mathbf{r}_i$ is the location of solute atom i.

The 3D-RISM integral eq 22 involves two correlation functions, $h_\alpha(\mathbf{r})$ and $c_\alpha(\mathbf{r})$, and to be complete has to be complemented with another relation between $h_\alpha(\mathbf{r})$ and $c_\alpha(\mathbf{r})$ called a closure which also involves the interaction potential $u_\alpha(\mathbf{r}_\alpha)$ specified at input with the molecular force field. The exact closure relation has a nonlocal functional form that can be represented as an infinite diagrammatic series in terms of multiple integrals of the total correlation function;[79] however, it is computationally intractable, as the series is poorly convergent and the higher-order diagrams are integrals extremely cumbersome to calculate. Therefore, the exact closure is replaced in practice with amenable approximations which should analytically ensure asymptotics of the correlation functions and features of the solvation structure and thermodynamics to properly represent the solvation physics. The approximation proposed by Kovalenko and Hirata (KH closure),[24,27,29] the 3D version of which reads as

$$g_\alpha(\mathbf{r}) = \begin{cases} \exp(-u_\alpha(\mathbf{r})/(k_BT) + h_\alpha(\mathbf{r}) - c_\alpha(\mathbf{r})) & \text{for } g_\alpha(\mathbf{r}) \le 1 \\ 1 - u_\alpha(\mathbf{r})/(k_BT) + h_\alpha(\mathbf{r}) - c_\alpha(\mathbf{r}) & \text{for } g_\alpha(\mathbf{r}) > 1 \end{cases} \tag{23}$$

couples in a nontrivial way the so-called hypernetted chain (HNC) and mean spherical approximation (MSA) closures,[79] the former automatically applied to spatial regions of density depletion $g_\alpha(\mathbf{r}) < 1$, including the repulsive core, and the latter to spatial regions of solvent density enrichment $g_\alpha(\mathbf{r}) > 1$, such as association peaks and long-range tails of near-critical fluid phases, while keeping the right long-range asymptotics of $c_\alpha(\mathbf{r})$ peculiar in both the HNC and MSA. (The distribution function and its first derivative are continuous at the joint boundary $g_\alpha(\mathbf{r}) = 1$ by construct.) The KH approximation consistently accounts for both electrostatic and nonpolar solvation forces, such as hydrogen bonding and other associative effects, hydrophobic hydration and interaction, preferential solvation, desolvation and other steric effects for macromolecules and supramolecules in simple and complex liquids, solvent mixtures, nonelectrolyte and electrolyte solutions in various chemical,[24−27,29,80−86] soft matter,[87] synthetic organic supramolecular,[29,88−90] biopolymeric,[91−93] and biomolecular[27,29,40,41,43,65,87,94−104] systems.

The radially dependent site−site susceptibility of solvent $\chi_{\gamma\alpha}(r)$ determines nonlocal response of solvent to an external field, statistically mechanically averaged over orientations and arrangements of solvent molecules in the solvation shells. In the context of eq 22, the solvent density change given by the total correlation function $h_\alpha(\mathbf{r})$ comes from the insertion of a solute molecule characterized with the direct correlation function $c_\alpha(\mathbf{r})$ which propagates across the solvation shells through the effective solvent−solvent correlations given by the solvent susceptibility $\chi_{\gamma\alpha}(r)$. The latter breaks up into the intra- and intermolecular terms

$$\chi_{\gamma\alpha}(r) = \omega_{\gamma\alpha}(r) + \rho_\gamma h_{\gamma\alpha}(r) \tag{24}$$

where the intramolecular correlation function $\omega_{\gamma\alpha}(r)$ normalized as $\int d\mathbf{r}\omega_{\gamma\alpha}(r) = 1$ represents the geometry of solvent molecules (i.e., $\omega_{\gamma\alpha}(r) = 0$ for sites γ and α on different species). For rigid molecular species with site separations $l_{\gamma\alpha}$ it has the form $\omega_{\gamma\alpha}(r) = \delta(r-l_{\gamma\alpha})/(4\pi l_{\gamma\alpha}^2)$ specified in the reciprocal k-space as $\omega_{\gamma\alpha}(k) = j_0(kl_{\gamma\alpha})$, where $j_0(x)$ is the zeroth-order spherical Bessel function. ρ_γ is the average number density of solvent interaction site γ in the solution bulk. The radially dependent site−site total correlation function $h_{\gamma\alpha}(r)$ for

DOI: 10.1021/ct5010438
J. Chem. Theory Comput. 2015, 11, 1875−1895

all pairs of sites on all species of the solvent are obtained in advance to the 3D-RISM-KH calculations from the dielectrically consistent RISM theory[105,106] coupled with the KH closure relation for the radial correlation functions (DRISM-KH approach).[27−29] The DRISM-KH theory can be applied to solution systems of a given composition in a wide range of thermodynamic conditions, including different solvents,[107,108] solvent mixtures,[109,110] polymeric solutions,[87,111] and electrolyte solutions.[27,29,112]

An important feature of the KH closure 23, is that the solvation free energy μ_{solv} as determined by Kirkwood's thermodynamic integration gradually switching the interactions on from 0 to the full potential $u_\alpha(\mathbf{r})$ is obtained analytically in a closed form of a single spatial integral in terms of the correlation functions[24,27,29]

$$\mu_{\text{solv}} = k_B T \sum_\alpha \rho_\alpha \int d\mathbf{r}\left(\frac{1}{2}(h_\alpha(\mathbf{r}))^2\Theta(-h_\alpha(\mathbf{r})) - \frac{1}{2}h_\alpha(\mathbf{r})c_\alpha(\mathbf{r}) - c_\alpha(\mathbf{r})\right) \tag{25}$$

where the sum goes over all the sites of all solvent species, and $\Theta(x)$ is the Heaviside step function. Other thermodynamic quantities can also be obtained analytically by taking the corresponding derivaties. In particular, mean solvation forces acting on each atom of the solute macromolecule or supramolecule are readily obtained as a simple 3D spatial integral in terms of the 3D site distribution functions $g_\alpha(\mathbf{r})$ calculated by converging the 3D-RISM-KH integral eqs 22 and 23 (see the next section).

To properly treat electrostatic forces in electrolyte solution with polar molecular solvent and ionic species, the long-range electrostatic asymptotics of both the 3D direct and total correlation functions in the 3D-RISM integral eq 22 are separated out and handled analytically.[25−28,82,97,113] The remaining short-range parts of the 3D site correlation functions are discretized on a uniform rectangular 3D grid in a box large enough to accommodate the solvation structure, typically 2 to 3 solvation shell oscillations. The spatial convolution of the short-range therm in eq 22 is calculated by means of 3D fast Fourier transform. Note that even though the solvent susceptibility $\chi_{\gamma\alpha}(r)$ has a long-range electrostatic part, no aliasing occurs in the backward 3D-FFT of the short-range part of $h_\alpha(k)$ on the 3D box supercell since, for the physical reason, it typically contains merely 2−3 oscillations and thus vanishes at the box boundaries.[28] The same analytical treatment with separation of the electrostatic asymptotics is applied also to the radial site−site correlation functions in the DRISM-KH integral equations that produce the water susceptibility 24, as well as to the 3D site correlation functions in the solvation free energy integral 25 which is reduced to a 3D integral of the short-range terms on the 3D box and one-dimensional integrals of the asymptotics easy to compute.[25−28,82,97]

The 3D-RISM-KH integral eqs 22 and 23 are converged by using the modified algorithm of direct inversion in the iterative subspace (MDIIS).[25−28,114,115] The MDIIS numerical solver accelerates convergence of integral equations of liquid state theory by optimizing each iterative solution in a Krylov subspace of typically last 10−20 successive iterations and then making the next iterative guess by mixing the optimized solution with the approximated optimized residual.

The computational expenses of converging the 3D-RISM-KH equations can be significantly reduced with several strategies, including a high-quality initial guess for the 3D direct correlation functions $c_\alpha(r)$; pre- and postprocessing of the 3D solute−solvent potentials $u_\alpha(r)$, the long-range asymptotics of the 3D correlation functions $c_\alpha(r)$, and $h_\alpha(r)$, and forces; several cutoff schemes and an adaptive solvation box.[31] Further, memory and corresponding CPU load in the MDIIS numerical solver are decreased by up to an order of magnitude using the core−shell-asymptotics treatment of solvation shells.[28]

3.2. Combining MD with 3D-RISM-KH. Unlike conventional MD dealing with trajectories of explicit solvent molecules, the hybrid MD/3D-RISM-KH approach[30−32] contracts them to 3D site density distribution functions $g_\alpha(\mathbf{r})$ of quasiequilibrium solvent at successive conformations of the biomolecule and thus performs quasidynamics of the biomolecule steered with mean solvation forces. The latter can be determined in a general case from Kirkwood's thermodynamic charging integral by differentiation with respect to solute atomic coordinates. For the solvation free energy not dependent on a thermodynamic integration path (which is true for the exact solvation free energy but not necessarily for a given closure approximation of integral equation theory of liquids), the mean solvation force is immediately obtained as the "detailed" solute−solvent site interaction potential force averaged over the solvation shells with the 3D solute−solvent site density distribution function[30,31]

$$\mathbf{f}_i \equiv \mathbf{f}(\mathbf{r}_i) = -\frac{\partial \mu_{\text{solv}}}{\partial \mathbf{r}_i} = \sum_\alpha \rho_\alpha \int d\mathbf{r} g_\alpha(\mathbf{r}) \frac{\partial u_{i\alpha}(\mathbf{r}-\mathbf{r}_i)}{\partial \mathbf{r}_i} \tag{26}$$

where $u_{i\alpha}(\mathbf{r}-\mathbf{r}_i)$ is the pairwise isotropic interaction potential between solute atom i located at position $\mathbf{r}_i$ and solvent site α at $\mathbf{r}$. With the solute−solute forces evaluated directly as $-\partial U_1/\partial \mathbf{r}_i$ (see Section 2.1) and the effect of solvent accounted with mean solvation forces 26, the equations of quasidynamic motion are solved only for solute atoms. In the approach of an adaptive box, the 3D-RISM-KH integral eqs 22 and 23 are discretized and converged on a grid in a nonperiodic box of size and shape that includes about 2−3 solvation shells around the biomolecule to minimize boundary effects and varies during the simulation adjusting to solute conformational changes so as to optimize computational load.[31] This is different from conventional MD which typically uses a periodic rectangular box and the Ewald summation technique[116,117] to evaluate long-range electrostatic interactions.

3.3. MTS-MD in OIN Ensemble Steered with Extrapolated 3D-RISM-KH Mean Solvation Forces. The equations of motion for solute atoms in hybrid MD/3D-RISM-KH simulation in the canonical-isokinetic OIN ensemble steered with 3D-RISM-KH mean solvation forces which are extrapolated with the GSFE technique can be cast in the compact form[32]

$$\frac{d\mathbf{\Gamma}}{dt} = L\mathbf{\Gamma}(t) \tag{27}$$

where $\mathbf{\Gamma} = \{\mathbf{r},\mathbf{v};\varsigma,\mathbf{w}\}$ denotes the extended phase space and L is the Liouville operator. The extended space, apart from the full set of coordinates $\mathbf{r} \equiv \{\mathbf{r}_i\}$ and velocities $\mathbf{v} \equiv \{\mathbf{v}_i\}$ of all solute atoms, includes also all thermostat frequencies $\omega \equiv \{\omega_{\kappa,i}\}$ with $\kappa = 1, ..., \mathcal{K}$ and their conjugated dynamical variables $\varsigma \equiv \{\varsigma_i\}$. The latter are introduced by means of the relation $d\varsigma_i/dt =$

DOI: 10.1021/ct5010438
J. Chem. Theory Comput. 2015, 11, 1875−1895

$(\tau_i^2\omega_{1,i}^2\omega_{2,i} - \sum_{\kappa=2}^{\mathcal{K}}\omega_{\kappa,i})$, where $\mathcal{K}$ is the number of chains per thermostat. The Liouvillian can be split up as

$$L = \sum_{i=1}^{M}(\mathcal{A}_i + B_i + C_{v,\omega,i} + C_{\omega,i} + C_{\varsigma,i}) \tag{28}$$

into the kinetic $\mathcal{A}_i = \mathbf{v}_i\cdot\partial/\partial\mathbf{r}_i$, potential

$$B_i = \left(\frac{\boldsymbol{f}_i}{m_i} - \mathbf{v}_i\frac{\mathbf{v}_i\cdot\boldsymbol{f}_i}{2\mathrm{T}_i}\right)\cdot\frac{\partial}{\partial\mathbf{v}_i} - \frac{\mathbf{v}_i\cdot\boldsymbol{f}_i}{2\mathrm{T}_i}\omega_{1,i}\frac{\partial}{\partial\omega_{1,i}} \tag{29}$$

and chain-thermostat parts with

$$C_{v,\omega,i} = \frac{\tau_i^2\omega_{1,i}^2}{4}\omega_{2,i}\mathbf{v}_i\cdot\frac{\partial}{\partial\mathbf{v}_i} + \left(\frac{\tau_i^2\omega_{1,i}^2}{4} - 1\right)\omega_{1,i}\omega_{2,i}\frac{\partial}{\partial\omega_{1,i}} \tag{30}$$

$$C_{\omega,i} = \sum_{\kappa=2}^{\mathcal{K}}\left(\omega_{\kappa-1,i}^2 - \frac{1}{\tau_i^2} - \omega_{\kappa+1,i}\omega_{\kappa,i}\right)\frac{\partial}{\partial\omega_{\kappa,i}} \tag{31}$$

$$C_{\varsigma,i} = -(\tau_i^2\omega_{1,i}^2\omega_{2,i} - \sum_{\kappa=2}^{\mathcal{K}}\omega_{\kappa,i})\frac{\partial}{\partial\varsigma_i} \tag{32}$$

In the canonical OIN ensemble[32] each atom is coupled with its own thermostat by imposing the constraint $\mathrm{T}_i = 3k_\mathrm{B}T/2$, where

$$\mathrm{T}_i = \frac{m_i\mathbf{v}_i^2}{2} + \frac{3k_\mathrm{B}T}{4}\frac{\tau_i^2\omega_{1,i}^2}{2} \tag{33}$$

is the full kinetic energy of the ith subsystem. The quantity τ_i is related to the relaxation time, determining the strength of coupling of atom i with its thermostat.

The total forces $\boldsymbol{f}_i = \mathbf{f}_{i(\mathrm{f})} + \mathbf{f}_i$ are now divided into the fast solute−solute component $\mathbf{f}_{i(\mathrm{f})}$ and slow 3D-RISM-KH solute−solvent one $\mathbf{f}_i$. In view of eq 29, this results in the corresponding splitting of the potential operator as $B_i(\{\boldsymbol{f}_i\}) = B_i(\{\mathbf{f}_{i(\mathrm{f})}\}) + B_i(\{\mathbf{f}_i\}) \equiv B_\mathrm{f} + B_\mathrm{s}$. Mention that the solute−solute forces $\mathbf{f}_{i(f)}$ are calculated always directly (by $-\partial U_1/\partial\mathbf{r}_i$, see Section 2.1), while the 3D-RISM-KH solute−solvent mean forces $\mathbf{f}_i$ are either evaluated explicitly in the form of eq 26 or approximated with $\tilde{\mathbf{f}}_i$ using the transformation 16, as described in Section 2.2. Then $B_\mathrm{s}(\{\mathbf{f}_i\})$ transforms to $B_\mathrm{s}(\{\tilde{\mathbf{f}}_i\}) \equiv \tilde{B}_\mathrm{s}$.

Thus, using the MTS decomposition method,[55,61,62,75] the solution $\boldsymbol{\Gamma}(h) = e^{Lh}\boldsymbol{\Gamma}(0)$ to eq 27 over the outer time interval h from an initial state $\boldsymbol{\Gamma}(0)$ can be presented[32] as the following product of exponential operators:

$$\boldsymbol{\Gamma}(h) = \prod_{n'=1}^{n} e^{C\delta t/2}e^{B_\mathrm{fs}^{(n')}\delta t/2}e^{\mathcal{A}\delta t}e^{B_\mathrm{fs}^{(n')}\delta t/2}e^{C\delta t/2}\boldsymbol{\Gamma}(0) + O(\delta t^2) \tag{34}$$

Here, $n = h/\delta t \gg 1$ is the total number of subinner time steps with length $\delta t \ll \Delta t$ each, $C = C_{v,\omega} + C_\omega + C_\varsigma$,

$$e^{B_\mathrm{fs}^{(n')}\delta t/2} = \begin{cases} e^{B_\mathrm{f}\delta t/2}e^{B_\mathrm{s}\Delta t/2} & \text{only once per } h \text{ when } n'/\dfrac{\Delta t}{\delta t} = 1 \\ e^{B_\mathrm{f}\delta t/2}e^{\tilde{B}_\mathrm{s}\Delta t/2} & \text{for other inner steps, } n' = 2\dfrac{\Delta t}{\delta t}, \ldots, n \\ e^{B_\mathrm{f}\delta t/2} & \text{for all other } n' \end{cases} \tag{35}$$

is the generalized velocity propagator, Δt is the inner ($\delta t \ll \Delta t \ll h$) time step, $O(\delta t^2)$ is the accuracy of the decomposition, and the subscript i is omitted for the sake of simplicity. Note that we should first update (by $e^{C\delta t/2}$ and $e^{B\delta t/2}$) the complete set of velocities $\mathbf{v}_i$ and frequencies $\omega_{\kappa,i}$ belonging to all atoms (i = 1, 2, ..., M) and thermostat chains (κ = 1,..., $\mathcal{K}$) before to change the coordinates $\mathbf{r}_i$ of all particles by $e^{\mathcal{A}\delta t}$. A nice feature of the OIN decomposition is that the action of all the single-exponential operators which arise in eqs 34 and 35 on $\boldsymbol{\Gamma}$ can be handled analytically using elementary functions.[32]

Therefore, the propagation $\boldsymbol{\Gamma}(t) = [\boldsymbol{\Gamma}(h)]^{t/h}$ of dynamical variables from their initial values $\boldsymbol{\Gamma}(0)$ to arbitrary future time t can be performed by consecutively applying the single exponential transformations of a phase space point $\boldsymbol{\Gamma}$ in the order defined in eqs 34 and 35. As can be seen, the fastest B_f-component of motion is integrated most frequently, namely, $n = h/\delta t$ times per outer interval h with the smallest (subinner) time step δt, while the (original or approximated) slow 3D-RISM-KH forces are applied impulsively only every $\Delta t/\delta t$ subinner step, i.e., $h/\Delta t < n$ times. Note that almost all these impulses (when $\Delta t \ll h$) are obtained by employing the extrapolated 3D-RISM-KH forces 16 in terms of the operator $\tilde{B}_\mathrm{s}$, and the explicit 3D-RISM-KH calculations 26 are used in B_s only once per outer time interval h. Taking into account that the solute−solute forces are much cheaper to evaluate than the solute−solvent ones, obvious speedup is achieved as compared to single time step propagation (n = 1, $\delta t = h$) without extrapolation ($\Delta t = h$). Furthermore, the existence of the impulsive inner time steps of length $\Delta t > \delta t$ gives a possibility of reducing the number of (either extrapolative or direct) 3D-RISM-KH evaluations from $h/\delta t$ to $h/\Delta t$. Finally, applying the GSFE approach allows further significant improvement of the overall efficiency, since 3D-RISM-KH calculations which are the most expensive are performed just once per outer step h.

It is worth emphasizing that the presence of OIN thermostatting terms in B and C [see eqs 29, 30, 31, and 32] eliminates MTS resonance instabilities. The latter appear in conventional MD simulation (in the microcanonical and canonical ensembles) already at relatively small values of outer time steps (see Introduction in Section 1). In hybrid MD/3D-RISM-KH simulation, additional instability can come from uncertainties caused by the approximate character of solvent force extrapolation. The canonical-isokinetic OIN propagation 34 efficiently eliminates both these types of instabilities by imposing the individual isokinetic constraint 33 on each solute atom. Moreover, the accurate GSFE approach makes it possible to drastically increase the sizes of inner and especially outer time steps compared to standard integration.

In view of the above, the following hierarchy of time steps

$$\delta t \ll \Delta t \ll h \ll Nh \ll N'h = \Delta H \ll H \tag{36}$$

should be set in order to achieve optimal performance of hybrid MTS-MD/OIN/3D-RISM-KH simulation using the GSFE approach. We thus have up to five time scales. This completes coupling of GSFE with the MTS-MD/OIN/3D-RISM-KH method. We will refer to the resulting scheme as hybrid MTS-MD/OIN/GSFE/3D-RISM-KH simulation or OIN/GSFE/3D-RISM-KH for brevity.

Strictly speaking, quasidynamic behavior obtained in MTS-MD/OIN/GSFE/3D-RISM-KH simulation differs from true dynamics in conventional MD with explicit solvent. In particular, such quasidynamics does not obey the Maxwell velocity distribution, and thus, unlike microcanonical MD, cannot get us real time correlation functions. However, as we have proven rigorously,[32] the configurational part of the

DOI: 10.1021/ct5010438
J. Chem. Theory Comput. 2015, 11, 1875−1895

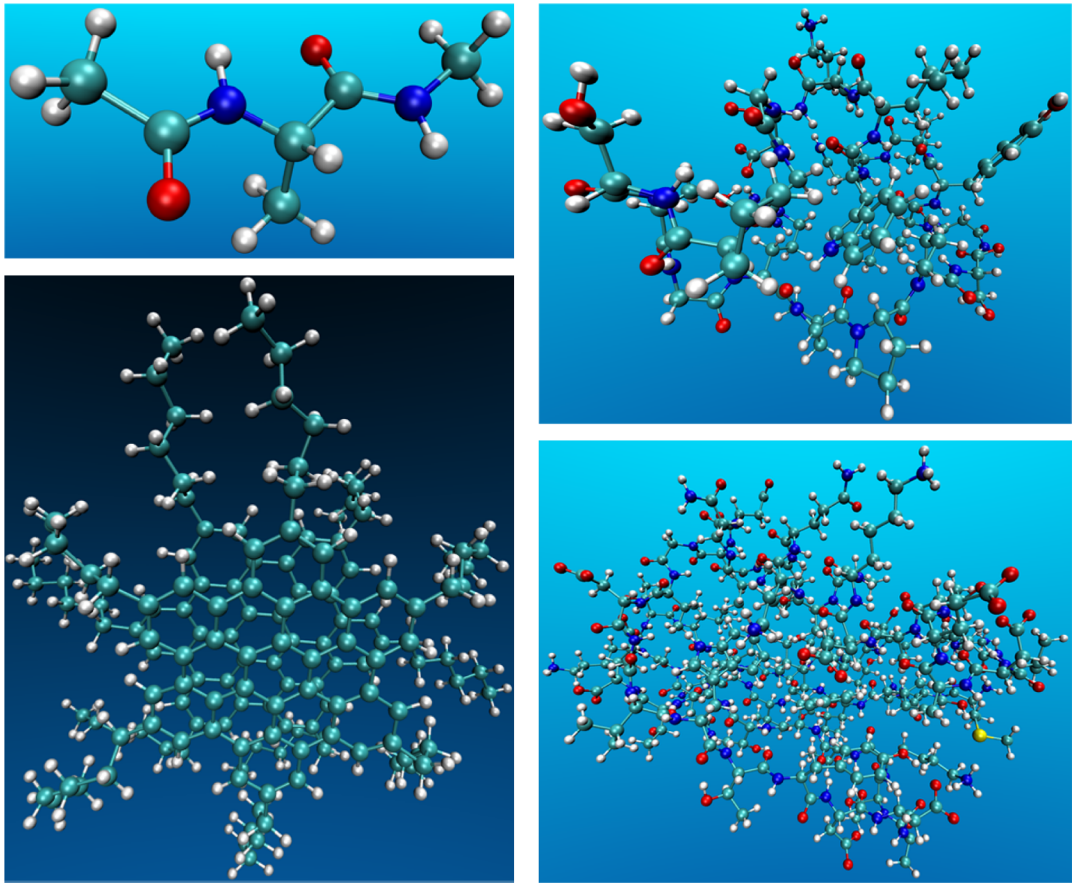


Figure 1. Ball-and-stick representation of the molecular structures of alanine dipeptide (22 atoms, upper left part), asphaltene dimer (336 atoms, lower left), miniprotein 1L2Y (304 atoms, upper right), and protein G (862 atoms, lower right). Color scheme: white (hydrogen), red (oxygen), green (carbon), blue (nitrogen), yellow (sulfur).

extended partition function obtained in hybrid MTS-MD/OIN/GSFE/3D-RISM-KH simulation at targeted temperature T does coincide with the true canonical distribution of the physical system in coordinate space. This is a very important feature because the original conformational properties, including spatial atom−atom density distribution functions of the solute macromolecule (or supramolecule), are readily reproduced in MTS-MD/OIN/GSFE/3D-RISM-KH simulation. Such quasidynamic sampling appears to be much more efficient than that following from "real-time" (microcanonical or canonical) MD simulations.

4. VALIDATION OF THE HYBRID MTS-MD/OIN/GSFE/3D-RISM-KH INTEGRATOR ON DIFFERENT SOLUTE−SOLVENT SYSTEMS

4.1. Simulation Details. We will now examine the proposed OIN/GSFE/3D-RISM-KH approach in actual simulations for fully flexible models of the alanine dipeptide (M = 22) in water solvent, asphaltene (M = 336) in toluene solvent, miniprotein 1L2Y (M = 304), and protein G (M = 862) in water solvent. The structures of the four solute molecules considered are shown in the ball-and-stick representation in Figure 1. The Amber03,[118] Amber99SB,[119] and general Amber[120] force fields were used to describe the interactions in the alanine dipeptide and miniprotein 1L2Y, in protein G, and in the asphaltene dimer, respectively. Water was represented with the modified cSPC/E model.[27,30,31] The parameters for toluene solvent were taken from the optimized potentials of the general Amber force field.[121] We applied free boundary conditions and an adaptive solvation box with varying sizes along all the three axes determined by the current size of the solute molecule plus a buffer space of width r_b = 10 Å. Note that the mean diameters of the alanine-dipeptide, asphaltene, miniprotein 1L2Y, and protein G molecules are about 9, 28, 26, and 42 Å, respectively. The cutoff radius of the solute−solvent interactions was set to r_c = 14 Å. No truncation was made for the solute−solute forces. The 3D-RISM-KH integral equations were discretized on a rectangular grid of resolution δr = 0.5 Å and converged to a relative root-mean-square residual tolerance of $\delta\epsilon = 10^{-4}$ by using the MDIIS accelerated numerical solver of integral equations.[27] Further increase of the cutoff radius r_c and the buffer size r_b, as well as refinement of the grid resolution δr and the accuracy $\delta\epsilon$, did not affect the results noticeably. The three chains ($\mathcal{K}$ = 3) per each atom have been employed in the OIN ensemble. The correlation times in all the OIN thermostats were set to the same value $\tau_i \equiv \tau$ = 50 fs for asphaltene and 25 fs for proteins. The subinner and inner time steps were chosen to be always equal to δt = 1 fs and Δt = 8 fs, respectively.

Eight OIN/GSFE/3D-RISM-KH simulation series with outer time steps of h = 12, 24, 96, 200, 400, 1000, 2000, and 4000 fs were carried out to obtain the whole pattern (Section 4.2) on generalized solvation force extrapolation (GSFE) accuracy. In each of these series, the number of points of the basis and extended sets varied in the ranges $1 \le N \le 96$ and $N \le N' \le 4000$, respectively. Several frequency values from the

DOI: 10.1021/ct5010438
J. Chem. Theory Comput. 2015, 11, 1875−1895

interval $1 \le p \le 25$ were also utilized. The balancing parameter was set to $\varepsilon^2 = 0.005$ for alanine dipeptide and proteins in water and 0.015 for asphaltene in toluene. An optimal truncation radius of $R = 7$ Å was used for the weighting function. For comparison with the GSFE scheme, simulations were performed also with the original SFE technique[31] and the ASFE approach.[33] The main runs of OIN/GSFE/3D-RISM-KH simulations were performed at $h = 1000$ and 2000 ps with $N = 56$ and $N' = 1000$ (or $N' = 4000$) as well as $p = 25$ to investigate conformational properties (Sections 4.3 and 5).

The source files `SANDER.F`, `RUNMD.F`, `MDREAD.F`, `MD.H`, `AMBER_RISM_INTERFACE.F`, and `FCE_C.F` from the original Amber 11 package[122] were modified to run hybrid MTS-MD/OIN/GSFE/3D-RISM-KH simulation in parallel by the compiled module `SANDER.RISM.MPI`. In the first two Fortran modules, we implemented a program code for solving the canonical-isokinetic OIN equations of motion by adapting the velocity-Verlet like version of the decomposition integration 34 to its leapfrog-like counterpart. (The differences between the standard velocity-Verlet and leapfrog schemes are discussed elsewhere.[66,123]) The third and fourth modules were altered to organize the input/output for new parameters and observable quantities. The last two Fortran modules were rewritten to implement the GSFE procedures.

For comparison, we also performed the conventional canonical MD simulations of the miniprotein and protein G in explicit water described with the TIP3P[124] and SPC/E[125] models, respectively. We deal accordingly with 16895 and 4526 water molecules with 8 and 14 Å cutoffs in the direct space for nonbonded electrostatic interactions. The truncation terms were handled by the particle-mesh Ewald summation method[116] with periodic boundary conditions. The equations of motion were solved with a single time step of $\Delta t = 2$ fs and no extrapolation ($h = \Delta t$) exploiting the Langevin algorithm[56] at a friction viscosity of $\gamma = 1$ ps^{-1} as well as SHAKE[126,127] to fix bonds involving hydrogen.

All the runs were carried out at temperature $T = 300$ K, water solvent density 1 g/cm^3, and toluene solvent density 0.87 g/cm^3. The total duration of the MD simulations was extended to 25 and 50 ns. The simulations of miniprotein and protein G started from the folded crystal conformations obtained in NMR experiment, taken from PDB (protein data bank) structures 1L2Y[128] and 1P7E,[129] respectively. The initial structure of the asphaltene dimer was based on the full geometry optimization using density functional theory at the ωB97X-D/6-31G* level.[130]

4.2. Accuracy of Mean Solvation Force Extrapolation.

The accuracy of mean solvation force extrapolation was estimated by measuring the relative mean square deviations

$$\Sigma = \frac{1}{2}\frac{\langle \sum_{i=1}^{M} (\tilde{\mathbf{f}}_i - \mathbf{f}_i)^2 \rangle^{1/2}}{\langle \sum_{i=1}^{M} \mathbf{f}_i^2 \rangle^{1/2}} \tag{37}$$

of the extrapolated forces $\tilde{\mathbf{f}}_i$ from their original values $\mathbf{f}_i$ calculated directly (i.e., from the 3D-RISM-KH integral equations) at each outer time step, where $\langle\ \rangle$ denotes the statistical averaging along the whole simulation length. Note that between two successive outer time steps the deviations increase from zero when the inner coordinates coincide with those of the first basis point to maximal values at the end of the current outer time interval, and so a factor of 1/2 appears in the average value. It is worth remarking also that such an estimate does not require any extra computational efforts since it operates on outer forces which are already known when updating the basis pair set.

The relative mean square deviations Σ obtained during the MTS-MD/OIN/3D-RISM-KH simulations of the asphaltene in toluene, hydrated miniprotein 1L2Y, and hydrated protein G by using various extrapolation approaches at most characteristic outer time steps $h = 1$, 2, and 4 ps are depicted in Figure 2

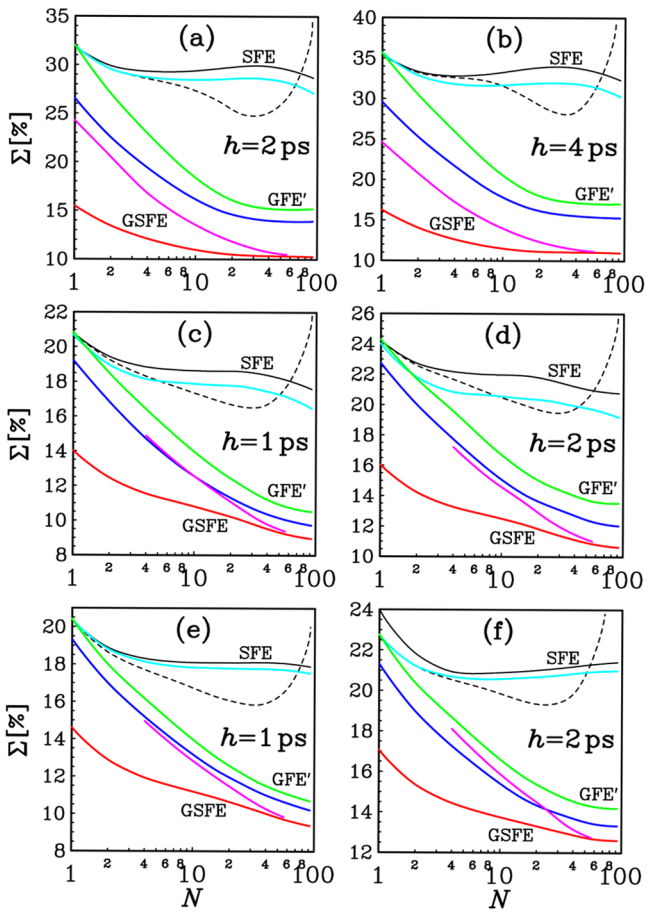


Figure 2. Uncertainty Σ versus the number of basis points N in the OIN/3D-RISM-KH simulations of asphaltene in toluene solvent [parts (a), (b)], hydrated miniprotein 1L2Y [(c), (d)], and hydrated protein G [(e), (f)] by using different extrapolation approaches [see the text for color line nomenclature] at outer time steps $h = 1$, 2, and 4 fs.

versus the number of basis points N. (Note the N-logarithmic scale was utilized for presentation convenience.) The approaches tested are the new GSFE method versus the original SFE scheme[31] as well as the intermediate version, ASFE.[33] The SFE, ASFE, and GSFE functions $\Sigma(N)$ are plotted with black, cyan, and red curves, respectively; the other methods used are described below. The corresponding dependencies of $\Sigma(h)$ on the size of the outer time step at a fixed optimal $N = 56$ are presented in Figure 3 in the whole range of h varying up to 2 or 4 ps using the same color scheme. For comparison, the case of the constant force extrapolation (CFE), i.e., SFE at $N = 1$, is included, too.

As can be seen, the SFE and CFE approaches lead to the worst accuracy of the force extrapolation with the largest deviations Σ for any values of N and h. Unlike the case of hydrated alanine dipeptide,[33] the ASFE method only slightly improves the SFE results for asphaltene and proteins. A reason

DOI: 10.1021/ct5010438
J. Chem. Theory Comput. 2015, 11, 1875−1895

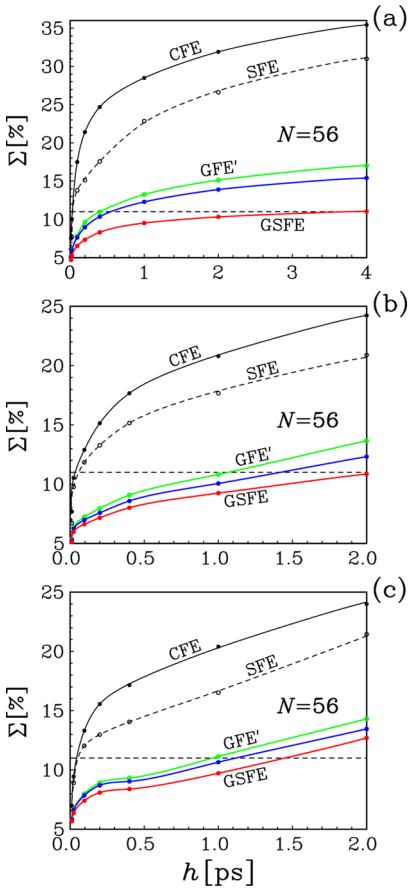


Figure 3. Uncertainty Σ against the outer time step h in the OIN/3D-RISM-KH simulations of asphaltene in toluene solvent [part (a)], hydrated miniprotein 1L2Y [(b)], and hydrated protein G [(c)] by using different extrapolation approaches [see the text for color line nomenclature] at fixed $N = 56$. Upper black solid curve: constant force extrapolation (CFE), i.e. SFE with $N = 1$.

is that both the schemes operate on all the atoms M of the solute molecule ($R = \infty$, no truncation) when constructing the deviations between the exact and extrapolated coordinates. Obviously, then for large systems with $M \gtrsim 300$, such as asphaltene and proteins, we should come to a poor force extrapolation because of $N \ll 3M$, where $3M$ is the total number of degrees of freedom. The uncertainties could be decreased to some extent with increasing N up to $3M \gtrsim 1000$, but too large numbers of $N \gtrsim 100$ are unacceptable in view of the drastic increase of the computational costs in this case [see Section 2.2.4]. The fact that the ASFE can be used with a great success[33] for relatively small solutes, such as alanine dipeptide ($M = 22$), is merely explained by their small numbers of degrees of freedom, enabling to achieve a good extrapolation already at $N \sim 3M = 66$. For much larger $3M \gtrsim 1000$ and the same $N \lesssim 100$ this is impossible even within ASFE, despite the usage of much better techniques than in SFE. On the other hand, applying only a simple truncation (at $R = 7$ Å) within SFE somewhat lowers the uncertainties at intermediate $N \sim 30$ (dashed curves in Figure 2). However, then the function $\Sigma(N)$ exhibits a singular behavior for $N \sim 3M' \sim 100$, where $M' = \langle M_i \rangle$ is the averaged number of neighbors on the distance R with respect to the reference atoms.

A significant improvement of force extrapolation accuracy for large solute molecules can be made with the following first steps [see Sections 2.2.2 and 2.2.3]: (i) when minimizing the residuals, using local rather than global coordinates, i.e. atomic positions with respect to each reference site rather than molecule center of mass; (ii) using the normalized weighting function with an optimal truncation [at $R = 7$ Å]; and (iii) balancing the normal equations by $\varepsilon^2 > 0$ to exclude the singularity at $N \sim 3M_i$. The corresponding results which take into account only these three components of GSFE are plotted in Figures 2 and 3 with green curves marked as GFE′. A further decrease of the extrapolation uncertainties is observed on the following steps: (iv) the local non-Eckart-like transformations are additionally included during the extrapolation (blue curves). Furthermore, (v) the basis set can be extended from N to $N' > N$ pairs followed by selecting the best subset with N points. We used $N' = 4000$ for the asphaltene and $N' = 1000$ for the proteins. With all these five techniques, we come to the GSFE approach which provides the best accuracy (red curves). Finally, applying the frequency scheme for $p = 25$ (magenta curves in Figure 2) at $N = 56$ gives nearly the same accuracy as for $p = 1$ but with appreciably less computational effort. Indeed, the red and magenta curves practically coincide at $N = 56$.

Computations show that the simple truncation, i.e., when $w(r) = 1$ at $r \le R$ and $w(r) = 0$ if $r > R$, without using the weighting function $w(r) = 1/r^2$ for the rotational transformations 2 and 3 and minimization of coordinate residuals 11 leads to a considerably worse accuracy in the force extrapolation. This confirms our theoretical arguments (see Sections 2.2.1 and 2.2.2) about the importance of introducing the smooth weighting $w(r) = 1/r^2$ for $r \le R$ and $w(r) = 0$ if $r > R$. The truncation radiuses $R_{opt} = 6-8$ Å appeared to be optimal in the sense that then the extrapolation uncertainties accept minimal values. Out of these radiuses we observed an increase of Σ. The reason is that for $R \gtrsim R_{opt}$ the angles of local rotations decrease, because larger molecular segments having larger mass and moments of inertia are less sensitive to the torques acting on these segments due to the interactions. This does not allow one to make the transformed solvation forces varying smooth enough with the atomic coordinates. Moreover, with increasing R, the number of degrees of freedom $3M_i$ of molecular segments around a given reference atom i increases, too. In turn, this requires a larger number N of basis coordinate-force pairs to achieve the same accuracy of the extrapolation, which increases computational cost. On the other hand, the truncation radius R cannot be chosen too small, since at $R \lesssim R_{opt}$ the boundary effects in eqs 2, 3, and 11 become too large and cannot be neglected.

Note that the formula 37 estimates, in fact, the upper limit of the extrapolation uncertainties. Indeed, it involves scalar deviations $(\tilde{\mathbf{f}}_i - \mathbf{f}_i)^2$ at the end of each outer interval h without taking into account that the force $\mathbf{f}$ is a vector which can change its direction during inner time steps. Such a change may compensate to some extent the uncertainties in velocities which are determined as a vector sum of force deviations $(\tilde{\mathbf{f}}_i - \mathbf{f}_i)$ over the inner time steps. This in turn decreases the uncertainties in coordinates as well. The fact that eq 37 overestimates the extrapolation errors is confirmed in Figure 3 where we see that $\Sigma(h)$ does not fall down to small enough values even at a tiny outer time step of $h = 12$ fs, whereas $\lim_{h\to 0}\Sigma(h) = 0$ by definition. Instead, all the dependencies $\Sigma(h)$ in Figure 3 tend to a some finite value of $\Sigma_0 \sim 5\%$ while h approaches very small steps. Thus, the most simplest way to correct the estimate given by eq 37 is merely to extract Σ_0 from Σ, i.e., $\Delta\Sigma = \Sigma - \Sigma_0$. A more accurate estimate could be to calculate the deviations at

DOI: 10.1021/ct5010438
J. Chem. Theory Comput. 2015, 11, 1875−1895

each inner time step. However, this would require enormous computational costs significantly larger than those needed for extrapolation of forces itself, making no sense to perform such kind of estimation.

Note also that the set (k = 1, 2, ..., N) of the coefficients $A_k^{(i)}$ in the extrapolation 16 is individual for each atom i. Therefore, from the equality $\sum_{i=1}^{M}\mathbf{f}_{i,k} = 0$ it does not necessarily follow that the net force $\tilde{\mathbf{f}}_{\text{net}} = \sum_{i=1}^{M}\tilde{\mathbf{f}}_i$ acting on the molecule as a whole will also be equal to zero. Calculations showed, however, that the quantity $\tilde{\mathbf{f}}_{\text{net}}$ is very small and never exceeds 0.5−1% for any h considered, indicating a high accuracy of the extrapolation. The possible effect of the nonzero net force on the results can be minimized by subtracting the term $\tilde{\mathbf{f}}_{\text{net}}/M$ for each atom, i.e., replacing $\tilde{\mathbf{f}}_i$ by $\tilde{\mathbf{f}}_i - \tilde{\mathbf{f}}_{\text{net}}/M$ every outer time step h. We used this procedure in all our OIN/GSFE/3D-RISM-KH simulations.

In view of the above, we see that the generalized extrapolation approach we propose gives a possibility of providing a much better accuracy than the previous schemes. Taking into account that the conformational properties are sensitive enough to any uncertainties, not only to the choice of force fields but to inaccuracy of the extrapolation as well, a severe criterion of $\Delta\Sigma(h_{\text{m}}) \sim 6\%$ has been imposed on the maximal allowed outer time steps h_{m}. The level $\Delta\Sigma(h_{\text{m}}) = 6\%$ is marked in Figure 3 by the horizontal dashed lines. It follows from Figure 3 that the above criterion can be satisfied by the original SFE only at very small outer time steps of order of $h \lesssim$ 24 fs. On the other hand, the same level $\Delta\Sigma(h_{\text{m}}) = 6\%$ of accuracy can be provided with our GSFE at huge sizes of outer time step up to h = 4, 2, or 1.5 ps for the asphaltene, miniprotein 1L2Y, and protein G, respectively. Such outer time steps are by 2 orders of magnitude longer than the maximal allowed outer time steps in SFE.

Even larger outer time steps up to tens of picoseconds can be achieved for simpler systems, such as hydrated alanine dipeptide. This conclusion made in our previous paper within the ASFE scheme[33] has been confirmed for GSFE as well. Note that the present results on Σ for hydrated alanine dipeptide obtained with the GSFE approach are only slightly better than those with ASFE,[33] and so both are not shown here. The possibility of using longer outer steps for smaller systems is explained by two main reasons. First, for a small solute molecule (almost) all atoms are located inside the truncation sphere (with an optimal radius R = 6−8 Å) in the rotational transformation, thus reducing all possible boundary effects to zero. Second, such systems are characterized with a small number of equilibrium states with relatively short lifetimes of order of a nanosecond or less. As a result, the extended basis set from the broad time interval $\Delta H = N'h$ can sample almost all important conformations already at N' = 1000 with $h \geq 1$ ps (then $\Delta H \geq 1$ ns), providing a high accuracy of the force extrapolation even at $h \gtrsim 10$ ps. This is very different from complex protein systems characterized with a huge number of degrees of freedom and local equilibrium states, where the lifetime of the most stable conformations can exceed many micro- and milliseconds in real time.

4.3. Conformational Properties. We checked the conformational properties in terms of atomic root-mean-square deviations (RMSD), radius of gyration R_g, and tertiary structure of the solute molecules. The RMSDs were calculated with respect to the initial conformations mentioned in Section 4.1, which included only the most massive atoms, C for the asphalatene and C_α for the proteins. The radius of gyration was computed for all solute atoms. Figure 4 shows the backbone

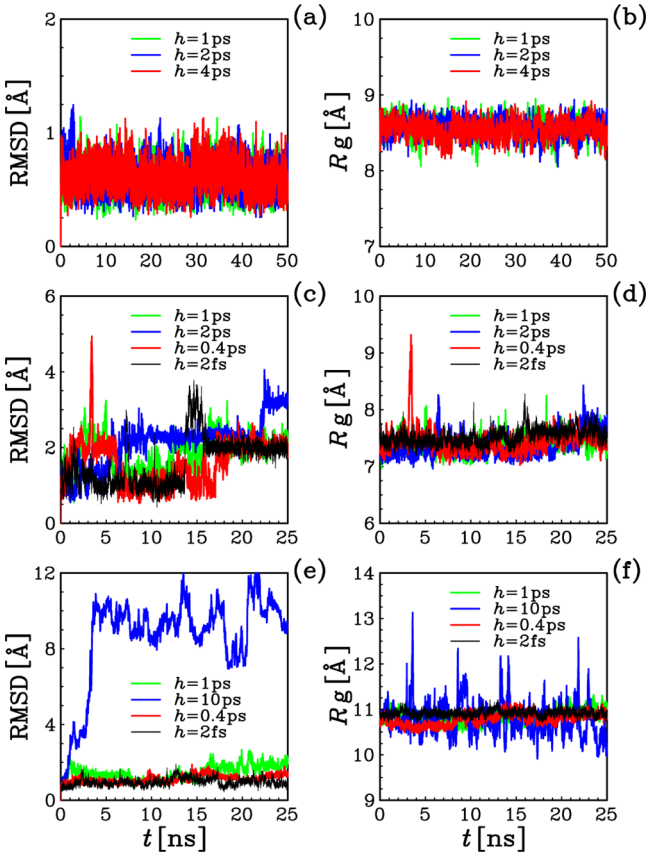


Figure 4. Root mean square deviation (RMSD) of atomic coordinates and radius of gyration R_g of the asphaltene [parts (a), (b)], miniprotein 1L2Y [(c), (d)], and protein G [(e), (f)] against of the duration of the OIN/GSFE/3D-RISM-KH simulation at different outer time step h. Black curves in parts (c)-(f): explicit solvent MD.

RMSD and R_g of these systems against the duration t of the OIN/GSFE/3D-RISM-KH simulations at different outer time steps h. The black curves in parts (c)−(f) represent the results of conventional MD simulations with explicit water. As we can see, both the OIN/GSFE/3D-RISM-KH and explicit MD atomic deviations starting from 0 quickly exhibit an equilibrium behavior with oscillations around 0.6 Å, 1 Å levels or higher for the asphaltene, miniprotein 1L2Y, and protein G, respectively. The magnitude of these oscillations does not exceed about 0.5 Å in all the cases (except for h = 10 ps, see below). For the asphaltene [part (a)], the mean values of RMSD almost do not change with the simulation duration t and remain practically the same even at the end of the 50 ns runs. We can thus interpret that as a very stable asphaltene conformation.

A somewhat different situation is for the miniprotein and protein G where transitions from one to another local equilibrium state take place [see parts (c) and (e) of Figure 4]. However, the variations in the RMSD are not large and do not exceed 1 to 2 Å (at $h \leq 2$ ps), meaning that the proteins remain in their folded conformation [see also Figures 5 and 6]. The results obtained for all the three systems with different sizes of the outer time step h are quite similar, confirming the high quality of the GSFE approach even at huge h of order of 1 to 4 ps. In addition, they are very close to those in the explicit MD simulations (black curves). Note that the latter were carried out at a small single time step Δt = 0.002 ps = 2 fs with no extrapolation ($h = \Delta t$) and should be treated as an "exact"

DOI: 10.1021/ct5010438
J. Chem. Theory Comput. 2015, 11, 1875−1895

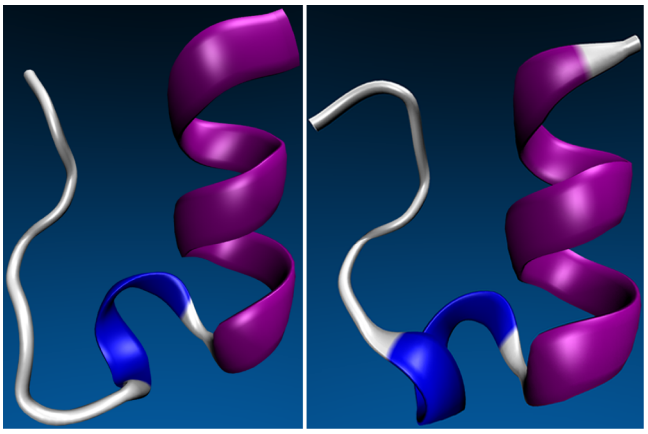


Figure 5. Tertiary structure of miniprotein 1L2Y taken from NMR experiment[128] before the simulation (left part) and after OIN/GEFE/3D-RISM-KH quasidynamics of duration t = 25 ns with outer time step h = 1 ps (right part).

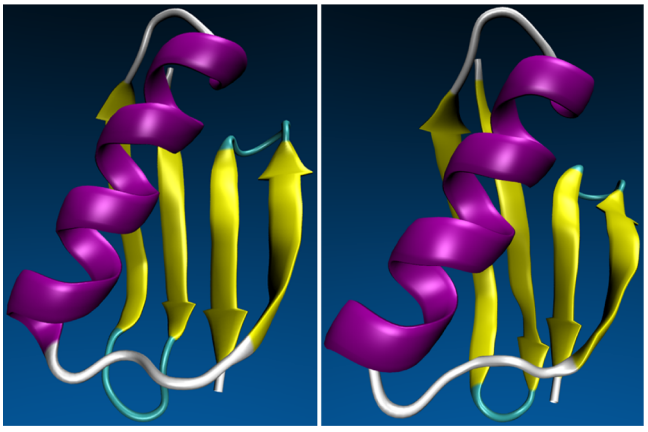


Figure 6. Same as in Figure 5 but for protein G.[129]

(or rather, "expected") results. Too long outer steps (h > 4 ps) are not recommended to use, as the accuracy of the extrapolation then becomes worse. This can affect the proper conformational behavior. As a demonstration, we increased the outer time step significantly over the maximum acceptable level up to h = 10 ps (where $\Sigma \sim 25\%$) and so obtained an artificial unfolding of the protein G from its initial folded state during already 5 ns [see Figure 4(e)]. A similar pattern to that described for the RMSD is observed in the case of the gyration radius [parts (b), (d), and (f) of Figure 4].

Figures 5 and 6 present the tertiary structures of, respectively, miniprotein 1L2Y and protein G obtained at the end of the OIN/GSFE/3D-RISM-KH simulations of duration t = 25 ns with h = 1 ps in comparison to their initial conformations taken from NMR experiment.[128,129] Note that we used the new cartoon representation with STRIDE[131] in the VMD (Visual Molecular Dynamics) package,[132] in which the secondary structure formations are assigned as follows: α-helix (purple), β-sheet (yellow), turn (cyan), coil (white), and 3_{10}-helix (blue). As is seen, the OIN/GSFE/3D-RISM-KH quasidynamics maintains the secondary and tertiary structures very well even at a huge outer time step of h = 1 ps. It is an excellent result, taking into account how many numerical techniques have been involved in this hybrid approach. Note that the main sources of possible uncertainties here are the approximate characters of the force fields, their extrapolation, and the 3D-RISM integral equation with the KH closure approximation. A comparison of the tertiary structures shows that these uncertainties do not affect the method ability to reproduce the true conformational behavior. Some difference between the initial and final conformations in Figures 5 and 6 is explained by the fact that they correspond to crystal and liquid states. Moreover, in our conventional MD simulations with explicit solvent, we obtained the folded conformations very similar to those presented in the right-hand parts of Figures 5 and 6.

4.4. Simulation Speedup. The SANDER module in which the MTS-MD/OIN/GSFE/3D-RISM-KH algorithm has been implemented is relatively fast for single CPU core run but not the best for parallel execution, compared to the PMEMD module, top parallel performer in the Amber MD package. However, SANDER has been available in the GPL version from the Amber Tools and thus presents a convenient platform to test the method, including scaling up in parallelization. All the MD simulations were carried out with up to 48 CPU cores per parallel job on the Grex and Parallel clusters with 4x Infiniband Interconnect, part of the WestGrid − Compute Canada national advanced computing platform. Figure 7(a) shows the

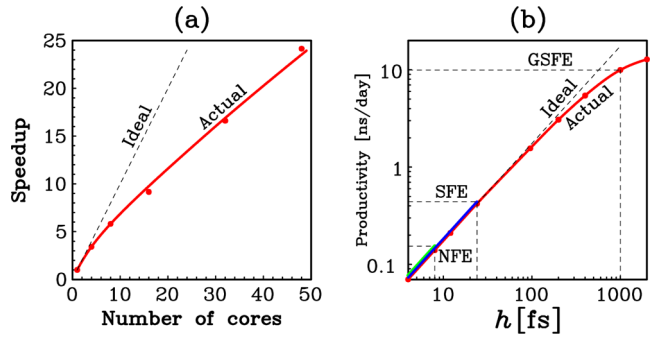


Figure 7. (a) Total parallel speedup of OIN/GSFE/3D-RISM-KH against the number of CPU cores. (b) Productivity achieved with 48 cores against the outer time step size.

speedup with the number of parallel CPU cores utilized in the hybrid OIN/GSFE/3D-RISM-KH simulation of hydrated protein G at the outer time step h = 1 ps. A similar behavior was observed for hydrated protein G at other values of h, as well as for hydrated miniprotein 1L2Y and for the asphaltene dimer in toluene. The efficiency constitutes 80% to 60% for 4 to 8 CPU cores and then levels out to 50% for up to 48 CPU cores—a typical picture for interprocessor communication toll in the SANDER module. The efficiency staying at 50% is far from the saturation regime, and so utilizing much more CPU cores in parallel is possible.

The productivity achieved in MTS-MD/OIN/GSFE/3D-RISM-KH quasidynamics of hydrated protein G on 48 CPU cores in parallel is presented in Figure 7(b) against the size of the outer time step h (red curve). The productivity with the original SFE scheme[31] is shown, too (SFE, blue curve). For comparison, we included also the result of the MD/3D-RISM-KH with no extrapolation of solvation force (NFE, green curve). The NFE and SFE curves terminate at h = 8 and 24 fs, respectively, related to the maximal allowed outer time steps in these schemes. As distinct, the new GSFE scheme extends up to h = 2 ps, whereas the maximal allowed outer time step here is 1 to 1.5 ps [see Figure 3(c)]. The 8 fs, 24 fs, and 1 ps upper limits of h marked in Figure 7(b) with dashed vertical lines correspond to the possible maximal productivity of 0.15 ns/

DOI: 10.1021/ct5010438
J. Chem. Theory Comput. 2015, 11, 1875−1895

day with no extrapolation, 0.44 ns/day with SFE, and 10 ns/day with GSFE.

For outer time step up to $h \sim 400$ fs, the main computing effort is spent on converging the 3D-RISM-KH integral equations at each outer time step, and the productivity of OIN/GSFE/3D-RISM-KH quasidynamics increases linearly with h [Figure 7(b)]. The growth slows down for larger h due to the other computing load for solving the OIN equations, the square-least and eigenvalue problems, and handling the extension-selection procedures. The portion on converging the 3D-RISM-KH equations drops inversely with h to about 88%, 75%, and 50% at h = 1, 2, and 4 ps. It then goes below the expenses invariably required at each inner time step for calculation of intramolecular solute atomic forces, propagation of coordinates and velocities, and solvation force extrapolation, and the productivity saturates making no sense to further increase h above 4 ps. For the present examples of OIN/GSFE/3D-RISM-KH simulations using the SANDER module of the Amber MD package, the productivity saturation crossover is observed at outer time step $h \sim 1$ ps.

The acceleration achieved with the GSFE scheme compared to no force extrapolation (NFE) in OIN/3D-RISM-KH quasidynamics determined as productivities ratio thus reaches a factor of 10 ns/day: 0.15 ns/day = 67 times, which is about half of the ideal acceleration calculated as the ratio 1000/8 = 125 of outer time steps. This provides a very good speedup in terms of absolute productivity. For comparison, we ran conventional MD simulation of protein G in explicit water using the standard SANDER module on 48 CPU cores in parallel with all the same simulation setup parameters like the cutoff radii for the solute−solvent interactions and so on and obtained a productivity of about 1 ns/day of explicit solvent MD. The OIN/GSFE/3D-RISM-KH quasidynamics of hydrated protein G thus achieved 10-fold speedup in terms of a "direct" comparison of the simulation rate (productivity) of protein evolution time (number of inner time steps) to that in explicit solvent MD. Furthermore, as demonstrated in Section 5 below, quasidynamics steered with 3D-RISM-KH mean solvation forces provides 5- to 100-fold time scale compression of protein conformational changes coupled with solvent exchange, thus achieving a huge overall effective speedup of protein conformational sampling for these systems by a factor of 50 to 1000 times compared to conventional MD with explicit solvent or real time dynamics.

5. PROTEIN FOLDING

We will now illustrate the ability of our hybrid MTS-MD/OIN/GSFE/3D-RISM-KH integrator algorithm to produce self-organized, native conformations of proteins, starting from denatured states. For this purpose, we picked miniprotein 1L2Y that was designed more than ten years ago[128] and up to date is the smallest protein to display folding properties. Its 304 atoms constitute a 20-residue amino acid sequence within the so-called tryptophan (Trp) cage (TC5b, PDB code: 1L2Y). The small size and stability of this miniprotein at room temperature make it an ideal candidate for computer simulation tests.

So far, there have been no results on folding of this simplest 1L2Y protein by conventional (unbiased) MD simulations in explicit solvent. The reason is that a time interval of order of 4 to 9 μs is still practically unreachable in one MD run even for modern supercomputers. There are a lot of challenges hindering protein folding simulations.[1] The main difficulty is that protein energy landscape is characterized by a vast number of local minima separated by energy barriers hard to overcome. One of the ways to obviate this problem lies in applying biased replica exchange MD simulation.[133−135] Mention that in the replica exchange approach, a great number of short simulations (replicas) are performed in parallel at different temperatures. After certain periods, the conformations are exchanged with a Metropolis probability. As a result, the necessary simulation length corresponding to each replica may be much shorter than the real folding time. However, the whole simulation must span over a wide temperature range with levels spaced closely enough to enable exchanges with high acceptance ratios. This significantly increases the total computational load.

Being of much less challenge, high-temperature unfolding MD simulation of the Trp cage in explicit water has been performed.[136] Replica exchange MD simulations of reversible unfolding/folding of the miniprotein in explicit water have been done, too.[134,135]

An efficient way to overcome the difficulties inherent in conventional MD of biomolecules is to contract detailed degrees of freedom of individual solvent molecules and perform quasidynamics of the biomolecule steered with mean solvation forces,[63,64] as described in Section 2.1. This has several advantages over explicit solvent including (i) much lower computational load with no necessity to treat explicit dynamics and slow exchange and localization of solvent molecules and (ii) significantly enhanced sampling of protein conformational space. The latter follows from the fact that averaging out solvent degrees of freedom to mean solvation forces eliminates an astronomical number of local minima in the energy landscape arising from local fluctuations in the solvation structure. Without explicit solvent treatment, mean solvation forces can be obtained either from continuum solvation models constructed empirically or from the integral equations of molecular theory of solvation derived from the first-principles of statistical mechanics. In the context of hydration of biomolecules, continuum solvation methods reproduce polar solvation forces with either the Poisson−Boltzmann (PB)[34] or the Generalized Born (GB)[35−37] models and nonpolar solvation forces with the solvent accessible surface area (SASA, or SA) model supplemented with additional volume and dispersion integral terms.[38,39] These empirical methods work well for the hydration free energy but have inherent disadvantages of being nontransferable to other solvents, cosolvents, and solvent systems, in particular, electrolyte solutions, missing solvent size effects such as a desolvation barrier in protein aggregation, and being inadequate to reproduce solvation of internal cavities such as narrow channels. As distinct, the 3D-RISM-KH molecular theory of solvation[24−29] is transferable and yields solvation structure and mean solvation forces with proper account of chemical specificities of a biomolecule and solvent system of various composition, including in a single framework both electrostatic associative forces and steric and entropic nonpolar effects, such as polar solvent and ionic screening, hydrophobicity, hydrogen bonding, salt bridges, preferential solvation, etc.

Many works have been devoted to study folding and unfolding pathways in the Trp-cage system by using MD simulations with GBSA implicit solvation. For example, Simmerling et al.[119] obtained the folded conformation of the miniprotein beginning from an unfolded state. Laser temperature jump relaxation experiments have shown that it is the most rapidly folding protein known with a folding time of order of 4 to 9 μs.[137] These experiments were supplemented with

DOI: 10.1021/ct5010438
J. Chem. Theory Comput. 2015, 11, 1875−1895

stochastic MD/GBSA simulations of the kinetics that indicated a folding time between 1.5 and 8.7 μs.[138] Replica exchange MD in implicit solvent has also been carried out to determine the folding thermodynamics of the miniprotein.[133] Quite recently, the experimentally suggested intermediate and unfolded states in the folding pathway of the Trp-cage miniprotein have been identified in enhanced MD/GBSA simulations.[139] Until now, there have been no successful attempts to fold any protein with hybrid MD/3D-RISM-KH simulations due to the low productivity of calculation and account of mean solvation forces in the previous integrator algorithms.[30,31] Inspired by the high efficiency of the new MTS-MD/OIN/GSFE/3D-RISM-KH approach as demonstrated with the promising results for the case studies in the previous section, we applied it to protein folding. The simulation of miniprotein 1L2Y using the same force field and parameters as those described in Section 4.1 was now carried out for 60 ns with 1 ps outer time step, starting from a well denatured, almost fully extended conformation. Another difference was that similarly to the MD/GBSA simulation,[119] we increased the solution temperature from T = 300 to 325 K so as to make the folding time less dependent on initial conditions.

In Figure 8, we present four conformational states of hydrated miniprotein 1L2Y obtained in the MTS-MD/OIN/GSFE/3D-RISM-KH simulation on 15, 18, 20, and 60 ns of quasidynamics. As is seen, the miniprotein remains almost completely unfolded like in the initial denatured conformation until 15 ns of simulation and, after that, drastic conformational changes appear quickly. An α-helix (purple) begins forming from 18 ns on and expands to its nearly full size by 20 ns of quasidynamics. Furthermore, the 3_{10}-helix (blue) secondary structure arises. This looks very similar to the miniprotein native folded states (Figure 5). Some difference apparently means that the simulation time length of 20 ns is still not long enough to achieve the lowest energy state. During the next 40 ns of quasidynamics totaling to 60 ns simulation, the miniprotein partially unfolded and folded up again several times but each time in a somewhat different way, confirming the conclusions from MD/GBSA simulations.[139] Finally, by 60 ns of quasidynamics, the most folded lowest energy conformation of hydrated miniprotein 1L2Y has been attained. This demonstrates for the first time that protein folding can be achieved with the unbiased approach of hybrid OIN/GSFE/3D-RISM-KH quasidynamics. (This was previously considered feasible rather with biased methods like replica exchange MD.) A more detailed analysis of these preliminary results will be presented in our next paper.

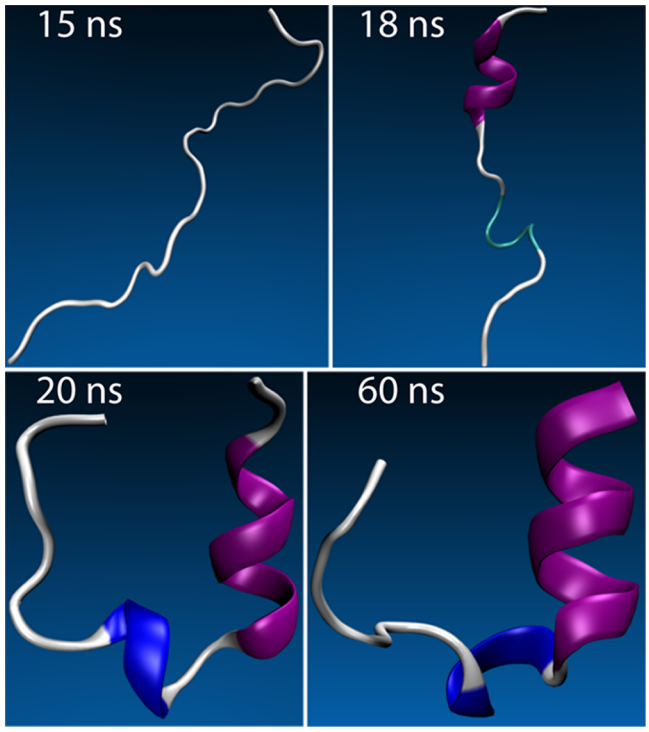


Figure 8. Conformational states of hydrated miniprotein 1L2Y in the OIN/GSFE/3D-RISM-KH folding simulation.

For comparison, no folding activity was observed in the 60 ns conventional MD simulation of miniprotein 1L2Y in explicit water (with the simulation setup analogous to that in Section 4.1). This is quite reasonable because, as mentioned above, the folding time of the miniprotein in real experiment is about 4 to 9 μs,[137,138] and so the same time scale should be expected in explicit water MD simulation. With the miniprotein folding observed in 60 ns quasidynamics of the hybrid OIN/GSFE/3D-RISM-KH simulation, we thus can claim time scale compression by a factor of 100 compared to conventional MD with explicit solvent. (Note that the time scale compression obtained with the earlier versions of solvation force extrapolation for hydrated alanine dipeptide was only about 5 times.[33]) We could expect the intrinsic acceleration inherent in OIN/GSFE/3D-RISM-KH quasidynamics to further increase with the complexity of proteins since 3D-RISM-KH mean solvation forces properly account for effects of chemical specificities of both solute and solvent molecules in protein confinement, as discussed in Section 2.1. Thus, the hybrid MTS-MD/OIN/GSFE/3D-RISM-KH integrator can efficiently sample phase space of solvated biomolecules for essential events with rare statistics such as protein conformational changes coupled with solvation exchange and localization. This provides a substantial gain over conventional MD with explicit solvent which requires an enormous number of time steps and computational time in such cases. In this context, the use of the 3D-RISM-KH molecular theory of solvation appears to be similar to some extent to other techniques enhancing statistical convergence. Besides replica exchange MD already mentioned,[133−135] these techniques also include umbrella sampling, weighted histogram, parallel tempering, and other biased methods[18] that were originally developed for conventional MD or Monte Carlo simulations. The advantage of the hybrid MTS-MD/OIN/GSFE/3D-RISM-KH method is that it can dramatically improve conformational sampling while using much lower computational cost.

6. CONCLUSION

In this paper, we have developed a new method of generalized solvation force extrapolation (GSFE) for the hybrid approach of multitime step molecular dynamics (MTS-MD) of biomolecules steered with 3D-RISM-KH mean solvation forces. By applying the non-Eckart-like transformation of coordinate space separately to each solute atom rather than the whole molecule, it modifies and generalizes our previous method of advanced solvation force extrapolation (ASFE)[33] to the case of arbitrary biomolecular solutes, including proteins, and significantly improves the extrapolation accuracy compared to the originally proposed method of solvation force extrapolation

DOI: 10.1021/ct5010438
J. Chem. Theory Comput. 2015, 11, 1875−1895

(SFE).[31] The whole procedure of MTS-MD steered with mean solvation forces calculated with GSFE at inner time steps based on those obtained from the 3D-RISM-KH molecular theory of solvation at outer time steps is efficiently stabilized with our optimized isokinetic Nosé−Hoover chain (OIN) thermostat.[32] The computational speed and accuracy of GSFE and the stabilization efficiency of OIN allowed us to use huge outer time steps up to 4 ps and thus to achieve dramatic acceleration compared to conventional MD simulation with explicit solvent.

The 3D-RISM-KH theory yields the solvation structure in terms of 3D maps of density distribution functions of solvent interaction sites around a solute molecule with full and consistent account for effects of chemical functionalities of all solution species. The solvation free energy and subsequent thermodynamics is then obtained at once as a simple integral of the correlation functions by performing thermodynamic integration analytically. The latter allows analytical differentiation of the free energy functional and thus self-consistent field coupling with MD.

It should be emphasized that 3D-RISM-KH mean solvation forces are based on the first-principles of statistical mechanics and consistently reproduce, at the level of *fully converged* molecular simulation, both electrostatic forces (hydrogen bonding, other association, salt bridges, dielectric and Debye screening, ion localization) and nonpolar solvation effects (desolvation, hydrophobic hydration, hydrophobic interaction), as well as subtle interplays of these such as preferential solvation,[110] molecular recognition,[40] and ligand binding.[40−43] This is very distinct from the continuum solvation schemes such as the Poisson−Boltzmann (PB)[34] and Generalized Born (GB)[35−37] models combined with the solvent accessible surface area (SASA) empirical nonpolar terms and additional volume and dispersion integral corrections,[38,39] which are parametrized for hydration free energy of biomolecules but are neither really applicable to solvation structure effects in complex confined geometries nor transferable to solvent systems with cosolvent or electrolyte solutions at physiological concentrations.

It should be noted again that in MD steered with 3D-RISM-KH mean solvation forces, the solvent dynamics is averaged out, and the protein quasidynamics, strictly speaking, differs from its true dynamics. However, as we have demonstrated, the contraction of the solvent dynamics does not affect the equilibrium conformational properties of the system. Moreover, as mean solvation forces is a statistical average of detailed solvation forces quickly varying in explicit solvent, this quasidynamics automatically filters out fast and short movements of the protein due to detailed interactions with solvent molecules but keeps the overall damping and steering effect of solvation forces. In fact, it performs essential dynamics of protein in solution (in the presence of solvent, cosolvent, counterions, and other possible components), with full effect of molecular steric forces and chemical specificities, such as desolvation barrier in hydrophobic interaction and hydrogen bonding (as distinct from implicit solvent MD missing such effects).

Another point is that generally speaking, mean solvation forces obtained by statistical-mechanical averaging depend on the protein dynamics as well as conformation, which results in mutual coupling of the protein and solvent dynamics. However, the protein dynamics, in particular, essential dynamics realizing protein functions, is usually much slower that the solvent dynamics. (Mind that solvent localization and exchange have slow rate−related to diffusion and partitioning in protein confined geometries, protein conformational transitions, and binding strength−but fast local solvent motion dynamics.) Therefore, except for particular processes where the solvent dynamics is important, its coupling with the protein is weak enough and can be neglected, resulting in a quasiequilibrium solvent description. Much in this way, 3D-RISM-KH mean solvation forces are obtained from statistically averaged correlations of quasiequilibrium solvent around the protein at successive conformational snapshots, and so the solvent dynamics is entirely decoupled from the protein conformational dynamics and is contracted in the mean solvation forces. (For comparison, in the continuum solvation models, mean solvation forces are also calculated for successive frozen conformational snapshots of the protein and so do not include coupling of the solute and solvent dynamics. However, constructed empirically, they are missing important molecular effects of solvent such as desolvation barrier and hydrogen bonding directionality, which are naturally represented in the 3D-RISM-KH theory.)

For systems where solvent dynamics is important or is a target property, these shortcomings can be overcome by advancing to the recently developed method of the generalized Langevin equation (GLE) in the 3D site formalism which, using an input of equilibrium correlation functions from 3D-RISM-KH, gives time-dependent correlation functions of both protein and solvent in a single formalism.[140−143] The GLE theory is involved computationally and also requires formulation of so-called memory kernels that adequately represent both relaxation physics and effects of chemical specificities of solvent and protein.[140] Nevertheless, it constitutes a promising approach to include the solvent dynamics and its coupling with the protein in the framework of MTS-MD steered with mean solvation forces obtained from molecular theory of solvation.

On the examples of an asphaltene dimer in toluene solvent, and alanine dipeptide, miniprotein 1L2Y, and protein G in water solvent, we have demonstrated that the hybrid MTS-MD/OIN/GSFE/3D-RISM-KH integrator algorithm even with huge outer time steps of up to 4 ps accurately reproduces conformational properties of biomolecular systems against the reference simulations using conventional MD with explicit solvent. The OIN/GSFE/3D-RISM-KH quasidynamics of hydrated protein G showed 10-fold speedup of productivity in terms of a "direct" comparison of the simulation rate of protein evolution time (number of inner time steps) to that in explicit solvent MD.

Moreover, 5- to 100-fold time scale compression achieved in OIN/GSFE/3D-RISM-KH quasidynamics of solvated protein due to the use of 3D-RISM-KH mean solvation forces result in further significant acceleration of protein conformational sampling compared to experimental real time dynamics and so to conventional MD with explicit solvent. The overall productivity of OIN/GSFE/3D-RISM-KH quasidynamics in protein sampling has been estimated on hydrated miniprotein 1L2Y as 50- to 1000-fold compared to conventional MD with explicit water. As an illustration, we have been able for the first time to fold miniprotein 1L2Y from a fully denatured state in 60 ns of our quasidynamics, whereas the folding duration in explicit solvent MD is expected to be similar to the 4−9 μs folding time observed experimentally.[137,138]

The intrinsic acceleration inherent in this quasidynamics is expected to further increase with the complexity of proteins since 3D-RISM-KH mean solvation forces properly and

DOI: 10.1021/ct5010438
J. Chem. Theory Comput. 2015, 11, 1875−1895

efficiently account for effects of chemical specificities of both solute and solvent molecules on slow processes of function related solvent localization and exchange in protein confinement. The hybrid MTS-MD/OIN/GSFE/3D-RISM-KH integrator can be applied to various biomolecular systems such as large proteins and DNA strands and to biomaterials such as cellulose nanocrystals, in different solvents, solvent systems, and electrolyte solutions. Illustration of these capabilities will be a subject of our next investigations.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: omelyan@icmp.lviv.ua.
*E-mail: andriy.kovalenko@nrc-cnrc.gc.ca.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We gratefully acknowledge the support by the ArboraNano—the Canadian Forest NanoProducts Network, by the University of Alberta, and by the National Institute for Nanotechnology (NINT), National Research Council (NRC) of Canada. The computations were carried out on the high performance computing resources provided by the WestGrid — Compute/Calcul Canada national advanced computing platform. I.O. is thankful for the hospitality during his stay at the University of Alberta and the National Institute for Nanotechnology. The authors are grateful to Dr. Blinov, Dr. Luchko, Dr. Stoyanov, and Dr. Gusarov for fruitful discussions.

REFERENCES

(1) Freddolino, P. L.; Harrison, C. B.; Liu, Y.; Schulten, K. *Nat. Phys.* **2010**, *6*, 751−758.
(2) Alder, B. J.; Wainwright, T. E. *J. Chem. Phys.* **1959**, *31*, 459.
(3) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Press Oxford University Press: Oxford, England, New York, 1989.
(4) Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed.; Academic Press: New York, **1996**.
(5) Leimkuhler, B.; Reich, S. *Simulating Hamiltonian Dynamics*; Cambridge University Press: 2005; Vol. *14*.
(6) McCammon, J. A.; Gelin, B. R.; Karplus, M. *Nature* **1977**, *267*, 585−590.
(7) Brooks, C. L.; Karplus, M.; Pettitt, B. M. *Adv. Chem. Phys.* **1988**, *71*, 1−6.
(8) Rojnuckarin, A.; Kim, S.; Subramaniam, S. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 4288−4292.
(9) Duan, Y. *Science* **1998**, *282*, 740−744.
(10) Hernandez, G.; Jenney, F. E.; Adams, M. W. W.; LeMaster, D. M. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 3166−3170.
(11) Karplus, M.; McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9*, 646−652.
(12) Zhang, Y.; Peters, M. H.; Li, Y. *Proteins* **2003**, *52*, 339−348.
(13) Adcock, S. A.; McCammon, J. A. *Chem. Rev.* **2006**, *106*, 1589−1615.
(14) Freddolino, P. L.; Liu, F.; Gruebele, M.; Schulten, K. *Biophys. J.* **2008**, *94*, L75−L77.
(15) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. *Curr. Opin. Struct. Biol.* **2009**, *19*, 120−127.
(16) Service, R. F. *Science* **2010**, *330*, 308−309.
(17) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. *Science* **2010**, *330*, 341−346.
(18) Tuckerman, M. *Statistical Mechanics: Theory and Molecular Simulation*; Oxford University Press: New York, 2010.
(19) Genheden, S.; Ryde, U. *Phys. Chem. Chem. Phys.* **2012**, *14*, 8662.
(20) Chandler, D.; McCoy, J. D.; Singer, S. J. *J. Chem. Phys.* **1986**, *85*, 5971−5976.
(21) Chandler, D.; McCoy, J. D.; Singer, S. J. *J. Chem. Phys.* **1986**, *85*, 5977−5982.
(22) Beglov, D.; Roux, B. *J. Phys. Chem. B* **1997**, *101*, 7821−7826.
(23) Kovalenko, A.; Hirata, F. *Chem. Phys. Lett.* **1998**, *290*, 237−244.
(24) Kovalenko, A.; Hirata, F. *J. Chem. Phys.* **1999**, *110*, 10095−10112.
(25) Kovalenko, A.; Hirata, F. *J. Chem. Phys.* **2000**, *112*, 10391−10402.
(26) Kovalenko, A.; Hirata, F. *J. Chem. Phys.* **2000**, *112*, 10403−10417.
(27) Kovalenko, A. In *Molecular Theory of Solvation*; Hirata, F., Ed.; Understanding Chemical Reactivity; Kluwer Academic Publishers: Norwell, MA, USA, 2003; Vol. *24*, Chapter 4, pp 169−275.
(28) Gusarov, S.; Pujari, B. S.; Kovalenko, A. *J. Comput. Chem.* **2012**, *33*, 1478−1494.
(29) Kovalenko, A. *Pure Appl. Chem.* **2013**, *85*, 159−199.
(30) Miyata, T.; Hirata, F. *J. Comput. Chem.* **2008**, *29*, 871−882.
(31) Luchko, T.; Gusarov, S.; Roe, D. R.; Simmerling, C.; Case, D. A.; Tuszynski, J.; Kovalenko, A. *J. Chem. Theory Comput.* **2010**, *6*, 607−624.
(32) Omelyan, I.; Kovalenko, A. *Mol. Simul.* **2013**, *39*, 25−48.
(33) Omelyan, I.; Kovalenko, A. *J. Chem. Phys.* **2013**, *139*, 244106.
(34) Antosiewicz, J. M.; Shugar, D. *Mol. Biosyst.* **2011**, *7*, 2923.
(35) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.
(36) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383−394.
(37) Onufriev, A. *Modeling Solvent Environments*; Wiley-Blackwell: 2010; pp 127−165.
(38) Mongan, J.; Simmerling, C.; McCammon, J. A.; Case, D. A.; Onufriev, A. *J. Chem. Theory Comput.* **2007**, *3*, 156−169.
(39) Wagoner, J. A.; Baker, N. A. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 8331−8336.
(40) Yoshida, N.; Imai, T.; Phongphanphanee, S.; Kovalenko, A.; Hirata, F. *J. Phys. Chem. B* **2009**, *113*, 873−886.
(41) Imai, T.; Oda, K.; Kovalenko, A.; Hirata, F.; Kidera, A. *J. Am. Chem. Soc.* **2009**, *131*, 12430−12440.
(42) Imai, T.; Miyashita, N.; Sugita, Y.; Kovalenko, A.; Hirata, F.; Kidera, A. *J. Phys. Chem. B* **2011**, *115*, 8288−8295.
(43) Nikolić, D.; Blinov, N.; Wishart, D.; Kovalenko, A. *J. Chem. Theory Comput.* **2012**, *8*, 3356−3372.
(44) Tuckerman, M.; Berne, B. J.; Martyna, G. J. *J. Chem. Phys.* **1992**, *97*, 1990−2001.
(45) Stuart, S. J.; Zhou, R.; Berne, B. J. *J. Chem. Phys.* **1996**, *105*, 1426−1436.
(46) Kopf, A.; Paul, W.; Dünweg, B. *Comput. Phys. Commun.* **1997**, *101*, 1−8.
(47) Schlick, T.; Barth, E.; Mandziuk, M. *Annu. Rev. Biophys. Biomol. Struct.* **1997**, *26*, 181−222.
(48) Watanabe, M.; Karplus, M. *J. Phys. Chem.* **1995**, *99*, 5680−5697.
(49) Mandziuk, M.; Schlick, T. *Chem. Phys. Lett.* **1995**, *237*, 525−535.
(50) Barth, E.; Schlick, T. *J. Chem. Phys.* **1998**, *109*, 1633−1642.
(51) Schlick, T.; Mandziuk, M.; Skeel, R. D.; Srinivas, K. *J. Comput. Phys.* **1998**, *140*, 1−29.
(52) Ma, Q.; Izaguirre, J. A.; Skeel, R. D. *SIAM J. Sci. Comput.* **2003**, *24*, 1951−1973.
(53) Omelyan, I. P. *Phys. Rev. E* **2008**, *78*, 026702.
(54) Omelyan, I. P. *J. Chem. Phys.* **2009**, *131*, 104101.
(55) Omelyan, I. P.; Kovalenko, A. *J. Chem. Phys.* **2011**, *135*, 114110.
(56) Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. *Biopolymers* **1992**, *32*, 523−535.
(57) Barth, E.; Schlick, T. *J. Chem. Phys.* **1998**, *109*, 1617−1632.
(58) Minary, P.; Tuckerman, M.; Martyna, G. *Phys. Rev. Lett.* **2004**, *93*, 150201.

DOI: 10.1021/ct5010438
J. Chem. Theory Comput. 2015, 11, 1875−1895

(59) Abrams, J.; Tuckerman, M.; Martyna, G. *Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology*; Springer-Verlag: Berlin, 2006; Vol. 1, pp 139−192.
(60) Minary, P.; Martyna, G. J.; Tuckerman, M. E. *J. Chem. Phys.* **2003**, *118*, 2510.
(61) Omelyan, I. P.; Kovalenko, A. J. *Chem. Phys.* **2011**, *135*, 234107.
(62) Omelyan, I. P.; Kovalenko, A. *J. Chem. Theory Comput.* **2012**, *8*, 6−16.
(63) Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300−313.
(64) McQuarrie, D. A. *Statistical Mechanics*; University Science Books: Sausalito, CA, 2000.
(65) Stumpe, M. C.; Blinov, N.; Wishart, D.; Kovalenko, A.; Pande, V. S. *J. Phys. Chem. B* **2011**, *115*, 319−328.
(66) Omelyan, I. P. *Mol. Simul.* **1999**, *22*, 213−236.
(67) Lawson, C. *Solving Least Squares Problems*; SIAM: Philadelphia, PA, 1995.
(68) Quintana-Ortí, G.; Quintana-Ortí, E. S.; Petitet, A. *SIAM J. Sci. Comput.* **1998**, *20*, 1155−1163.
(69) Kneller, G. R. *J. Chem. Phys.* **2008**, *128*, 194101.
(70) Eckart, C. *Phys. Rev.* **1935**, *47*, 552−558.
(71) Louck, J. D.; Galbraith, H. W. *Rev. Mod. Phys.* **1976**, *48*, 69−106.
(72) Janežič, D.; Praprotnik, M.; Merzel, F. *J. Chem. Phys.* **2005**, *122*, 174101.
(73) Praprotnik, M.; Janežič, D. *J. Chem. Phys.* **2005**, *122*, 174102.
(74) Praprotnik, M.; Janežič, D. *J. Chem. Phys.* **2005**, *122*, 174103.
(75) Omelyan, I.; Kovalenko, A. *Phys. Rev. E* **2012**, *85*, 026706.
(76) Coutsias, E. A.; Seok, C.; Dill, K. A. *J. Comput. Chem.* **2004**, *25*, 1849−1857.
(77) Liu, P.; Agrafiotis, D. K.; Theobald, D. L. *J. Comput. Chem.* **2009**, 1561−1563.
(78) Chevrot, G.; Calligari, P.; Hinsen, K.; Kneller, G. R. *J. Chem. Phys.* **2011**, *135*, 084110.
(79) Hansen, J.-P.; McDonald, I. *Theory of Simple Liquids*, 3rd ed.; Elsevier Academic Press: London, Burlington, MA, 2006.
(80) Gusarov, S.; Ziegler, T.; Kovalenko, A. *J. Phys. Chem. A* **2006**, *110*, 6083−6090.
(81) Casanova, D.; Gusarov, S.; Kovalenko, A.; Ziegler, T. *J. Chem. Theory Comput.* **2007**, *3*, 458−476.
(82) Kaminski, J. W.; Gusarov, S.; Wesolowski, T. A.; Kovalenko, A. *J. Phys. Chem. A* **2010**, *114*, 6082−6096.
(83) Yamazaki, T.; Kovalenko, A. *J. Chem. Theory Comput.* **2009**, *5*, 1723−1730.
(84) da Costa, L. M.; Hayaki, S.; Stoyanov, S. R.; Gusarov, S.; Tan, X.; Gray, M. R.; Stryker, J. M.; Tykwinski, R.; de M. Carneiro, J. W.; Sato, H.; Seidl, P. R.; Kovalenko, A. *Phys. Chem. Chem. Phys.* **2012**, *14*, 3922.
(85) Stoyanov, S. R.; Gusarov, S.; Kovalenko, A. In *Industrial applications of molecular simulations*; Meunier, M., Ed.; CRC Press: Boca Raton, FL, 2012; Chapter 14.
(86) Fafard, J.; Lyubimova, O.; Stoyanov, S. R.; Dedzo, G. K.; Gusarov, S.; Kovalenko, A.; Detellier, C. *J. Phys. Chem. C* **2013**, *117*, 18556−18566.
(87) Kovalenko, A.; Kobryn, A. E.; Gusarov, S.; Lyubimova, O.; Liu, X.; Blinov, N.; Yoshida, M. *Soft Matter* **2012**, *8*, 1508−1520.
(88) Moralez, J. G.; Raez, J.; Yamazaki, T.; Motkuri, R. K.; Kovalenko, A.; Fenniri, H. *J. Am. Chem. Soc.* **2005**, *127*, 8307−8309.
(89) Johnson, R. S.; Yamazaki, T.; Kovalenko, A.; Fenniri, H. *J. Am. Chem. Soc.* **2007**, *129*, 5735−5743.
(90) Yamazaki, T.; Fenniri, H.; Kovalenko, A. *ChemPhysChem* **2010**, *11*, 361−367.
(91) Silveira, R. L.; Stoyanov, S. R.; Gusarov, S.; Skaf, M. S.; Kovalenko, A. *J. Am. Chem. Soc.* **2013**, *135*, 19048−19051.
(92) Kovalenko, A. *Nord. Pulp Pap. J.* **2014**, *29*, 144−155.
(93) Silveira, R. L.; Stoyanov, S. R.; Gusarov, S.; Skaf, M. S.; Kovalenko, A. *J. Phys. Chem. Lett.* **2015**, *6*, 206−211.
(94) Phongphanphanee, S.; Yoshida, N.; Hirata, F. *J. Am. Chem. Soc.* **2008**, *130*, 1540−1541.
(95) Kiyota, Y.; Hiraoka, R.; Yoshida, N.; Maruyama, Y.; Imai, T.; Hirata, F. *J. Am. Chem. Soc.* **2009**, *131*, 3852−3853.
(96) Li, Q.; Gusarov, S.; Evoy, S.; Kovalenko, A. *J. Phys. Chem. B* **2009**, *113*, 9958−9967.
(97) Genheden, S.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Ryde, U. *J. Phys. Chem. B* **2010**, *114*, 8505−8516.
(98) Blinov, N.; Dorosh, L.; Wishart, D.; Kovalenko, A. *Biophys. J.* **2010**, *98*, 282−296.
(99) Maruyama, Y.; Yoshida, N.; Hirata, F. *J. Phys. Chem. B* **2010**, *114*, 6464−6471.
(100) Phongphanphanee, S.; Rungrotmongkol, T.; Yoshida, N.; Hannongbua, S.; Hirata, F. *J. Am. Chem. Soc.* **2010**, *132*, 9782−9788.
(101) Kovalenko, A.; Blinov, N. *J. Mol. Liq.* **2011**, *164*, 101−112.
(102) Yamazaki, T.; Kovalenko, A. *J. Phys. Chem. B* **2011**, *115*, 310−318.
(103) Yamazaki, T.; Fenniri, H. *J. Phys. Chem. C* **2012**, *116*, 15087−15092.
(104) Kovalenko, A. *Partial Molar Vol.s of Proteins in Solution: Prediction by Statistical-Mechanical 3D-RISM-KB Molecular Theory of Solvation*; Royal Society of Chemistry: 2015; Chapter 22, pp 575−610.
(105) Perkyns, J.; Pettitt, B. M. *Chem. Phys. Lett.* **1992**, *190*, 626−630.
(106) Perkyns, J.; Pettitt, B. M. *J. Chem. Phys.* **1992**, *97*, 7656−7666.
(107) Kovalenko, A.; Hirata, F. *Chem. Phys. Lett.* **2001**, *349*, 496−502.
(108) Kovalenko, A.; Hirata, F. *J. Theor. Comput. Chem.* **2002**, *01*, 381−406.
(109) Yoshida, K.; Yamaguchi, T.; Kovalenko, A.; Hirata, F. *J. Phys. Chem. B* **2002**, *106*, 5042−5049.
(110) Yamazaki, T.; Kovalenko, A.; Murashov, V. V.; Patey, G. N. *J. Phys. Chem. B* **2010**, *114*, 613−619.
(111) Kobryn, A. E.; Nikolić, D.; Lyubimova, O.; Gusarov, S.; Kovalenko, A. *J. Phys. Chem. B* **2014**, *118*, 12034−12049.
(112) Schmeer, G.; Maurer, A. *Phys. Chem. Chem. Phys.* **2010**, *12*, 2407.
(113) Perkyns, J. S.; Lynch, G. C.; Howard, J. J.; Pettitt, B. M. *J. Chem. Phys.* **2010**, *132*, 064106.
(114) Kovalenko, A.; Ten-no, S.; Hirata, F. *J. Comput. Chem.* **1999**, *20*, 928−936.
(115) Truchon, J.-F.; Pettitt, B. M.; Labute, P. *J. Chem. Theory Comput.* **2014**, *10*, 934−941.
(116) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577−8593.
(117) Omelyan, I. P. *Comput. Phys. Commun.* **1997**, *107*, 113−122.
(118) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999−2012.
(119) Simmerling, C.; Strockbine, B.; Roitberg, A. E. *J. Am. Chem. Soc.* **2002**, *124*, 11258−11259.
(120) Jorgensen, W. L.; Laird, E. R.; Nguyen, T. B.; Tirado-Rives, J. *J. Comput. Chem.* **1993**, *14*, 206−215.
(121) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(122) 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. P.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; 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.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *AMBER 11*; University of California: San Francisco, 2010.
(123) Omelyan, I. P.; Mryglod, I. M.; Folk, R. *Phys. Rev. E* **2002**, *65*, 056706.
(124) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.
(125) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269−6271.
(126) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. *J. Comput. Phys.* **1977**, *23*, 327−341.
(127) Ciccotti, G.; Ferrario, M.; Ryckaert, J.-P. *Mol. Phys.* **1982**, *47*, 1253−1264.

DOI: 10.1021/ct5010438
J. Chem. Theory Comput. 2015, 11, 1875−1895

(128) Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H. *Nat. Struct. Biol.* **2002**, *9*, 425−430.
(129) Ulmer, T. S.; Ramirez, B. E.; Delaglio, F.; Bax, A. *J. Am. Chem. Soc.* **2003**, *125*, 9179−9191.
(130) Chai, J.-D.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615.
(131) Fischer, D.; Eisenberg, D. *Protein Sci.* **1996**, *5*, 947−955.
(132) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33−38.
(133) Pitera, J. W.; Swope, W. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 7587−7592.
(134) Paschek, D.; Nymeyer, H.; García, A. E. *J. Struct. Biol.* **2007**, *157*, 524−533.
(135) Day, R.; Paschek, D.; Garcia, A. E. *Proteins* **2010**, *78*, 1889−1899.
(136) Seshasayee, A. S. N. *Theor. Biol. Med. Modell.* **2005**, *2*, 7.
(137) Qiu, L.; Pabit, S. A.; Roitberg, A. E.; Hagen, S. J. *J. Am. Chem. Soc.* **2002**, *124*, 12952−12953.
(138) Snow, C. D.; Zagrovic, B.; Pande, V. S. *J. Am. Chem. Soc.* **2002**, *124*, 14548−14549.
(139) Shao, Q.; Shi, J.; Zhu, W. *J. Chem. Phys.* **2012**, *137*, 125103.
(140) Kim, B.; Hirata, F. *J. Chem. Phys.* **2013**, *138*, 054108.
(141) Hirata, F.; Kim, B. Theoretical formulae useful for determining fluctuation and dynamics of biopolymer, density fluctuation and dynamics of solution, variance-covariance matrix of structural fluctuation of biopolymer from free energy respectively. Patent WO2014115339-A1, 2014.
(142) Hirata, F.; Kim, B. Theoretical determination of variance-covariance matrix of structural fluctuation of biopolymer, involves analyzing theoretical formulae denoting fluctuation and dynamics of biopolymer, and density fluctuation and dynamics of solvent. Patent WO2014115416-A1, 2014.
(143) Hirata, F.; Akasaka, K. *J. Chem. Phys.* **2015**, *142*, 044110.

DOI: 10.1021/ct5010438
J. Chem. Theory Comput. 2015, 11, 1875−1895