Gaussian Accelerated Molecular Dynamics (GaMD)

1. Gaussian Accelerated Molecular Dynamics (GaMD)

Gaussian Accelerated Molecular Dynamics (GaMD) is a biomolecular enhanced sampling method that works by adding a harmonic boost potential to smoothen the system potential energy surface. The boost potential follows Gaussian distribution, which allows for accurate reweighting using cumulant expansion to the second order. In a previous study[1], GaMD has been demonstrated on simulations of alanine dipeptide, chignolin folding and ligand binding to the T4-lysozyme. Without the need to set predefined reaction coordinates, GaMD enables unconstrained enhanced sampling of these biomolecules. Furthermore, the free energy profiles obtained from reweighting of the GaMD simulations help us to identify distinct low energy states of the biomolecules and characterize the protein folding and ligand binding pathways quantitatively.

Consider a system with *N* atoms at positions $r \equiv \{\vec{r}_1, \cdots, \vec{r}_N\}$. When the system potential $V(r)$ is lower than a reference energy *E*, the modified potential $V^*(r)$ of the system is calculated as:

$$V^*(r) = V(r) + \Delta V(r),$$

$$\Delta V(r) = \begin{cases} \frac{1}{2}k\big(E - V(r)\big)^2, & V(r) < E \\ 0, & V(r) \geq E \end{cases} \tag{1.1}$$

where *k* is the harmonic force constant. The two adjustable parameters *E* and *k* are automatically determined based on three enhanced sampling principles. The reference energy needs to be set in the following range:

$$V_{max} \leq E \leq V_{min} + \frac{1}{k}, \tag{1.2}$$

where V_{max} and V_{min} are the system minimum and maximum potential energies. To ensure that **Eqn. (1.2)** is valid, *k* has to satisfy: $k \leq \frac{1}{V_{max} - V_{min}}$ Let us define $k \equiv k_0 \cdot \frac{1}{V_{max} - V_{min}}$, then $0 < k_0 \leq 1$. The standard deviation of ΔV needs to be small enough (i.e., narrow distribution) to ensure proper energetic reweighting: $\sigma_{\Delta V} = k\big(E - V_{avg}\big)\sigma_V \leq \sigma_0$ where V_{avg} and σ_V are the average and standard deviation of the system potential energies, $\sigma_{\Delta V}$ is the standard deviation of ΔV with σ_0 as a user-specified upper limit (e.g., $10k_B$T) for proper reweighting. When *E* is set to the lower bound $E=V_{max}$, k_0 can be calculated as:

$$k_0 = \min(1.0, k_0') = \min\left(1.0, \frac{\sigma_0}{\sigma_V} \cdot \frac{V_{max} - V_{min}}{V_{max} - V_{avg}}\right). \tag{1.3}$$

Alternatively, when the threshold energy *E* is set to its upper bound $E = V_{min} + \frac{1}{k}$, k_0 is set to:

$$k_0 = k_0'' \equiv \left(1 - \frac{\sigma_0}{\sigma_V}\right)\frac{V_{max} - V_{min}}{V_{avg} - V_{min}}, \tag{1.4}$$

if k_0'' is found to be between *0* and *1*. Otherwise, k_0 is calculated using **Eqn. (1.3)**.

2. Ligand Gaussian Accelerated Molecular Dynamics (LiGaMD)

Based on GaMD, a new algorithm called ligand GaMD or "LiGaMD" has been developed to simulate ligand binding and unbinding[2]. It works by selectively boosting the ligand non-bonded interaction potential energy. Another boost potential could be applied to the remaining potential energy of the entire system in a dual-boost algorithm (LiGaMD_Dual) to facilitate ligand binding. LiGaMD has been demonstrated on host-guest and protein-ligand binding model systems. Repetitive guest binding and unbinding in the β-cyclodextrin host were observed in hundreds-of-nanosecond LiGaMD simulations. The calculated binding free energies of guest molecules with sufficient sampling agreed excellently with experimental data (< 1.0 kcal/mol error). In comparison with previous microsecond-timescale conventional molecular dynamics simulations, accelerations of ligand kinetic rate constants in LiGaMD simulations were properly estimated using Kramers' rate theory. Furthermore, LiGaMD allowed us to capture repetitive dissociation and binding of the benzamidine inhibitor in trypsin within 1 μs simulations. The calculated ligand binding free energy and kinetic rate constants compared well with the experimental data. Therefore, LiGaMD provides a promising approach for characterizing ligand binding thermodynamics and kinetics simultaneously.

In LiGaMD, we consider a system of ligand *L* binding to a protein *P* in a biological environment *E*. We decompose the potential energy into the following terms:

$$\begin{aligned} V(r) = {} & V_{P,b}(r_P) + V_{L,b}(r_L) + V_{E,b}(r_E) \\ & + V_{PP,nb}(r_P) + V_{LL,nb}(r_L) + V_{EE,nb}(r_E) \\ & + V_{PL,nb}(r_{PL}) + V_{PE,nb}(r_{PE}) + V_{LE,nb}(r_{LE}). \end{aligned} \tag{2.1}$$

where $V_{P,b}$, $V_{L,b}$ and $V_{E,b}$ are the bonded potential energies in protein *P*, ligand *L* and environment *E*, respectively. $V_{PP,nb}$, $V_{LL,nb}$ and $V_{EE,nb}$ are the self non-bonded potential energies in protein *P*, ligand *L* and environment *E*, respectively. $V_{PL,nb}$, $V_{PE,nb}$ and $V_{LE,nb}$ are the non-bonded interaction energies between *P*-*L*, *P*-*E* and *L*-*E*, respectively. According to classical molecular mechanics force fields, the non-bonded potential energies are usually calculated as:

$$V_{nb} = V_{elec} + V_{vdW}. \tag{2.2}$$

Where V_{elec} and V_{vdW} are the system electrostatic and van der Waals potential energies. Presumably, ligand binding mainly involves the non-bonded interaction energies of the ligand, $V_{L,nb}(r) = V_{LL,nb}(r_L) + V_{PL,nb}(r_{PL}) + V_{LE,nb}(r_{LE})$. Therefore, we add a boost potential selectively to the ligand non-bonded potential energy according to the GaMD algorithm:

$$\Delta V_{L,nb}(r) = \begin{cases} \frac{1}{2}k_{L,nb}\left(E_{L,nb} - V_{L,nb}(r)\right)^2, & V_{L,nb}(r) < E_{L,nb} \\ 0, & V_{L,nb}(r) \geq E_{L,nb} \end{cases} \tag{2.3}$$

where $E_{L,nb}$ is the threshold energy for applying boost potential and $k_{L,nb}$ is the harmonic constant.

Next, one can add multiple ligand molecules in the solvent to facilitate ligand binding to proteins in MD simulations. This is based on the fact that the ligand binding rate constant k_{on} is inversely proportional to the ligand concentration. The higher the ligand concentration, the faster the ligand binds, provided that the ligand concentration is still within its solubility limit. In addition to selectively boosting the bound ligand, another boost potential could thus be applied on the unbound ligand molecules, protein and solvent to facilitate both ligand dissociation and rebinding. The second boost potential is calculated using the total system potential energy other than the non-bonded potential energy of the bound ligand as:

$$\Delta V_D(r) = \begin{cases} \frac{1}{2}k_D\left(E_D - V_D(r)\right)^2, & V_D(r) < E_D \\ 0, & V_D(r) \geq E_D \end{cases} \tag{2.4}$$

Where V_D is the total system potential energy other than the non-bonded potential energy of the bound ligand, E_D is the corresponding threshold energy for applying the second boost potential and k_D is the harmonic constant. This leads to dual-boost LiGaMD (LiGaMD_Dual) with the total boost potential $\Delta V(r) = \Delta V_{L,nb}(r) + \Delta V_D(r)$.

3. Peptide Gaussian Accelerated Molecular Dynamics (Pep-GaMD)

Peptides often undergo large conformational changes during binding to the target proteins, being distinct from small-molecule ligand binding or protein-protein interactions. We have developed another algorithm called peptide GaMD or "Pep-GaMD" that enhances sampling of peptide-protein interactions[3].

In Pep-GaMD, we consider a system of ligand peptide L binding to a target protein P in a biological environment E. We decompose the potential energy into the following terms:

$$\begin{aligned} V(r) = {} & V_{P,b}(r_P) + V_{L,b}(r_L) + V_{E,b}(r_E) \\ & + V_{PP,nb}(r_P) + V_{LL,nb}(r_L) + V_{EE,nb}(r_E) \\ & + V_{PL,nb}(r_{PL}) + V_{PE,nb}(r_{PE}) + V_{LE,nb}(r_{LE}). \end{aligned} \tag{3.1}$$

where $V_{P,b}$, $V_{L,b}$ and $V_{E,b}$ are the bonded potential energies in protein P, peptide L and environment E, respectively. $V_{PP,nb}$, $V_{LL,nb}$ and $V_{EE,nb}$ are the self non-bonded potential energies in protein P, peptide L and environment E, respectively. $V_{PL,nb}$, $V_{PE,nb}$ and $V_{LE,nb}$ are the non-bonded interaction energies between P-L, P-E and L-E, respectively.

Presumably, peptide binding mainly involves in both the bonded and non-bonded interaction energies of the peptide since peptides often undergo large conformational changes during binding to the target proteins. Thus, the essential peptide potential energy is $V_L(r) = V_{LL,b}(r_L) + V_{LL,nb}(r_L) + V_{PL,nb}(r_{PL}) + V_{LE,nb}(r_{LE})$. In Pep-GaMD, we add boost potential selectively to the essential peptide potential energy according to the GaMD algorithm:

$$\Delta V_L(r) = \begin{cases} \frac{1}{2}k_L\left(E_L - V_L(r)\right)^2, & V_L(r) < E_L \\ 0, & V_L(r) \geq E_L \end{cases} \quad (3.2)$$

where E_L is the threshold energy for applying boost potential and k_L is the harmonic constant.

In addition to selectively boosting the peptide, another boost potential is applied on the protein and solvent to enhance conformational sampling of the protein and facilitate peptide rebinding. The second boost potential is calculated using the total system potential energy other than the peptide potential energy as:

$$\Delta V_D(r) = \begin{cases} \frac{1}{2}k_D\left(E_D - V_D(r)\right)^2, & V_D(r) < E_D \\ 0, & V_D(r) \geq E_D \end{cases} \quad (3.3)$$

Where V_D is the total system potential energy other than the peptide potential energy, E_D is the corresponding threshold energy for applying the second boost potential and k_D is the harmonic constant. This leads to dual-boost Pep-GaMD (Pep-GaMD_Dual) with the total boost potential $\Delta V(r) = \Delta V_L(r) + \Delta V_D(r)$.

4. Protein-Protein Interaction - Gaussian Accelerated Molecular Dynamics (PPI-GaMD)

In PPI-GaMD [4], we selectively boost interaction potential energy between protein partners to facilitate their slow dissociation. Meanwhile, another boost potential is applied to the remaining potential energy of the entire system to effectively model the protein's flexibility and rebinding. We consider a system of ligand protein *L* binding to a target protein *P* in a biological environment *E*. The system comprises of *N* atoms with their coordinates $r \equiv \{\vec{r}_1, \cdots, \vec{r}_N\}$ and momenta $p \equiv \{\vec{p}_1, \cdots, \vec{p}_N\}$. The system Hamiltonian can be expressed as:

$$H(r,p) = K(p) + V(r), \quad (4.1)$$

where $K(p)$ and $V(r)$ are the system kinetic and total potential energies, respectively. Next, we decompose the potential energy into the following terms:

$$\begin{aligned} V(r) = {} & V_{P,b}(r_P) + V_{L,b}(r_L) + V_{E,b}(r_E) \\ & + V_{PP,nb}(r_P) + V_{LL,nb}(r_L) + V_{EE,nb}(r_E) \\ & + V_{PL,nb}(r_{PL}) + V_{PE,nb}(r_{PE}) + V_{LE,nb}(r_{LE}), \end{aligned} \quad (4.2)$$

where $V_{P,b}$, $V_{L,b}$ and $V_{E,b}$ are the bonded potential energies in protein *P*, protein *L* and environment *E*, respectively. $V_{PP,nb}$, $V_{LL,nb}$ and $V_{EE,nb}$ are the self non-bonded potential energies in protein *P*, protein *L* and environment *E*, respectively. $V_{PL,nb}$, $V_{PE,nb}$ and $V_{LE,nb}$ are the non-bonded interaction energies between *P*-*L*, *P*-*E* and *L*-*E*, respectively. According to classical molecular mechanics force fields,[5, 6] the non-bonded potential energies are usually calculated as $V_{nb} = V_{elec} + V_{vdW}$, where V_{elec} and V_{vdW} are the system electrostatic and van der Waals potential energies. The interaction energy between the protein binding partners is $V_{PL,nb}(r_{PL})$. In PPI-GaMD, we add

boost potential selectively to the protein-protein interaction energy according to the GaMD algorithm:

$$\Delta V_{PL.nb}(r) = \begin{cases} \frac{1}{2}k_{PL,nb}\left(E_{PL,nb} - V_{PL,nb}(r_{PL})\right)^2, & V_{PL,nb}(r_{PL}) < E_{PL,nb} \\ 0, & V_{PL,nb}(r_{PL}) \geq E_{PL,nb}, \end{cases} \tag{4.3}$$

where $E_{PL,nb}$ is the threshold energy for applying boost potential and $k_{PL,nb}$ is the harmonic constant. The PPI-GaMD simulation parameters are derived similarly as in the previous GaMD algorithm.

In addition to selectively boosting the interaction energy between proteins P and L, another boost potential is applied on the remaining potential energy of the system to enhance conformational sampling of the proteins and facilitate protein diffusion and rebinding. The second boost potential is calculated using the total system potential energy other than the interaction potential between the proteins as:

$$\Delta V_D(r) = \begin{cases} \frac{1}{2}k_D\left(E_D - V_D(r)\right)^2, & V_D(r) < E_D \\ 0, & V_D(r) \geq E_D \end{cases} \tag{4.4}$$

where V_D is the total system potential energy other than the interaction potential between the proteins, E_D is the corresponding threshold energy for applying the second boost potential and k_D is the harmonic force constant. This leads to dual-boost PPI-GaMD with the total boost potential $\Delta V(r) = \Delta V_{PL,nb}(r_{PL}) + \Delta V_D(r)$.

5. Energetic Reweighting using Cumulant Expansion to the Second Order (Gaussian Approximation)

For energetic reweighting of GaMD simulations to calculate potential of mean force (PMF), the probability distribution along a reaction coordinate is written as $p^*(A)$. Given the boost potential $\Delta V(r)$ of each frame, $p^*(A)$ can be reweighted to recover the canonical ensemble distribution, $p(A)$, as:

$$p(A_j) = p^*(A_j)\frac{\langle e^{\beta\Delta V(r)}\rangle_j}{\sum_{i=1}^{M}\langle p^*(A_i)e^{\beta\Delta V(r)}\rangle_i}, \; j = 1, \ldots, M, \tag{5.1}$$

where M is the number of bins, $\beta = k_B T$ and $\langle e^{\beta\Delta V(r)}\rangle_j$ is the ensemble-averaged Boltzmann factor of $\Delta V(r)$ for simulation frames found in the j^{th} bin. The ensemble-averaged reweighting factor can be approximated using cumulant expansion:

$$\langle e^{\beta\Delta V(r)}\rangle = exp\left\{\sum_{k=1}^{\infty}\frac{\beta^k}{k!}C_k\right\}, \tag{5.2}$$

where the first two cumulants are given by:

$$\begin{aligned} C_1 &= \langle\Delta V\rangle, \\ C_2 &= \langle\Delta V^2\rangle - \langle\Delta V\rangle^2 = \sigma_v^2. \end{aligned} \tag{5.3}$$

The boost potential obtained from GaMD simulations usually follows near-Gaussian distribution. Cumulant expansion to the second order thus provides a good approximation for computing the reweighting factor. The reweighted free energy $F(A) = -k_B T \ln p(A)$ is calculated as:

$$F(A) = F^*(A) - \sum_{k=1}^{2} \frac{\beta^k}{k!} C_k + F_c, \quad (5.4)$$

where $F^*(A) = -k_B T \ln p^*(A)$ is the modified free energy obtained from GaMD simulation and F_c is a constant.

6. Kinetic reweighting of GaMD simulations with Kramers' Rate Theory

Reweighting of biomolecular kinetics from GaMD simulations can be obtained by applying Kramers rate theory[2, 7]. For a particle climbing over potential energy barriers, Kramers showed that the reaction rate depends on temperature and viscosity of the host medium. The reaction rates were derived for both limiting cases of small and large viscosity. In the context of biomolecular simulations in aqueous medium, it is relevant for us to focus on the large viscosity limiting case. Biomolecules move in the high friction ("overdamping") regime and energy barriers are much greater than k_BT (k_B is the Boltzmann's constant and T is temperature). In this case, the reaction rate is calculated as:

$$k_R \cong \frac{2\pi w_m w_b}{\xi} e^{-\Delta F / k_B T}, \quad (6.1)$$

where w_m and w_b are frequencies of the approximated harmonic oscillators (also referred to as curvatures of free energy surface) near the energy minimum and barrier, respectively, ξ is the apparent friction coefficient and ΔF is the free energy barrier of transition.

Without the loss of generality, we consider a 1D potential of mean force (PMF) free energy profile of a reaction coordinate $F(A)$. Near minimum at A_m, the free energy can be approximated by a harmonic oscillator of frequency w_m, i.e., $F(A) = \frac{1}{2}(2\pi w_m)^2 (A - A_m)^2$. Near barrier at A_b, the free energy is approximated as $F(A) = F_b - \frac{1}{2}(2\pi w_b)^2 (A - A_b)^2$, where F_b is the free energy at A_b and w_b is the frequency of the approximated harmonic oscillator. Then we can calculate w_m and w_b as:

$$w = \sqrt{\frac{|F''(A)|}{2\pi}}, \quad (6.2)$$

where $F''(A)$ is the second-order derivative of the PMF profile.

The apparent friction coefficient ξ or diffusion coefficient D with $\xi = k_B T / D$ can be estimated as follows. First, we calculate a survival function $S(t)$ as the probability that the system remains in an energy well longer than time t. In a direct approach[8], we count the events that the system visits the energy well throughout a simulation. We record and measure the time intervals of each visiting event until the system escapes over an energy barrier. Then we

have a time series T_i, where i=1, 2, …, N, and N is the total number barrier transitions observed in the simulation. The time series is subsequently ordered such that $\hat{T}_1 \le \hat{T}_2 \le \cdots \le \hat{T}_N$. With that, the survival function is estimated as $S(\hat{T}_i) \approx 1 - i/N$, which is the probability that the system is trapped in the energy well for time longer than $\hat{T}_i$. Alternatively, we can numerically calculate the time-dependent probability density of reaction coordinate A, $\rho(A,t)$ by solving the Smoluchowski equation along 1D PMF profile of the reaction coordinate:

$$\frac{\partial \rho(A,t)}{\partial t} = D\frac{\partial}{\partial A}\left[e^{-F(A)/k_BT}\frac{\partial}{\partial A}\left(e^{F(A)/k_BT}\rho\right)\right]. \quad (6.3)$$

Then the survival function is calculated as $S(t) = \int_t^{\infty}\int_{A_{b1}}^{A_{b2}} \rho(A,t)dAdt$, where A_{b1} and A_{b2} are two boundaries of the energy well. The initial condition is often set as the Boltzmann distribution of reaction coordinate A in the energy well, i.e., $\rho(A,0) = e^{-F(A)/k_BT}$.

Second, using the above survival functions, we estimate the effective kinetic rates as the negative of the slopes in linear fitting of the ln[$S(t)$] versus t, i.e., $k = -d\ln[S(t)]/dt$. This is based on the assumption that the survival function exhibits exponential decay as observed in earlier studies. Finally, the apparent diffusion coefficient D is obtained by dividing the kinetic rate calculated directly using the transition time series collected from the simulation by that using the probability density solution of the Smoluchowski equation.

The curvatures and energy barriers of the reweighted and modified free energy profiles, as well as the apparent diffusion coefficients, are calculated and used in Kramers' rate equation to determine accelerations of biomolecular kinetics in the GaMD simulations. This allows us to recover the original biomolecular kinetic rate constants from the GaMD simulations.

References:

1. Miao, Y., V.A. Feher, and J.A. McCammon, *Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation.* Journal of Chemical Theory and Computation, 2015. **11**(8): p. 3584-3595.
2. Miao, Y., A. Bhattarai, and J. Wang, *Ligand Gaussian accelerated molecular dynamics (LiGaMD): Characterization of ligand binding thermodynamics and kinetics.* Journal of Chemical Theory and Computation, 2020. **16**(9): p. 5526–5547.
3. Wang, J. and Y. Miao, *Peptide Gaussian accelerated molecular dynamics (Pep-GaMD): Enhanced sampling and free energy and kinetics calculations of peptide binding.* The Journal of Chemical Physics, 2020. **153**(15): p. 154109.
4. Wang, J. and Y. Miao, *Protein-Protein Interaction-Gaussian Accelerated Molecular Dynamics (PPI-GaMD): Characterization of Protein Binding Thermodynamics and Kinetics.* J Chem Theory Comput, 2022. **18**(3): p. 1275-1285.
5. Vanommeslaeghe, K. and A.D. MacKerell, Jr., *CHARMM additive and polarizable force fields for biophysics and computer-aided drug design.* Biochim. Biophys. Acta, Gen. Subj., 2015. **1850**(5): p. 861-871.
6. Duan, Y., et al., *A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations.* J. Comput. Chem., 2003. **24**(16): p. 1999-2012.
7. Miao, Y., *Acceleration of Biomolecular Kinetics in Gaussian Accelerated Molecular Dynamics.* Journal of Chemical Physics, 2018. **149**(7): p. 072308.
8. Hamelberg, D., T. Shen, and J.A. McCammon, *Relating kinetic rates and local energetic roughness by accelerated molecular-dynamics simulations - art. no. 2411003.* Journal of Chemical Physics, 2005. **122**(24).