Supporting Information to “Serial generalized ensemble simulations of biomolecules with self-consistent determination of weights”

Riccardo Chelli* and Giorgio F. Signorini

Dipartimento di Chimica, Università di Firenze, Via della Lastruccia 3, I-50019 Sesto Fiorentino, Italy and

European Laboratory for Non-linear Spectroscopy (LENS),

Via Nello Carrara 1, I-50019 Sesto Fiorentino, Italy

I. CALCULATION OF THERMODYNAMICAL AVERAGES IN SERIAL GENERALIZED ENSEMBLE SIMULATIONS

Here, we report the basic equations of an estimator devised to compute thermodynamic expectations and free energy differences in a system sampled from multiple equilibrium states via computer simulations[1] (such as serial generalized ensemble and replica exchange simulations). Following Shirts and Chodera[1], we call the estimator multistate Bennett acceptance ratio estimator (MBAR), even if Kumar *et al.*[2] presented it in 1992 as a straightforward extension of the weighted histogram analysis method (WHAM) developed by Ferrenberg and Swendsen[3, 4]. In fact, MBAR is equivalent to WHAM in the limit that histogram bin widths are shrunk to zero[1, 2]. However, unlike WHAM, MBAR avoids the errors introduced by the histogram discretization[1, 2, 5]. The MBAR equations follow.

Suppose to have a set of configurations of a system sampled from various equilibrium simulations characterized by different Hamiltonians. In the context of serial generalized ensemble methods, such configurations arise from sampling in different temperature or λ -ensembles (simulated tempering the former, simulated solute tempering or multi-window tempering the latter). Let us give some definitions

1. $h_n(x, p)$ is the dimensionless Hamiltonian of the (x, p) system configuration sampled in the ensemble n , where x and p are the atomic positions and momenta, respectively. Examples of dimensionless Hamiltonians can be found in Eqs. 11, 13 and 14 of the main text.
2. f_n is the dimensionless free energy of the n ensemble associated with the Hamiltonian $h_n(x, p)$.
3. N is the number of ensembles (in our simulations $N = 8$).
4. N_n is the number of system configurations sampled in the ensemble n .
5. (x_{nj}, p_{nj}) denotes the j th system configuration sampled in the ensemble n .
6. $H(x, p)$ is the physical dimensionless Hamiltonian for an arbitrary configuration (x, p) . The probability of (x, p) at equilibrium conditions is proportional to $\exp[-H(x, p)]$. In the cases of simulated solute tempering and multi-window tempering (see Sec. “Implementations of the serial generalized ensemble method (in the ORAC program)” of the main text), the physical Hamiltonian is $H(x, p) = \beta[V(x) + u(p)]$, where $V(x)$ is the potential energy and $u(p)$ the kinetic energy (we refer to the main text for a more detailed discussion on the role of $u(p)$ in molecular dynamics simulations).
7. $A(x)$ is some mechanical observable that depends only on the configuration x .

Given the above definitions, the equilibrium expectation of the observable A is

$$\hat{A} = \sum_{m=1}^N \sum_{j=1}^{N_m} W_{mj} A(x_{mj}), \quad (1)$$

*Electronic address: riccardo.chelli@unifi.it

where W_{mj} is the weight factor for the configuration (x_{mj}, p_{mj}) , the sum on m runs over the ensembles and the sum on j runs over the configurations sampled in the ensemble m . The generic weight W_{mj} is

$$W_{mj} = C^{-1} \frac{\exp[-H(x_{mj}, p_{mj})]}{\sum_{n=1}^N N_n \exp[f_n - h_n(x_{mj}, p_{mj})]}, \quad (2)$$

where C is the normalization factor such that

$$\sum_{m=1}^N \sum_{j=1}^{N_m} W_{mj} = 1. \quad (3)$$

In order to calculate W_{mj} , the dimensionless free energies f_n for each $n \in (1, 2, \dots, N)$ must be known. The estimating equations for the dimensionless free energies are

$$f_m = -\ln \left(\sum_{q=1}^N \sum_{j=1}^{N_q} \frac{\exp[-h_m(x_{qj}, p_{qj})]}{\sum_{n=1}^N N_n \exp[f_n - h_n(x_{qj}, p_{qj})]} \right), \quad (4)$$

which must be solved self-consistently for f_m . Because normalization factors are only determined up to a multiplicative constant, the estimated free energies f_m are determined uniquely only up to an additive constant, so only differences $\Delta f_{i \rightarrow j} = f_j - f_i$ will be meaningful. Efficient solution of Eq. 4 is reported in the appendix C of Ref. [1].

We now report Eqs. 2 and 4 for the specific cases treated in our work.

A. Simulated solute tempering

In the case of simulated solute tempering, Eqs. 2 and 4 can be recovered taking into account the expression of the dimensionless Hamiltonian of Eq. 11 and exploiting the equivalence $H(x, p) = h_1(x, p)$. With few simple algebra passages, Eq. 2 becomes

$$W_{mj} = C^{-1} \frac{\exp\{-\beta[\lambda_1^{\text{ss}} \cdot \mathbf{v}_{\text{ss}}(x_{mj}) + \lambda_1^{\text{se}} \cdot \mathbf{v}_{\text{se}}(x_{mj})]\}}{\sum_{n=1}^N N_n \exp\{f_n - \beta[\lambda_n^{\text{ss}} \cdot \mathbf{v}_{\text{ss}}(x_{mj}) + \lambda_n^{\text{se}} \cdot \mathbf{v}_{\text{se}}(x_{mj})]\}}, \quad (5)$$

where $\lambda_1^{\text{ss}} = \lambda_1^{\text{se}} = (1, 1, \dots, 1)$. For Eq. 4 we get

$$f_m = -\ln \left(\sum_{q=1}^N \sum_{j=1}^{N_q} \frac{\exp\{-\beta[\lambda_m^{\text{ss}} \cdot \mathbf{v}_{\text{ss}}(x_{qj}) + \lambda_m^{\text{se}} \cdot \mathbf{v}_{\text{se}}(x_{qj})]\}}{\sum_{n=1}^N N_n \exp\{f_n - \beta[\lambda_n^{\text{ss}} \cdot \mathbf{v}_{\text{ss}}(x_{qj}) + \lambda_n^{\text{se}} \cdot \mathbf{v}_{\text{se}}(x_{qj})]\}} \right), \quad (6)$$

To calculate the Φ - Ψ distribution function, $g(\Phi, \Psi)$, the $A(x_{mj})$ to be used in Eq. 1 is defined as

$$A(x_{mj}) = \delta(\Phi - \Phi_{mj}) \delta(\Psi - \Psi_{mj}), \quad (7)$$

where $\delta(x)$ is the Dirac delta function, while Φ_{mj} and Ψ_{mj} are the Φ and Ψ dihedral angles of the configuration j sampled in the ensemble m .

B. Multi-window tempering

For the multi-window tempering technique, the above equations are obtained substituting Eq. 14 into Eqs. 2 and 4 and considering that $H(x, p) = V(x) + u(p)$

$$W_{mj} = \left\{ C \sum_{n=1}^N N_n \exp[f_n - \beta k(r_{mj} - \lambda_n^{\text{dist}})^2] \right\}^{-1}, \quad (8)$$

$$f_m = -\ln \left(\sum_{q=1}^N \sum_{j=1}^{N_q} \frac{\exp[-\beta k(r_{qj} - \lambda_m^{\text{dist}})^2]}{\sum_{n=1}^N N_n \exp[f_n - \beta k(r_{qj} - \lambda_n^{\text{dist}})^2]} \right), \quad (9)$$

where r_{mj} is the distance between two selected atoms in the system configuration j sampled in the ensemble m .

-
- [1] M. R. Shirts and J. D. Chodera, J. Chem. Phys. **129**, 124105 (2008).
 - [2] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, J. Comput. Chem. **13**, 1011 (1992).
 - [3] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **61**, 2635 (1988).
 - [4] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989).
 - [5] M. N. Kobra, J. Comput. Chem. **24**, 1437 (2003).

II. FIGURES

FIG. 1S: Normalized autocorrelation functions of the dimensionless work as a function of time (see the definition of $W[n \rightarrow m]$ introduced in Eq. (16) of the main text). In particular, the curves refer to three representative work series, *i.e.* $W[1 \rightarrow 2]$, $W[4 \rightarrow 5]$ and $W[8 \rightarrow 7]$, recorded during standard molecular dynamics simulations realized in the ensembles λ_1 , λ_4 and λ_8 , respectively (see Table 1 of the main text for the definition of the ensembles). The vertical line is drawn to mark the work values after a period corresponding to the work storing time step (60 fs). The functions show that loss of correlation between consecutive stored works, *i.e.* after 60 fs, is about 50%.

FIG. 2S: Walks of replicas through the λ_m ensembles for the SGE-8 simulation (specifically, m versus time). Each panel refers to a replica. For completeness, the panel (b) of Figure 2 of the main text is also reported here (top left panel). For the sake of clarity, only the first nanosecond of simulation is reported.

FIG. 3S: Walks of replicas through the λ_m ensembles for the RE simulation (specifically, m versus time). Each panel refers to a replica. For completeness, the panel (a) of Figure 2 of the main text are reported here (top left panel). For the sake of clarity, only the first nanosecond of simulation is reported.

FIG. 4S: Population function $P_{i\alpha}(t)$ (Eq. 19 of the main text) for all replicas of the SGE-8 simulation. To highlight the early time behavior, only data for half simulation run are shown. The correspondences between colors and λ ensembles follow: black/ λ_1 , red/ λ_2 , green/ λ_3 , blue/ λ_4 , orange/ λ_5 , brown/ λ_6 , magenta/ λ_7 , violet/ λ_8 . The value corresponding to a homogeneous distribution, *i.e.* 12.5, is represented by the dashed lines. For completeness, the data drawn in the left panel of Figure 3 of the main text are reported here in the top left panel.

FIG. 5S: Population function $P_{i\alpha}(t)$ (Eq. 19 of the main text) for all replicas of the RE simulation. To highlight the early time behavior, only data for half simulation run are shown. The correspondences between colors and λ ensembles follow: black/ λ_1 , red/ λ_2 , green/ λ_3 , blue/ λ_4 , orange/ λ_5 , brown/ λ_6 , magenta/ λ_7 , violet/ λ_8 . The value corresponding to a homogeneous distribution, *i.e.* 12.5, is represented by the dashed lines. For completeness, the data drawn in the right panel of Figure 3 of the main text are reported in the top left panel.

FIG. 6S: Time series of dimensionless free energy differences between neighboring λ ensembles (see y axes) calculated using BAR-SGE. Red lines: SGE-1 simulation; black lines: SGE-8 simulation; dashed lines: reference value calculated from RE simulation using MBAR with all trajectory data (12 ns per replica). The curves are drawn until 6 ns to highlight the early time behavior.

FIG. 7S: Time series of dimensionless free energy differences between neighboring λ ensembles (see y axes). Red lines: RE simulation using MBAR; blue lines: SGE-1 simulation using MBAR; green lines: SGE-8 simulation using MBAR; black lines: SGE-8 simulation using BAR-SGE; dashed lines: reference value calculated from RE simulation using MBAR with all trajectory data (12 ns per replica). The curves are drawn until 6 ns to highlight the early time behavior.

FIG. 8S: Φ - Ψ projection of the free energy surface, $F(\Phi, \Psi)$, calculated from the SGE-8 and SGE-1 simulations using data till 1.5 ns per replica and 12 ns per replica, respectively (see labels on the panels). The chromatic scale on the right of the panels is in units of kJ mol^{-1} . The deep brown color marks regions with out of scale $F(\Phi, \Psi)$. The histogram channel size for Φ and Ψ is 7.2° .