THE JOURNAL OF PHYSICAL CHEMISTRY Letters

Letter

pubs.acs.org/JPCL

A Stochastic Solution to the Unbinned WHAM Equations

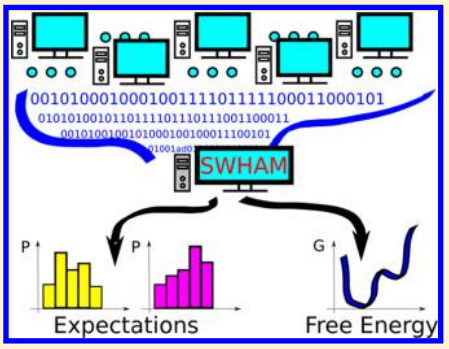
Bin W. Zhang,[†] Junchao Xia,[†] Zhiqiang Tan,[‡] and Ronald M. Levy*[,†]

[†]Center for Biophysics and Computational Biology, Department of Chemistry, and Institute for Computational Molecular Science, Temple University, Philadelphia, Pennsylvania 19122, United States

[‡]Department of Statistics, Rutgers University, Piscataway, New Jersey 08854, United States

Supporting Information

ABSTRACT: The weighted histogram analysis method (WHAM) and unbinned versions such as the multistate Bennett acceptance ratio (MBAR) and unbinned WHAM (UWHAM) are widely used to compute free energies and expectations from data generated by independent or coupled parallel simulations. Here we introduce a replica exchange-like algorithm (RE-SWHAM) that can be used to solve the UWHAM equations stochastically. This method is capable of analyzing large data sets generated by hundreds or even thousands of parallel simulations that are too large to be "WHAMMED" using standard methods. We illustrate the method by applying it to obtain free energy weights for each of the 240 states in a simulation of host−guest ligand binding containing ~3.5×10^7 data elements collected from 16 parallel Hamiltonian replica exchange simulations, performed at 15 temperatures. In addition to using much less memory, RE-SWHAM showed a nearly 80-fold improvement in computational time compared with UWHAM.



The weighted histogram analysis method (WHAM) is used to compute free energies and expectations from data generated by multicanonical simulations and independent simulations at multiple Hamiltonian or thermodynamic states.[1−13] For example, after running umbrella sampling, the standard procedure is to use WHAM to combine the pieces of the free energy profile along chosen reaction coordinates.[4] After running replica exchange (RE) simulations,[14,15] WHAM is used to estimate the free energy differences between Hamiltonian states, or to obtain expectation values at the targeted thermodynamic states from the simulation data at all of the thermodynamic states.[6,7] In this Letter we describe a RE-like algorithm we have developed that corresponds to a stochastic solution of the UWHAM equations. We refer to this algorithm as replica exchange stochastic WHAM or RE-SWHAM. Since running RE-SWHAM requires much less computing power compared with present WHAM-based analysis tools such as MBAR[8] and UWHAM,[5,9] we believe RE-SWHAM is a promising method for analyzing very large ensembles of data generated by multicanonical simulations or massively parallel but independent data sets generated on very large computer grids.[16,17]

Before introducing RE-SWHAM, we review the binless WHAM (UWHAM) equations.[9] In parallel simulations, if each simulation has different thermodynamic parameters such as temperature and pressure, they are referred to as simulations at different thermodynamic states; if each simulation has different potential energy functions, such as in Hamiltonian RE simulations, they can be referred to as simulations at different alchemical Hamiltonian states. Moreover, parallel simulations can differ in both thermodynamic parameters and potential energy functions such as in two-dimensional RE simulations. For simplicity, we refer to each of these states, which is characterized by a specific combination of thermodynamic parameters and potential energy functions, as a thermodynamic state, in order to avoid confusion with our description of the UP and DOWN conformational states. Consider a system of M parallel simulations labeled by Greek letters. Each thermodynamic state has a biased potential $w_\alpha(u)$, where u is the reduced coordinate of the observation. Suppose $u_{\alpha i}$ is the ith observation at the αth thermodynamic state, and N_α is the total number of observations at the αth thermodynamic state. The UWHAM estimates of the density of states $\Omega(u_{\gamma i})$ and the partition function Z_α are (up to a multiplicative constant):[6,9]

$$\hat{Z}_\alpha = \sum_{\gamma=1}^{M} \sum_{i=1}^{N_\gamma} c_\alpha(u_{\gamma i}) \hat{\Omega}(u_{\gamma i}) \tag{1}$$

$$\hat{\Omega}(u_{\gamma i}) = \frac{1}{\sum_{\kappa=1}^{M} N_\kappa \hat{Z}_\kappa^{-1} c_\kappa(u_{\gamma i})} \tag{2}$$

where

$$c_\alpha(u_{\gamma i}) = \exp[-\beta_\alpha w_\alpha(u_{\gamma i})] \tag{3}$$

is the bias factor. Eqs 1, 2, and 3 constitute a coupled set of equations which can by solved by Newton iteration[8] or optimization.[9,10] The UWHAM estimate of the probability of the observation $u_{\gamma i}$ at the αth state is

Received: August 12, 2015
Accepted: September 9, 2015

DOI: 10.1021/acs.jpclett.5b01771
J. Phys. Chem. Lett. 2015, 6, 3834−3840

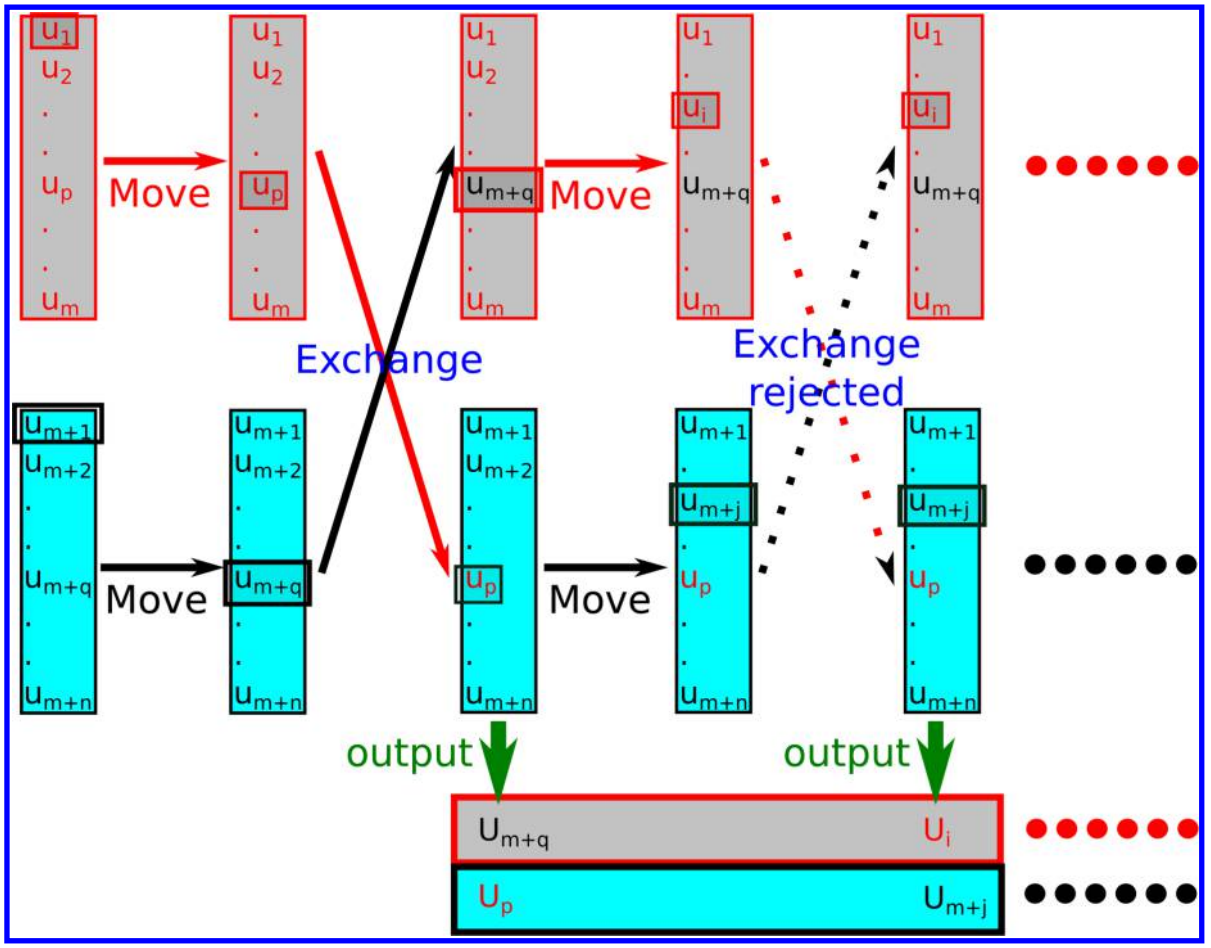


Figure 1. An illustration of the RE-SWHAM algorithm. This drawing illustrates two replica exchange cycles of the RE-SWHAM method, and shows only two thermodynamic states with "grey" or "cyan" color. In each cycle, one data element is chosen from each thermodynamic state first, then a replica exchange is performed. In the first cycle, since the swap is accepted, the data associated with the two replicas is swapped to the other thermodynamic state's data array. At the end of each cycle, the data associated with replicas are recorded as the output like explicit RE simulations.

$$\hat{p}_\alpha(u_{\gamma i}) = \hat{Z}_\alpha^{-1}\hat{\Omega}(u_{\gamma i})c_\alpha(u_{\gamma i}) \tag{4}$$

We note that although the solution to the UWHAM equations depends on the number of observations at each thermodynamic state N_α, it does not depend on the original thermodynamic state at which each sample $u_{\gamma i}$ is observed.

We now describe a stochastic Monte Carlo approach to solve the coupled eqs 1, 2, and 3 inspired by the replica exchange algorithm. Consider a system of M parallel simulations labeled by Greek letters. Suppose there are N_α observations at the αth thermodynamic state, and N is the total number of observations, namely $N = \sum_{\alpha=1}^{M} N_\alpha$. To initialize RE-SWHAM we first construct a weight array of N elements for each thermodynamic state, corresponding to the N observations. (A computationally efficient algorithm implementing RE-SWHAM is presented later and illustrated by Figure 1.) There are M arrays of N elements, one for each thermodynamic state. The individual elements of the weight array have value 0 or 1 at every iteration. Initially, we set all the elements corresponding to the data actually observed at this thermodynamic state to 1, and the others to 0. Considering the αth thermodynamic state, the initial weight array is

$$\begin{aligned} X_\alpha(t=0) = \{&\delta_\alpha(u_{11}, t=0) = 0,\cdots,\delta_\alpha(u_{\alpha 1}, t=0) \\ &= 1, \delta_\alpha(u_{\alpha 2}, t=0) = 1,\cdots,\delta_\alpha(u_{\alpha N_\alpha}, t=0) = 1,\cdots, \\ &\delta_\alpha(u_{\gamma i}, t=0) = 0,\cdots,\delta_\alpha(u_{MN_M}, t=0) = 0\} \end{aligned} \tag{5}$$

where $\gamma \neq \alpha$, and it has N_α nonzero elements. For simplicity, we reindex the N observations and omit the label of the original thermodynamic state at which each data element was observed since that information is not required. The weight array for the αth thermodynamic state at time t is referred to as $X_\alpha(t) = \{\delta_\alpha(u_1, t), \delta_\alpha(u_2, t), \cdots, \delta_\alpha(u_i, t), \cdots, \delta_\alpha(u_N, t)\}$, where $\delta_\alpha(u_i, t) = 1$ if the observation u_i occupies the αth thermodynamic state at t, $\delta_\alpha(u_i, t) = 0$ otherwise.

The Metropolis Monte Carlo (MC) RE-SWHAM algorithm resembles running explicit replica exchange simulations, which consists of cycles. There are two components in each cycle: (i) the molecular dynamics (MD) or MC simulation of each replica at a fixed thermodynamic state (the "move" process), which includes comformational relaxation at that thermodynamic state; (ii) the attempted swaps of replicas (the "exchange" process), which is the relaxation in the replica and thermodynamic state permutation space. The RE-SWHAM algorithm also runs by replica exchange cycles. At each cycle, a data element at each thermodynamic state is chosen based on the current normalized weights, $\delta_\alpha(u_i, t)/\sum_{j=1}^{N} \delta_\alpha(u_j, t) = \delta_\alpha(u_i,$

DOI: 10.1021/acs.jpclett.5b01771
J. Phys. Chem. Lett. 2015, 6, 3834−3840

$t)/N_\alpha$. Since in RE-SWHAM the elements of the weight array at each iteration represent the set of N_α out of N data elements that currently occupy that thermodynamic state, randomly choosing the next data element is analogous to the move process of an explicit RE simulation when its MD simulation period per cycle is so long that two adjacent configurations chosen for replica exchange have no correlations. Then a set of replica exchanges is performed which follows the same exchange criterion as in explicit replica exchange simulations (see the Metropolis function in eq 9). In this Letter we use the independence sampling algorithm as the proposal scheme,[18] which attempts to exchange replicas from two thermodynamic states chosen at random.

RE-SWHAM updates the weights of observations of thermodynamic states based on the exchange process. When an exchange attempt is accepted, RE-SWHAM swaps the thermodynamic states of the two replicas, and changes the weights of the observations associated with the replicas at both thermodynamic states. For example, if one replica associated with the observation u_i at the αth thermodynamic state hops to the γth thermodynamic state at time t, RE-SWHAM changes the weight of u_i to 0 at the αth thermodynamic state, namely, $\delta_\alpha(u_i, t) = 0$, and to 1 at the γth thermodynamic state, namely, $\delta_\gamma(u_i, t) = 1$. As part of this exchange move, a corresponding data element u_j moves from the γth to the αth thermodynamic state. Notice that in the RE-SWHAM algorithm the number of nonzero weights remains the same at each thermodynamic state and each observation has a nonzero weight at one and only one thermodynamic state. Following this procedure, the instantaneous weight of each data element at each thermodynamic state oscillates as 0 and 1, but the normalized time-average of the weight converges to the unbiased estimate of the probability of the data element, $\hat{p}_\alpha(u_i)$, in eq 4.

The RE-SWHAM algorithm performs a random walk in the space of the weight arrays of observations. In each cycle, RE-SWHAM moves from one set of weight arrays, or a configuration, to another following the same exchange criterion as that for multicanonical replica exchange simulations:

$$\mathbf{X}(0) \xrightarrow{\text{RE cycle}} \mathbf{X}(1) \xrightarrow{\text{RE cycle}} \cdots \mathbf{X}(t) \xrightarrow{\text{RE cycle}} \cdots \tag{6}$$

where $\mathbf{X}(t)$ is a set of weight arrays $\{X_1(t), X_2(t), \cdots, X_\alpha(t), \cdots, X_M(t)\}$ with $X_\alpha(t)$ corresponding to the weight array at the αth thermodynamic state as described earlier. The total number of configurations of this Markov chain corresponds to the total number of ways to assign N nonzero instantaneous weights to M thermodynamic states, with the number of observations at each thermodynamic state fixed at $\{N_1, N_2, \cdots, N_\alpha, \cdots, N_M\}$, which is $N!/(\prod_{\alpha=1}^{M} N_\alpha!)$. The equivalent Master equation perspective for this random walk is

$$\frac{d\mathbf{P}_\mathbf{X}(t)}{dt} = \mathbf{A}\cdot\mathbf{P}_\mathbf{X}(t) \tag{7}$$

where $\mathbf{A}$ is the rate matrix and $\mathbf{P}_\mathbf{X}(t)$ represents the vector probability of all the possible configurations. The solution of eq 7 can be written as

$$\mathbf{P}_\mathbf{X}(t + \Delta t) = \mathbf{P}_\mathbf{X}(t)\cdot\mathbf{T} \tag{8}$$

where $\mathbf{T}$ is the row-normalized transition matrix.

It is possible to write down the transition matrix for RE-SWHAM when one exchange is attempted per cycle. Suppose at the beginning of a cycle, RE-SWHAM is at the ith configuration (which corresponds to a specification of the N instantaneous weights at each of the M thermodynamic states), and two replicas are chosen at random to attempt an exchange. Consider the trial move in RE-SWHAM, which attempts to exchange one observation u_m at the αth thermodynamic state and the other observation u_n at the γth thermodynamic state. The new configuration will be called the jth configuration, which is the same as the ith configuration, except that the weights of u_m and u_n are exchanged in the weight arrays at the αth and the γth thermodynamic states. The probability that this trial move is accepted, namely, that the exchange from the ith configuration to the jth configuration is accepted, is

$$\begin{aligned} T_{ij} &= \frac{2}{M(M-1))}\frac{1}{N_\alpha}\frac{1}{N_\gamma}\min\left(1, \frac{\exp[-\beta_\alpha w_\alpha(u_n)]\exp[-\beta_\gamma w_\gamma(u_m)]}{\exp[-\beta_\alpha w_\alpha(u_m)]\exp[-\beta_\gamma w_\gamma(u_n)]}\right) \\ &= \frac{2}{M(M-1))}\frac{1}{N_\alpha}\frac{1}{N_\gamma}\Psi\left(\log\left[\frac{c_\alpha(u_m)c_\gamma(u_n)}{c_\alpha(u_n)c_\gamma(u_m)}\right]\right) \end{aligned} \tag{9}$$

where the first factor $2/(M(M-1)))$ is the probability of choosing the replicas at the αth and the γth thermodynamic states from the M replicas, and the factors $1/N_\alpha$ and $1/N_\gamma$ are the probabilities of choosing one observation from the respective thermodynamic states. Throughout, Ψ is the Metropolis function[19]

$$\Psi(x) = \min(1, \exp[-x]) \tag{10}$$

Consider the reverse trial exchange move from the jth configuration to the ith configuration, in other words, the swap of the observation u_n at the αth state and the observation u_m in at the γth state with all other weights in configuration i and j having the same values. The probability of this move is

$$T_{ji} = \frac{2}{M(M-1))}\frac{1}{N_\alpha}\frac{1}{N_\gamma}\Psi\left(\log\left[\frac{c_\alpha(u_n)c_\gamma(u_m)}{c_\alpha(u_m)c_\gamma(u_n)}\right]\right) \tag{11}$$

By construction, the stationary probabilities of the ith and jth configurations, denoted by p_i and p_j, satisfy the detailed balance condition:

$$p_i T_{ij} = p_j T_{ji} \tag{12}$$

Next consider a subgroup I including all the configurations that have the observation u_m at the αth thermodynamic state and the other observation u_n at the γth thermodynamic state, and another subgroup J including all the configurations that have the observation u_n at the αth thermodynamic state and the other observation u_m at the γth thermodynamic state. For each configuration k in the subgroup I, there exists one configuration l in the subgroup J for which the only difference between these two configurations are the thermodynamic state the observation u_m and u_n belongs to. Every pair of such configurations, (k, l) with $k \in I$ and $l \in J$, satisfies $p_k T_{ij} = p_l T_{ji}$, where T_{ij} and T_{ji} remain the same regardless of (k, l). Then the total probabilities of these two subgroups satisfy

$$\left(\sum_{k\in I} p_k\right) T_{ij} = \left(\sum_{l\in J} p_l\right) T_{ji} \tag{13}$$

Theoretically the correlation is nonzero between the occurrence of the mth observation at the αth thermodynamic state and the occurrence of the nth observation at the γth thermodynamic state (or the occurrence of the mth observation

DOI: 10.1021/acs.jpclett.5b01771
J. Phys. Chem. Lett. 2015, 6, 3834−3840

at the γth thermodynamic state and the occurrence of the nth observation at the αth thermodynamic state). However, such correlations become negligible when the total number of observations at each thermodynamic state is large. Therefore, the total probabilities of the subgroup I and J are

$$\begin{aligned}\sum_{k\in I} p_k &= \langle \delta_\alpha(u_m, t)\delta_\gamma(u_n, t)\rangle \approx \langle \delta_\alpha(u_m, t)\rangle\langle \delta_\gamma(u_n, t)\rangle \\ &= N_\alpha \tilde{p}_\alpha(u_m) N_\gamma \tilde{p}_\gamma(u_n) \\ \sum_{l\in J} p_l &= \langle \delta_\alpha(u_n, t)\delta_\gamma(u_m, t)\rangle \approx \langle \delta_\alpha(u_n, t)\rangle\langle \delta_\gamma(u_m, t)\rangle \\ &= N_\alpha \tilde{p}_\alpha(u_n) N_\gamma \tilde{p}_\gamma(u_m),\end{aligned} \tag{14}$$

where $\langle \delta_\alpha(u_m, t)\rangle$ is the time-average weight of the observation u_m at the αth thermodynamic state, and $\tilde{p}_\alpha(u_m)$ is the normalized RE-SWHAM time-average weight, namely $\tilde{p}_\alpha(u_m) = \langle \delta_\alpha(u_m, t)\rangle/N_\alpha$. Combining eqs 9, 11, 13−14 and applying the property $\Phi(x)/\Phi(-x) = \exp(-x)$ led to the detailed balance relation

$$\frac{\tilde{p}_\alpha(u_m)/c_\alpha(u_m)}{\tilde{p}_\gamma(u_m)/c_\gamma(u_m)} = \frac{\tilde{p}_\alpha(u_n)/c_\alpha(u_n)}{\tilde{p}_\gamma(u_n)/c_\gamma(u_n)}, \quad \text{for all } (u_m, u_n) \text{ and } (\alpha, \gamma) \tag{15}$$

Denote the common value of the ratios in eq 15 by $f(\alpha, \gamma)$, which depends on (α, γ), but not (u_m, u_n), and fix γ at a baseline thermodynamic state, say α^0. Then eq 15 gives

$$\tilde{p}_\alpha(u_m)/c_\alpha(u_m) = f(\alpha, \alpha^0)(\tilde{p}_{\alpha^0}(u_m)/c_{\alpha^0}(u_m)) \tag{16}$$

That is, the probability $\tilde{p}_\alpha(u_m)$ can be expressed in the form

$$\tilde{p}_\alpha(u_m) = \tilde{Z}_\alpha^{-1}\tilde{\Omega}(u_m)c_\alpha(u_m), \quad \text{for all } u_m \text{ and } \alpha \tag{17}$$

where $\tilde{Z}_\alpha = 1/f(\alpha, \alpha^0)$ and $\tilde{\Omega}(u_m) = \tilde{p}_{\alpha^0}(u_m)/c_{\alpha^0}(u_m)$. Since $\sum_{m=1}^{N} \tilde{p}_\alpha(u_m) = 1$ at each thermodynamic state, summing both sides of eq 17 over m yields

$$\tilde{Z}_\alpha = \sum_{m=1}^{N} c_\alpha(u_m)\tilde{\Omega}(u_m) \tag{18}$$

As mentioned previously, the RE-SWHAM algorithm keeps the total number of observations, N_α, unchanged at each thermodynamic state; the normalized time-average weight of every observation $\tilde{p}_\alpha(u_m)$ satisfies $\sum_{\alpha=1}^{M} N_\alpha \tilde{p}_\alpha(u_m) = \sum_{\alpha=1}^{M} \langle \delta_\alpha(u_m, t)\rangle = 1$, because each observation appears at one and only one thermodynamic state at one time. Multiplying both sides of eq 17 by N_α and summing over α yields

$$\tilde{\Omega}(u_m) = \frac{1}{\sum_{\alpha=1}^{M} N_\alpha \tilde{Z}_\alpha^{-1} c_\alpha(u_m)} \tag{19}$$

Thus, the RE-SWHAM estimates $\{\tilde{\Omega}(u_m)\colon m = 1,..., N\}$ and $\{\tilde{Z}_\alpha\colon \alpha = 1, ..., M\}$ satisfy eqs 18−19, which are equivalent to the UWHAM equations eqs 1−2. However, the UWHAM estimates $\{\hat{\Omega}(u_m)\colon m = 1,..., N\}$ and $\{\hat{Z}_\alpha\colon \alpha = 1,..., M\}$are, up to a multiplicative constant, unique solutions to eq 1−2.[5,9] Then $\tilde{\Omega}(u_m) = \hat{\Omega}(u_m)$ for all $m = 1,...,N$ and $\tilde{Z}_\alpha = \hat{Z}_\alpha$ for all $\alpha = 1,..., M$ provided that the same baseline thermodynamic state α^0 is used in UWHAM and RE-SWHAM. Therefore, by eq 4 and 17, the estimated probabilities from RE-SWHAM agree with those from UWHAM: $\tilde{p}_\alpha(u_m) = \hat{p}_\alpha(u_m)$.

In the previous discussion, we showed how to write any element of the transition matrix $\mathbf{T}$ for RE-SWHAM with one exchange attempt per cycle. Theoretically, one can obtain the UWHAM solution by diagonalizing the transition matrix $\mathbf{T}$ and solving the equilibrium probabilities of all the possible configurations. This is impractical because of the overwhelming size of the transition matrix; however, RE-SWHAM provides a Markov Chain Monte Carlo solution to the problem.

RE-SWHAM, described conceptually above, can be implemented in practice as a computationally efficient algorithm. See Figure 1 for an illustration of RE-SWHAM for a problem containing two thermodynamic states. To initialize RE-SWHAM, instead of constructing a weight array, we construct a current data array for each thermodynamic state using all the data elements observed from that thermodynamic state. In Figure 1, there are m observations originally from the "grey" state and n observations from the "cyan" state. Then RE-SWHAM is run by cycles: first, one of the data elements is chosen with equal probability from each thermodynamic state. Second, a replica exchange attempt is performed following the multicanonical exchange criterion. When an exchange attempt is accepted, the thermodynamic states of the two replicas are swapped, and the observations (data elements) associated with these two replicas are also swapped to the other thermodynamic state's data array. As shown by the first replica exchange cycle in Figure 1, the observation u_p is swapped to the "cyan" state and the observation u_{m+q} is swapped to the "grey" state. At the end of each cycle, no matter whether the exchange attempt is accepted, the observation associated with the replica at each thermodynamic state is recorded as the output like explicit RE simulations. Notice our implementation does not track the weight array but tracks the explicit observations with nonzero weights for each thermodynamic state to save the memory, since the total number of weights, $(M \times \sum_{\alpha=1}^{M} N_\alpha)$, increases rapidly with the number of thermodynamic states M. Although the illustration shows only one exchange attempt per cycle, in practice performing multiple exchange attempts per cycle accelerates the convergence of RE-SWHAM.[18]

The output of RE-SWHAM at each thermodynamic state is a sample of all observations according to their UWHAM weights from eq 4 or RE-SWHAM weights from eq 17. The free energy difference between two adjacent thermodynamic states can be obtained by the standard "free energy perturbation formula" (FEP),[20] but note that unlike standard FEP, in eq 20 the reweighted (i.e., maximum likelihood) density of states appears. For instance, the free energy difference between the αth and $(\alpha + 1)$th thermodynamic states is

$$\begin{aligned}-k_B T \ln\left(\frac{\tilde{Z}_{\alpha+1}}{\tilde{Z}_\alpha}\right) &= -k_B T \ln \frac{\sum_{i=1}^{N} \tilde{\Omega}(u_i)c_{\alpha+1}(u_i)}{\sum_{i=1}^{N} \tilde{\Omega}(u_i)c_\alpha(u_i)} \\ &= -k_B T \ln \frac{\sum_{i=1}^{N} \tilde{\Omega}(u_i)c_\alpha(u_i)\left(\frac{c_{\alpha+1}(u_i)}{c_\alpha(u_i)}\right)}{\sum_{i=1}^{N} \tilde{\Omega}(u_i)c_\alpha(u_i)} \\ &= -k_B T \ln\left\langle \frac{c_{\alpha+1}(u_i)}{c_\alpha(u_i)} \right\rangle_\alpha\end{aligned} \tag{20}$$

where the triangular brackets denote an average over the u_i values sampled at the αth thermodynamic state according to their converged RE-SWHAM probabilities $\tilde{p}_\alpha(u_i)$.

DOI: 10.1021/acs.jpclett.5b01771
J. Phys. Chem. Lett. 2015, 6, 3834−3840

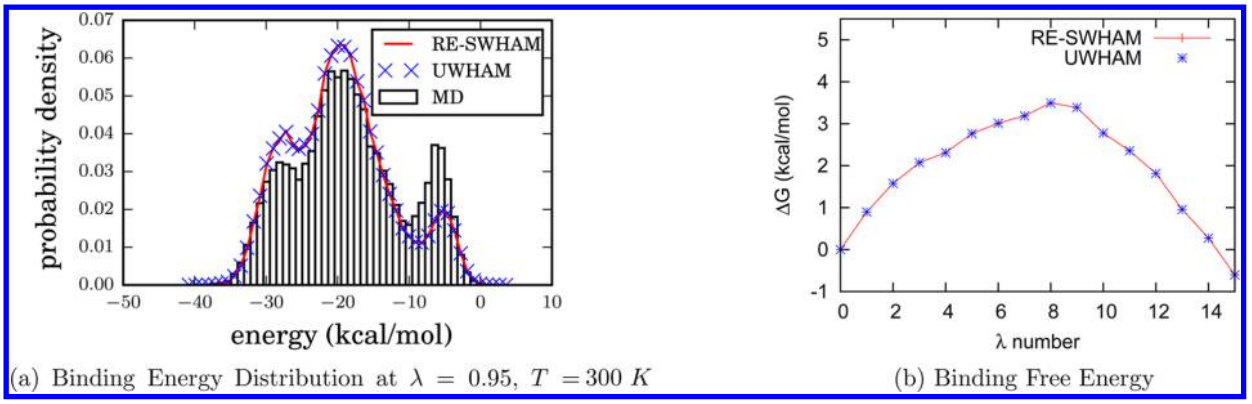


Figure 2. Numerical comparisons between UWHAM and RE-SWHAM. (a) Heptanoate-β-cyclodextrin binding energy distribution of $\lambda = 0.95$ state at 300 K obtained from a 12 ns independent MD simulation, and the UWHAM and RE-SWHAM estimates calculated from 16 independent 12 ns MD simulations run at different λ states. (b) Free energy differences between these 16 λ states at 300 K estimated by UWHAM and RE-SWHAM from a 72 ns RE simulation. In both plots, the UWHAM and RE-SWHAM results are indistinguishable.

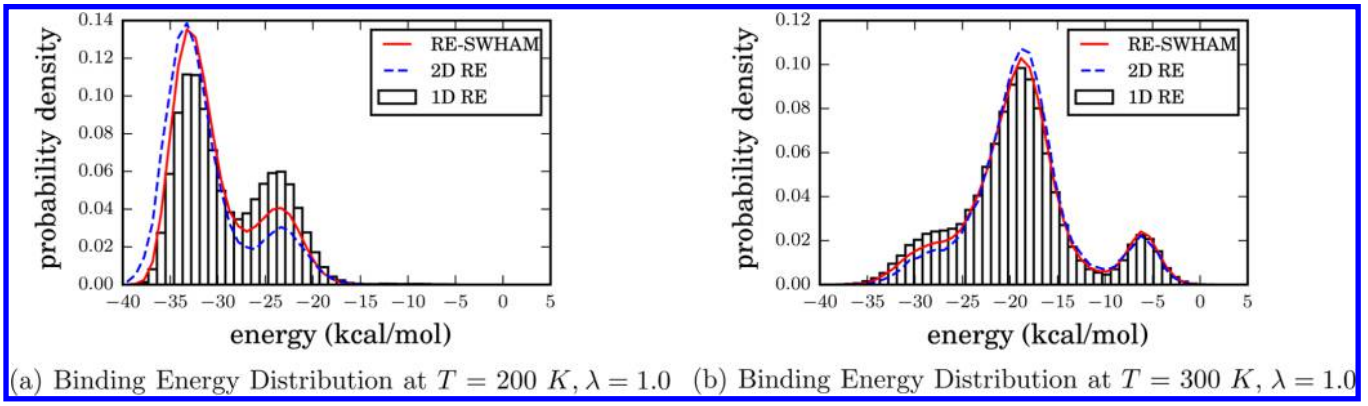


Figure 3. An application of RE-SWHAM to analyze a very large data set containing 240 thermodynamic states and 3.456×10^7 data elements. These plots show the binding energy distributions of $\lambda = 1.0$ state at 200 K and 300 K obtained from one-dimensional RE simulations and the respective RE-SWHAM estimates calculated from 15 independent one-dimensional RE simulations run at 15 different temperatures. The blue dash lines are the binding energy distributions obtained from an asynchronous two-dimensional 240 state RE simulation, which serve as a converged benchmark. The estimates from RE-SWHAM run for 12 min agree with those obtained from UWHAM (not shown), which required 15.7 h to converge. The RE-SWHAM results show significant improvements for the binding energy distribution at 200 K compared with the unconverged one-dimensional RE simulation data. At 300 K the one-dimensional RE simulation data are better converged, but even for this data set, reweighting with RE-SWHAM leads to some improvements in the estimates of the low energy tail of the distribution.

We include three numerical examples of the application of RE-SWHAM in this Letter. The first example is to estimate the binding of a guest molecule to a host at 300 K from 16 independent 12 ns long MD simulations with different Hamiltonians for which the coupling between the guest and the host is varied. (See Supporting Information and Simulation Methods section for more details.) We deliberately chose this set of short unconverged independent parallel simulations to guarantee the obvious differences between the raw data and their RE-SWHAM estimates. The second example is to calculate the free energy differences between the 16 λ states at 300 K from a 72 ns one-dimensional RE simulation using the same 16 Hamiltonians used in the first example. In both examples, we applied UWHAM to analyze the data as the benchmark. As shown in Figure 2a, b, the binding energy distribution at the $\lambda = 0.95$ thermodynamic state and the free energy differences estimated by UWHAM and RE-SWHAM are indistinguishable. (See the Supporting Information for the distribution of binding energy at each λ thermodynamic state.)

To illustrate the ability of RE-SWHAM to analyze a very large data set, which is difficult to analyze using standard UWHAM or MBAR methods, we applied RE-SWHAM to estimate the binding energy distributions from the data generated by a set of independent one-dimensional RE simulations (each with 16 replicas, using the same 16 Hamiltonians used in the first two examples) run at 15 different temperatures. The data ensemble in this example contains 240 thermodynamic states varying in λ values and temperatures, and a total of 3.456×10^7 data elements. Figure 3 shows the raw data and the RE-SWHAM estimates of the binding energy distributions of $\lambda = 1.0$ state at 200 K and 300 K. We also plotted the results obtained from an asynchronous two-dimensional 240 state RE simulation reported previously.[16] These results serve as a converged benchmark. As can be seen, at low temperature (200 K), even the one-dimensional RE simulations are not converged after 72 ns. However, compared with the two-dimensional asynchronous RE simulation results, applying RE-SWHAM to this unconverged two-dimensional (λ, T) data set led to significant improvements for the binding energy distribution at 200 K by using information from the simulations at higher temperature. At 300 K, RE-SWHAM estimates also show some improvements compared with the raw data. (see the Supporting Information for more discussion including the distribution of binding energy of $\lambda = 1.0$ state at

DOI: 10.1021/acs.jpclett.5b01771
J. Phys. Chem. Lett. 2015, 6, 3834−3840

each of 15 temperatures and the corresponding UWHAM estimate.) The RE-SWHAM estimates in Figure 3 are the statistical results from the output stream of RE-SWHAM run for 12 min (2.0×10^6 RE cycles); however, it took 15.7 h to obtain similar converged estimates with UWHAM.

As far as we aware, this is the first time that the WHAM equations have been solved stochastically. More importantly, we believe RE-SWHAM is a promising tool to handle very large data sets. Over the past decade, hardware and software developments made it possible to run enhanced sampling simulations with hundreds of thermodynamic states, or umbrella sampling with thousands of windows.[16,17,21−23] These kinds of large-scale simulations provide significant sampling power to study complex biological systems; however, they also pose a challenge for reweighting techniques like UWHAM to analyze the large raw data sets. For example, for the data obtained from a multicanonical simulation with M thermodynamic states and $\overline{N_\alpha}$ observations per state, UWHAM needs to manipulate a matrix with $\overline{N_\alpha} \times M^2$ elements, and the size of this matrix increases rapidly with the number of thermodynamic states M. Instead, the RE-SWHAM algorithm is a better choice for such scenario because it only manipulates $\overline{N_\alpha} \times M$ data elements, and performs simple replica exchange processes during the analysis. We have developed another algorithm called "local WHAM" (LWHAM) to analyze very large data sets, which is based on the idea of only "WHAMMing" the data elements in the local neighborhood for each thermodynamic state (manuscript in preparation). Finally, we note a recent study by Meng and Roux who proposed an algorithm based on a multivariate linear regression, to process large data sets generated by Umbrella Sampling along a chosen reaction coordinate without solving the WHAM equations.[23] RE-SWHAM can also be used to construct potentials of mean force from data generated by umbrella sampling.

SIMULATION METHODS

The biological system used to generate simulation data is the binding of a guest molecule (Heptanoate) to a host molecule (β-cyclodextrins (BCD))—a problem we studied previously.[24,25] The binding energy distribution analysis method (BEDAM) was applied to study the binding of Heptanoate/BCD complex.[26] BEDAM is a free energy method based on RE simulations in which the interaction between ligand and acceptor is scaled by the factor λ changing gradually from zero to one, namely, $H = H_0 + \lambda V$; $0 \le \lambda \le 1$. Here we chose 16 λ values (0.0, 0.001, 0.002, 0.004, 0.01, 0.04, 0.07, 0.1, 0.2, 0.4, 0.6, 0.7, 0.8, 0.9, 0.95, 1.0). For the two-dimensional problem we "RE-SWHAMED" data from 240 states with different (λ, T) values. One-dimensional replica exchange simulations each with 16 λ values were carried out independently at 15 temperatures (200 K, 206 K, 212 K, 218 K, 225 K, 231 K, 238 K, 245 K, 252 K, 260 K, 267 K, 275 K, 283 K, 291 K, 300 K). There are replica exchanges between simulations of different λ states at each temperature, but no replica exchanges were allowed between simulations at different temperatures. Each one-dimensional replica exchange simulation lasted 72 ns. (See Supporting Information for more details.)

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jpclett.5b01771.

The Heptanoate and β-cyclodextrins binding complex, additional plots to Figure 2a, and additional plots to Figure 3 (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: ronlevy@temple.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by grants from the National Science Foundation (CDI type II 1125332) and the National Institute of Health (GM30580). We would like to acknowledge valuable scientific discussions with Dr. Emilio Gallicchio, Peng He, and Wei Dai.

REFERENCES

(1) Ferrenberg, A.; Swendsen, R. Optimized Monte Carlo Data Analysis. *Phys. Rev. Lett.* **1989**, *63*, 1195−1198.
(2) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.* **1992**, *13*, 1011−1021.
(3) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. Multidimensional Free-energy Calculations Using the Weighted Histogram Analysis Method. *J. Comput. Chem.* **1995**, *16*, 1339−1350.
(4) Bartels, C.; Karplus, M. Multidimensional Adaptive Umbrella Sampling: Applications to Main Chain and Side Chain Peptide Conformations. *J. Comput. Chem.* **1997**, *18*, 1450−1462.
(5) Tan, Z. On a Likelihood Approach for Monte Carlo Integration. *J. Am. Stat. Assoc.* **2004**, *99*, 1027−1036.
(6) Gallicchio, E.; Andrec, M.; Felts, A. K.; Levy, R. M. Temperature Weighted Histogram Analysis Method, Replica Exchange, and Transition Paths. *J. Phys. Chem. B* **2005**, *109*, 6722−6731.
(7) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Seok, C.; Dill, K. A. Use of the Weighted Histogram Analysis Method for the Analysis of Simulated and Parallel Tempering Simulations. *J. Chem. Theory Comput.* **2007**, *3*, 26−41.
(8) Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* **2008**, *129*, 124105.
(9) Tan, Z.; Gallicchio, E.; Lapelosa, M.; Levy, R. M. Theory of Binless Multi-State Free Energy Estimation with Applications to Protein-Ligand Binding. *J. Chem. Phys.* **2012**, *136*, 144102.
(10) Zhu, F.; Hummer, G. Convergence and Error Estimation in Free Energy Calculations Using the Weighted Histogram Analysis Method. *J. Comput. Chem.* **2012**, *33*, 453−465.
(11) Law, S. M.; Ahlstrom, L. S.; Panahi, A.; Brooks, C. L., III Hamiltonian Mapping Revisited: Calibrating Minimalist Models to Capture Molecular Recognition by Intrinsically Disordered Proteins. *J. Phys. Chem. Lett.* **2014**, *5*, 3441−3444.
(12) Wu, H.; Mey, A. S. J. S.; Rosta, E.; Noé, F. Statistically Optimal Analysis of State-Discretized Trajectory Data from Multiple Thermodynamic States. *J. Chem. Phys.* **2014**, *141*, 214106.
(13) Rosta, E.; Hummer, G. Free Energies from Dynamic Weighted Histogram Analysis Using Unbiased Markov State Model. *J. Chem. Theory Comput.* **2015**, *11*, 276−285.
(14) Sugita, Y.; Okamoto, Y. Replica-exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314*, 141−151.

DOI: 10.1021/acs.jpclett.5b01771
J. Phys. Chem. Lett. 2015, 6, 3834−3840

(15) Lyman, E.; Ytreberg, F. M.; Zuckerman, D. M. Resolution Exchange Simulation. *Phys. Rev. Lett.* **2006**, *96*, 028105.
(16) Xia, J.; Flynn, W. F.; Gallicchio, E.; Zhang, B. W.; He, P.; Tan, Z.; Levy, R. M. Large-scale Asynchronous and Distributed Multi-dimensional Replica Exchange Molecular Simulations and Efficiency Analysis. *J. Comput. Chem.* **2015**, *36*, 1772−1785.
(17) Gallicchio, E.; Xia, J.; Flynn, W. F.; Zhang, B.; Samlalsingh, S.; Mentes, A.; Levy, R. M. Asynchronous Replica Exchange Software for Grid and Heterogeneous Computing. *Comput. Phys. Commun.* **2015**, DOI: 10.1016/j.cpc.2015.06.010.
(18) Chodera, J. D.; Shirts, M. R. Replica Exchange and Expanded Ensemble Simulations as Gibbs Sampling: Simple Improvements for Enhanced Mixing. *J. Chem. Phys.* **2011**, *135*, 194110.
(19) Bennett, C. H. Efficient Estimation of Free Energy Differences from Monte Carlo Data. *J. Comput. Phys.* **1976**, *22*, 245−268.
(20) Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22*, 1420−1426.
(21) Jiang, W.; Luo, Y.; Maragliano, L.; Roux, B. Calculation of Free Energy Landscape in Multi-Dimensions with Hamiltonian-Exchange Umbrella Sampling on Petascale Supercomputer. *J. Chem. Theory Comput.* **2012**, *8*, 4672−4680.
(22) Kokubo, H.; Tanaka, T.; Okamoto, Y. Two-dimensional Replica-Exchange Method for Predicting Protein-Ligand Binding Structures. *J. Comput. Chem.* **2013**, *34*, 2601−2614.
(23) Meng, Y.; Roux, B. Efficient Determination of Free Energy Landscapes in Multiple Dimensions from Biased Umbrella Sampling Simulations Using Linear Regression. *J. Chem. Theory Comput.* **2015**, *11*, 3523−3529.
(24) Wickstrom, L.; He, P.; Gallicchio, E.; Levy, R. M. Large Scale Affinity Calculations of Cyclodextrin Host-Guest Complexes: Understanding the Role of Reorganization in the Molecular Recognition Process. *J. Chem. Theory Comput.* **2013**, *9*, 3136−3150.
(25) Gallicchio, E.; Levy, R. M. Prediction of SAMPL3 Host-Guest Affinities with the Binding Energy Distribution Analysis Method (BEDAM). *J. Comput.-Aided Mol. Des.* **2012**, *26*, 505−516.
(26) Gallicchio, E.; Lapelosa, M.; Levy, R. M. The Binding Energy Distribution Analysis Method (BEDAM) for the Estimation of Protein-Ligand Binding Affinities. *J. Chem. Theory Comput.* **2010**, *6*, 2961−2977.

DOI: 10.1021/acs.jpclett.5b01771
J. Phys. Chem. Lett. 2015, 6, 3834−3840