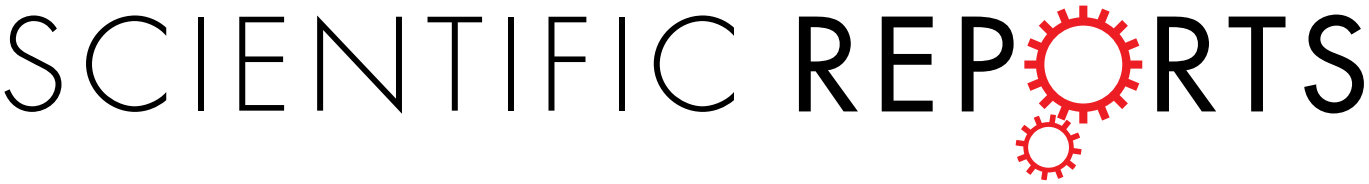

OPEN

The UWHAM and SWHAM Software Package

Bin W. Zhang, Shima Arasteh & Ronald M. Levy

Received: 25 October 2018
Accepted: 21 January 2019
Published online: 26 February 2019

We introduce the UWHAM (binless weighted histogram analysis method) and SWHAM (stochastic UWHAM) software package that can be used to estimate the density of states and free energy differences based on the data generated by multi-state simulations. The programs used to solve the UWHAM equations are written in the C++ language and operated via the command line interface. In this paper, first we review the theoretical bases of UWHAM, its stochastic solver RE-SWHAM (replica exchange-like SWHAM)and ST-SWHAM (serial tempering-like SWHAM). Then we provide a tutorial with examples that explains how to apply the UWHAM program package to analyze the data generated by different types of multi-state simulations: umbrella sampling, replica exchange, free energy perturbation simulations, etc. The tutorial examples also show that the UWHAM equations can be solved stochastically by applying the RE-SWHAM and ST-SWHAM programs when the data ensemble is large. If the simulations at some states are far from equilibrium, the Stratified RE-SWHAM program can be applied to obtain the equilibrium distribution of the state of interest. All the source codes and the tutorial examples are available from our group's web page: https://ronlevygroup.cst.temple.edu/software/UWHAM_and_SWHAM_webpage/index.html.

The weighted histogram analysis method (WHAM) algorithm[1,2] is widely applied to estimate the density of states and free energy differences based on the data generated by multi-state simulations. Multi-state simulations are popular advanced sampling algorithms that are applied in computational biophysics and computational chemistry. For example, the temperature replica exchange method is extensively applied to explore the configurational space of biomolecules; the umbrella sampling method is applied to construct free energy landscape of a system on chosen reaction coordinates; the free energy perturbation and Hamiltonian replica exchange method are powerful tools used to estimate the binding affinities of ligands and proteins for small-molecule drug discovery[3–5]. The WHAM algorithm is the standard tool to analyze the data generated by these multi-state simulations. Consider the simulation at each state as a measurement of density of states, the WHAM algorithm answers the question what the best estimate of density of states is if measurements have been taken at multiple states.

Since its introduction in 1992, the WHAM algorithm has been examined and studied by several research groups[6–11]. The most important improvement of WHAM is that an unbinned WHAM version named the multi-state Bennett acceptance ratio (MBAR) or the binless WHAM (UWHAM) was introduced[12–14]. Compared with the original WHAM, which coarse-grains observations into bins of a histogram, the binless WHAM provides the estimate of density of states for each data point therefore increasing the statistical precision and importantly, estimating the density of states provides a connection with the potential distribution theorem[15,16].

Complementary to the study of WHAM itself, how to solve WHAM equations efficiently in practice is another topic that has been an object of research[17–20]. In fact, this topic became more challenging and more urgent after the introduction of binless WHAM because of the dramatic increase of the number of variables without coarse-graining. In ref.14, Tan *et al.* proposed to solve the UWHAM equations by minimizing a convex function. To further remove the computational bottleneck in scaling up UWHAM, we developed methods called stochastic UWHAM (SWHAM) which solve the UWHAM equations stochastically by using generalized ensemble algorithms to resample the data collected at multiple states[21,22]. One important assumption of applying WHAM is that the data obtained from each state has already reached global equilibrium. However, sometimes this assumption does not hold if the barriers between free energy basins are high at some of the states and the simulation times are not long enough. We developed a method called Stratified-UWHAM[23] to analyze the data generated by multi-state simulations when the simulations at some states are far from equilibrium.

Center for Biophysics and Computational Biology, Department of Chemistry and Institute for Computational Molecular Science, Temple University, Philadelphia, Pennsylvania, 19122, United States. Correspondence and requests for materials should be addressed to B.W.Z. (email: bin.w.zhang@temple.edu)

The purpose of this paper is to introduce the UWHAM and SWHAM software package developed by our group. The programs used to solve the UWHAM equations are written in the C++ language and operated via the command-line interface. The basic solver solves the UWHAM equations by either a direct iteration method or minimization of a convex function. When the data ensemble is large, we show that the multi-state free energies can be obtained directly by running serial tempering-like SWHAM (ST-SWHAM), which resamples the raw data by applying the serial tempering (ST) protocol; the multi-state distributions can be obtained directly by running replica exchange-like SWHAM (RE-SWHAM), which resamples the raw data by applying the replica exchange (RE) protocol. If the simulations at some states are far from convergence, the multi-state distributions can be estimated by Stratified RE-SWHAM. Local WHAM[22], which is a variant of ST-SWHAM that couples the adjacent states by a stochastic resampling procedure, is also included in this software package. The remaining part of the paper proceeds as follows: First, we briefly review the theoretical basis of UWHAM and SWHAM. Then we introduce the tutorial examples on the web page of the UWHAM and SWHAM software package.

Methods and Discussion

UWHAM. Suppose M parallel (independent or coupled) simulations in the canonical ensemble are run at M states. Each state is characterized by a specific combination of thermodynamic parameters and potential energy functions. They are referred to as λ-states in the remaining part of this paper to avoid the confusion with the terms such as conformational states and microstates. Suppose $X_{\alpha i}$ is the ith microstate observed at the αth λ-state, the probability of observing $X_{\alpha i}$ at the γth λ-state is

$$P_\gamma(X_{\alpha i}) \sim \frac{q_\gamma(\{x\}_{\alpha i})}{Z_\gamma} = \frac{\exp\{-\beta_\gamma E_\gamma(\{x\}_{\alpha i})\}}{Z_\gamma}, \tag{1}$$

where $q_\gamma(\{x\}_{\alpha i}) = \exp\{-\beta_\gamma E_\gamma(\{x\}_{\alpha i})\}$ is the Boltzmann's factor of $X_{\alpha i}$ at the γth λ-state; $\{x\}_{\alpha i}$ is the coordinates of the microstate $X_{\alpha i}$; β_γ is the inverse temperature of the γth λ-state; $E_\gamma(\{x\}_{\alpha i})$ is the potential energy of the microstate $X_{\alpha i}$ at the γth λ-state; and Z_γ is the partition function of the γth λ-state. The likelihood of the observed data is proportional to[14]

$$\prod_{\alpha=1}^{M}\prod_{i=1}^{N_\alpha}\frac{\Omega(u_{\alpha i})q_\alpha(u_{\alpha i})}{Z_\alpha}, \tag{2}$$

where $u_{\alpha i}$ is the energy coordinate of the microstate $X_{\alpha i}$ that in general may be written as the sum of a reference energy plus perturbations (see ref.14). N_α is the total number of observations observed at the αth λ-state; and $\Omega(u_{\alpha i})$ is the density of states. Let $\hat{Z}_\alpha$ and $\hat{\Omega}(u_{\gamma i})$ denote estimates of the partition function of the αth λ-states and the density of states of $u_{\gamma i}$, respectively. These two estimates satisfy the equation

$$\hat{Z}_\alpha = \sum_{\gamma=1}^{M}\sum_{i=1}^{N_\gamma} q_\alpha(u_{\gamma i})\hat{\Omega}(u_{\gamma i}). \tag{3}$$

Maximizing the log likelihood function yields

$$\hat{\Omega}(u_{\gamma i}) = \frac{1}{\sum_{\kappa=1}^{M} N_\kappa \hat{Z}_\kappa^{-1} q_\kappa(u_{\gamma i})}. \tag{4}$$

Eqs (3) and (4) are the UWHAM (or MBAR) equations[13,14]. Note that the UWHAM estimates do not depend on the original λ-state at which each observation was observed. Therefore the UWHAM equations can be simplified as

$$\begin{aligned}\hat{Z}_\alpha &= \sum_{i=1}^{N} q_\alpha(u_i)\hat{\Omega}(u_i)\\ \hat{\Omega}(u_i) &= \frac{1}{\sum_{\kappa=1}^{M} N_\kappa \hat{Z}_\kappa^{-1} q_\kappa(u_i)},\end{aligned} \tag{5}$$

where $N = \sum_{\alpha=1}^{M} N_\alpha$ is the total number of observations.

The UWHAM estimate of the probability of observing the observation u_i at the αth λ-state is

$$\hat{p}_\alpha(u_i) = \hat{Z}_\alpha^{-1}\hat{\Omega}(u_i)q_\alpha(u_i) = \frac{\hat{w}_\alpha(u_i)}{\hat{Z}_\alpha}. \tag{6}$$

where $\hat{w}_\alpha(u_i) = \hat{\Omega}(u_i)q_\alpha(u_i)$ is the unnormalized probability. We can define one of the λ-states as the reference state, and the normalized probability of observing the observation u_i at the reference state is

$$\hat{w}_0(u_i) = \frac{\hat{\Omega}(u_i)q_0(u_i)}{\sum_{j=1}^{N}\hat{\Omega}(u_j)q_0(u_j)}; \tag{7}$$

and $\hat{w}_\alpha = \hat{w}_0(u_i)\Delta q_\alpha(u_i)$, where $\Delta q_\alpha(u_i) = q_\alpha(u_i)/q_0(u_i) = \exp\{-[\beta_\alpha E_\alpha(u_i) - \beta_0 E_0(u_i)]\}$ is the biasing factor. Then the equation array Eq. (5) can be rewritten as

$$\begin{aligned}\hat{Z}_\alpha &= \sum_{i=1}^{N} \hat{w}_0(u_i)\Delta q_\alpha(u_i)\\ \hat{w}_0(u_i) &= \frac{1}{\sum_{\kappa=1}^{M} N_\kappa \hat{Z}_\kappa^{-1}\Delta q_\kappa(u_i)}.\end{aligned} \tag{8}$$

In practice, the UWHAM program solves the equation array Eq. (8) instead of Eq. (5). Suppose A is a property of interest of the system. According to Eq. (6), the expectation value of the property A at the αth λ-state is calculated by the weighted average

$$\langle A\rangle_\alpha = \frac{\sum_{i=1}^{N} \hat{w}_\alpha(u_i)A(u_i)}{\sum_{i=1}^{N} \hat{w}_\alpha(u_i)} = \frac{\sum_{i=1}^{N} \hat{w}_0(u_i)\Delta q_\alpha(u_i)A(u_i)}{\sum_{i=1}^{N} \hat{w}_0(u_i)\Delta q_\alpha(u_i)}. \tag{9}$$

where $A(u_i)$ is the the property A measured by using the ith observation.

Currently, a self-consistent iteration solver and a solver that optimizes a convex function by using the Newton-Raphson algorithm[14] have been implemented in the UWHAM program to solve the UWHAM equations.

SWHAM. Suppose the raw data were generated from simulations at M λ-states, and the total number of observations is N. During the procedure of UWHAM analysis, the program needs to evaluate M biasing factors (or Boltzmann's factors) for each observation at the beginning. Namely, the UWHAM program evaluates a biasing matrix which contains $n \times M^2$ elements, where $n = N/M$ is the average number of observations observed at each λ-state. Then the UWHAM equations are solved by minimization of a convex function, which involves multiplication of matrices that contain $M \times N$ elements (as large as the biasing matrix) and diagonalization of matrices that contain $M \times M$ elements. The costs of memory and computational time of running UWHAM are proportional to the second order of the number of λ-states M. To remove this computational bottleneck in scaling up UWHAM, we developed methods which solve UWHAM equations stochastically by using the generalized ensemble algorithms.

RE-SWHAM. RE-SWHAM is an algorithm that we developed to solve the UWHAM equations stochastically by applying the replica exchange (RE) protocol to resample the raw data generated by multi-state simulations[21]. As shown in Fig. 1, the observations observed at each λ-state are collected as the database for that λ-state beforehand. Then RE-SWHAM analyses are run by performing cycles of RE simulations. Each cycle consists of a "move" procedure and an "exchange" procedure. During the move procedure of RE-SWHAM, an observation in the database of a λ-state is randomly chosen with equal probability to associate with the replica at that λ-state. During the exchange procedure of RE-SWHAM, we attempt to swap two random replicas based on the Metropolis criterion. If the swap is accepted, in addition to swapping the replicas, the observation associated with the replica is also swapped to the database of the other λ-state[21]. The exchange step should be repeated multiple times to approach the infinite swapping limit for the best sampling efficiency[24]. At the end of the exchange procedure, the observation associated with the replica at each λ-state is recorded as the output of that λ-state. Note the direct outputs of RE-SWHAM are the estimates of the equilibrium distribution at each λ-state.

In ref.21, we proved that the distribution of the outputs of RE-SWHAM at each λ-state are asymptotic to the UWHAM estimate when the number of observations observed at each λ-state is large by treating RE-SWHAM as a random walk in the space of the weight arrays of observations. Here we provide an alternative proof. Consider a trial exchange in RE-SWHAM which swaps one observation u_m at the αth λ-state and the other observation u_n at the γth λ-state. The probability that this trial exchange is accepted is

$$\begin{aligned}P_{ex} &= \widetilde{p}_\alpha(u_m)\widetilde{p}_\gamma(u_n)\,\min\left(1, \frac{\exp[-\beta_\alpha E_\alpha(u_n)]\,\exp[-\beta_\gamma E_\gamma(u_m)]}{\exp[-\beta_\alpha E_\alpha(u_m)]\,\exp[-\beta_\gamma E_\gamma(u_n)]}\right)\\ &= \widetilde{p}_\alpha(u_m)\widetilde{p}_\gamma(u_n)\Psi\left(\log\left[\frac{q_\alpha(u_m)q_\gamma(u_n)}{q_\alpha(u_n)q_\gamma(u_m)}\right]\right)\end{aligned}. \tag{10}$$

where $\widetilde{p}_X(u_Y)$ is the normalized time-average probability of choosing the observation u_Y to associate with the replica at the X th λ-state, and Ψ is the Metropolis function[25]

$$\Psi(x) = \min(1, \exp[-x]), \tag{11}$$

which has the property $\Psi(x)/\Psi(-x) = \exp\{-x\}$. Consider the reverse trial exchange that swaps the observation u_n at the αth λ-state and the observation u_m at the γ λ-state. The probability that this trial exchange is accepted is

$$\begin{aligned}P'_{ex} &= \widetilde{p}_\alpha(u_n)\widetilde{p}_\gamma(u_m)\,\min\left(1, \frac{\exp[-\beta_\alpha E_\alpha(u_m)]\,\exp[-\beta_\gamma E_\gamma(u_n)]}{\exp[-\beta_\alpha E_\alpha(u_n)]\,\exp[-\beta_\gamma E_\gamma(u_m)]}\right)\\ &= \widetilde{p}_\alpha(u_n)\widetilde{p}_\gamma(u_m)\Psi\left(\log\left[\frac{q_\alpha(u_n)q_\gamma(u_m)}{q_\alpha(u_m)q_\gamma(u_n)}\right]\right).\end{aligned} \tag{12}$$

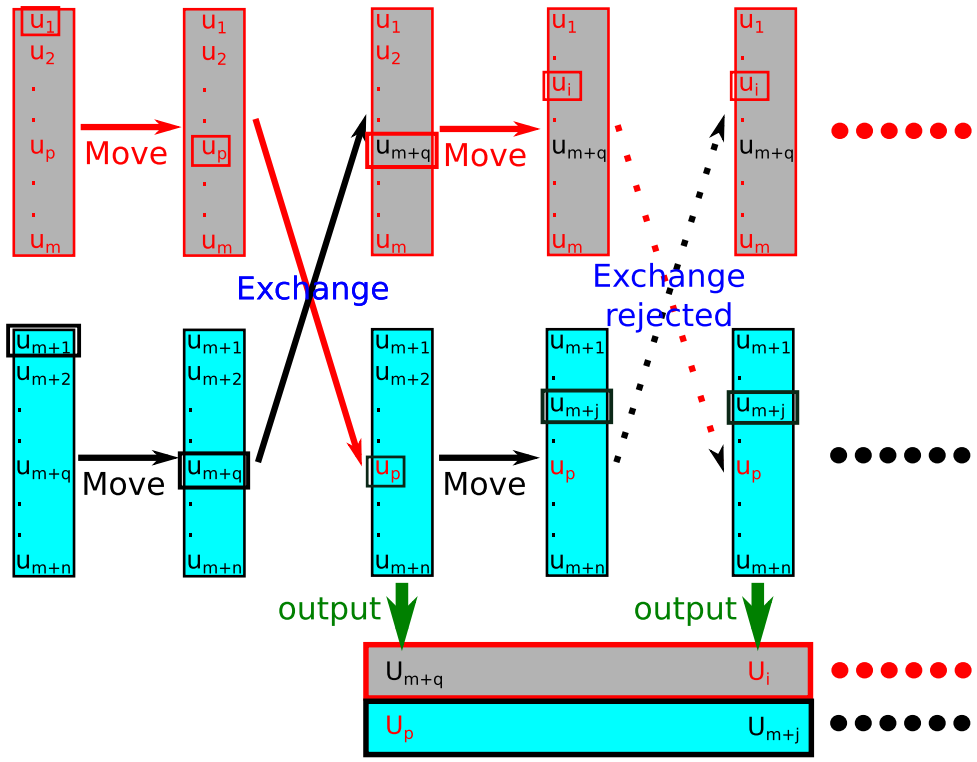

Figure 1. An illustration of the RE-SWHAM algorithm. This drawing illustrates two replica exchange cycles of the RE-SWHAM method, and shows only two λ-states with "gray" or "cyan" color. In each cycle one data element is chosen from λ-state first, then a replica exchange is performed. In the first cycle since the swap is accepted, the data associated with the two replicas is swapped to the other λ-state's data array. At the end of each cycle, the data associated with replicas are recorded as the output like explicit RE simulations. Reprinted (adapted) with permission from ref.21. Copyright (2015) American Chemical Society.

If the RE-SWHAM resampling procedure converges, P_{ex} and P'_{ex} will agree with each other for each pair of observations (u_m, u_n) and each pair of λ-states (α, γ), which leads to the detailed balance relation of RE-SWHAM:

$$\frac{\widetilde{p}_\alpha(u_m)/q_\alpha(u_m)}{\widetilde{p}_\alpha(u_n)/q_\alpha(u_n)} = \frac{\widetilde{p}_\gamma(u_m)/q_\gamma(u_m)}{\widetilde{p}_\gamma(u_n)/q_\gamma(u_n)}, \quad \text{for all } (u_m, u_n) \text{ and } (\boldsymbol{\alpha}, \gamma). \tag{13}$$

Eq. (13) can be rewritten as

$$\frac{\widetilde{p}_\alpha(u_m)/q_\alpha(u_m)}{\widetilde{p}_0(u_m)/q_0(u_m)} = \frac{\widetilde{p}_\alpha(u_n)/q_\alpha(u_n)}{\widetilde{p}_0(u_n)/q_0(u_n)} = \hat{Z}_\alpha^{-1}, \text{ for all } (u_m, u_n) \text{ and } \boldsymbol{\alpha}. \tag{14}$$

where subscript 0 denotes the reference state. Then the probability $\widetilde{p}_\alpha(u_m)$ can be expressed as

$$\widetilde{p}_\alpha(u_m) = \hat{Z}_\alpha^{-1}\hat{\Omega}(u_m)q_\alpha(u_m), \text{ for all } u_m \text{ and } \boldsymbol{\alpha}, \tag{15}$$

where $\hat{\Omega}(u_m) = \widetilde{p}_0(u_m)/q_0(u_m)$. Summing both sides of Eq. (15) over all the observations and applying the relationship $\sum_{m=1}^{N}\widetilde{p}_\alpha(u_m) = 1$ at each λ-state yields

$$\hat{Z}_\alpha = \sum_{m=1}^{N}\hat{\Omega}(u_m)q_\alpha(u_m). \tag{16}$$

Note that the probability of finding the observation u_m in the database of the αth λ-state is $N_\alpha\widetilde{p}_\alpha(u_m)$ and there is one copy of each observation in the databases of all λ-states, namely, $\sum_{\alpha=1}^{M}N_\alpha\widetilde{p}_\alpha(u_m) = 1$. Multiplying both side of Eq. (15) by N_α and summing over all the λ-states yields

$$\hat{\Omega}(u_m) = \frac{1}{\sum_{\alpha=1}^{M}N_\alpha\hat{Z}_\alpha^{-1}q_\alpha(u_m)}. \tag{17}$$

Thus, the RE-SWHAM estimates $\hat{Z}_\alpha$ and $\hat{\Omega}(u_m)$ satisfy Eqs (16) and (17), which are equivalent to the UWHAM equations (Eq. (5)).

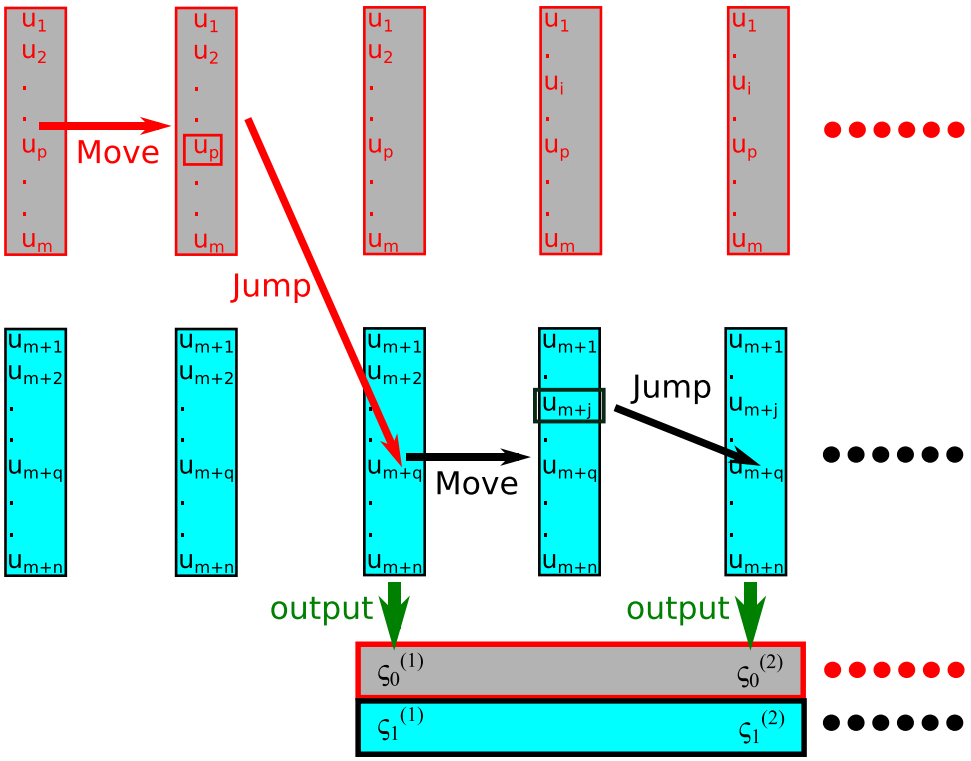

Figure 2. An illustration of the ST-SWHAM algorithm. This drawing illustrates two serial tempering cycles of the ST-SWHAM method, and shows only two λ-states with "gray" or "cyan" color. In each cycle one data element is chosen from the λ-state sampled by the replica with equal probability to associate with the replica. Then the replica jumps to one of the λ-states according to the probability calculated by Eq. [18]. At the end of each cycle, the free energy estimates $\{\zeta_k\}$ are adjusted to match the observed proportion of the replica being at the κth λ-state π_κ with the proportion of the κth λ-state of the raw data generated by the multi-state simulations π_κ^0.

ST-SWHAM. ST-SWHAM is an algorithm that we developed to solve the UWHAM equations stochastically by applying the serial tempering (ST) protocol to resample the raw data generated by multi-state simulations[22]. The procedure is illustrated in Fig. 2. Like the RE-SWHAM analysis, the observations observed at each λ-state are collected as the database for that λ-state beforehand. However, unlike resampling the data using replica exchanges, there is only one "simulation run" in the serial tempering resampling algorithm. For the sake of comparison and convenience, we still refer to this single simulation as a replica in this paper. Serial tempering simulations are also run by cycles, and each cycle consists of a "move" procedure and a "jump" procedure. During the move procedure of ST-SWHAM, an observation in the database of the λ-state sampled by the replica is randomly chosen with equal probability to associate with the replica. During the jump procedure, the replica jumps to the αth λ-state according to the probability[22]

$$p(\alpha|u_i;\,\zeta,\,\pi^0) = \frac{\pi_\alpha^0 e^{\zeta_\alpha} q_\alpha(u_i)}{\sum_{\kappa=1}^{M} \pi_\kappa^0 e^{\zeta_\kappa} q_\kappa(u_i)}, \tag{18}$$

where u_i is the ith observation associated with the replica; $\zeta_\kappa = -\ln Z_\kappa$ is the unitless free energy of the κth λ-state; $\pi_\kappa^0 = N_\kappa/N$ is the proportion of the κth λ-state of the raw data generated by the multi-state simulations; and $q_\kappa(u_i)$ is the biasing factor of the ith observation at the κth λ-state. Suppose π_κ is the observed proportion of the κth λ-state sampled by the replica during the ST-SWHAM analysis. The values of $\{\zeta_\kappa\}$ are adjusted during the analysis of ST-SWHAM until the observed proportion of the replica being at the κth λ-state π_κ agrees with π_κ^0[22]. Note the direct outputs of ST-SWHAM are the estimates of the free energies of different λ-states — $\{\zeta_\kappa\}$.

It can be shown that ζ_κ is the UWHAM estimate of the free energy of the κth λ-state when π_κ and π_κ^0 agree with each other for all λ-states. The details of the proof that ST-SWHAM solves the UWHAM equations stochastically can be found in ref.22. One brief rationale is as follows. First, if π_κ equals π_κ^0, the probability of each observation being chosen to associate with the replica during the ST-SWHAM analysis is $1/N$, where N is the total number of observations. Therefore, the observed proportion of the αth λ-state sampled by the replica during the ST-SWHAM analysis is

$$\pi_\alpha = \frac{1}{N}\sum_{i=1}^{N} p(\alpha|u_i;\,\zeta,\,\pi^0) = \frac{1}{N}\sum_{i=1}^{N} \frac{\pi_\alpha^0 e^{\zeta_\alpha} q_\alpha(u_i)}{\sum_{\kappa=1}^{M} \pi_\kappa^0 e^{\zeta_\kappa} q_\kappa(u_i)}. \tag{19}$$

On the other hand, note that

$$\frac{1}{N}\sum_{i=1}^{N}\frac{e^{\hat{\zeta}_\alpha}q_\alpha(u_i)}{\sum_{\kappa=1}^{M}\pi_\kappa^0 e^{\hat{\zeta}_\kappa}q_\kappa(u_i)}=\sum_{i=1}^{N}\frac{\hat{\Omega}(u_i)q_\alpha(u_i)}{Z_\alpha}=\sum_{i=1}^{N}\hat{p}_\alpha(u_i)=1. \tag{20}$$

Then π_α^0 can be rewritten as

$$\pi_\alpha^0=\frac{1}{N}\sum_{i=1}^{N}\frac{\pi_\alpha^0 e^{\hat{\zeta}_\alpha}q_\alpha(u_i)}{\sum_{\kappa=1}^{M}\pi_\kappa^0 e^{\hat{\zeta}_\kappa}q_\kappa(u_i)}. \tag{21}$$

Comparison between Eqs (19) and (21) leads to the conclusion that π_α and π_α^0 agree with each other if $\zeta_\alpha=\hat{\zeta}_\alpha$ for each λ-state.

The jump of the replica following Eq. (18) was referred to as the global jump proposal in ref.22 because the replica can reach any λ-state of the system by one jump. According to Eq. (18), every jump of the replica requires calculations of M exponential functions, where M is the total number of λ-states. When the total number of states is large, ST-SWHAM analyses using the global jump proposal take a long time to converge. In our software package, we implemented a much faster approximate solver of UWHAM–ST-SWHAM using a local jump proposal. This algorithm was referred to as local WHAM in ref.22 because the replica can only be at the λ-states that are the local neighbors of the initial λ-state at the end of the jump procedure if the number of jumps per cycle is finite. Suppose the replica that associates with the observation u_i is at the γth λ-state initially. The procedure of performing one jump in local WHAM is as follows[22]:

- select a trial λ-state with uniform probabilities from the nearest neighbors of the γth λ-state, suppose the chosen λ-state is the αth λ-state.
- accept the αth λ-state as the new λ-state to jump to according to the Metropolis probability

$$\min\left\{1,\ \frac{\Gamma(\alpha,\gamma)}{\Gamma(\gamma,\alpha)}\frac{p(\alpha|u_i;\zeta,\pi^0)}{p(\gamma|u_i;\zeta,\pi^0)}\right\}, \tag{22}$$

where $p(\alpha|u_i;\zeta,\pi^0)$ and $p(\gamma|u_i;\zeta,\pi^0)$ are defined by Eq. (18); and $\Gamma(\gamma,\alpha)$ is the probability of choosing the αth λ-state as the trial λ-state when the replica is at the γth λ-state originally. Namely, $\Gamma(\gamma,\alpha)=1/n_\gamma$, where n_γ is the total number of the nearest neighbors of the γth λ-state if the αth λ-state is one of the nearest neighbors of the γth λ-state; $\Gamma(\gamma,\alpha)=0$ otherwise.

As can be seen, the replica can only be at the original λ-state or one of its nearest neighbors after one jump. However, the replica can diffuse further away from the original λ-state by repeating this one jump procedure multiple times. As the number of jumps per cycle increases, the results of local WHAM converges asymptotically to the results of ST-SWHAM that uses the global jump proposal[22]. Therefore, the jump of the replica following Eq. (18) in the infinite jump limit in serial tempering simulations is analogous to the infinite swapping limit in replica exchange simulations[24].

In ST-SWHAM, the free energy estimates are adjusted during the analysis until the observed proportion of the replica being at the κth λ-state π_κ agrees with the proportion of the κth λ-state of the raw data generated by the multi-state simulations π_κ^0. So far a variant of the updating algorithm discussed in ref.22 is implemented in the ST-SWHAM program.

Stratified RE-SWHAM. When applying UWHAM and its stochastic solvers, the basic assumption is that the simulation at each λ-state is "approximately" equilibrated. However, this assumption might not always hold. To handle such situations, we developed an analysis tool called Stratified-UWHAM and its stochastic solver Stratified RE-SWHAM to compute free energy and expectations from a multi-state ensemble when the simulations at a subset of λ-states are far from global equilibrium[23]. In ref.23, we showed that the Stratified UWHAM equations can be solved in the form of the original UWHAM equations (Eq. (5)) with an expanded set of λ-states. The stochastic solver, Stratified RE-SWHAM, has been included in the UWHAM and SWHAM software package. See the Supporting Information for a brief review and discussion about Stratified UWHAM and Stratified RE-SWHAM.

Illustrative applications. So far the tutorial examples include how to analyze the data generated by "one dimensional umbrella sampling", "two dimensional umbrella sampling", "temperature replica exchange", "Hamiltonian replica exchange", "two dimensional replica exchange" and " free energy perturbation" simulations. The tutorials provide the raw data generated by different types of multi-state simulations and explain the corresponding analysis procedures and outputs in details.

One Dimensional Umbrella Sampling. We explain how to apply UWHAM or ST-SWHAM to analyze the raw data generated by one dimensional umbrella sampling simulations. The potential function of the system studied in this example is a one dimensional double well potential[26]

$$U(x) = \frac{H}{W^4}(x^2 - W^2)^2, \tag{23}$$

where $H = 20\,k_BT$ is the height of the barrier between the two wells; k_B is Boltzmann's constant; T is the temperature; and $W = 1$ is the half width between the two minima of the potential. This one dimensional potential can be explored by a Brownian particle simulated with the over-damped Langevin dynamics[26]. Here we applied 31 parabolic potentials in the region between $x = -3$ to $x = 3$ to perform the umbrella sampling simulations. Then UWHAM and ST-SWHAM are used to analyze the data and construct the potential energy profile.

Umbrella sampling simulations are usually applied to construct free energy profiles for systems with multiple degrees of freedom. Although the example that we used here is a Brownian particle governed by a one dimensional potential function, the analysis procedure is the same for applying UWHAM or ST-SWHAM to construct one dimensional free energy profiles of complex systems. In such cases, the position of the complex system projected on the chosen reaction coordinate is analogous to the position of the Brownian particle in this tutorial.

Two Dimensional Umbrella Sampling. This example explains how to apply UWHAM or ST-SWHAM to raw data generated by two dimensional umbrella sampling simulations (of ~100 degrees of freedom) to construct the free energy profile. The system studied in this example is an alanine dipeptide (AlaD) molecule in implicit solvent at 300 K. The simulations were performed by using the GROMACS 5.1.2 simulation package with the Amber99SB force field and the OBC GB model[27,28]. To explore the two dimensional free energy surface (the Ramachandran plot of AlaD), we applied 24×24 parabolic potentials by using the PLUMED plugin[29] to perform the umbrella sampling simulations. The Ramachandran plots of AlaD are constructed by using the UWHAM and ST-SWHAM estimates.

Temperature Replica Exchange. We explain how to apply UWHAM or RE-SWHAM to raw data generated by temperature replica exchange simulations to obtain the estimates of the equilibrium distribution at the λ-state of interest. The system studied in this example is the same as the previous example–an alanine dipeptide (AlaD) molecule in implicit solvent. The RE simulations were performed by using the GROMACS 5.1.2 simulation package with the Amber99SB force field and the OBC GB model[27,28]. The coupled simulations were run at 10 temperatures (300 K, 317.52 K, 336.063 K, 355.689, 376.462 K, 398.447 K, 421.716 K, 446.345 K, 472.411, 500 K). The Ramachandran plots of AlaD in implicit solvent at 300 K are constructed by using the UWHAM and RE-SWHAM estimates.

Free Energy Perturbation. This example shows how to analyze the data generated by free energy perturbation (FEP) simulations. Here we calculate the solvation free energy of a water molecule in pure solvent (TIP3P) at 300 K by using the slow-growth method. The simulations were performed by using the GROMACS 5.1.2[27] and the TIP3P water model. In this example, we ran 11 independent parallel simulations for a box of pure solvent with a fixed tagged water molecule inside. The interaction between the tagged water molecule and the environment were gradually turned off through 11 λ-states[30]. The chosen λ values are 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0. The UWHAM estimate of the solvation free energy of a water molecule in pure solvent is $-6.18\,kcal/mol$. One can obtain the same result by using ST-SWHAM. More details and discussion about measuring the excess chemical potential of water molecules in solution using UWHAM can be found in ref.30.

BEDAM: Hamiltonian Replica Exchange. We explain how to use UWHAM or RE-SWHAM to analyze the data generated by Hamiltonian replica exchange simulations. In this example, we study the binding affinity of a guest molecule (heptanoate) to a host molecule (β-cyclodextrin) in implicit solvent (OPLA-AA/AGBNP2)[31,32]. Here we apply the binding energy distribution analysis method (BEDAM)[33] to obtain the binding free energy and binding energy distributions of this complex. BEDAM is a free energy method based on the Hamiltonian replica exchange algorithm. Suppose there are M parallel simulations in BEDAM, the Hamiltonian (potential) function of the ith λ-state is

$$H_i = V_0 + \lambda_i u, \tag{24}$$

where V_0 is the effective potential energy of the complex without the direct and solvent-mediated ligand-receptor interactions, and u is the binding energy[33]. Namely, the λ factor in BEDAM simulations linearly scales the interaction between the ligand and acceptor. We ran BEDAM simulations at 300 K by using 16 λ-states. The chosen λ values are 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[21]. We applied UWHAM to estimate the binding free energy of the β-cyclodextrin Heptanoate Complex— $-0.603\,kcal/mol + G_{vsite}$, where G_{vsite} is a correction because of the restraint applied to the ligand during the BEDAM simulation. One can obtain the same result by applying ST-SWHAM to the raw data. We also show how to apply UWHAM or RE-SWHAM to estimate the equilibrium distribution of the binding energy at the $\lambda = 1$ state (full interaction state).

Two Dimensional (Temperature and Hamiltonian) Replica Exchange. This example shows how to use SWHAM to analyze the data generated by two dimensional (Hamiltonian and temperature) replica exchange simulations. We study the binding affinity of a guest molecule (heptanoate) to a host molecule (β-cyclodextrin) in implicit solvent (OPLA-AA/AGBNP2)[32] at different temperatures. The raw data used in this example were generated by 15 separated BEDAM[33] simulations at temperatures 200 K, 206 K, 212 K, 218 K, 225 K, 231, 238 K, 245 K, 252 K, 260 K, 267 K, 275, 283 K, 291 K, 300 K. The chosen λ values are the same as the previous example–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. There are totally $16 \times 15 = 240$ states, and each state has 144,000 data points[21,22]. Although there are no exchanges between replicas at different temperatures, the procedure described in this tutorial can be applied to two dimensional (Hamiltonian and temperature) replica

exchange simulations without any alteration. The goal of this practice is to obtain the best estimates of the binding affinity at 200 K, which is the most difficult for BEDAM simulation to converge. Because the raw data ensemble is large, UWHAM is not suitable to analyze them directly. Here, we applied RE-SWHAM to estimate the equilibrium distribution of binding energies of each λ-state at 200 K. And the RE-SWHAM results are compared with the corresponding distributions calculated from the raw data. See ref.21 and 22 for more discussion about this tutorial example. The equilibrium distributions constructed by the RE-SWHAM output can be used as the input for UWHAM to estimate the binding free energy at the temperature of interest. The binding free energy of the β-cyclodextrin Heptanoate Complex is about $-6.3\,kcal/mol + G_{vsite}$ at 200 K, which is much stronger compared with its binding free energy at 300 K. This result can also be obtained by applying ST-SWHAM with the local jump algorithm to the raw data directly.

Two Binding Modes of the β-cyclodextrin Heptanoate Complex. The β-cyclodextrin heptanoate complex has two binding states depending on the orientation of the heptanoate molecule[23]. The two binding modes are referred to as the UP and DOWN macrostates. We ran two sets of independent MD simulations at 300 K of the β-cyclodextrin heptanoate complex in implicit solvent (AGBNP GB model[32]) at 16 λ-states. The initial structures of the complex in the first and the second sets of simulations were chosen from the UP and Down macrostates, respectively. The interaction between the ligand and the receptor was scaled by a λ factor like BEDAM[33], However, all the simulations are independent. The chosen λ values are (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). In this example, the λ-states with the largest seven λ values ($\lambda = 1.0, 0.95, 0.9, 0.8, 0.7, 0.6, 0.4$) are considered as the partially connected states because it is difficult for the binding complex to switch its binding mode and the simulations have not converged at these λ-states; the other nine λ-states are the fully connected states[23]. This tutorial shows how to apply Stratified RE-SWHAM to estimate the equilibrium distribution at the $\lambda = 1$ state (full interaction state) when some simulations are far from convergence. See ref.23 for more discussion about this tutorial example.

Data Availability

The UWHAM and SWHAM software package and its tutorials are available from the web page: https://ronlevy-group.cst.temple.edu/software/UWHAM_and_SWHAM_webpage/index.html. The UWHAM and SWHAM software package is distributed using the MIT license. In the future, we will keep adding more examples of the application of UWHAM and SWHAM to the web page. For instance, free energy perturbation (FEP) is one popular method that is applied to measure the relative ligand binding potency[34,35]. Currently we are applying UWHAM to analyze the FEP data and extract a density of states that can be used to estimate the relative binding free energy differences for multiple ligands simultaneously to solve the cycle closure challenge[34]. We will continue optimizing the code and plan to introduce parallelism to the software package.

References

1. 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.* **13**, 1011–1021, https://doi.org/10.1002/jcc.540130812 (1992).
2. 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.* **16**, 1339–1350, https://doi.org/10.1002/jcc.540161104 (1995).
3. Zuckerman, D. M. Equilibrium sampling in biomolecular simulations. *Annu. Rev. Biophys.* **40**, 41–62, https://doi.org/10.1146/annurev-biophys-042910-155255 (2011).
4. Gallicchio, E. & Levy, R. M. Advances in all atom sampling methods for modeling protein, ligand binding affinities. *Curr. Opin. Struct. Biol.* **21**, 161–166, https://doi.org/10.1016/j.sbi.2011.01.010 (2011).
5. Maximova, T., Moffatt, R., Ma, B., Nussinov, R. & Shehu, A. Principles and overview of sampling methods for modelling macromolecular structure and dynamics. *PLoS Comput. Biol.* **12**, e1004619, https://doi.org/10.1371/journal.pcbi.1004619 (2016).
6. Roux, B. The calculation of the potential of mean force using computer-simulations. *Comput. Phys. Commun.* **91**, 275–282, https://doi.org/10.1016/0010-4655(95)00053-I (1995).
7. Bartels, C. & Karplus, M. Multidimensional adaptive umbrella sampling: Applications to main chain and side chain peptide conformations. *J. Comput. Chem.* 18, 1450–1462, https://doi.org/10.1002/(sici)1096-987x (1997).
8. Bartels, C. Analyzing biased Monte Carlo and molecular dynamics simulations. *Chem. Phys. Lett.* **331**, 446–454, https://doi.org/10.1016/s0009-2614(00)01215-x (2000).
9. Souaille, M. & Roux, B. Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Comput. Phys. Commun.* **135**, 40–57, https://doi.org/10.1016/s0010-4655(00)00215-0 (2001).
10. Gallicchio, E., Andrec, M., Felts, A. K. & Levy, R. M. Temperature weighted histogram analysis method, replica exchange, and transition paths. *J. Phys. Chem. B* **109**, 6722–6731, https://doi.org/10.1021/jp045294f (2005).
11. 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.* **3**, 26–41, https://doi.org/10.1021/ct0502864 (2007).
12. Tan, Z. On a likelihood approach for Monte Carlo integration. *J. Am. Stat. Assoc.* **99**, 1027–1036, https://doi.org/10.1198/016214504000001664 (2004).
13. Shirts, M. R. & Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **129**, 124105, https://doi.org/10.1063/1.2978177 (2008).
14. 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.* **136**, 144102, https://doi.org/10.1063/1.3701175 (2012).
15. Widom, B. Some topics in the theory of fluids. *J. Chem. Phys.* **39**, 2808, https://doi.org/10.1063/1.1734110 (1963).
16. Beck, T. L., Paulaitis, M. E. & Pratt, L. R. *The Potential Distribution Theorem and Models of Molecular Solutions* (Cambridge University Press, 2006).
17. Bereau, T. & Swendsen, R. H. Optimized convergence for multiple histogram analysis. *J. Comput. Phys.* **228**, 6119–6129, https://doi.org/10.1016/j.jcp.2009.05.011 (2009).
18. Kim, J., Keyes, T. & Straub, J. E. Communication: Iteration-free, weighted histogram analysis method in terms of intensive variables. *J. Chem. Phys.* **135**, 061103, https://doi.org/10.1063/1.3626150 (2011).
19. Zhu, F. & Hummer, G. Convergence and error estimation in free energy calculations using the weighted histogram analysis method. *J. Comput. Chem.* **33**, 453–465, https://doi.org/10.1002/jcc.21989 (2012).

20. Zhang, C., Lai, C.-L. & Pettitt, B. M. Accelerating the weighted histogram analysis method by direct inversion in the iterative subspace. *Mol. Simul.* **42**, 1079–1089, https://doi.org/10.1080/08927022.2015.1110583 (2016).
21. Zhang, B. W., Xia, J., Tan, Z. & Levy, R. M. A stochastic solution to the unbinned wham equations. *J. Phys. Chem. Lett.* **6**, 3834–3840, https://doi.org/10.1021/acs.jpclett.5b01771 (2015).
22. Tan, Z., Xia, J., Zhang, B. W. & Levy, R. M. Locally weighted histogram analysis and stochastic solution for large-scale multi-state free energy estimation. *J. Chem. Phys.* **144**, 034107, https://doi.org/10.1063/1.4939768 (2016).
23. Zhang, B. W., Deng, N., Tan, Z. & Levy, R. M. Stratified UWHAM and its stochastic approximation for multicanonical simulations which are far from equilibrium. *J. Chem. Theory Comput.* **13**, 4660–4674, https://doi.org/10.1021/acs.jctc.7b00651 (2017).
24. Zhang, B. W. *et al*. Simulating replica exchange: Markov state models, proposal schemes, and the infinite swapping limit. *J. Phys. Chem. B* **120**, 8289–8301, https://doi.org/10.1021/acs.jpcb.6b02015 (2016).
25. Bennett, C. H. Efficient estimation of free energy differences from monte carlo data. *J. Comput. Phys.* **22**, 245–268, https://doi.org/10.1016/0021-9991(76)90078-4 (1976).
26. Zhang, B. W., Jasnow, D. & Zuckerman, D. M. Transition-event durations in one-dimensional activated processes. *J. Chem. Phys.* **126**, 074504, https://doi.org/10.1063/1.2434966 (2007).
27. Abraham, M. J. *et al*. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25, https://doi.org/10.1016/j.softx.2015.06.001 (2015).
28. Onufriev, A., Bashford, D. & Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Struct., Funct., Bioinf.* **55**, 383–394, https://doi.org/10.1002/prot.20033 (2004).
29. Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C. & Bussi, G. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **185**, 604–613, https://doi.org/10.1016/j.cpc.2013.09.018 (2014).
30. Zhang, B. W., Cui, D., Matubayasi, N. & Levy, R. M. The excess chemical potential of water at the interface with a protein from end point simulations. *J. Phys. Chem. B* **122**, 4700–4707, https://doi.org/10.1021/acs.jpcb.8b02666 (2018).
31. 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.* **9**, 3136–3150, https://doi.org/10.1021/ct400003r (2013).
32. Gallicchio, E., Paris, K. & Levy, R. M. The agbnp2 implicit solvation model. *J. Chem. Theory Comput.* **5**, 2544–2564, https://doi.org/10.1021/ct900234u (2009).
33. 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.* **6**, 2961–2977, https://doi.org/10.1021/ct1002913 (2010).
34. Wang, L. *et al*. Modeling local structural rearrangements using fep/rest: Application to relative binding affinity predictions of CDk2 inhibitors. *J. Chem. Theory Comput.* **9**, 1282–1293, https://doi.org/10.1021/ct300911a (2013).
35. Wang, L. *et al*. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.* **137**, 2695–2703, https://doi.org/10.1021/ja512751q (2015).

Acknowledgements

This work was supported by NIH grant (GM30580), NSF grant (1665032) and by an NIH computer equipment grant OD020095. This work also used Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation (ACI-1053575). The authors acknowledge invaluable discussions with Zhiqiang Tan at Rutgers University.

Author Contributions

B.W.Z. and R.M.L. wrote the manuscript. B.W.Z. wrote the program codes and prepared the tutorial examples. B.W.Z. and S.A. constructed the web page. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-39420-x.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.