Reweighting from a mixture distribution with WHAM and MBAR

The purpose of this document is to provide a simple explanation of the relationship between MBAR and WHAM using the same formalism for both methods.

Definitions

Configuration (**x**): Vector containing the Cartesian coordinates for a molecule

Full configuration space (Γ): Defines all possible **x**

Thermodynamic state index (*k*): index for a thermodynamic state

Unnormalized density for thermodynamic state *k* ($q_k(\mathbf{x})$): A Boltzmann distribution used as the basic functional form for each thermodynamic state

Thermodynamic state: A region of configuration space that contributes significantly to the mixture distribution. Each thermodynamic state is defined by a single Boltzmann weight multiplied by $q_k(\mathbf{x})$. The product of all normalized thermodynamics state distribution gives the mixture distribution

Partition function for state *k* (c_k, also called the "weight"): A normalization constant that determines the "weight" (or contribution) of each thermodynamic state, *k*, to the total mixture distribution, $c_k = \int_\Gamma d\mathbf{x}\, q_k(\mathbf{x})$

Probability distribution for thermodynamic state *k*: $p_k(\mathbf{x}) = \dfrac{q_k(\mathbf{x})}{c_k}$

Total thermodynamic states (*K*): Index defining the total number of estimator equations (one for each thermodynamic state)

Mixture distribution (***P***): Product of distribution functions for all thermodynamic states $\boldsymbol{P} = \prod_k^K p_k(\mathbf{x})$

Simulation index (*i*): Indexes the simulations used to generate an ensemble

Total number of simulations (*M*): Total number of simulations (or replicas) used to generate the mixture distribution

Total samples counted for thermodynamic state *k* in simulation *i* (N_k^i): Contains the counts for thermodynamic state *k* in simulation *i*

Total samples counted for thermodynamic state *k* (N_k): Contains the sum of counts from all simulations (replicas) that sample state *K:* $N_k = \sum_i^M N_k^i$

Total samples counted for thermodynamic state k (N_k): Contains the sum of counts from all simulations that sample state k

Unbiased potential energy of thermodynamic state k at $\mathbf{x}$ ($U_k^0\left(\mathbf{x}\right)$):

Biasing potential of state k at $\mathbf{x}$ ($U_k^{bias}\left(\mathbf{x}\right)$):

WHAM

The purpose of WHAM is to define the smallest possible number of thermodynamic states for whom a suitable linear combination produces a low-variance mixture distribution. In the ideal implementation of WHAM, the resulting mixture distribution allows prediction of the free energy and/or other observables for unknown configurations.

One of the important concepts in WHAM is the definition of a functional relationship between the biased and unbiased probability distributions:

$$\boldsymbol{P}^{unbias}\left(\mathbf{x}\right)=\exp\left[\beta\left(\boldsymbol{U}^{bias}\left(\eta\left(\mathbf{x}\right)\right)-\Delta f\right)\right]\boldsymbol{P}^{bias}\left(\mathbf{x}\right), \qquad (1.1)$$

where Δf is the change in free energy resulting from the biasing potential $\boldsymbol{U}^{bias}\left(\eta\left(\mathbf{x}\right)\right)$, and $\eta\left(\mathbf{x}\right)$ is a reaction coordinate. The unnormalized ($\tilde{p}$), unbiased probability distribution is defined:

$$\tilde{\boldsymbol{P}}^0\left(\mathbf{x}\right)=\sum_k^K\tilde{p}_k^0\left(\mathbf{x}\right). \qquad (1.2)$$

For a given $\eta\left(\mathbf{x}\right)$ we can define the unbiased potential energy for thermodynamic state k as $U_k^0\left(\eta\left(\mathbf{x}\right)\right)$, and the applied biasing potential as $U_k^{bias}\left(\eta\left(\mathbf{x}\right)\right)$.

$$\tilde{p}_k^0\left(\eta\left(\mathbf{x}\right)\right)=\exp\left[\beta\left(U_k^{bias}\left(\eta\left(\mathbf{x}\right)\right)-\Delta f_k\right)\right]\tilde{p}_k^{bias}\left(\eta\left(\mathbf{x}\right)\right). \qquad (1.3)$$

The goal of WHAM is to define the mixture distribution in the coordinate range for $\eta\left(\mathbf{x}\right)$ with the smallest *K* possible, and to normalize these distributions as well. The normalized, unbiased mixture distribution is defined as:

$$\boldsymbol{P}^0\left(\mathbf{x}\right)=\sum_{k=1}^{K}p_k\left(\mathbf{x}\right)=\sum_{k=1}^{K}\frac{q_k^0\left(\mathbf{x}\right)}{c_k}\,. \tag{1.4}$$

We know the functional form of all $q_k^0\left(\mathbf{x}\right)$ (within a multiplicative constant). Thus, in order to find ***P*** we will need to solve for all c_k. More specifically, we solve for c_k subject to the condition that the statistical error in our mixture distribution is minimal:

$$\frac{\partial\left(\sigma^2\left[\boldsymbol{P}\left(\eta\left(\mathbf{x}\right)\right)\right]\right)}{\partial p_k}=0\,. \tag{1.5}$$

Thus, we have a set of *k* unknown variables, c_k, whose solution requires evaluation of a system of *k* differential equations. Importantly, c_k are coupled and subject to a normalization condition.

It has been shown[1] that:

$$c_k=\int_{\Gamma}d\mathbf{x}\;q_k\left(\mathbf{x}\right)=\frac{n_k\exp\left[-\beta\left(U_k^{bias}\left(\eta\left(\mathbf{x}\right)\right)-\Delta f_k\right)\right]}{\sum_{j=1}^{K}n_j\exp\left[-\beta\left(U_j^{bias}\left(\eta\left(\mathbf{x}\right)\right)-\Delta f_j\right)\right]}\,. \tag{1.6}$$

Through substitution of equation (1.1) and (1.4) into equation (1.6), we can also show that:

$$c_k=\frac{N_k}{\sum_{j=1}^{K}N_j\exp\left[-\beta\left(U_j^{bias}\left(\eta\left(\mathbf{x}\right)\right)-\Delta f_j\right)\right]}\,. \tag{1.7}$$

Another way to interpret/understand the equivalency of equations (1.6) and (1.7) is this: because of normalization conditions we know that any sampling reduction in one thermodynamic state must be compensated by increased sampling in another thermodynamic state. Thus, the weights, c_k, are coupled. Computationally, it is efficient to represent a set of coupled thermodynamic states like this using a *K* x *K* matrix, whose matrix elements are $c_{jk=}c_j{\cdot}c_k$. But we need not discuss this further unless we want to have a rigorous understanding of the numerical methods used to solve for the weights. In order to define Δf_k we may use the relationship

$$\exp\left[-\beta\Delta f_k\right]=\int d\eta \boldsymbol{P}^0\left(\eta\right)\exp\left[-\beta U_k^{bias}\left(\eta\right)\right]. \tag{1.8}$$

Equations (1.5)-(1.8) can be solved iteratively, and self-consistently with respect to Δf_k and c_k. This self-consistent procedure is what most people would refer to as WHAM. In the limit $K\rightarrow\infty$ this is the lowest variance approach to bridge sampling[2]. This limit is often called "un-binned" WHAM (U-WHAM). In practice, however, we have to use bins for most chemical reaction coordinates, because we can't afford to solve for an infinite number of c_k.

MBAR

In the event that the observable we are evaluating is the free energy, the equations solved in MBAR and U-WHAM are identical. The formalism for MBAR[3] allows evaluation of observables other than the free energy. However, in principle WHAM does also, in the event that the formalism is adapted appropriately. This new manuscript, while it has some problems, discusses the relationship between WHAM and MBAR in more detail, and cites several other papers that demonstrate the equivalence of U-WHAM and MBAR[4].

(1) Souaille, M.; Roux, B. Extension to the Weighted Histogram Analysis Method: Combining Umbrella Sampling with Free Energy Calculations. *Computer Physics Communications* **2001**, *135* (1), 40–57.

(2) On a Likelihood Approach for Monte Carlo Integration: Journal of the American Statistical Association: Vol 99, No 468 https://www.tandfonline.com/doi/abs/10.1198/016214504000001664 (accessed Feb 1, 2019).

(3) Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *The Journal of chemical physics* **2008**, *129* (12), 124105–124105.

(4) Ding, X.; Vilseck, J. Z.; Brooks, C. L. A Fast Solver for Large Scale Multistate Bennett Acceptance Ratio Equations. *J. Chem. Theory Comput.* **2019**. https://doi.org/10.1021/acs.jctc.8b01010.