JCTC Journal of Chemical Theory and Computation

Letter

pubs.acs.org/JCTC

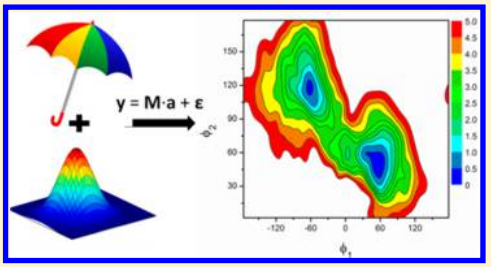
# Efficient Determination of Free Energy Landscapes in Multiple Dimensions from Biased Umbrella Sampling Simulations Using Linear Regression

Yilin Meng and Benoît Roux*

Department of Biochemistry and Molecular Biology, The University of Chicago, 929 East 57th Street, Chicago, Illinois 60637, United States

Supporting Information

**ABSTRACT:** The weighted histogram analysis method (WHAM) is a standard protocol for postprocessing the information from biased umbrella sampling simulations to construct the potential of mean force with respect to a set of order parameters. By virtue of the WHAM equations, the unbiased density of state is determined by satisfying a self-consistent condition through an iterative procedure. While the method works very effectively when the number of order parameters is small, its computational cost grows rapidly in higher dimension. Here, we present a simple and efficient alternative strategy, which avoids solving the self-consistent WHAM equations iteratively. An efficient multivariate linear regression framework is utilized to link the biased probability densities of individual umbrella windows and yield an unbiased global free energy landscape in the space of order parameters. It is demonstrated with practical examples that free energy landscapes that are comparable in accuracy to WHAM can be generated at a small fraction of the cost.




## INTRODUCTION

Molecular dynamics (MD) simulations of detailed atomic models provide a virtual microscope to examine a wide range of complex molecular processes that can play an important role in chemistry, biochemistry, physics, and material science. While a broad range of systems can be investigated computationally, the usefulness of MD is mainly limited by the accuracy of physical approximations used to derive intermolecular forces and our ability to computationally sample the configurational space adequately. The most straightforward sampling strategy relies on brute-force simulations, assuming that the evolution of an unbiased trajectory will be sufficient to generate a Boltzmann weighted sample of the configurational space **R** of interest. To correctly determine the relative statistical weight of different regions of configurational space, it is critical that the unbiased trajectory should be sufficiently long in order for the system to travel back-and-forth in the configurational space **R**. Nevertheless, the perception is that such back-and-forth fluctuations of a trajectory evolving freely according to Newton's classical equation of motions are inefficient and undesirable, because the system spends a large fraction of its time returning to regions that were previously visited. This has motivated a number of special strategies designed to enhance sampling efficiency by trying to prevent excessive return to previously explored regions.

Among the approaches designed for calculating the potential of mean force (PMF) over subspace **Z**, the most commonly used is perhaps the umbrella sampling (US) method.[1,2] This was pioneered by Torrie and Valleau[2] in the 1970s to perform Monte Carlo simulations of systems containing large energy barriers. Umbrella sampling introduces the concept of a biased "window" simulation, a theoretical object aimed at producing an enhanced sampling over a focused region of configurational space. Biasing is typically achieved by introducing an additional (artificial) potential for each window simulation that is referred to as "umbrella potential" or "window potential". Perhaps the most straightforward implementation of this approach is "stratified" US, in which a collection of simulations with narrowly defined biasing potentials (often of quadratic form) are carried out to cover the relevant region of **Z**. Multiple windows simulations are required in order to obtain a sufficiently complete sampling by covering all relevant regions within the subspace of interest. The information from these different biased simulations is converted into local probability histograms, which are then pieced together to produce an unbiased Boltzmann statistical probability.

The weighted histogram analysis method (WHAM),[3] which was developed on the basis of multiple histograms reweighting,[4] has become the standard protocol to combine all of the time series from umbrella windows and to generate the unbiased probability of each bin. Those unbiased probabilities are further processed to yield a potential of mean force (also called free energy landscape). WHAM has not only been applied to umbrella sampling simulations but also been employed to process data from replica-exchange molecular dynamics (REMD)[5,6] and string method simulations.[7] Intrinsically a maximum likelihood estimation of free energy,[6,8,9] the conven-

**Received:** December 15, 2014
**Published:** June 25, 2015





DOI: 10.1021/ct501130r
J. Chem. Theory Comput. 2015, 11, 3523−3529

tional way to solve the WHAM equations is to satisfy a self-consistent condition through an iterative procedure. As a result, the procedure can converge very slowly and the processing time can become very long in some cases, especially when there are multiple dimensions and a very tight convergence criterion is used.[9] Poorly converged WHAM postprocessing of US data can give rise to a quantitatively incorrect PMF. A previous study of ion permeation through gramicidin with umbrella sampling reported 100,000 iterations for WHAM to achieve a satisfactory convergence of the PMF.[10] Further compounding these issues, it is important to note that the bin size for the biased histogram adds an unwanted source of error and uncertainty in postprocessing umbrella sampling data. If the histogram bins are too small, there is a large statistical error in the count of events (particularly in multidimensions), while there is a large error in estimating the biasing potential when the bins are too large because the histogram is coarsely represented. Error estimation and convergence in the iterations are important issues when using WHAM, and several works have been published to address those.[6,9,11,12] However, one should notice that alternative methodologies such as single-sweep,[13] umbrella integration,[14] and a variational method based on maximum likelihood estimation[15] are also able to combine umbrella windows to produce an unbiased free energy landscape. More recently, a Gaussian process regression method was developed to reconstruct the free energy landscape from umbrella sampling.[16] In this approach, a Bayesian model with Gaussian prior and likelihood functions is used to combine the observed data.

In this work, we present a simple and efficient strategy to address these issues to determine an unbiased free energy landscape in a multidimensional space of order parameters $Z$ without employing WHAM. Inspired by the single-sweep method,[13] a multivariate linear regression model is utilized to link the biased probability densities of individual umbrella windows and to yield an unbiased global free energy landscape over the subspace $\mathbf{Z}$. It is demonstrated that free energy landscapes in multidimension that are of comparable accuracy to those obtained with WHAM can be produced with this method at a much reduced computational cost.

## METHODOLOGY

In this section, the WHAM methodology is first briefly explained. Then we propose a method to construct the free energy landscape without employing WHAM. In umbrella sampling simulations, the PMF along an order parameter can be written in the following forms:

$$W(x) = -k_B T \ln P^{(0)}(x) \tag{1}$$

$$W(x) = -k_B T \ln P^{(b)}(x) - U^{(b)}(x) + F \tag{2}$$

where $x$ represents the multidimensional order parameter, $W$ is the PMF, $k_B$ is the Boltzmann constant, $T$ is the temperature of the canonical/isothermal−isobaric ensemble, $P^{(0)}$ is the unbiased probability distribution function (PDF), $P^{(b)}$ is the biased PDF from an umbrella window, $U^{(b)}$ is the biasing potential applied to that umbrella window, and $F$ is an undetermined factor. This factor depends on biasing potential and hence varies from window to window. Solving eqs 1 and 2 simultaneously outputs a PMF as a function of $x$. In most cases, WHAM is used to combine all windows and to optimally estimate $F$ for each window. The WHAM equations are given as follows:

$$P^{(0)}(x) = \frac{\sum_{l=1}^{N} h_l(x)}{\sum_{k=1}^{N} n_k \exp([F_k - U^{(b)}(x)]/k_B T)} \tag{3}$$

$$F_k = (-k_B T) \ln \sum_x P^{(0)}(x) \exp(-U^{(b)}(x)/k_B T) \tag{4}$$

where $N$ is the number of umbrella windows, $l$,$k$ are indices of umbrella windows, $h(x)$ is the counts at bin $x$, $n_k$ is the number of data points from window $k$, and $F_k$ denotes the undetermined factor for window $k$. Those two equations are coupled and will be solved in an iterative manner until self-consistent. Detailed descriptions of the WHAM iterative method were presented by Kumar et al.[3] and by Roux.[1]

To avoid constructing a PMF by iteratively solving the WHAM equations, the PMF is assumed to be a linear combination of radial-basis Gaussian functions,

$$W(x) = \sum_{m=1}^{M} a_m g_m(x) \tag{5}$$

$$g_m(x) = \exp(-(x - x_m)^2/2{\sigma_m}^2) \tag{6}$$

where $g_m(x)$ is a Gaussian function centered at $x_m$ and with a variance of ${\sigma_m}^2$ and $a_m$ is the weight (amplitude) of $g_m$. The rationale for expressing the PMF by a linear combination of radial-basis functions (in our case, Gaussian functions) can be found in the work by Maragliano and Vanden-Eijnden.[13] In the case of multidimensional umbrella sampling calculations, each multivariate $g_m$ is assumed to be simply the product of one-dimensional Gaussian functions, $g_m(u_1,u_2,...,u_N) = \prod_{n=1}^{N} g_m(u_n)$ as in the case of metadynamics.[17] Since eq 2 has an undetermined offset factor $F$ that depends on the particular umbrella sampling window, a direct fitting to the absolute value of the PMF is not feasible. However, the undetermined factor $F$ cancels out if two points ($x_1$ and $x_2$) are selected from the same umbrella window and the difference in $W(x)$ is considered. The difference in PMF ($\Delta W$) between $x_1$ and $x_2$ can be written as

$$\Delta W = W(x_2) - W(x_1) = \sum_{m=1}^{M} a_m(g_m(x_2) - g_m(x_1)) \tag{7}$$

$$W(x_2) - W(x_1) = -k_B T \ln[P^{(b)}(x_2)/P^{(b)}(x_1)] - [U^{(b)}(x_2) - U^{(b)}(x_1)] \tag{8}$$

Therefore, the actual $\Delta W$ (response variable) and the basis functions are associated through the following equation:

$$-k_B T \ln[P^{(b)}(x_2)/P^{(b)}(x_1)] - [U^{(b)}(x_2) - U^{(b)}(x_1)] = \sum_{m=1}^{M} a_m(g_m(x_2) - g_m(x_1)) + \varepsilon_m \tag{9}$$

where $\varepsilon_m$ is the residual error. If the means and the variances of Gaussian basis functions are not included in the fitting, a multivariate linear regression model of the form $\mathbf{y} = \mathbf{M}\cdot\mathbf{a} + \boldsymbol{\varepsilon}$ is obtained. In this model, $\mathbf{y}$ is a vector that holds the values of response variables (the left-hand side of eq 9), $\mathbf{a}$ is a vector of the coefficients $\{a_m\}$, and $\mathbf{M}$ is a matrix whose element $M_{mk}$ is $g_m(x_k)$. The selection of $x_1$ and $x_2$ in eq 9 is arbitrary for any umbrella window. The sampling is commonly maximum near the center of a given window, and the uncertainty on the biased histogram $P^{(b)}(x)$ becomes larger as $x$ moves away from the


DOI: 10.1021/ct501130r
J. Chem. Theory Comput. 2015, 11, 3523−3529

center. For simplicity, one point (e.g., $x_1$) is set to be the center of an umbrella window. The least-squares estimator is used to obtain the coefficients $\{a_m\}$. In the least-squares estimator, the sum of residuals ($\chi^2 = \sum_{m=1}^{M} \varepsilon_m{}^2$) is minimized with respect to $\{a_m\}$. In the matrix form, $\chi^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}$ where the letter T denotes the transpose. $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{Ma}$ because the linear regression model is being used. Therefore,

$$\chi^2 = (\mathbf{y}^T - \mathbf{a}^T\mathbf{M}^T)(\mathbf{y} - \mathbf{Ma}) \tag{10}$$

$$\chi^2 = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{Ma} - \mathbf{a}^T\mathbf{M}^T\mathbf{y} + \mathbf{a}^T\mathbf{M}^T\mathbf{Ma} \tag{11}$$

The first term on the right-hand side (RHS) of eq 11 is independent of **a**, the second and the third terms on the RHS are equal and could be replaced by $2\mathbf{a}^T\mathbf{M}^T\mathbf{y}$, and the last term on the RHS is in a quadratic form of **a**. To determine the optimal solution, the first derivative of $\chi^2$ with respect to **a** is set to zero:

$$\frac{\partial\chi^2}{\partial\mathbf{a}} = -2\mathbf{M}^T\mathbf{y} + 2\mathbf{M}^T\mathbf{Ma} = 0 \tag{12}$$

$$\mathrm{M}^T\mathrm{Ma} = \mathrm{M}^T\mathrm{y} \tag{13}$$

$$\mathrm{a} = (\mathrm{M}^T\mathrm{M})^{-1}\mathrm{M}^T\mathrm{y} \tag{14}$$

In practice, a singular value decomposition (SVD) method[18] is employed to achieve a robust and stable least-squares estimation of the coefficients $\{a_m\}$. Once the value of the coefficients $a_m$ has been determined, the free energy landscape can be reconstructed using eq 5.

## RESULTS AND DISCUSSION

Exploring the folding free energy landscape of a solvated peptide is a realistic task that is often used to demonstrate the efficiency of enhanced sampling approaches.[19−22] Met-enkephalin is a small pentapeptide with the sequence YGGFM (see Supporting Information Figure S1) and was used as a test case here. We previously utilized a self-learning adaptive US strategy to explore the folding free energy landscape, and 263 umbrella windows were determined to be essential in constructing the free energy landscape.[22] Each one of the 263 umbrella windows was propagated for 1 ns. Those 263 umbrella sampling simulations formed the basis of the analyses presented in this work. More details on system construction and simulation parameters can be found in the Supporting Information.

**Free Energy Landscape of Met-enkephalin Folding Obtained from Linear Regression of $\Delta W$.** As mentioned earlier, the center and the width of a basis function were kept fixed in order to apply a linear regression model. In the current scheme of fitting, the number of Gaussians and their centers were chosen to be the same as those of the umbrella windows. To further simplify our model, we used the same width for all basis functions. For each umbrella window, the $x_1$ in eq 9 was selected as the center of the window and another 50 data points were randomly selected to serve as $x_2$ values, in order to build the response vector **y** and the design matrix **M**. Therefore, **y** is a column vector with 13150 rows, while **M** is a 13150 by 263 matrix. **a** is a column vector containing the weights of 263 Gaussian functions. To have a coarse investigation of the effect of width of Gaussian ($\sigma$) on regression, a scan of $\sigma$ from 3° to 14° was performed with an increment of 1°. The mean of squared residuals (sum of squared residuals divided by number of data points and will be referred to as residuals for short in future discussion), the condition number of singular values, and the accuracy of the fitted PMF were considered in order to evaluate the choice of $\sigma$.

Residual and condition number with respect to $\sigma$ (illustrated in Supporting Information Figure S2) were examined first. By investigating residual and condition number, one would identify how $\sigma$ affected regression analysis. Ideally, a $\sigma$ value producing a smaller residual and condition number should be used. As shown in Supporting Information Figure S2, both residual and condition number increase rapidly from $\sigma = 10° = 1.0d_{us}$, where $d_{us}$ is the size of umbrella windows in each dimension. When $\sigma$ is greater than 13°, the smallest singular value becomes zero, resulting in an infinite condition number. The accuracy of the PMFs generated from the linear model as a function of $\sigma$ was also evaluated. The PMF obtained from WHAM calculation (using all 263 windows) was used as the reference (labeled as $W_{ref}$). The RMSE of the fitted PMF relative to $W_{ref}$ as a function of $\sigma$ is plotted in Figure 1A. Since PMFs from both WHAM and the linear model revealed two stable conformations, the free energy difference ($\Delta G$) between those two conformations was also calculated to evaluate the performance of the linear model (the definition of each conformation can be found in ref 21). Those $\Delta G$ values along with the $\Delta G$ from $W_{ref}$ are displayed in Figure 1B. According to Figure 1, the RMSE values are quite large at $\sigma = 3°$ and 4° and stabilize around 0.5 kcal/mol when $\sigma \geq 5°$. The large differences in PMFs at $\sigma = 3°$ and 4° indicate that basis functions being too local would cause a problem in producing an accurate PMF and should not be used even though they perform well in both condition number and residual tests. Surprisingly, all of the $\Delta G$ values yielded from the linear model are close to the WHAM $\Delta G$, even though the RMSE values were quite large for $\sigma = 3°$ and 4°. Selected PMFs from fitting and $W_{ref}$ are plotted in Figure 2. The quantitative evaluations and visual inspections of PMFs (Figures 1 and 2) demonstrate that fitting $\Delta W$ with a linear model is able to generate PMFs resembling $W_{ref}$. Results of free energies, condition numbers, and residuals suggest that our model is not very sensitive on the choice of $\sigma$ values, in the case of Met-enkephalin. The range of $\sigma$ values that would give satisfactory PMFs, condition numbers, and residuals is between 50% and 100% of the size of umbrella windows (0.5−1.0 times $d_{us}$). Our results indicate that the range of $\sigma$ values is important for obtaining an accurate PMF: $\sigma$ values that are too small tend to generate PMFs with poor quality, and $\sigma$ values that are too large yield condition numbers that approach infinity.

In practical application, it is necessary to pick an optimal value for the width of the Gaussian basis function ($\sigma$) without any prior knowledge of the PMF. A number of factors may affect the information gained from US simulations (e.g., the force constant, the spacing between windows, and so on), emphasizing the need for a robust estimate for the width of the Gaussian basis function. As can be observed in the present analysis, inaccuracies in the PMF (e.g., Figure 2B) are caused by the basis function being too narrow. Simply put, the width is too small and each basis function is adjusted based on very local and limited information, leaving large inaccurate gaps in the PMF. To provide an objective criterion to pick a reasonable value for the width $\sigma$, it is useful to consider the overlap between neighboring basis functions. The normalized overlap coefficient between two Gaussian basis functions is defined as


DOI: 10.1021/ct501130r
J. Chem. Theory Comput. 2015, 11, 3523−3529

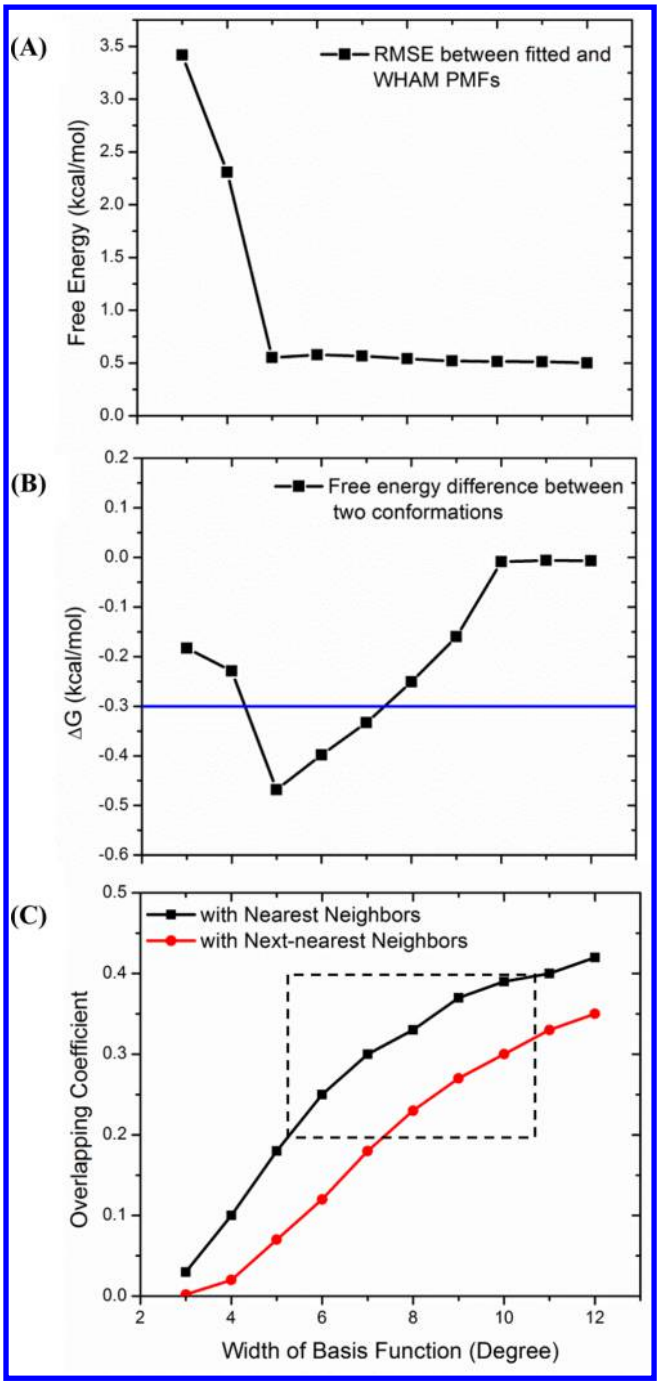



**Figure 1.** (A) Root-mean-squared error (RMSE) between the fitted PMF and $W_{ref}$, as a function of the width of basis functions. $W_{ref}$ is the PMF produced by WHAM. 263 umbrella windows were employed in the calculation of all PMFs. (B) $\Delta G$ between two conformations with respect to the width of the basis functions. The blue horizontal line in B represents the $\Delta G$ value yielded from $W_{ref}$. (C) Overlapping coefficient between a Gaussian basis function and its nearest neighbors and its next-nearest neighbors as a function of the width $\sigma$.

$$\text{overlap} = \int g_m(\mathbf{R})\, g_n(\mathbf{R})\, d\mathbf{R} \Big/ \int g_m(\mathbf{R})\, d\mathbf{R}$$

where $g_m$ and $g_n$ are the Gaussian basis function and are as defined in this work, and $\mathbf{R}$ is the collective variable (CV) space. In Figure 1C, we show the overlapping coefficient between a Gaussian and its nearest neighbors and its next-nearest neighbors as a function of $\sigma$ to clarify its relationship to the accuracy of the PMF. As expected, the overlapping coefficient increases monotonically as a function of $\sigma$. By comparison with Figure 1A, it is observed that the overlapping coefficient between the nearest neighbors needs to be at least on the order of 0.2 or larger in order to achieve an accurate PMF. A value of 0.4 is a reasonable upper bound based on an analysis of residuals and condition numbers (see Supporting Information Figure S2). For maximum confidence in practical applications, a scan of $\sigma$ values should be performed. To further explore these issues, we have also applied the linear model to the activating conformational transition of the Src kinase domain and achieved an accurate PMF (the smallest RMSE and $\Delta\Delta G$ are 0.6 and 0.1 kcal/mol and occur simultaneously). Accurate PMFs can also be obtained in approximately the same range of overlapping coefficients for the conformational changes in the Src kinase domain (data not shown).

We further investigated the impact of the number of umbrella windows on the accuracy of the fitted PMFs. To evaluate this effect, a subset of the original 263 windows was selected and the PMFs were reconstructed. Two cases were considered here: using 128 and 67 windows, representing placing a window every 20° on the $x$-axis and 10° on the $y$-axis, and every 20° on both axes. This approximately corresponds to 50% and 25% of the original 263 windows. In the case where 128 windows were used, two trial values of $\sigma$ (8° and 10°) were employed in the fitting (see Supporting Information Figure S3 for the resulting free energy landscapes). Moreover, WHAM calculation was performed using the same 128 windows and the free energy landscape is illustrated in Supporting Information Figure S3C. The RMSE values of all three free energy landscapes relative to $W_{ref}$ and the $\Delta G$ values are listed in Table 1. First of all, the 128-window WHAM PMF is rather noisy, likely to be caused by insufficient overlap among time series (a scatter plot of the time series is displayed in Supporting Information Figure S3D). This behavior confirms that WHAM is sensitive to the distribution of data points in the configurational space. In this case, reducing the number of windows greatly decreases the quality of the free energy landscape: the 128-window WHAM PMF not only displays the largest RMSE relative to $W_{ref}$ but also yields the wrong relative stability of the two conformations, according to Table 1. Comparison of RMSE and $\Delta G$ values from two schemes of generating PMFs reveals that the linear model outperforms WHAM, in the case of using 128 umbrella windows. Next, Table 1 demonstrates that the linear model is able to produce PMFs similar to $W_{ref}$ (RMSE is ~0.6 kcal/mol for both $\sigma$ values) and to yield accurate $\Delta G$ between the two conformations, even with only ~50% of the original windows. This indicates that the computational cost of umbrella sampling would reduce by 50% but without losing much accuracy in free energy landscapes, when our method is employed to process US data. In the case where 67 windows were used, $\sigma$ = 10° and 15° (0.5 and 0.75 times the new $d_{us}$) were employed in the fitting. The resulting PMFs along with the 67-window WHAM PMF are shown in Supporting Information Figure S4. The RMSE of each PMF relative to $W_{ref}$ and the $\Delta G$ value from each PMF can also be found in Table 1. In this case, all three RMSEs and $\Delta G$ values display large discrepancy (>1 kcal/mol) relative to the reference. However, the more stable conformation can be correctly identified in the PMFs generated from the linear model, while WHAM using 67 windows failed to do so.


DOI: 10.1021/ct501130r
J. Chem. Theory Comput. 2015, 11, 3523−3529

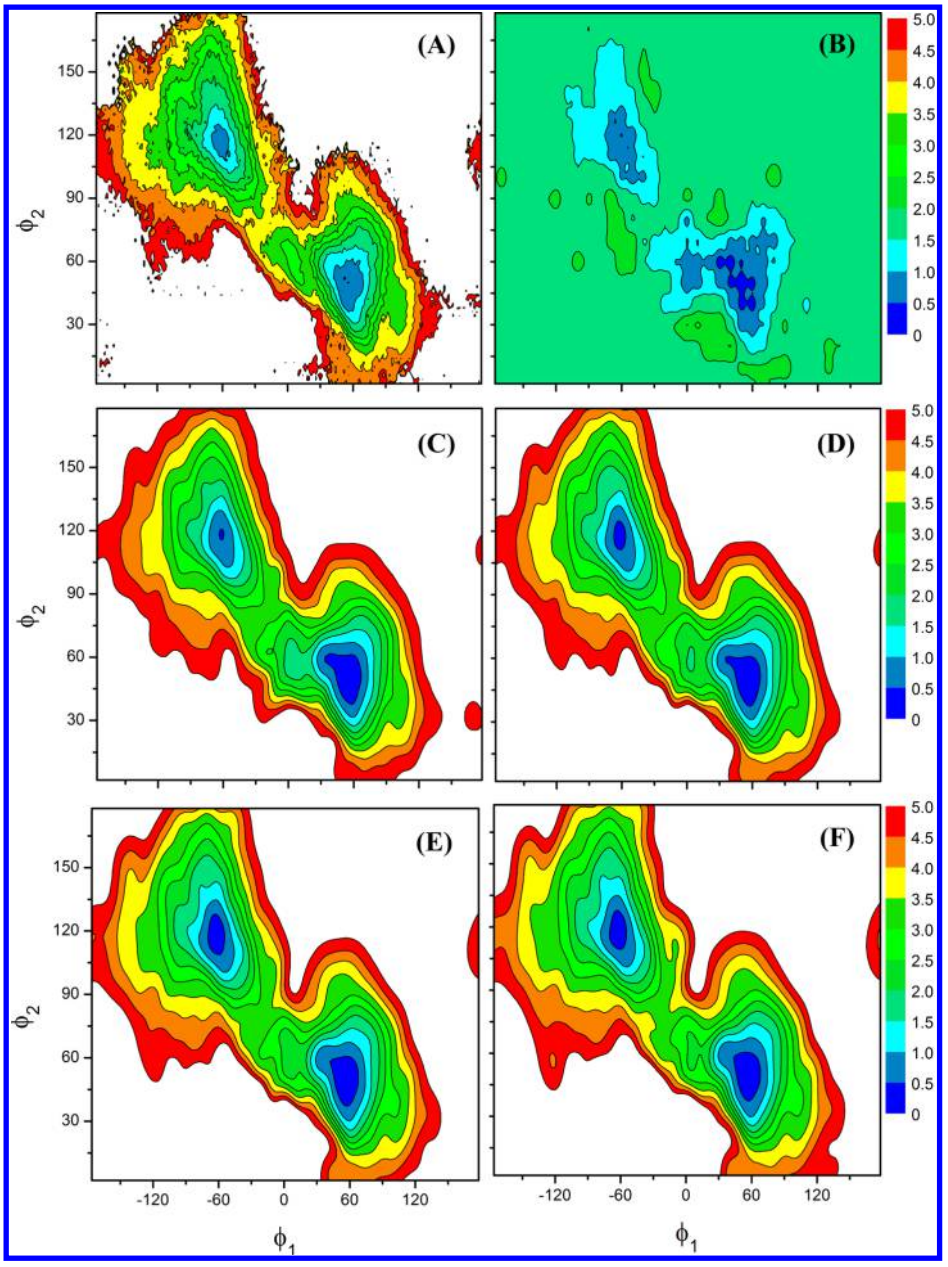



**Figure 2.** $W_{ref}$ and selected PMFs obtained from the linear model using 263 umbrella windows. (A) WHAM ($W_{ref}$). (B) Linear model with $\sigma = 4°$. (C) Linear model with $\sigma = 6°$. (D) Linear model with $\sigma = 8°$. (E) Linear model with $\sigma = 10°$. (F) Linear model with $\sigma = 12°$. The unit of all free energy landscapes is in kilocalories per mole.

**Table 1. RMSE and $\Delta G$ Values Generated from Time Series of 128 and 67 Windows**[a]

| | 128 windows | | | 67 windows | | |
|---|---|---|---|---|---|---|
| | WHAM | $\sigma = 8°$ | $\sigma = 10°$ | WHAM | $\sigma = 10°$ | $\sigma = 15°$ |
| RMSE | 1.06 | 0.64 | 0.58 | 1.41 | 1.61 | 1.11 |
| $\Delta G$ | 0.30 | −0.17 | −0.29 | 0.83 | −1.26 | −1.31 |

[a]All energetic quantities have the unit of kilocalories per mole. RMSE represents the root-mean-squared error of a fitted PMF relative to $W_{ref}$ (PMF generated from WHAM with 263 windows). $\Delta G$ is the Gibbs free energy difference between the two conformations shown in PMF (the definition of each conformation can be found in ref 21). The $\Delta G$ calculated from $W_{ref}$ is −0.3 kcal/mol.

It is worth noting that one could avoid building any histograms with one additional approximation. Under this approximation, each local histogram for a given umbrella sampling window will be replaced by the product of a set of univariate normal PDFs, where each univariate normal distribution represents the PDF of one order parameter. The rationale of using this approximation is that order parameters are often independent random variables and are distributed normally when harmonic potentials are applied to an umbrella window, provided that the force constants are large enough. In such a case, $W(u_1,u_2,...,u_N)$ is a slowly varying function in comparison with biasing potentials. Therefore, the joint biased PDF (multivariate normal distribution) can be approximately expressed as a product of the PDF of each individual order parameter. Approximating a local histogram to the product of independent normal PDFs would allow a better sampling of the tails and a more flexible choice of data points when constructing the design matrix. It could also further accelerate


DOI: 10.1021/ct501130r
J. Chem. Theory Comput. 2015, 11, 3523−3529

data analysis by avoiding building histograms: only the mean and the variance of an order parameter need to be estimated from the time series, when generating a univariate normal PDF. In the case of Met-enkephalin simulation, Pearson's correlation coefficient ($\rho$) between $\phi_1$ and $\phi_2$ should be zero for all umbrella windows. For a bivariate normal distribution, a zero correlation coefficient is equivalent to independence. A histogram of 263 correlation coefficients is given in Supporting Information Figure S5A. Kolmogorov−Smirnov statistical test was performed to examine whether each order parameter is normally distributed. Histograms of the $p$-values obtained from the Kolmogorov−Smirnov test are demonstrated in Supporting Information Figure S5B,C. Out of 263 $p$-values for each order parameter, 11 and 5 of them were smaller than 0.05 for $\phi_1$ and $\phi_2$, respectively. The results of correlation coefficient and $p$-value suggested that the approximation holds for most of the windows even though the force constant in our simulation is not large (0.02 kcal/(mol·deg$^2$)). A larger force constant should be used so that replacing the local histograms of umbrella windows with a product of univariate normal PDFs is more rigorous. One may note that a large biasing harmonic potential is also one of the underlying assumptions of the single-sweep mean force method of Maragliano and Vanden-Eijnden.[13] If building a histogram can be avoided, the memory usage can be considerably reduced because there is no need to store the bin locations and counts. Decreasing memory usage and avoiding the iterative procedure are useful in postprocessing US calculations in high dimensions ($N > 3$).

**Fitting $\Delta W$ versus Fitting Mean Forces $\langle F \rangle$.** An alternative strategy of constructing a free energy landscape based on least-squares fitting is the single-sweep method.[13] In the single-sweep method, the conformational space is sampled by temperature-accelerated molecular dynamics (TAMD).[23] TAMD also generates an irregular grid of points to which Gaussian radial functions centered. Restrained simulations are performed to estimate the mean force (first derivatives of the PMF) at those grid points, which are fitted by a linear expansion of Gaussian basis functions. This force-matching strategy can be applied to umbrella sampling calculations as well. The mean force applied to the center of an umbrella window could be estimated from the time series.[24] In an ideal case (when TAMD simulation is long enough), our linear model is equivalent to the single-sweep method when a uniform grid of Gaussian centers is generated by TAMD. However, one must pay attention to two issues when utilizing the force-matching scheme to process US data. One issue is, unlike fitting $\Delta W$, the number of mean forces (the number of independent variables used to build the design matrix) is equal to the number of umbrella windows. The amount of data that can be used by regression is smaller than what is used in fitting $\Delta W$. Increasing the number of data points requires performing more umbrella sampling simulations which increases the computational cost. The other issue is calculating the mean force $\langle \mathbf{F} \rangle$ from a biasing potential. Computing $\langle \mathbf{F} \rangle$ from umbrella potentials results in a restraining force instead of a constraining force. In order to better approximate the constraining force, the force constant of the harmonic potential needs to be large which could cause a narrow distribution of data points in an umbrella window. Therefore, more windows are required to cover the configurational space, even the essential regions.

In this case study, 263 umbrella windows and three $\sigma$ values (5°, 8°, and 10°) were employed in fitting mean forces (the resulting free energy landscapes are shown in Supporting Information Figure S6). The RMSEs relative to $W_{ref}$ and $\Delta G$ values are listed in Table 2. Comparison of RMSE and $\Delta G$ values from two strategies of choosing independent variables reveals that regression on $\Delta W$ shows superior performance to fitting $\langle \mathbf{F} \rangle$. However, one must notice that this better performance may be caused by the two issues mentioned previously. For example, the force constant used in our US calculations is small: 0.02 kcal/(mol·deg$^2$) which is ∼65 kcal/(mol·rad$^2$). This is much smaller than the force constant used in the mean force calculations by Maragliano et al.[24]

**Table 2. Comparison of Results Generated from Linear Regression of $\Delta W$ and $\langle \mathbf{F} \rangle$ (Mean Force)**

| | fitting $\Delta W$ | | | fitting $\langle \mathbf{F} \rangle$ | | |
|---|---|---|---|---|---|---|
| | $\sigma$ = 5° | $\sigma$ = 8° | $\sigma$ = 10° | $\sigma$ = 5° | $\sigma$ = 8° | $\sigma$ = 10° |
| RMSE | 0.55 | 0.54 | 0.51 | 1.96 | 1.77 | 1.83 |
| $\Delta G$ | −0.47 | −0.25 | −0.009 | −0.33 | −0.098 | −0.017 |

[a]All energetic quantities have the unit of kilocalories per mole. RMSE and $\Delta G$ values were calculated using the same scheme as those listed in Table 1.

## CONCLUSION

A simple and efficient model based on a multivariate linear regression is proposed for processing the time series generated from biased US simulations to reconstruct the unbiased free energy landscape. The basic idea of the method is to express the PMF as a linear combination of Gaussian basis functions, and the $\Delta W$ values are associated with basis functions through a multivariate linear regression model. By employing the linear regression model, the PMF can be computed without solving WHAM equations iteratively. This model is applied to the study of conformational equilibrium of Met-enkephalin in explicit solution. When all 263 umbrella windows are used, our model is able to generate PMFs with comparable accuracy to the PMF yielded by WHAM. When only a subset of the umbrella windows is used (128 windows), the PMF determined from the linear regression model is actually superior to the one obtained by solving the self-consistent WHAM equations. In this case, the PMFs generated by the linear regression model are still comparable with 263-window WHAM PMF, suggesting that a significantly smaller number of umbrella windows is required to construct the free energy landscape without loss in accuracy. When an insufficiently small subset of umbrella windows is used (67 windows), neither the linear regression model nor WHAM yields accurate free energy estimation. This behavior confirms the critical role of sampling in constructing a free energy landscape. Nevertheless, the overall performance of the linear regression to determine $\Delta W$ suggests that this approach has the ability to yield accurate PMFs at a much smaller computational cost, with respect to both the postprocessing analysis and the number of biased US simulations.

## ASSOCIATED CONTENT

### Supporting Information

Text discussing computation details, Figure S1 showing the structure of Met-enkephalin in stick-and-ball representation, Figure S2 showing plots of the condition number and the mean of the squared residuals with respect to the width of the basis functions ($\sigma$), Figure S3 illustrating the free energy landscape


DOI: 10.1021/ct501130r
J. Chem. Theory Comput. 2015, 11, 3523−3529

yielded from WHAM using 128 umbrella windows and a scatter plot of the time series, Figure S4 displaying free energy landscapes obtained from 67 umbrella windows, Figure S5 demonstrating the results from statistical analyses of time series, and Figure S6 showing the PMFs obtained from fitting mean force. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/ct501130r.

## AUTHOR INFORMATION

### Corresponding Author

*E-mail: roux@uchicago.edu. Tel.:1-773-834-3557. Fax: 1-773-702-0439.

### Funding

This work was supported by the National Cancer Institute (NCI) of the National Institutes of Health (NIH) through Grant CAO93577. The computations were supported in part by the Extreme Science and Engineering Discovery Environment (XSEDE) Grant No. OCI-1053575, and by NIH through resources provided by the Computation Institute and the Biological Sciences Division of the University of Chicago and Argonne National Laboratory, under Grant S10 RR029030-01.

### Notes

The authors declare no competing financial interest.

## REFERENCES

(1) Roux, B. The Calculation of the Potential of Mean Force Using Computer Simulations. *Comput. Phys. Commun.* **1995**, *91* (1−3), 275−282.

(2) Torrie, G. M.; Valleau, J. P. Monte-Carlo Free-Energy Estimates Using Non-Boltzmann Sampling - Application to Subcritical Lennard-Jones Fluid. *Chem. Phys. Lett.* **1974**, *28* (4), 578−581.

(3) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules 0.1. The Method. *J. Comput. Chem.* **1992**, *13* (8), 1011−1021.

(4) Ferrenberg, A. M.; Swendsen, R. H. Optimized Monte-Carlo Data-Analysis. *Phys. Rev. Lett.* **1989**, *63* (12), 1195−1198.

(5) 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* (1), 26−41.

(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* (14), 6722−6731.

(7) Rosta, E.; Nowotny, M.; Yang, W.; Hummer, G. Catalytic Mechanism of RNA Backbone Cleavage by Ribonuclease H from Quantum Mechanics/Molecular Mechanics Simulations. *J. Am. Chem. Soc.* **2011**, *133* (23), 8934−8941.

(8) Bartels, C. Analyzing biased Monte Carlo and molecular dynamics simulations. *Chem. Phys. Lett.* **2000**, *331* (5−6), 446−454.

(9) Zhu, F. Q.; Hummer, G. Convergence and error estimation in free energy calculations using the weighted histogram analysis method. *J. Comput. Chem.* **2012**, *33* (4), 453−465.

(10) Allen, T. W.; Andersen, O. S.; Roux, B. Ion permeation through a narrow channel: Using gramicidin to ascertain all-atom molecular dynamics potential of mean force methodology and biomolecular force fields. *Biophys. J.* **2006**, *90* (10), 3447−3468.

(11) Hub, J. S.; de Groot, B. L.; van der Spoel, D. g_wham-A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates. *J. Chem. Theory Comput.* **2010**, *6* (12), 3713−3720.

(12) Kastner, J.; Thiel, W. Analysis of the statistical error in umbrella sampling simulations by umbrella integration. *J. Chem. Phys.* **2006**, *124* (23), 234106.

(13) Maragliano, L.; Vanden-Eijnden, E. Single-sweep methods for free energy calculations. *J. Chem. Phys.* **2008**, *128* (18), 184110.

(14) Kastner, J.; Thiel, W. Bridging the gap between thermodynamic integration and umbrella sampling provides a novel analysis method: "umbrella integration". *J. Chem. Phys.* **2005**, *123* (14), 144104.

(15) Lee, T. S.; Radak, B. K.; Pabis, A.; York, D. M. A New Maximum Likelihood Approach for Free Energy Profile Construction from Molecular Simulations. *J. Chem. Theory Comput.* **2013**, *9* (1), 153−164.

(16) Stecher, T.; Bernstein, N.; Csanyi, G. Free Energy Surface Reconstruction from Umbrella Samples Using Gaussian Process Regression. *J. Chem. Theory Comput.* **2014**, *10* (9), 4079−4097.

(17) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (20), 12562−12566.

(18) Press, W. H. *Numerical recipes in C: The art of scientific computing*, 2nd ed.; Cambridge University Press: Cambridge, U.K. and New York, NY, USA, 1992; p xxvi, 994 p.

(19) Henin, J.; Fiorin, G.; Chipot, C.; Klein, M. L. Exploring Multidimensional Free Energy Landscapes Using Time-Dependent Biases on Collective Variables. *J. Chem. Theory Comput.* **2010**, *6* (1), 35−47.

(20) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314* (1−2), 141−151.

(21) Sutto, L.; D'Abramo, M.; Gervasio, F. L. Comparing the Efficiency of Biased and Unbiased Molecular Dynamics in Reconstructing the Free Energy Landscape of Met-Enkephalin. *J. Chem. Theory Comput.* **2010**, *6* (12), 3640−3646.

(22) Wojtas-Niziurski, W.; Meng, Y. L.; Roux, B.; Berneche, S. Self-Learning Adaptive Umbrella Sampling Method for the Determination of Free Energy Landscapes in Multiple Dimensions. *J. Chem. Theory Comput.* **2013**, *9* (4), 1885−1895.

(23) Maragliano, L.; Vanden-Eijnden, E. A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. *Chem. Phys. Lett.* **2006**, *426* (1−3), 168−175.

(24) Maragliano, L.; Fischer, A.; Vanden-Eijnden, E.; Ciccotti, G. String method in collective variables: Minimum free energy paths and isocommittor surfaces. *J. Chem. Phys.* **2006**, *125* (2), 024106.



DOI: 10.1021/ct501130r
*J. Chem. Theory Comput.* 2015, 11, 3523−3529