Improved estimation of density of states for Monte Carlo sampling via MBAR

Yuanwei Xu[*,†] and P. Mark Rodger[*,‡]

Centre for Scientific Computing, University of Warwick, Coventry, UK, and Centre for Scientific Computing and Department of Chemistry, University of Warwick, Coventry, UK

E-mail: Yuanwei.Xu@warwick.ac.uk; p.m.rodger@warwick.ac.uk

Abstract

We present a new method to calculate the density of states using the Multistate Bennett Acceptance Ratio (MBAR) estimator. We use a combination of parallel tempering (PT) and multicanonical simulation to demonstrate the efficiency of our method in a statistical model of sampling from a two-dimensional normal mixture and also in a physical model of aggregation of lattice polymers. While MBAR has been commonly used for final estimation of thermodynamic properties, our numerical results show that the efficiency of estimation with our new approach, which uses MBAR as an intermediate step, often improves upon conventional use of MBAR. We also demonstrate that it can be beneficial in our method to use full PT samples for MBAR calculations in cases where simulation data exhibit long correlation.

[*]To whom correspondence should be addressed
[†]Centre for Scientific Computing, University of Warwick, Coventry, UK
[‡]Centre for Scientific Computing and Department of Chemistry, University of Warwick, Coventry, UK

1 Introduction

Systems with rugged energy landscapes are very common in physics, biology and chemistry. Crucial to understanding such systems is to have a comprehensive and efficient sampling of the configuration space that makes optimal use of available computational resources. Monte Carlo (MC) methods have been widely used to study such systems. The basic Metropolis algorithm,[1] although commonly used, will most likely be trapped in a local energy basin when temperature is low. Many advanced algorithms have been proposed to improve sampling efficiency, and the choice of which algorithm to use largely depends on the nature of the system in question.

Generally speaking, the MC methods used to study physical and chemical systems can be broadly classified into two categories: temperature-based and energy-based. In temperature-based methods, the system is simulated at one or several predefined temperatures and the Boltzmann weight is used; examples of such methods include the classical Metropolis algorithm, simulated tempering[2,3] and parallel tempering.[4,5] One deficiency of these methods, apart from the actual sampling, is that quantities like the free energy and entropy are not directly accessible. Fortunately, advances in free energy calculation greatly facilitate analysis of simulation data with much higher statistical efficiency than earlier methods. For example, in the Weighted Histogram Analysis Method[6] (WHAM), the density of states (DoS) are estimated by optimally combining estmates from each simulation after discretizing energy with a suitable resolution; and in the Multistate Bennett Acceptance Ratio (MBAR) method,[7] the problem of calculating free energy differences is treated as a problem in estimating the ratio of normalizing constants and it uses extended bridge sampling theory to derive statistical estimators that are proven to be optimal. MBAR removes discretization of energy, is capable of directly producing estimates of free energy and equilibrium expectations, and thus obviates the need to calculate the density of states.

In contrast to temperature-based methods, energy-based methods usually work in a generalized ensemble which is independent of temperature. The multicanonical sampling,[8] Wang-

Landau algorithm,[9] and transition matrix Monte Carlo[10,11] are examples of such methods. A common feature of the aforementioned algorithms is that the DoS is refined iteratively so that the resulting energy histogram is approximately flat. One could either reweight these biased samples to obtain properties with respect to the canonical ensemble or use the density of states produced from the final iteration to calculate thermodynamic properties, in which case the histogram should be sufficiently flat to provide acceptable accuracy.

While it is generally considered a merit that methods like the Wang-Landau algorithm allow for a quick exploration of the whole energy spectrum, there are situations where this is not always desirable. For example, the range of potential energy in complex systems could span several orders of magnitude. Two implications are that computationally, the time needed to traverse all energy levels in a random walk increases as the square of energy range; and that practically, it might be the case that only part of the configuration space, hence a subset of all accessible energy levels, are of interest. This suggests that instead of directly applying energy-based methods, which assumes no prior knowledge about the system, we may initially run a temperature-based simulation and then, based on the information we have collected, apply one of the energy-based methods to a reduced range of energy.

In this paper, we report an approach to derive estimates of the density of states from the MBAR estimator. In WHAM, these estimates come out naturally because histograms are used. However, since WHAM solves a self-consistent equation concerning the DoS and free energy, this discretization will introduce error to free energy estimates which in turn causes DoS estimates to be inaccurate. In contrast, MBAR does not require discretization of energy so this error in free energy is removed.

To illustrate how this idea can be applied in practice, we use a combination of parallel tempering (PT) and multicanonical sampling (MUCA) in two examples: a statistical example to demonstrate the correctness of the method, and a lattice-polymer example to demonstrate the efficiency and utility of the method in real physical models.

Use of both PT and MUCA, rather than either method alone, is beneficial to the kind of

model we are interested in. Although parallel tempering is a powerful algorithm to simulate bead-polymer systems, which are often characterized by many local energy minima, it can become inefficient in situations where the system undergoes a phase change that resembles a first-order phase transition. Because there is a steep change in energy, the transition rate between low and high energy states can be low even if the chosen temperature difference is small. In contrast, by sampling from the multicanonical ensemble, the sampler can move freely in energy space because a flat energy histogram will be produced with good estimates of the density of states. As the weights used in multicanonical simulation are *a priori* unknown, they are normally set to be equal to one at the start of the simulation, indicating that the system is started from the disordered state with all configurations equally likely. This means that the sampler may have difficulty sampling low energy configurations whose phase space volume is proportionally smaller than high energy configurations, and so it may take some time to produce "working estimates" of the density of states.

This suggests that there can be merit in using an estimated DoS from PT as the weights of a MUCA simulation. However, it is not straightforward to obtain reliable estimates of DoS given data of a PT simulation due to correlated sampling. A natural approach would be using the WHAM method by following the advice of Ref. 12. In this paper we show that an efficient alternative is to use the MBAR estimator instead.

This paper is organized as follows. Section 2 shows how MBAR can be used to produce optimal estimates of DoS. Section 3 discusses two approaches to applying the idea to parallel tempering simulations. Section 4 demonstrates the method with two models. We also compare both approaches of Section 3 with the common practice of using MBAR directly to analyse simulation data. We conclude with a summary in Section 5.

2 Estimating density of states using MBAR

Suppose independent canonical simulations have been carried out at K temperatures. The probability density function of energy U at inverse temperature β is given by

$$p(U|\beta) = Z(\beta)^{-1}\Omega(U)e^{-\beta U},$$

where $Z(\beta) = \int \Omega(U)e^{-\beta U}dU$ is the configurational partition function and $\Omega(U)$ is the density of states.

Consider a discretization of U in the energy range sampled from the K simulations. Instead of including all energy levels that have been seen, it is possible to ignore those that are close to the high energy end of the spectrum and so are rarely sampled. In this way one can reduce the range of the spectrum to include only interesting system events, e.g. phase transitions, although results associated with the highest temperature distribution are then likely to be inaccurate. Let $\{U_m\}_{m=1}^{M}$ be the midpoints of energy bins and $\hat{\Omega}_{km}$ be the estimate of $\Omega(U_m)$ from the simulation at temperature level k, given by

$$\hat{\Omega}_{km} = Z(\beta_k)e^{\beta_k U_m}H_{km}/(N_k\Delta U), \tag{1}$$

where H_{km} is the histogram count of energy bin m at temperature level k, N_k is the number of samples from simulation k, and ΔU is the bin width. We can write down estimates of $\log\Omega(U_m)$ from each temperature by taking the logarithm of (1):

$$\begin{aligned}
\log\hat{\Omega}_{1m} &= \log Z(\beta_1) + \beta_1 U_m + \log H_{1m} - \log(N\Delta U) \\
\log\hat{\Omega}_{2m} &= \log Z(\beta_1) + \log\frac{Z(\beta_2)}{Z(\beta_1)} + \beta_2 U_m + \log H_{2m} - \log(N\Delta U) \\
&\vdots \\
\log\hat{\Omega}_{Km} &= \log Z(\beta_1) + \log\frac{Z(\beta_K)}{Z(\beta_1)} + \beta_K U_m + \log H_{Km} - \log(N\Delta U).
\end{aligned} \tag{2}$$

Because it is only needed to determine $\log \Omega$ up to an additive constant, we see that the terms $\log Z(\beta_1)$ and $\log(N\Delta U)$ can be ignored, and only $\log Z(\beta_k)/Z(\beta_1),\ k = 2, \ldots, K$ need to be estimated; but these are precisely the dimensionless free energy differences. To best estimate these quantities, we use the Multistate Bennett Acceptance Ratio(MBAR) estimator,[7] which is a generalization of the Bennett Acceptance Ratio(BAR) method[13] to multiple thermodynamic states. MBAR has been proven to be optimal in the sense that it has the lowest variance among all reweighting estimators and is asymptotically unbiased.[14]

Note that there will be K independent estimates of the density of states in (2). Since MBAR also yields uncertainty estimates of the free energy differences, it followes naturally that the estimates $\log \hat{\Omega}_{km}$ should then be weighted inversely proportional to their variances.

3 Working with parallel tempering simulation

The parallel tempering simulation[4,5] simultaneously simulates the system at multiple temperatures that form a temperature ladder. A key step in PT is the exchange of configurations between neighbouring temperatures to speed up the mixing of chains simulated at low temperatures, thus enabling the lowest temperature chain to escape from local energy basins with the help of high temperature chains, whose distributions are more flat.

Because a PT simulation yields data from all temperatures, it is natural to think of applying the method in Section 2. However, the exchange step that makes PT effective also introduces correlations between temperature trajectories. This violates the independence assumption in Section 2, so subsampling is needed to remove the correlation (Section 3.1). In Section 3.2, however, we show that there are situations in which it is justifiable to use this method to estimate the DoS from the full PT dataset.

3.1 Using a subsampled PT trajectory

To deal with the correlation introduced in PT, a reordering of temperature trajectories by so called replicas may be applied if we have recorded the history of temperature swaps. Here each replica contains blocks of configurations sampled at different temperatures and are nearly independent.[12] In doing so, the main contribution to the correlation now comes from within each replica, and this is the correlation that results from correlated sampling in MCMC simulations. Subsampling with a suitable statistical inefficiency $g > 1$ can then be applied to each replica to obtain effectively uncorrelated data. We point out that once g is known, we can use it to subsample the original temperature trajectory because it is equivalent to first subsampling the replicas and then permuting the subsampled replica back by temperature.

Instead of constructing multiple replicas, we use a subsampling strategy suggested by Chodera.[1] We construct a new time series defined by $u_t = \sum_{k=1}^{K} \beta_k U(\mathbf{x}_{kt})$, from each temperature trajectory in the PT simulation, and then use the statistical inefficiency of $\{u_t\}_{t=1}^{N}$ for subsampling. The rationale is that if the reduced potential of a single-temperature simulation provides a practical estimate for the relaxation time of the trajectory, as suggested by Shirts and Chodera, then an extension to the multi-temperature case should be given by the u_t defined above, where the exponential of $-u_t$ effectively gives the overall relative probability of observing a sample in the product state space of the parallel tempering simulation.

3.2 Using a full PT trajectory

We note that the estimating equations of MBAR can still be applied to correlated datasets, but the estimated uncertainties will no longer be valid.[7] Thus if one wishes to report statistical uncertainties of any MBAR estimator, subsampling is required whenever simulation data are correlated. However, there are reasons why using full PT samples may still be an option here. First, our DoS estimates are not used to produce thermodynamic properties; rather,

[1] Personal communication.

they are the weights to be used in the subsequent MUCA simulation. Second, MBAR is not used as a final step to obtain free energy differences or other thermodynamic quantities; it is only used as an intermediate step to estimate the log ratio of partition functions in (2). Evidently, the optimal combination of $\log \Omega$ estimators based on their variances will no longer apply if full PT samples are used. However, if we had subsampled data with long correlations, the resulting subsample size would be small and the uncertainty estimates may still be unreliable. This is because MBAR estimators are derived under the asymptotic limit, and so the estimated standard deviation will only reflect its true value when the sample size is large. Furthermore, a result from statistics[15] states that the variance of the Monte Carlo estimator for the expectation of some function of state, constructed from a subsampled Markov chain, is no smaller than the variance of the estimator constructed from the full Markov chain. An implication of this result is that variance of a full sample MBAR estimator may be smaller than the variance of a subsample MBAR estimator.

From the above observations, a simple alternative to using a subsampled PT trajectory, which also avoids complications associated with computing statistical inefficiency, is to feed into MBAR the full dataset from the PT simulation, ignore uncertainty estimates of the computed free energy differences, and average over the resulting estimates in (2). In the last step, if bin width ΔU is small, then we may only average over those $\log \hat{\Omega}_{km}$ whose corresponding entries in H_{km} are greater than some small integer, say 1 or 2, because the calculation is likely to be unreliable if there is only one observation in the bin.

4 Numerical Study

In this section, we present two examples where PT samples interesting regions of configuration space but suffers from low swap rate in the vicinity of a transition temperature. We apply the method in Section 2 to accurately estimate the density of states from the PT simulation and then run MUCA simulations with those estimated weights. We show

that, for these two examples, the additional MUCA simulation is trivial in the sense that it will produce a more or less flat energy historgram sufficient for analysis, without needing to implement multicanonical recursions as in Ref. 16.

The first example (Section 4.1) is designed to mimic a phase transition by sampling from a mixture of two-dimensional normal distributions with suitably chosen parameters. The advantage of using a statistical model is that exact results can be obtained relatively easily through numerical integration, so that we know whether different simulation methods perform correctly. The second example (Section 4.2), inspired by reality, concerns the simulation of the aggregation of lattice-polymers in an implicit membrane and water environment. Since this model exhibits a phase transition between aggregated and dispersed states, the statistical model may be viewd as a simplified, low-dimensional analog of the physical model, with the benefit of knowing correct solutions.

To facilitate our study, we have used the `pymbar` module[7] for MBAR calculations. The PT and MUCA simulations were run with our own code. Details of the simulation protocol not mentioned in text can be found in Supporting Information.

4.1 A statistical example

Consider a mixture of bivariate normal distributions defined by

$$\pi \sim 0.5N(\boldsymbol{\mu}_1, \Sigma_1) + 0.5N(\boldsymbol{\mu}_2, \Sigma_2),$$

where the mean vector and covariance matrix of the two distributions are given by $\boldsymbol{\mu}_1 = [0,0]^T$, $\Sigma_1 = \text{diag}[0.01, 0.01]$ and $\boldsymbol{\mu}_2 = [2,2]^T$, $\Sigma_2 = \text{diag}[2,2]$. We define the energy function to be $U(\mathbf{x}) = -\log \pi(\mathbf{x})$ and implement a PT sampler that samples from $\pi_T \propto \exp(-U(\mathbf{x})/T)$ with temperatures listed in Table 1. We then followed the procedure described in Section 2 and Section 3 to estimate the density of states from the PT simulation. Specifically, we used both subsampled PT (Section 3.1) and full PT (Section 3.2) to obtain

these estimates. Once this was done, MUCA simulations were run with weights proportional to the inverse of the estimated density of states. In this example, 10 independent MUCA simulations with different initial configurations randomly generated from π were run. We refer the approach that uses subsampled PT as PTMBARMUCA, and the approach that uses full PT as FPTMBARMUCA.

Table 1: Temperatures used and associated PT swap rates observed in the statistical model.

Temperature	T_1	T_2	T_3	T_4
	0.4	0.5	2.0	3.0
Swap rates	0.29	0.01	0.28	

The swap rate between T_2 and T_3 in the PT simulation was observed to be only 1% (Table 1). Because the interval $[T_2, T_3]$ contains $T = 1$, which is when $\pi_T = \pi$, the observed frequency implies that transitions between the two modes in π are rare under the currrent parameter setting.

The parameters of π were chosen to mimic a broad high energy state and a narrow low energy state. A plot of the energy surface is shown in Figure 1. Because we are interested in quantities that vary with temperature, our goal is not just to sample from π, which corresponds to $T = 1$, but to ensure efficient crossing between the two states. This is different from common practice in statistics where one would use a temperature ladder $1 = T_1 < \ldots < T_K$ and only the lowest temperature distribution is of interest. Instead, in our setting the temperatures can be chosen as any positive values, and the relatively large difference between T_2 and T_3 was chosen to test if our method can sustain large gaps in the temperature ladder across a phase transition.

Two properties were calculated: the mean potential energy $\langle U(\mathbf{x})\rangle$ and heat capacity C_{v}. For a given inverse temperature $\beta = 1/T$,

$$\langle U(\mathbf{x})\rangle_T = \frac{-\int \pi(\mathbf{x})^\beta \log \pi(\mathbf{x}) d\mathbf{x}}{\int \pi(\mathbf{x})^\beta d\mathbf{x}},$$

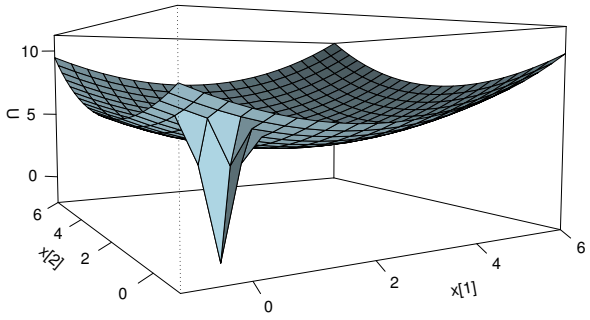


Figure 1: The energy surface in the statistical example: $U(\mathbf{x}) = -\log \pi(\mathbf{x})$.

and

$$(C_{\mathrm{v}})_T = \frac{\langle U^2(\mathbf{x})\rangle_T - \langle U(\mathbf{x})\rangle_T^2}{T^2}.$$

We note that both $\langle U(\mathbf{x})\rangle_T$ and $(C_{\mathrm{v}})_T$ can be calculated through numerical integration, and so correct results are known. The `R` package `cubature` was used to perform the integration. For all integration results, the estimated relative errors were of order 10^{-5}. In Figure 2, we show separately the estimated potential energy and heat capacity across a series of temperatures between $T_1 = 0.4$ and $T_4 = 3.0$ using both the subsampled and full PT trajectory, along with exact integration results. Clearly, the estimates show good agreement with the correct values of potential energy and heat capacity, whether or not subsampling is used.

For comparison, we also used MBAR to obtain directly the estimates as well as uncertainties of thermodynamic quantities of interest, which was the original purpose of MBAR when it was introduced in Ref. 7. We note that although its use for PT simulations was not addressed there, a subsampling strategy as mentioned in Section 3.1 can be applied to

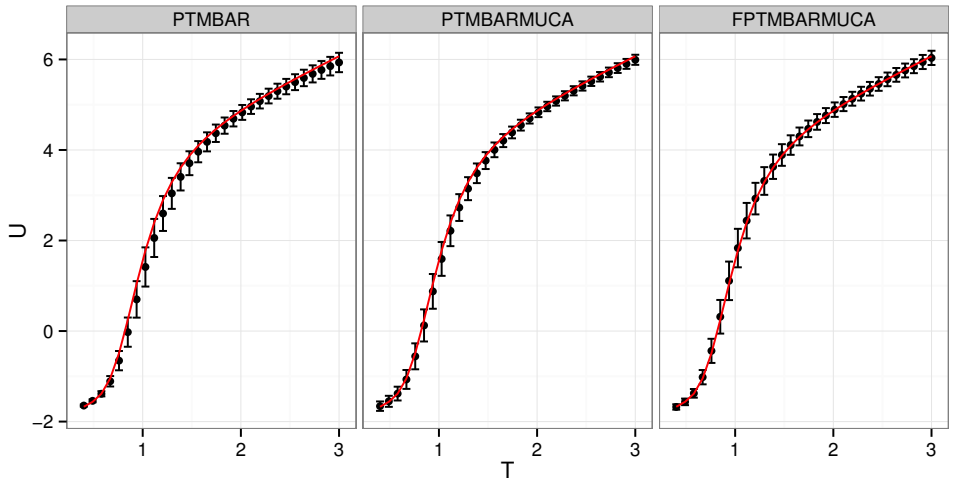
obtain effectively uncorrelated data. This approach will be refered to as PTMBAR.

Since exact numerical integration results are known, a detailed error analysis can be conducted to investigate the quality of estimation for different methods. We used the Mean Squared Error (MSE) as a quality measure. The MSE of an estimator is defined as the average squared deviation between the estimator and its true value, and so takes into account both the variance and the bias of the estimator. To inspect the performance of the methods in different temeprature ranges, we show pictorially the MSE of potential energy (Figure 3a) and heat capacity (Figure 3b) across all temperatures for all methods.

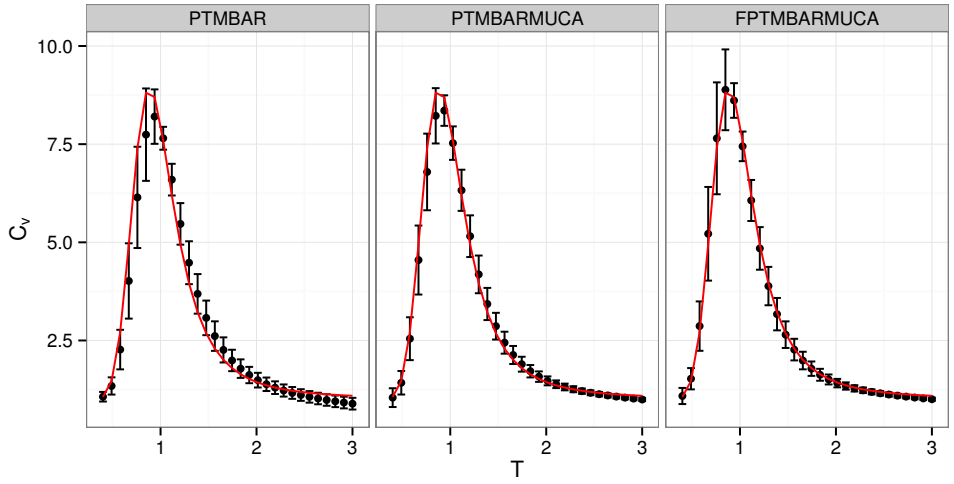
The MSE plots suggest that both PTMBARMUCA and FPTMBARMUCA have consistently smaller MSEs than PTMBAR across almost all temperatures, in particular, the MSEs are significantly smaller when T is around 1. In addition, larger MSE was observed in both energy and heat capacity estimates of PTMBAR when T is close to 3; as is also reflected in the PTMBAR plots in Figure 2, where relatively large error bars are observed near $T = 3$ and, in particular, deviations from the exact values of heat capacity exceed one standard deviation.

4.2 A lattice-polymer study

The work in this paper was originally motivited by a special protein transport mechanism known as the Twin-Arginine Translocation (Tat) pathway.[17] It is involved in the export of proteins in bacteria cytoplasmic membrane and in thylakoid membrane in plants. The Tat mechanism holds promising applications in the area of bioengineering due to the unique twin-arginine motif within a signal peptide that, when attached to a protein, allows the protein to translocate across the membrane without having to unfold. Despite much experimental progress, many aspects of this complex process remain unknown. One key step in Tat is the formation of the translocation pore, a channel which comprises variable amounts of the membrane protein TatA and that can vary its diameter to allow certain types of substrates to be transported.[18] It is therefore essential to understand what triggers the association

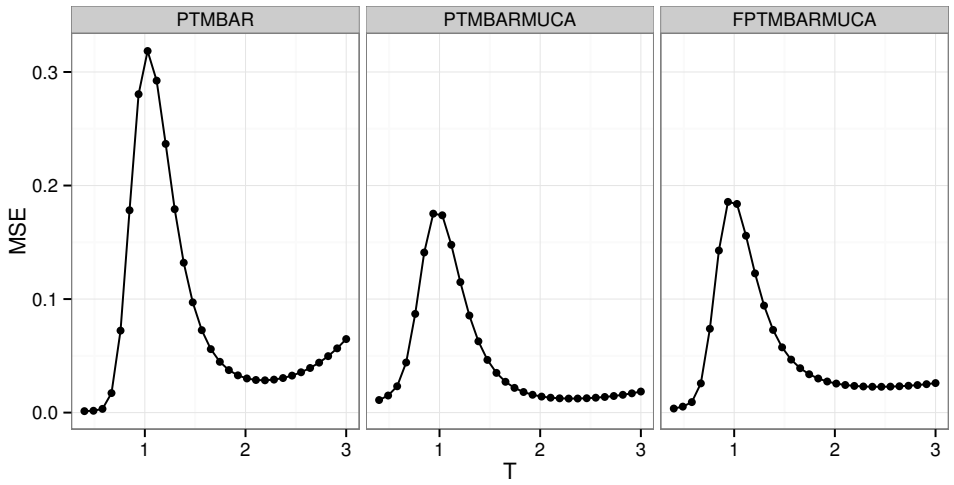


(a)

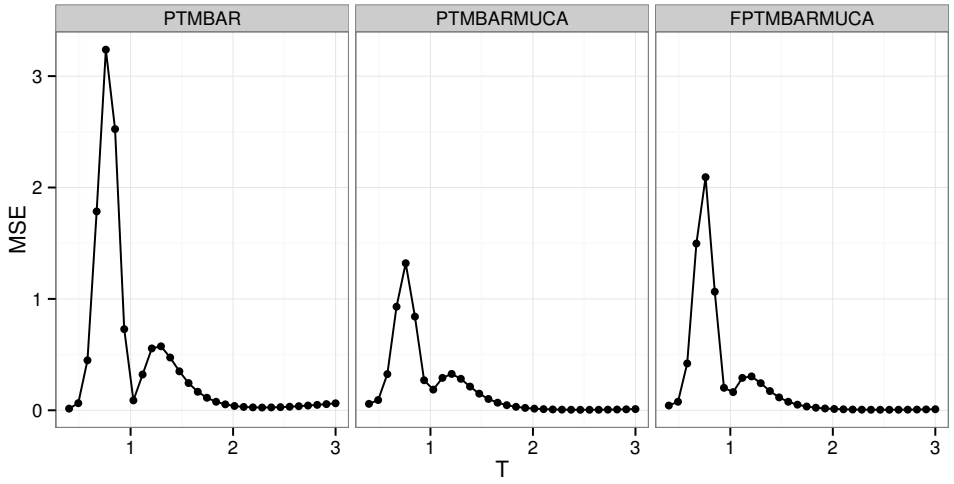


(b)

Figure 2: Estimated potential energy (Figure 2a) and heat capacity (Figure 2b) from three methods. The exact numerical integration results are shown in red curves. In FPTMBAR-MUCA, averages of $\log \hat{\Omega}_{km}$ for each bin m over all temperatures k with $H_{km} > 2$ were used. The error bar was calculated as follows: For PTMBAR, this was the analytical standard deviation of the MBAR estimator; For PTMBARMUCA and FPTMBARMUCA, this was the sample standard deviation of the estimates obtained from the 10 independent MUCA runs.



(a)



(b)

Figure 3: MSE of potential energy (Figure 3a) and heat capacity (Figure 3b) calculated from the corresponding plot in Figure 2. It can be seen that the MSEs of potential energy are comparable between both versions of our new method, although the heat capacity MSE is larger near $T = 0.75$ in FPTMBARMUCA. Comparing the corresponding plots in Figure 2b, however, we see that the bias of the estimates near $T = 0.75$ is smaller in FPTMBARMUCA than in PTMBARMUCA. Both methods have smaller MSE than PTMBAR.

and dissociation of TatA; for simulations, this means that an effective sampling of both aggregated and dispersed states becomes crucial.

Our model is essentially an adaptation of the H-P model.[19] A protein is represented as a sequence of non-overlapping beads on a three-dimensional cubic lattice with periodic boundary, and the type of each bead can be either hydrophobic (H) or hydrophilic (P). The environment is represented by a one-particle lattice site energy term that is either hydrophobic (membrane) or hydrophilic (water). For the current method-validation study we restrict our attention to a simpler lattice polymer which exhibits no secondary structure. Details of the interaction potential can be found in Supporting Information.

Since TatA consists of a transmembrane region and another region that is water-soluble,[20] we have studied two polymers with 34 beads each, the first 12 of which are H-beads and the rest are P-beads. Pull moves[21] and translation moves were used as trial moves. A PT simulation was run with a total of 3×10^7 iterations, the temperature ladder and associated swap rates for this model are listed in Table 2.

Table 2: Temperatures used and associated PT swap rates observed in the lattice-polymer model.

Temperature	T_1	T_2	T_3	T_4	T_5
	0.3	0.48	0.85	1.3	2.0
Swap rates	0.43	0.02	0.54	0.52	

With the potential energy function used in this study, the hydrophobic section of the polymer was observed to remain in the membrane in the temperature range studied whereas the hydrophilic section was able to explore both water and membrane regions. While it is entropically favorable for the polymers to move independently, there is an energetic tendency for the hydrophilic tails of both polymers to interact with each other within the membrane and form a dimer. The equilibrium of the system is a balance between these two driving forces. At high temperatures, the entropy dominates, and at low temperatures, the energy dominates. This monomer-dimer transition was our main interest of investigation. We

mention that, when multiple polymers are included, our model may also be adapted to study aggregate assembly in general protein aggregation processes.

In Section 4.1 we were mainly concerned with the correctness of the methods, and hence used a simple model which could be solved exactly. In this section we turn to the question of efficiency by using a model with more realistic complexity. As a result, while we continue to compare three approaches (PTMBAR, PTMBARMUCA and FPTMBARMUCA), we proceed as follows: we used all 3×10^7 iterations of PT for analysis with PTMBAR, whereas for the latter two approaches only the first 2×10^7 iterations were used to estimate the DoS, and this was then followed by an additional 1×10^7 iterations of MUCA. Hence the computational effort is roughly equal for all three methods.

In order to obtain error estimates, we ran 10 independent MUCA simulations using different initial configurations that belonged to both aggregated and dispersed states. The results of potential energy and heat capacity calculated from the three methods are shown in Figure 4. Since exact values are not known *a priori* in this case, the results are overlayed to check for self-consistency. Uncertainty estimates are also compared for four temperature points which cover the peak region of the heat capacity curve, i.e. the transition between aggregated and dispersed states.

The first point to notice is that energy and heat capacity estimates agree very well for all three methods; in particular, the peak of the heat capacity occurs around $T = 0.66$, and differences between the mean values calculated by the methods are not statistically significant at any of the four temperature points considered. Note that these temperatures were purposely chosen around the transition temperature since property estimates at these temperatures are most likely to suffer from incomplete sampling of the two states. The uncertainty plots suggest that the method of PTMBARMUCA has larger statistical errors than PTMBAR; however, with FPTMBARMUCA, that is, with the DoS estimated using full PT trajectory, we obtain errors that are no worse, and often smaller, than PTMBAR.

Whether or not subsampling should be performed to derive the DoS appears to be affected

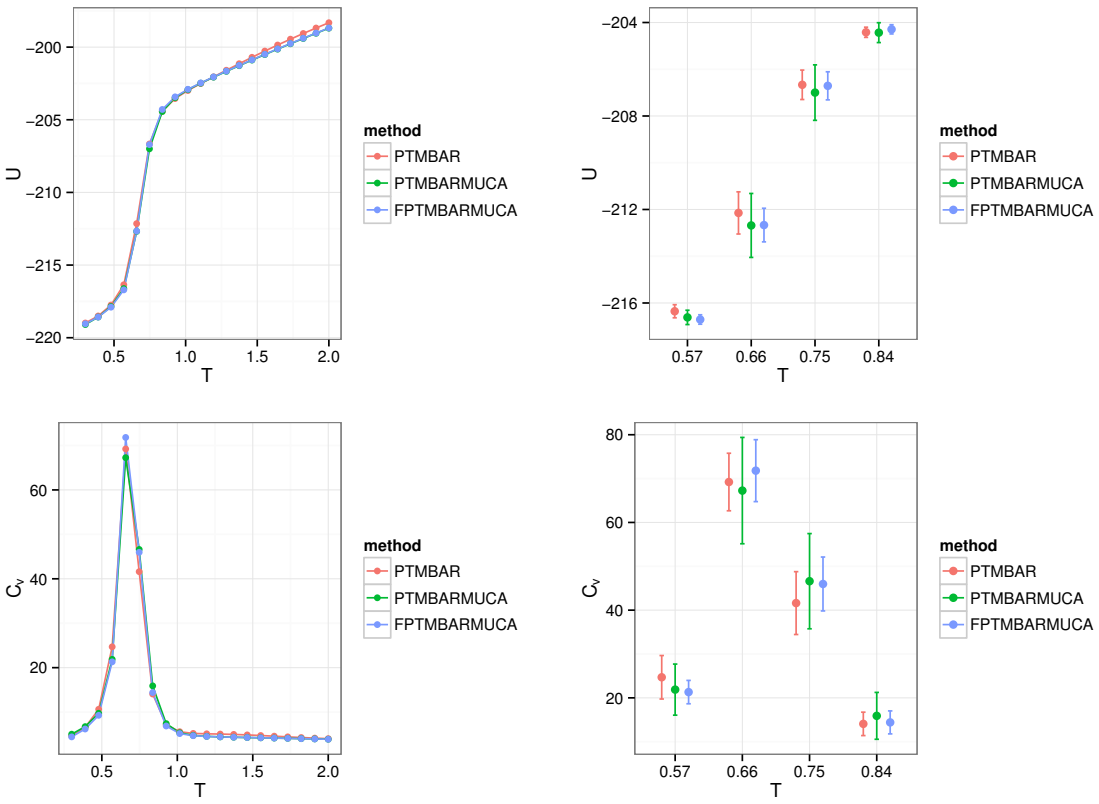


Figure 4: Estimated mean potential energy (top row) and heat capacity (bottom row) along with their uncertainties for the lattice polymer aggregation model. Results are overlayed. The methods with additional MUCA simulations only used half as many PT samples as the method of PTMBAR. In FPTMBARMUCA, an average of the equations in (2) over non-zero entries of H_{km} was used. The plots in the second column show the uncertainties (one standard deviation) of the corresponding property estimates on the left. The error bars were computed in exactly the same way as shown in Figure 2, and they are manually shifted left and right ($\pm\delta$) to avoid obstruction.

by the strength of correlation in the PT trajectory. The statistical inefficiency in the lattice polymer model is about 12 times larger than that in the statistical model of Section 4.1. Although more iterations were used, the polymer model posed a more significant sampling problem because the dimension of the configuration space got much bigger. Hence, for PTMBARMUCA, the subsample size may still be inadequate for obtaining reliable estimates of the uncertainties of the terms in (2). On the other hand, taking averages over the estimates would seem to be a better option here, but would be too arbitrary in the statistical model.

Lastly, we mention that although we used multiple MUCA simulations and calculated standard deviation of the estimates obtained from each run, a resampling technique known as bootstrap[22] can be used to obtain error estimates and replace multiple runs which, however, do serve a useful purpose for convergence check. In bootstrap methods, one generates by sampling with replacement many sets of bootstrapped samples called *resamples* from the original data, and use the distribution of the resamples to approximate the distribution of the population. When we have dependent data, as is the case for most Monte Carlo simulations, the resamples need also preserve the dependence structure and the so called block bootstrap methods[23] can be used. We refer interested readers to Ref. 24 for a comprehensive acount of bootstrap methods.

5 Conclusion

The MBAR estimator exhibits superior statistical properties and has been widely used in free energy calculations invloving multiple equilibrium states. In this paper we proposed an approach that makes use of MBAR to calculate the density of states, and showed how this could be applied to data from parallel tempering simulations. Subsequent MUCA simulations which use this estimated density of states were shown to converge rapidly, without the need for multicanonical recursions. In this way, MBAR "optimally connects" PT and MUCA simulations and constitutes an important and integrated part of the simulation stage, rather

than being confined to its more usual role as a post-simulation analysis tool. Our numerical study of a statistical model showed that the method was formally correct when compared with exact numerical integration results. We then used the method to study polymer aggregation in a lattice model and compared it with the traditional method of using MBAR to analyse simulation data. We observed that even when we applied our method to the first half of generated PT data, we were able to obtain comparable and even better results than the traditional method. Our results therefore suggests that it can be more beneficial and efficient to do analytical calculations, e.g. deriving MUCA weights through MBAR, than simply running longer Monte Carlo simulations.

Because in our method MBAR is not used to produce final estimates, nor the associated uncertainties, of physical properties, there is some leeway in how it can be applied. In particular, it is possible to use full PT samples for MBAR DoS calculations, i.e. without first subsampling to remove the intrinsic correlations in the MC trajectory. This aspect was explored in both of our examples. Clearly, it is natural to subsample the data because we can then properly combine the estimators of the log density of states. However, the optimality of such estimators decreases as subsample size shrinks, and hence if correlation is long, the full sample strategy of Section 3.2 may be preferred. The conventional usage of MBAR to report statistical uncertainties would preclude such possibility.

Acknowledgement

The authors thank the Centre for Scientific Computing at Warwick University for providing the computational resources. We would also like to acknowledge useful discussions with Prof. John D. Chodera, Memorial Sloan-Kettering Cancer Center.

Description of Supporting Information

The SI includes details of the simulation protocols in S4.1 and S4.2. In particular, it contains a full description of the energy function used in S4.2, a table of the associated parameter values chosen for the model and a description of trial moves. The information is available free of charge via the Internet at http://pubs.acs.org

References

(1) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087–1092.

(2) Marinari, E.; Parisi, G. *Europhys. Lett.* **1992**, *19*, 451.

(3) Geyer, C. J.; Thompson, E. A. *J. Am. Stat. Assoc.* **1995**, *90*, 909–920.

(4) Geyer, C. J. Markov Chain Monte Carlo Maximum Likelihood. In Computer Science and Statistics: Proceedings of the 23rd Symposium on the Interface; Keramidas, E. M., Kaufman, S. M., Eds.; Interface Foundation of North America: Fairfax Station, VA, 1991; pp 156-163.

(5) Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604–1608.

(6) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 1011–1021.

(7) Shirts, M. R.; Chodera, J. D. *J. Chem. Phys.* **2008**, *129*, 124105.

(8) Berg, B. A.; Neuhaus, T. *Phys. Lett. B* **1991**, *267*, 249–253.

(9) Wang, F.; Landau, D. P. *Phys. Rev. Lett.* **2001**, *86*, 2050.

(10) Fitzgerald, M.; Picard, R.; Silver, R. *Europhys. Lett.* **1999**, *46*, 282.

(11) Fitzgerald, M.; Picard, R.; Silver, R. *J. Stat. Phys.* **2000**, *98*, 321–345.

(12) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Seok, C.; Dill, K. A. *J. Chem. Theory Comput.* **2007**, *3*, 26–41.

(13) Bennett, C. H. *J. Comput. Phys.* **1976**, *22*, 245–268.

(14) Tan, Z. *J. Am. Stat. Assoc.* **2004**, *99*, 1027–1036.

(15) Robert, C. P.; Casella, G. *Monte Carlo Statistical Methods*, 2nd ed.; Springer: New York, 2004.

(16) Berg, B. A. *Nucl. Phys. B (Proc. Suppl.)* **1998**, *63*, 982–984.

(17) Lee, P.; Tullman-Ercek, D.; Georgiou, G. *Annu. Rev. Microbiol* **2006**, *60:373-395*.

(18) Gohlke, U.; Pullan, L.; McDevitt, C.; Porcelli, I.; de Leeuw, E.; Palmer, T.; Saibil, H.; Berks, B. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 10482–10486".

(19) Lau, K. F.; Dill, K. A. *Macromolecules* **1989**, *22*, 3986–3997.

(20) De Leeuw, E.; Porcelli, I.; Sargent, F.; Palmer, T.; Berks, B. C. *FEBS Lett.* **2001**, *506*, 143–148.

(21) Lesh, N.; Mitzenmacher, M.; Whitesides, S. A Complete and Effective Move Set for Simplified Protein Folding. In Proceedings of the 7th Annual International Conference on Research in Computational Molecular Biology, Berlin, Germany, April 10-13, 2003; ACM Press: New York, 2003.

(22) Efron, B. *Ann. Stat.* **1979**, *7*, 1–26.

(23) Politis, D. N.; Romano, J. P. *J. Am. Stat. Assoc.* **1994**, *89*, 1303–1313.

(24) Davison, A. C.; Hinkley, D. V. *Bootstrap Methods and Their Application*; Cambridge University Press: Cambridge, 1997.

Graphical TOC Entry