Computer Physics Communications 191 (2015) 125–131

Contents lists available at ScienceDirect

Computer Physics Communications

journal homepage: www.elsevier.com/locate/cpc

Microcanonical thermostatistics analysis without histograms: Cumulative distribution and Bayesian approaches

Nelson A. Alves [a], Lucas D. Morero [a], Leandro G. Rizzi [a,b,*]

[a] Departamento de Física, FFCLRP, Universidade de São Paulo, Avenida Bandeirantes, 3900, 14040-901, Ribeirão Preto, SP, Brazil

[b] School of Chemistry, University of Leeds, LS2 9JT, Leeds, UK

ARTICLE INFO

Article history:
Received 10 October 2014
Accepted 17 February 2015
Available online 25 February 2015

Keywords:
Weighted histogram analysis method
ST-WHAM
Microcanonical temperature
Cumulative distribution function
Bayesian analysis

ABSTRACT

Microcanonical thermostatistics analysis has become an important tool to reveal essential aspects of phase transitions in complex systems. An efficient way to estimate the microcanonical inverse temperature $\beta(E)$ and the microcanonical entropy $S(E)$ is achieved with the statistical temperature weighted histogram analysis method (ST-WHAM). The strength of this method lies on its flexibility, as it can be used to analyse data produced by algorithms with generalised sampling weights. However, for any sampling weight, ST-WHAM requires the calculation of derivatives of energy histograms $H(E)$, which leads to non-trivial and tedious binning tasks for models with continuous energy spectrum such as those for biomolecular and colloidal systems. Here, we discuss two alternative methods that avoid the need for such energy binning to obtain continuous estimates for $H(E)$ in order to evaluate $\beta(E)$ by using ST-WHAM: (i) a series expansion to estimate probability densities from the empirical cumulative distribution function (CDF), and (ii) a Bayesian approach to model this CDF. Comparison with a simple linear regression method is also carried out. The performance of these approaches is evaluated considering coarse-grained protein models for folding and peptide aggregation.

1. Introduction

Fundamental aspects of phase transitions in complex systems can be revealed by the analysis of its microcanonical thermostatistics [1,2], which is characterised by the well known entropy $S(E) = k_B \ln \Omega(E)$, where $\Omega(E)$ denotes the density of states of a system with energy E. In particular, the analysis of inflection points of the microcanonical inverse temperature $\beta(E) = dS(E)/dE$ plays an important role in the identification of stable, unstable and metastable regions in the phase diagram [3–5], providing alternative insights to the usual canonical analysis. Also, free-energy profiles can be obtained from the caloric curves β vs. E, from where one can easily evaluate the values of barrier heights and latent heats. In this way, the microcanonical thermostatistics analysis has been incorporated in many studies in the literature, *e.g.* Refs. [6–12] to name a few.

Nevertheless, any analysis must rely on data obtained from efficient exploration of the configurational space. It is well known that numerical simulations performed with Monte Carlo (MC) and molecular dynamics (MD) methods pose limitations to the achievement of reliable data sampling [13]. Such limitations are related to the critical slowing down [14], which is observed in studies of continuous phase transitions, and to the entrapment in local minima, in the case of systems with rugged energy landscapes. In both cases, the configurational space is poorly explored in a reasonable computational simulation time, which may produce biased physical averages. To overcome the trapping problem, it has been suggested that configurations must be sampled using algorithms based on generalised ensembles, where the updates are performed with non-Boltzmann statistical weights $\omega(E)$. For instance, the multicanonical algorithm (MUCA) [15,16], the extended Gaussian ensemble (EGE) [17–19], Tsallis statistical weight [20,21], and replica exchange method (REM) [22] either use a series of Boltzmann weights or any convenient generalised sampling weight [23].

MUCA simulations sample configurations with a weight $\omega_{mu}(E)$ in such a way that the energy distribution is uniform, $H_{mu}(E) \propto \Omega(E)\,\omega_{mu}(E) = constant$. Thus, a precise determination of $\omega_{mu}(E)$ is equivalent to obtain an estimate for the density of states $\Omega(E)$, *i.e.* $\omega_{mu}(E) \propto 1/\Omega(E)$. The weights $\omega_{mu}(E) = \exp[-b(E)E + a(E)]$ follows from the parameterisation of the entropy $S(E) = b(E)E - a(E)$, where $a(E)$and $b(E)$ are the so-called

* Corresponding author at: School of Chemistry, University of Leeds, LS2 9JT, Leeds, UK.

E-mail addresses: alves@ffclrp.usp.br (N.A. Alves), lucas.morero@usp.br (L.D. Morero), lerizzi@usp.br (L.G. Rizzi).

http://dx.doi.org/10.1016/j.cpc.2015.02.010
0010-4655/

multicanonical parameters. The iterative procedure to obtain the MUCA parameters is described in detail in Refs. [15,16], and can be read as,

$$a^n(E_{m-1}) = a^n(E_m) + [b^n(E_{m-1}) - b^n(E_m)]E_m, \quad (1)$$

$$b^n(E_m) = b^{n-1}(E_m) + [\ln H_{mu}^{n-1}(E_{m+1}) - \ln H_{mu}^{n-1}(E_m)]/\varepsilon, \quad (2)$$

where n stands for the nth multicanonical simulation. The recursion steps require the discretisation of the energy for continuous energy models. Therefore, it is convenient to define $E_m = E_0 + m\varepsilon$, where ε is the binsize, m is an integer, and E_0 is a constant that defines a reference energy. All the energies E in the interval $[E_m, E_{m+1}[$ contribute to the histogram $H_{mu}(E_m)$.

Methods to improve sampling based on simulations at different temperatures have been proposed to either be conducted in parallel (REM) or as a random walk between different temperatures. In REM, N_{rep} non-interacting replicas of the system are simultaneously simulated by the usual MC or MD algorithms, and from time to time, pairs of replicas at neighbouring temperatures are exchanged with a transition probability. From the data produced by simulations performed at a single temperature T_1 or at a set of temperatures T_α, with $\alpha = 1, 2, \ldots, N_{\text{rep}}$, it is necessary to employ a reweighing scheme to evaluate physical averages at a given temperature T. Reweighting techniques [24–26] use data from either a single histogram or multiple histograms obtained from MC or MD simulations.

Recently, a simple method called statistical weighted histogram analysis method (ST-WHAM) [27] has been proposed as an iteration-free procedure to obtain an estimate for the microcanonical inverse temperature. In this method the usual WHAM equations [24,25] are converted into a weighted average of the individual densities of states obtained from simulations carried out with different sampling weights $\omega(E)$. From energy histograms produced by multiple simulations, ST-WHAM yields a statistical temperature $\tilde{T}(E) = 1/\tilde{\beta}(E)$, which is an estimate of the inverse microcanonical temperature $\beta(E) = dS(E)/dE$. Interestingly, there is a numerical procedure based on the multicanonical recursion relations (1) and (2), which is called ST-WHAM-MUCA [28], that can be used to replace the direct integration in order to evaluate the entropy $S(E)$. Although both ST-WHAM and ST-WHAM-MUCA have the advantage of *a posteriori* discretisation of energies, their naive implementations may lead to biased evaluations of physical quantities for continuous energy models just like all the aforementioned algorithms.

As described in Section 2, the estimate $\tilde{\beta}(E)$ for inverse microcanonical temperature $\beta(E)$ depends on the derivatives of the energy histograms $H(E)$ (see Eq. (4)). Here, we analyse how the estimates $\tilde{\beta}(E)$ are energy binning dependent and, in Section 3, we present two alternative approaches that avoid the need for energy binning to evaluate the microcanonical caloric curve for continuous energy models: (i) a proposal by Berg and Harris [29], which involves an empirical cumulative distribution (CDF) and uses discrete Fourier series; and (ii) a Bayesian approach [30] to model this CDF with the assumption that the thermodynamic phase transitions are well described by the coexistence of two phases. A comparative analysis between these approaches is made in order to characterise $\beta(E)$ for two systems that undergo first-order-like phase transitions: the folding transition of a coarse-grained protein model for Ubiquitin and the aggregation transition of two heteropolymers that follows a Fibonacci sequence. These examples allow us to describe the statistical and systematic errors involved in the numerical calculations of $H(E)$ and $\tilde{\beta}(E)$, which are presented in Section 4. Conclusions on this comparative analysis are presented in Section 5.

2. Statistical temperature weighted histogram method

The ST-WHAM [27] yields a direct estimate of the inverse microcanonical temperature $\beta(E) = d\ \ln \Omega(E)/dE$ by considering the statistical inverse temperature

$$\tilde{\beta}(E) = \sum_\alpha f_\alpha^*(\beta_\alpha^H + \beta_\alpha^\omega), \quad (3)$$

where $f_\alpha^* = H_\alpha / \sum_\gamma H_\gamma$, $\beta_\alpha^H = d\ \ln H_\alpha/dE$, and $\beta_\alpha^\omega = -d\ \ln \omega_\alpha/dE$. It is preferable to write Eq. (3) as

$$\tilde{\beta}(E) = \frac{1}{\sum_\gamma H_\gamma(E)} \sum_\alpha H_\alpha(E) \left(\frac{d\ \ln H_\alpha(E)}{dE} - \frac{d\ \ln \omega_\alpha(E)}{dE} \right). \quad (4)$$

Note that $\beta_\alpha^\omega = 1/T_\alpha$ for simulations with the canonical weight. With the set of estimates $\tilde{\beta}(E_m)$, MUCA recurrence relations (1) and (2) can be applied to obtain estimates $\tilde{S}(E_m)$ for the microcanonical entropy $S(E_m)$,

$$\tilde{S}(E_m) = \tilde{\beta}(E_m)E_m - a(E_m). \quad (5)$$

This ST-WHAM-MUCA algorithm is quite simple if one has $\tilde{\beta}(E_m)$.

3. Numerical evaluation of derivatives

3.1. Linear regression

We can numerically evaluate the derivatives in Eq. (4) in a naive way, where the derivatives $d\ln H(E)/dE$ at energies E_m follow from a linear regression around this point. For instance, we use a linear regression with $k = 15$ points; selecting k points means that the derivative at E_m is calculated with the values of $H(E_\ell)$, where $\ell = m-(k-1)/2, m+1-(k-1)/2, \ldots, m, \ldots, m+(k-1)/2$. We chose a value for k according to the energy binsize ε. Consequently, we calculate the derivatives in the energy range $\Delta E = (k-1)\varepsilon$. In this method, it is more convenient to directly calculate the derivative of $\ln H(E_m)$ than the derivative of $H(E_m)$. We calculate the linear regression with a subroutine easily adapted from the linear fit subroutine in [31].

3.2. Cumulative distribution method

Another approach can be devised by considering an algorithm based on the cumulative distribution function (CDF) [29]. The advantage of such approach is that it avoids histogramming when describing probability densities $P(E)$, dismissing the need for any *ad hoc* energy discretisation. The method defines an estimator $\tilde{F}(E)$ for the CDF $F(E)$, where the function $\tilde{F}(E)$ is an empirical cumulative distribution function (ECDF) for the probability density $P(E)$. The algorithm sorts the energy time series of length NDAT in an ascending order ($E_1 < E_2 < \cdots < E_{\text{NDAT}}$), so any outliers can be eliminated by constructing a restricted ECDF $\tilde{F}_{ab}(E)$ in the range between two meaningful points a and b (in general one takes $a = E_1$ and $b = E_{\text{NDAT}}$). The basic idea is to propose an approximating function $F_0(E)$ to describe $\tilde{F}_{ab}(E)$, from where the difference function is defined,

$$R(E) = \tilde{F}_{ab}(E) - F_0(E). \quad (6)$$

This function can be expanded in Fourier series,

$$R(E) = \sum_{m=1}^{\text{MMAX}} d(m)\ \sin\left(\frac{m\pi(E-a)}{b-a}\right), \quad (7)$$

which gives the Fourier coefficients [29],

$$d(m) = \sqrt{\frac{2}{b-a}} \int_a^b R(E)\ \sin\left(\frac{m\pi(E-a)}{b-a}\right) dE. \quad (8)$$