Combining the multicanonical ensemble with generative probabilistic models of local biomolecular structure

Jes Frellsen[1,2,†], Thomas Hamelryck[2], Jesper Ferkinghoff-Borg[3]

[1]Department of Engineering, University of Cambridge, United Kingdom
[2]Department of Biology, University of Copenhagen, Denmark
[3]Department of Systems Biology, Technical University of Denmark, Denmark
[†]Corresponding author: Jes Frellsen, e-mail: jf519@cam.ac.uk

Abstract

Markov chain Monte Carlo is a powerful tool for sampling complex systems such as large biomolecular structures. However, the standard Metropolis-Hastings algorithm suffers from a number of deficiencies when applied to systems with rugged free-energy landscapes. Some of these deficiencies can be addressed with the multicanonical ensemble. In this paper we will present two strategies for applying the multicanonical ensemble to distributions constructed from generative probabilistic models of local biomolecular structure. In particular, we will describe how to use the multicanonical ensemble efficiently in conjunction with the reference ratio method.

Keywords: Markov chain Monte Carlo, multicanonical ensemble, generative probabilistic models, biomolecular structure.

1 Introduction

Simulating the folding of biomolecules, such as proteins and RNAs, remains one of the largest open problems in molecular biology. One approach to simulating biomolecules is to use Markov chain Monte Carlo (MCMC). In MCMC based structure simulations, one uses MCMC to sample the state of the system, and subsequently the samples are used for statistical inference on the systems. This includes calculations of key integrals and expectations, which can be used for analyzing the thermodynamics of the system.

Let $P_\beta(\boldsymbol{x})$ be the probability of finding the system in configuration $\boldsymbol{x} \in \mathbf{X}$. We will assume that this distribution can be written on the Boltzmann form

$$P_\beta(\boldsymbol{x}) = \frac{w_\beta(\boldsymbol{x})}{Z_\beta} = \frac{\exp(-\beta\mathcal{E}(\boldsymbol{x}))}{Z_\beta},$$

where β depends on the temperature and $\mathcal{E} : \mathbf{X} \to \mathbb{R}$ is the energy function.

For a biomolecular system, one can normally not sample directly from $P_\beta(\boldsymbol{x})$ and instead one uses the Metropolis-Hastings algorithm (Hastings, 1970). In this algorithm, a sequence of states $\{\boldsymbol{x}_t\}_{t=0}^T$ is generated by sampling a candidate point from a proposal distribution $\boldsymbol{x}' \sim q(\cdot|\boldsymbol{x}_{t-1})$ at time step $t > 0$ and accepting the point as a realization of $P_\beta(\boldsymbol{x})$ with probability

$$\alpha(\boldsymbol{x}'|\boldsymbol{x}_{t-1}) = \min\left(1, \frac{P_\beta(\boldsymbol{x}')q(\boldsymbol{x}_{t-1}|\boldsymbol{x}')}{P_\beta(\boldsymbol{x})q(\boldsymbol{x}'|\boldsymbol{x}_{t-1})}\right).$$

If the point is rejected, the chain stays in the previous state, that is $\boldsymbol{x}_t = \boldsymbol{x}_{t-1}$.

In this paper we will first give a short introduction the multicanonical ensemble and generative probabilistic models of local biomolecular structure. Subsequently, we will discuss two strategies for applying the multicanonical ensemble to distributions constructed from these generative probabilistic models.

1.1 Multicanonical ensemble

For complex systems, such as biomolecular structures, the standard Metropolis-Hasting algorithm suffers from slow convergence and poor mixing[1]. A number of methods has been suggested to address these deficiencies, including the tempering based methods and the multicanonical ensemble, see Iba (2001) or Ferkinghoff-Borg (2012) for a comprehensive survey. The multicanonical ensemble (Berg and Neuhaus, 1991, 1992; Lee, 1993; Berg et al., 1995) works by chaining the target distribution so that the Markov chain mixes better. Using the energy function as a reaction coordinate, the target distribution in the multicanonical ensemble is

$$P_{\text{MUCA}}(\boldsymbol{x}) \propto w_{\text{MUCA}}(\mathcal{E}(\boldsymbol{x})) = \frac{1}{g(\mathcal{E}(\boldsymbol{x}))}\,, \tag{1}$$

where w_{MUCA} is the multicanonical weights and the *density of states* $g : \mathbb{R} \to \mathbb{R}$ is given by

$$g(E) = \int_{\boldsymbol{x}\in\mathbf{X}} \delta(\mathcal{E}(\boldsymbol{x}) - E)\,\mathrm{d}\boldsymbol{x}\,.$$

Here, δ is the Dirac delta function.

In the multicanonical ensemble, samples are drawn from P_{MUCA} instead of the original target distribution of interest P_β. Normally the *density of states* $g(E)$ is inferred iteratively from samples, see for instance Kumar et al. (1992), Berg (1998), Wang and Landau (2001) or Ferkinghoff-Borg (2002). Based on an estimate of the *density of states* $\hat{g}$ and set of samples from P_{MUCA}, the expectation of a function $k : \mathbf{X} \to \mathbb{R}$ under the original Boltzmann distribution can be calculated by

$$\langle k(\boldsymbol{x})\rangle_{P_\beta} = \int_{E\in\mathbb{R}} \langle k(\boldsymbol{x})\rangle_E \hat{P}'_\beta(E)\,\mathrm{d}E\,, \tag{2}$$

where $\hat{P}'_\beta(E) \propto w_\beta(E)\hat{g}(E)$ and $\langle k(\boldsymbol{x})\rangle_E = \int_{\boldsymbol{x}\in\mathbf{X}} k(\boldsymbol{x})\delta(\mathcal{E}(\boldsymbol{x}) - E)\,\mathrm{d}\boldsymbol{x}$. In practice $\langle k(\boldsymbol{x})\rangle_E$ can be calculated as a sample mean.

1.2 Generative probabilistic models

Hamelryck et al. (2006) suggested to use generative probabilistic models (GPM) for describing the conformational space of biomolecular. In later publications full atomic models of the protein backbone (Boomsma et al., 2008) and side chains (Harder et al., 2010) were developed, as well as a full atomic model of RNA conformational space (Frellsen et al., 2009). These models describe a distribution over the dihedral angles in the molecule and the dependencies between the angles on a local length scale along the chain of monomers. We refer to Boomsma et al. (2012) for an elaborate review of these types of models.

As suggested by the authors, these models can be used as proposal distributions for MCMC simulations of biomolecules (Boomsma et al., 2008; Frellsen et al., 2009). By using such an informative well-designed proposal distribution together with an appropriate target distribution, it is expected that the Markov chain will mix better and converges

[1] See the introduction to MCMC by Gilks et al. (1995) for a description of these problems.

faster. This is in particular expected when the GPM also is a factor in the target distribution. In the following we will consider two examples of this.

The GPMs can be used as priors when modeling biomolecular structures based on experimental data. Rieping et al. (2005) formulated biomolecular structure determination as a Bayesian inference problem and denoted their approach inferential structure determination (ISD). Olsson et al. (2011) suggested to use GPMs as priors in the ISD method, and in this case we are interested in the posterior distribution

$$P(\boldsymbol{x}, \boldsymbol{n}|\boldsymbol{d}) = P(\boldsymbol{d}|\boldsymbol{x}, \boldsymbol{n})P(\boldsymbol{n})P_{\text{GPM}}(\boldsymbol{x})\,, \tag{3}$$

where $P(\boldsymbol{d}|\boldsymbol{x}, \boldsymbol{n})$ is the likelihood of the experimental data $\boldsymbol{d}$, $P(\boldsymbol{n})$ is the prior over the data model nuisance parameters $\boldsymbol{n}$ and $P_{\text{GPM}}(\boldsymbol{x})$ is the GPM.

Using the reference ratio formulation (Hamelryck et al., 2010), one can also construct a target distribution by combining the GPM with a distribution over some coarse grained variable $y \in \mathbf{Y}$, where the coarse grained variable $y = m(\boldsymbol{x})$ is a function of the state of the system, $m : \mathbf{X} \to \mathbf{Y}$. As exemplified by Hamelryck et al. (2010), the coarse grained variable could be the radius of gyration of a protein or represent a protein's hydrogen bonding network. For a given distribution over the course grained variable $\tilde{P}(y)$, the reference ratio distribution $P_{\text{RR}}(\boldsymbol{x})$ is the Kullback-Leibler minimal modification to $P_{\text{GPM}}(\boldsymbol{x})$ (Frellsen et al., 2012), such that the distribution over y becomes $\tilde{P}(y)$. Using this notation, the reference ratio distribution is given by

$$P_{\text{RR}}(\boldsymbol{x}) = \frac{\tilde{P}(y)}{\tilde{P}_{\text{GPM}}(y)} P_{\text{GPM}}(\boldsymbol{x})\,, \tag{4}$$

where $\tilde{P}_{\text{GPM}}(y) = \int_{\boldsymbol{x}\in\mathbf{X}} P_{\text{GPM}}(\boldsymbol{x})\delta(m(\boldsymbol{x}) - y)\,\mathrm{d}\boldsymbol{x}$. For details on the reference ratio method, we refer to Hamelryck et al. (2010) and Frellsen et al. (2012).

In both the ISD formulation in equation (3) and the reference ratio method in equation (4), the target distribution has the general form

$$P_f(\boldsymbol{x}) = f(\boldsymbol{x})P_{\text{GPM}}(\boldsymbol{x})\,. \tag{5}$$

For the sake of simplicity, we will leave out the nuisance parameters from ISD. In the following sections, we will discuss two strategies for sampling from a target distribution of this form using the multicanonical ensemble. We call these the explicit strategy and the implicit strategy.

2 Explicit strategy

In the explicit case, the GPM distribution is explicitly included in the energy function. This means that we will write the original distribution of interest as

$$P_f(\boldsymbol{x}) = \exp(-\mathcal{E}_{\text{Ex}}(\boldsymbol{x}))$$

where

$$\mathcal{E}_{\text{Ex}}(\boldsymbol{x}) = -\log(f(\boldsymbol{x})P_{\text{GPM}}(\boldsymbol{x}))\,.$$

In this case it is straightforward to apply the multicanonical ensemble, and from equation (1) we get the multicanonical target distribution

$$P_{\text{Ex}}(\boldsymbol{x}) \propto w_{\text{MUCA}}(\mathcal{E}_{\text{Ex}}(\boldsymbol{x}))\,.$$

When sampling from $P_{\text{Ex}}(\boldsymbol{x})$ using the Metropolis-Hastings algorithm with the GPM as proposal distribution, the acceptance probability is (Boomsma et al., 2012)

$$\alpha_{\text{Ex}}(\boldsymbol{x}'|\boldsymbol{x}) = \min\left(1, \frac{w_{\text{MUCA}}(\mathcal{E}_{\text{Ex}}(\boldsymbol{x}'))}{w_{\text{MUCA}}(\mathcal{E}_{\text{Ex}}(\boldsymbol{x}))} \frac{P_{\text{GPM}}(\boldsymbol{x})}{P_{\text{GPM}}(\boldsymbol{x}')}\right)\,.$$

See Boomsma et al. (2012) for details on how to use a GPM as proposal distribution.

Based on sample from such a simulation and an estimate of the *density of state*, expectation under the original target distribution $P_f(\boldsymbol{x})$ can be calculated directly using equation (2) with $\beta = 1$.

3 Implicit strategy

In the implicit case, the GPM distribution is not included in the energy function and we write the original distribution of interested from equation (5) as

$$P_f(\boldsymbol{x}) = \exp(-\mathcal{E}_{\mathrm{Im}}(\boldsymbol{x}))P_{\mathrm{GPM}}(\boldsymbol{x})$$

where

$$\mathcal{E}_{\mathrm{Im}}(\boldsymbol{x}) = -\log(f(\boldsymbol{x}))\,.$$

In this case we do not apply the multicanonical weights to the GPM, which means that the multicanonical target distribution is

$$P_{\mathrm{Im}}(\boldsymbol{x}) \propto \tilde{w}_{\mathrm{MUCA}}(\mathcal{E}_{\mathrm{Im}}(\boldsymbol{x}))P_{\mathrm{GPM}}(\boldsymbol{x})\,,$$

where $\tilde{w}_{\mathrm{MUCA}}(E) = 1/g_{\mathrm{GPM}}(E)$. This also means that the *density of states* is measured with respect to GPM distribution

$$g_{\mathrm{GPM}}(E) = \int_{\boldsymbol{x}\in\mathbf{X}} P_{\mathrm{GPM}}(\boldsymbol{x})\delta(\mathcal{E}_{\mathrm{Im}}(\boldsymbol{x}) - E)\,\mathrm{d}\boldsymbol{x}\,.$$

This redefinition of the *density of states* corresponds to a change of integration measure equivalent to replacing the uniform distribution with P_{GPM} as the reference distribution, for further details see Ferkinghoff-Borg (2012).

As before, we can use the Metropolis-Hastings algorithm to sample from $P_{\mathrm{Im}}(\boldsymbol{x})$ with the GPM as proposal distribution. Here the acceptance probability is

$$\begin{aligned}\alpha_{\mathrm{Im}}(\boldsymbol{x}'|\boldsymbol{x}) &= \min\left(1, \frac{\tilde{w}_{\mathrm{MUCA}}(\mathcal{E}_{\mathrm{Im}}(\boldsymbol{x}'))P_{\mathrm{GPM}}(\boldsymbol{x}')}{\tilde{w}_{\mathrm{MUCA}}(\mathcal{E}_{\mathrm{Im}}(\boldsymbol{x}))P_{\mathrm{GPM}}(\boldsymbol{x})}\frac{P_{\mathrm{GPM}}(\boldsymbol{x})}{P_{\mathrm{GPM}}(\boldsymbol{x}')}\right)\\ &= \min\left(1, \frac{\tilde{w}_{\mathrm{MUCA}}(\mathcal{E}_{\mathrm{Im}}(\boldsymbol{x}'))}{\tilde{w}_{\mathrm{MUCA}}(\mathcal{E}_{\mathrm{Im}}(\boldsymbol{x}))}\right)\,.\end{aligned}$$

Note, that in this case we do not need to evaluate $P_{\mathrm{GPM}}(\boldsymbol{x})$ in the acceptance probability. In practice, this can speed up the simulation significantly, see Olsson et al. (2011).

In the implicit case, the expectation of function $k : \mathbf{X} \to \mathbb{R}$ under the original distribution can be calculated from a set of samples and an estimate of *density of states* $\hat{g}_{\mathrm{GPM}}(E)$ using an expression similar to equation (2). In this case the expression becomes

$$\langle k(\boldsymbol{x})\rangle_{P_f} = \int_{E\in\mathbb{R}} \langle k(\boldsymbol{x})\rangle_E \hat{P}'_{\mathrm{Im}}(E)\,\mathrm{d}E\,,$$

where $\hat{P}'_{\mathrm{Im}}(E) \propto \exp(-E)\hat{g}_{\mathrm{GPM}}(E)$.

4 Discussion

In this paper we have presented two strategies for combining the multicanonical ensemble with GPMs for simulating the folding biomolecular structures. The explicit strategy is the straightforward application of the multicanonical ensemble, where the both $f(\boldsymbol{x})$ and $P_{\mathrm{GPM}}(\boldsymbol{x})$ are evaluated in each Monte Carlo step. Contrary, in the implicit case we

only need to evaluate $f(\boldsymbol{x})$ in each step, which reduces the required computations and can significantly speed up the simulation in practice. This implicit strategy has successfully been applied to simulations of RNA structures (Frellsen et al., 2009), simulations using the ISD framework (Olsson et al., 2011) and for simulations based on the reference ratio method (Hamelryck et al., 2010). From a computational point, the implicit approach was one of the key elements in the drastic speed-up of the IDS calculation obtained by Olsson et al. (2011).

However, for other proposals than the GPM, the term $P_{\text{GPM}}(\boldsymbol{x})$ has to be evaluated in both the explicit and implicit case. Accordingly, there are no direct computational advantages of the implicit approach for other proposal than the GPM. In that case the main question is which strategy gives rise to the best mixing of the Markov chain? We suggest that this question is investigated in future publications.

References

Berg BA (1998) Algorithmic aspects of multicanonical simulations. Nuclear Physics B – Proceedings Supplements 63: 982–984.

Berg BA, Hansmann UHE, Okamoto Y (1995) Comment on "Monte carlo simulation of a First-Order transition for protein folding". The Journal of Physical Chemistry 99: 2236–2237.

Berg BA, Neuhaus T (1991) Multicanonical algorithms for first order phase transitions. Physics Letters B 267: 249–253.

Berg BA, Neuhaus T (1992) Multicanonical ensemble: A new approach to simulate first-order phase transitions. Physical Review Letters 68: 9.

Boomsma W, Frellsen J, Hamelryck T (2012) Probabilistic models of local biomolecular structure and their applications. In Hamelryck T, Mardia K, Ferkinghoff-Borg J, editors, Bayesian Methods in Structural Bioinformatics, Statistics for Biology and Health, Springer Berlin Heidelberg. pp. 233–254.

Boomsma W, Mardia KV, Taylor CC, Ferkinghoff-Borg J, Krogh A, Hamelryck T (2008) A generative, probabilistic model of local protein structure. Proceedings of the National Academy of Sciences of the United States of America 105: 8932–8937.

Ferkinghoff-Borg J (2002) Optimized Monte Carlo analysis for generalized ensembles. The European Physical Journal B 29: 481–484.

Ferkinghoff-Borg J (2012) Monte carlo methods for inference in high-dimensional systems. In Hamelryck T, Mardia K, Ferkinghoff-Borg J, editors, Bayesian Methods in Structural Bioinformatics, Statistics for Biology and Health, Springer Berlin Heidelberg. pp. 49–93.

Frellsen J, Mardia KV, Borg M, Ferkinghoff-Borg J, Hamelryck T (2012) Towards a general probabilistic model of protein structure: The reference ratio method. In Hamelryck T, Mardia K, Ferkinghoff-Borg J, editors, Bayesian Methods in Structural Bioinformatics, Statistics for Biology and Health, Springer Berlin Heidelberg. pp. 125–134.

Frellsen J, Moltke I, Thiim M, Mardia KV, Ferkinghoff-Borg J, Hamelryck T (2009) A probabilistic model of RNA conformational space. PLoS Computational Biology 5: e1000406.

Gilks WR, Richardson S, Spiegelhalter D (1995) Introducing Markov chain Monte Carlo. In Gilks WR, Richardson S, Spiegelhalter D, editors, Markov Chain Monte Carlo in Practice, Interdisciplinary Statistics, Chapman & Hall/CRC.

Hamelryck T, Borg M, Paluszewski M, Paulsen J, Frellsen J, Andreetta C, Boomsma W, Bottaro S, Ferkinghoff-Borg J (2010) Potentials of mean force for protein structure prediction vindicated, formalized and generalized. PLoS ONE 5: e13714.

Hamelryck T, Kent JT, Krogh A (2006) Sampling realistic protein conformations using local structural bias. PLoS Computational Biology 2: e131.

Harder T, Boomsma W, Paluszewski M, Frellsen J, Johansson K, Hamelryck T (2010) Beyond rotamers: a generative, probabilistic model of side chains in proteins. BMC Bioinformatics 11: 306.

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57: 97–109.

Iba Y (2001) Extended ensemble Monte Carlo. International Journal of Modern Physics C 12: 623.

Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PA (1992) The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. Journal of Computational Chemistry 13: 1011–1021.

Lee J (1993) New Monte Carlo algorithm: Entropic sampling. Physical Review Letters 71: 211.

Olsson S, Boomsma W, Frellsen J, Bottaro S, Harder T, Ferkinghoff-Borg J, Hamelryck T (2011) Generative probabilistic models extend the scope of inferential structure determination. Journal of Magnetic Resonance 213: 182–186.

Rieping W, Habeck M, Nilges M (2005) Inferential structure determination. Science 309: 303–306.

Wang F, Landau DP (2001) Efficient, Multiple-Range random walk algorithm to calculate the density of states. Physical Review Letters 86: 2050.