

# Direct Reweighting Strategies in Conformation Dynamics

Martina Klimm, Alexander Bujotzek, Marcus Weber

Zuse Institute Berlin, Takustraße 7, D-14195 Berlin, Germany

e-mail: klimm@zib.de, bujotzek@zib.de, weber@zib.de

(Received April 27, 2010)

## Abstract

The aim of conformation dynamics is the identification of metastable sets in molecular conformation space and the computation of their statistical weights. In this article, we present two novel strategies that permit a direct calculation of the statistical weights, both based on the estimation of free energy differences. The first approach requires an inverse balance condition that leads to a direct calculation of the statistical weights as eigenvalues of a transition matrix. The second approach approximates entropy differences between metastable sets, which is again used to derive according free energy differences and, hence, the statistical weights. In conclusion, we present two examples that illustrate the applicability of these strategies.

## 1 Introduction

During the last years, many efforts have been made to find an efficient method to calculate free energy differences [1] for molecular systems defined in terms of classical mechanics and with given potential energy function. However, as yet there is no algorithm that is able to calculate the weighting without producing excess data, approximating functions or knowing the shape of the target function in advance.

Considering a molecular system consisting of  $N$  atoms, the probability of finding a state  $x = (q, p) \in (\Omega, \Gamma) \subset \mathbb{R}^{6N}$  in a canonical ensemble is given by the Boltzmann distribution

$$\pi(q, p) = \frac{\exp(-\beta H(q, p))}{\int_{\Omega} \int_{\Gamma} \exp(-\beta H(\tilde{q}, \tilde{p})) \, d\tilde{x}}, \quad (1)$$

with the thermodynamic beta  $\beta = 1/(k_B T)$ , the Boltzmann constant  $k_B$ , the temperature  $T$ , the differential  $d\bar{x} = d\bar{q} d\bar{p}$  and the total energy  $H(q, p) = E_k(p) + E_p(q)$  as the sum of the kinetic energy  $E_k(p)$  and the potential energy  $E_p(q)$ . Since  $H$  is obviously separable, it yields

$$\begin{aligned} \pi(q, p) &= \frac{\exp(-\beta E_k(p))}{\int_{\Gamma} \exp(-\beta E_k(\bar{p})) d\bar{p}} \cdot \frac{\exp(-\beta E_p(q))}{\int_{\Omega} \exp(-\beta E_p(\bar{q})) d\bar{q}} \\ &= \pi_p(p) \cdot \pi_q(q) \end{aligned}$$

While the first term is a rather simple quadratic function, the second term is too complex to be computed analytically. In the following, we want to consider free energy differences in order to approximate this term.

The free energy  $A$  of a set  $B \subset \Omega$  is defined as

$$A = -\frac{1}{\beta} \ln \left( \int_B \exp(-\beta H(q, p)) dx \right). \quad (2)$$

Due to the fact that the potential energy is determinable only except for a constant and, hence, the free energy as well, we are going to use the difference between two sets, where the unknown factor cancels out. Assuming that  $\Omega = \bigcup_k B_k$ , the free energy difference between sets  $B_i$  and  $B_j$  can be computed by

$$\begin{aligned} \Delta A &= -\frac{1}{\beta} \ln \left( \int_{B_i} \exp(-\beta H(q, p)) dx \right) + \frac{1}{\beta} \ln \left( \int_{B_j} \exp(-\beta H(q, p)) dx \right) \\ &= -\frac{1}{\beta} \ln \left( \frac{\int_{B_j} \exp(-\beta H(q, p)) dx}{\int_{B_i} \exp(-\beta H(q, p)) dx} \right) \stackrel{(*)}{=} -\frac{1}{\beta} \ln \left( \frac{\int_{B_j} \exp(-\beta E_p(q)) dq}{\int_{B_i} \exp(-\beta E_p(q)) dq} \right) \\ &= -\frac{1}{\beta} \ln \left( \frac{\int_{B_j} \frac{\exp(-\beta E_p(q))}{\int_{\Omega} \exp(-\beta E_p(\bar{q})) d\bar{q}} dq}{\int_{B_i} \frac{\exp(-\beta E_p(q))}{\int_{\Omega} \exp(-\beta E_p(\bar{q})) d\bar{q}} dq} \right) = -\frac{1}{\beta} \ln \left( \frac{w_j}{w_i} \right). \end{aligned} \quad (3)$$

In the process, we benefit from the fact that the kinetic energy  $E_k(p)$  is equal in each set (compare step (\*)). The weightings  $w_k = \int_{B_k} \exp(-\beta E_p(q)) \left[ \int_{\Omega} \exp(-\beta E_p(\bar{q})) d\bar{q} \right]^{-1} dq$ , which are indicating the probability to be in set  $B_k$ , are exponential and, hence, rather difficult to calculate.

As long-running deterministic dynamical systems are often chaotic and, consequently, do not converge to the wanted distribution (1), stochastic approaches were developed.

## 2 New Reweighting Strategies

In this section we want to present two new reweighting strategies. Both are based on a decomposition of the state space  $\Omega$  into non-intersecting sets  $B_k$ ,  $k = 1, \dots, b$ , with  $\Omega = \bigcup_{k=1}^b B_k$ . Let us assume that inside these subsets  $B_k$ , separate samplings have generated  $\tilde{n}_k$  sampling points  $q_l^{(k)}$ ,  $l = 1, \dots, \tilde{n}_k$ . This can be achieved by various dynamical models, i. e. the Hybrid Monte Carlo method [2], Nosé-Hoover dynamics [3], the Berendsen thermostat [4], Smoluchowski dynamics [5], Langevin dynamics [6], or others [7]. A restriction of the sampling to the according set  $B_k$  can be achieved by constraining the proposal step to  $B_k$ , or by rejecting any proposed state that does not belong to  $B_k$ . By doing so, local approximations of the Boltzmann distribution

$$\pi_q^{(k)} = \frac{\exp(-\beta E_p(q))}{\int_{B_k} \exp(-\beta E_p(\tilde{q})) d\tilde{q}}, \quad k = 1, \dots, b, \quad (4)$$

on the basis of the decomposition are obtained. Based on these separate samplings, we want to establish two reweighting strategies named Direct Free Energy Reweighting (DFER) and Estimation of Entropy Differences Reweighting (EEDR). By reweighting the local distributions obtained from the separate samplings, we intend to obtain a good approximation of the global Boltzmann distribution.

### 2.1 Direct Free Energy Reweighting

In contrast to the methods for the calculation of free energy differences presented in [1], we present a direct ansatz that is feasible without producing excess data. In the first paragraph, we want to describe the two-phase virtual sampling scheme of this approach.

*Jump phase.* Starting in state  $q_i^{(k)} \in B_k$ , a proposal step is made by selecting a target set  $B_l$  at random with uniform distribution in  $l \in \{1, \dots, b\}$  and by choosing a state  $q_j^{(l)} \in B_l$  out of the  $\tilde{n}_l$  sampling points at random with uniform distribution in  $j \in \{1, \dots, \tilde{n}_l\}$ . The proposed state is accepted with probability

$$P_{ij}^{(a)} = \min \left\{ 1, \frac{\exp(-\beta E_p(q_i^{(k)}))}{\exp(-\beta E_p(q_j^{(l)}))} \right\}, \quad (5)$$

otherwise the previous sampling point  $q_i^{(k)} \in B_k$  is retained. Note that the fraction in the acceptance probability is inverted with regard to the standard Metropolis. This leads to the *inverse balance condition*, to which we will come back later on.

*Evaluation phase.* The second phase of the virtual sampling scheme consists of the selection of evaluation regions  $\mathcal{E}_k \subset B_k$  of equal volume. A small subset of each set  $B_k$  is chosen where the sampling of the jump phase is rather dense, e. g. near a local minimum of the potential energy function  $E_p$ , or within a region of average potential energy. Finally, the statistical weight of set  $B_k$  can be estimated by the inverse of the number of states  $n_k$  in set  $\mathcal{E}_k$ , i. e.  $w_k \propto n_k^{-1}$ .

In the following, we want to show that the virtual sampling scheme in fact leads to an estimation of free energy differences between the sets  $B_k$ . The conditional transition probability  $P_{ij}$  from state  $q_i^{(k)} \in B_k$  to state  $q_j^{(l)} \in B_l$  is defined as a product of the proposition probability  $P_{ij}^{(p)}$  and the acceptance probability  $P_{ij}^{(a)}$ . As separate samplings correspond to local Boltzmann distributions (4), the proposal transition probability yields

$$\frac{P_{ij}^{(p)}}{P_{ji}^{(p)}} = \frac{\pi_q^{(l)}(q_j^{(l)})}{\pi_q^{(k)}(q_i^{(k)})} = \frac{\exp(-\beta E_p(q_j^{(l)}))}{\exp(-\beta E_p(q_i^{(k)}))} \cdot \frac{w_k^{(i)}}{w_l^{(j)}}.$$

The acceptance transition probability is constructed analogously to equation (5) of the jump phase

$$\frac{P_{ij}^{(a)}}{P_{ji}^{(a)}} = \frac{\exp(-\beta E_p(q_i^{(k)}))}{\exp(-\beta E_p(q_j^{(l)}))}.$$

Hence, the overall transition probability matrix  $P$  results in

$$\frac{P_{ij}}{P_{ji}} = \frac{w_k^{(i)}}{w_l^{(j)}}. \tag{6}$$

As this is the reciprocal of the standard balance condition, we call it the *inverse balance condition*.

A sampling according to this transition probability matrix is generating a sequence of sampling points whose distribution converges to the dominant left eigenvector  $\gamma$  of  $P$  to the eigenvalue 1, see [8], i. e.

$$\gamma^T P = \gamma^T.$$

It is easy to check that  $P$  is irreducible and positive, which is necessary for the uniqueness of  $\gamma$  because of the theorem of Frobenius and Perron [9]. By using equation (6) the eigenvalue problem can be written as

$$\sum_i \gamma_i \frac{w_k^{(i)}}{w_l^{(j)}} P_{ji} = \gamma_j \quad \forall j.$$

Due to the row-stochastic characteristic of  $P$ , i. e.  $\sum_i P_{ji} = 1 \quad \forall j$ , we obtain

$$\sum_i \left( \gamma_i \frac{w_k^{(i)}}{w_l^{(j)}} - \gamma_j \right) P_{ji} = 0 \quad \forall j \text{ and } \forall P_{ji} \in [0, 1].$$

The left eigenvector to the eigenvalue 1 of  $P$  results in  $\gamma_i \propto (w_k^{(i)})^{-1}$ . Due to this relation and the equal volume of the evaluation regions, the quantity  $n_k$  of all sampling points of the jump phase that are elements of the evaluation region  $\mathcal{E}_k$  is proportional to  $w_k^{-1}$ .

As trapping is the main reason for an ill-conditioned computation of the global distribution of states [10], we want to offer a modification of the above virtual sampling scheme that lessens trapping effects. To this end, we make use of the disaggregation method based on the ideas of Simon and Ando [11]. As points outside the evaluation regions typically have higher potential energy values than those inside, sampling in the jump phase may be trapped outside the evaluation regions. This applies especially when outlier states with very high potential energy values occur. A trapped sampling may yield only a very small number of states in the evaluation regions, which consequently would lead to broken statistics. The limit distribution of the jump sampling is given by the dominant left eigenvector  $\gamma$  of a stochastic matrix  $P$ . As we are only interested in the part of  $\gamma$  that corresponds to the evaluation regions, we intend to compute the required components of the vector only. This is realized by rejecting states that are outside of the evaluation regions during the jump phase. In this context, rejection means repetition of the current state in the Markov chain if the proposed state is not part of the evaluation region. Assuming that  $P$  is the original transition probability matrix, the disaggregated transition probability matrix  $\tilde{P}$  looks as follows:

$$\tilde{P}_{ij} = \begin{cases} P_{ij}, & \text{if } i \neq j \\ P_{ii} + \sum_{l: q^{(l)} \notin \mathcal{E}_k} P_{il}, & \text{if } i = j \end{cases}$$

With the identity  $\gamma_i P_{ij} = \gamma_j P_{ji}$ , one can easily show that  $\tilde{P}$  has the same left eigenvector as  $P$ , but is restricted to the components that correspond to evaluation regions.

In practice the algorithm is much shorter, because we can compute the statistical weights  $w_k$  directly and, hence, do not need the transition probability matrix  $P$ .

- (1) Define an evaluation region  $\mathcal{E}_k \subset B_k$  where the sampling is rather dense, e. g. near the minimum, the mean value or the median of the potential energy  $E_p$  in state space  $B_k$ . Keep in mind that all evaluation regions have to be of the same size.
- (2) Compute the statistical weights as

$$w_k \propto \left( \frac{1}{\tilde{n}_k} \sum_{q_i^{(k)} \in \mathcal{E}_k} \exp(\beta E_p(q_i^{(k)})) \right)^{-1}. \quad (7)$$

- (3) Finally, normalise the statistical weights to

$$\sum_{k=1}^b w_k = 1.$$

We tested this method on two examples that are presented in section 3.

In order to avoid the aforementioned trapping problems, we also want to offer a modification for the direct computation of the statistical weights. Typically, the energy levels of the subsets  $B_k$  are varying, i. e. the mean potential energy values  $E_p^{(k)} = \langle E_p^{(k)} \rangle$  differ. As a consequence, sets  $B_k$  with high potential energy levels can lead to a trapping of the sampling (compare (5)). As we are aiming at an equilibrated sampling among the different sets, we shift the potential energy functions according to their mean values, i. e.  $\tilde{E}_p(q_i^{(k)}) = E_p(q_i^{(k)}) - E_p^{(k)}$  with  $q_i^{(k)} \in B_k$ . Therefore, we have to rewrite expression (7) as

$$w_k \propto \left( \frac{1}{\tilde{n}_k} \sum_{q_i^{(k)} \in \mathcal{E}_k} \exp(\beta \tilde{E}_p(q_i^{(k)})) \right)^{-1} \cdot \exp(-\beta E_p^{(k)}).$$

That way, a set  $B_k$  with high mean potential energy  $E_p^{(k)}$  is assigned a reduced statistical weight  $w_k$ .

## 2.2 Estimation of Entropy Differences Reweighting

A simple method for the estimation of entropy differences has been developed by Weber and Andrae [12]. They applied their method to differentiate between varying polymer structures

from thermostated molecular dynamics simulations. We want to modify this approach in order to use it for the calculation of statistical weights.

The entropy of a system is a measure of the distribution's "disorder" or multiplicity. In the context of molecular conformation space, a low entropy system is rather uniform, i. e. sampled conformations look similar, whereas in a high entropy system, sampled conformations are of high variety. In the following, we are interested in the entropy difference

$$\Delta S = \frac{\Delta U - \Delta A}{T}, \quad (8)$$

which is defined in terms of the thermodynamic state functions internal energy  $U$  and free energy  $A$ . The internal energy difference between two systems with potential energy  $E_p^{(i)} : B_i \rightarrow \mathbb{R}$  and  $E_p^{(j)} : B_j \rightarrow \mathbb{R}$  is

$$\begin{aligned} \Delta U &= \frac{\int_{B_j} E_p^{(j)}(q) \exp(-\beta E_p^{(j)}(q)) \, dq}{\int_{B_j} \exp(-\beta E_p^{(j)}(q)) \, dq} - \frac{\int_{B_i} E_p^{(i)}(q) \exp(-\beta E_p^{(i)}(q)) \, dq}{\int_{B_i} \exp(-\beta E_p^{(i)}(q)) \, dq} \\ &= \langle E_p^{(j)} \rangle - \langle E_p^{(i)} \rangle. \end{aligned} \quad (9)$$

In practice, the value of the internal energy difference can be approximated by the mean potential energy value over a Boltzmann distributed sampling of position states.

The free energy  $A$  and its difference  $\Delta A$  have already been defined in equations (2) and (3). In order to get a numerical approximation of the free energy difference, we want to use a special Monte Carlo quadrature approach of an integrand  $f : \Omega \rightarrow \mathbb{R}^+$ , see [12],

$$I = \int_{\Omega} f(x) \, dx \approx f(\bar{x}) \, \text{vol}(\Omega) \frac{\mathcal{U}_1(\bar{x})}{\mathcal{U}_{\rho}(\bar{x})},$$

where  $\mathcal{U}_1(\bar{x})$  is the number of equally distributed states in an environment of  $\bar{x}$ , and  $\mathcal{U}_{\rho}(\bar{x})$  is the quantity of  $\rho$ -distributed states with  $\rho \propto f$ . Hence, the free energy difference can be written

as

$$\begin{aligned} \Delta A &= -\frac{1}{\beta} \ln \left( \frac{\int_{B_j} \exp(-\beta E_p^{(j)}(q)) \, dq}{\int_{B_i} \exp(-\beta E_p^{(i)}(q)) \, dq} \right) \\ &\approx -\frac{1}{\beta} \ln \left( \frac{\exp(-\beta E_p^{(j)}(q^{(j)})) \operatorname{vol}(B_j) \frac{\mathcal{U}_1^{(j)}(q^{(j)})}{\mathcal{U}_{\pi_q}^{(j)}(q^{(j)})}}{\exp(-\beta E_p^{(i)}(q^{(i)})) \operatorname{vol}(B_i) \frac{\mathcal{U}_1^{(i)}(q^{(i)})}{\mathcal{U}_{\pi_q}^{(i)}(q^{(i)})}} \right). \end{aligned}$$

Requiring a comparable structure of the position spaces  $B_i$  and  $B_j$ , i. e.  $\operatorname{vol}(B_i) \mathcal{U}_1^{(i)}(q^{(i)}) = \operatorname{vol}(B_j) \mathcal{U}_1^{(j)}(q^{(j)})$ , the estimation of the free energy difference simplifies to

$$\Delta A \approx -\frac{1}{\beta} \ln \left( \frac{\mathcal{U}_{\pi_q}^{(i)}(q^{(i)})}{\mathcal{U}_{\pi_q}^{(j)}(q^{(j)})} \right) + E_p^{(j)}(q^{(j)}) - E_p^{(i)}(q^{(i)}). \quad (10)$$

Using equations (9) and (10), the entropy difference can be written as

$$\Delta S \approx \frac{[\langle E_p^{(j)} \rangle - E_p^{(j)}(q^{(j)})] - [\langle E_p^{(i)} \rangle - E_p^{(i)}(q^{(i)})]}{T} + k_B \ln \left( \frac{\mathcal{U}_{\pi_q}^{(i)}(q^{(i)})}{\mathcal{U}_{\pi_q}^{(j)}(q^{(j)})} \right).$$

Considering representatives  $q^{(i)}$ ,  $q^{(j)}$  with mean potential energy value, i. e.  $E_p^{(i)}(q^{(i)}) = \langle E_p^{(i)} \rangle$  and  $E_p^{(j)}(q^{(j)}) = \langle E_p^{(j)} \rangle$ , respectively, the approximation of the entropy difference can be reduced to

$$\Delta S \approx k_B \ln \left( \frac{\mathcal{U}_{\pi_q}^{(i)}(q^{(i)})}{\mathcal{U}_{\pi_q}^{(j)}(q^{(j)})} \right).$$

As a single state might not provide a good approximation, we want to take the mean of several representatives with approximately mean potential energy value. Assuming that the number of the different representatives is equal in each set, that is  $R_i = R_j = R$ , the final approximation yields

$$\Delta S \approx k_B \ln \left( \frac{\sum_{l=1}^R (\mathcal{U}_{\pi_q}^{(j)}(q_l^{(j)}))^{-1}}{\sum_{l=1}^R (\mathcal{U}_{\pi_q}^{(i)}(q_l^{(i)}))^{-1}} \right).$$

Alternatively, instead of the arithmetic mean, the median can be used. Being able to estimate the entropy difference between two position spaces  $B_i$  and  $B_j$ , we can apply relation (8) to



determine the free energy difference  $\Delta A$  and, using equation (3), the statistical weights  $w_i$  and  $w_j$ .

For a better comprehension of the algorithm, we want to outline it shortly.

- (1) In subset  $B_i$  of position space  $\Omega$ , choose  $R$  reference points  $q_l^{(i)}$ ,  $l = 1, \dots, R$ , with approximately mean potential energy value.
- (2) Using some distance measure, count the number  $n_p^{(i)}$  of sampling points that are *near*, i. e. within a certain distance around each reference point  $q_l^{(i)}$ , and compute its inverse

$$\left(\mathcal{U}_{\pi_q}^{(i)}(q_l^{(i)})\right)^{-1} \approx \frac{1}{n_p^{(i)}/\tilde{n}_i + 1} =: \frac{1}{n_l^{(i)} + 1}$$

with  $\tilde{n}_i$  being the total number of sampling points in subset  $B_i$ .

- (3) Compute the entropy

$$S_i = k_B \ln \left( \frac{1}{R} \sum_{l=1}^R \left(\mathcal{U}_{\pi_q}^{(i)}(q_l^{(i)})\right)^{-1} \right),$$

the free energy

$$A_i = \langle E_p^{(i)} \rangle - T S_i$$

and the statistical weights

$$w_i = w_{i-1} \cdot \exp(-\beta (A_i - A_{i-1}))$$

with  $w_1 = 1$ . If necessary, the free energy values have to be ordered by size before calculating the statistical weights.

- (4) Finally, the statistical weights have to be normalised to

$$\sum_{i=1}^b w_i = 1.$$

### 3 Application

After having derived and explained new strategies for calculating statistical weights on the basis of separate samplings, we want to illustrate the methods with the help of two examples.

### 3.1 An abstract example: A tetrahedral molecule

The first example is meant to be a simple model, as it is rather abstract and not to be found in practice. Assume a molecule consisting of four atoms is arranged like a regular tetrahedron, see figure 1.

Vertices  $Q_i = (Q_{ix}, Q_{iy}, Q_{iz})^T$ ,  $i = 1, \dots, 4$ , represent the atoms and edges  $r_{ij} = \|Q_i - Q_j\|$ ,

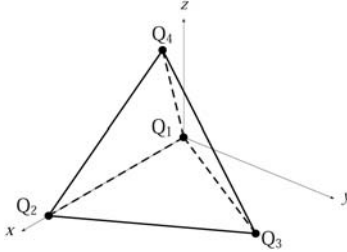


Figure 1: A regular tetrahedron

$i \neq j$ , form the bonds. Using generalized coordinates  $q = (Q_1^T, Q_2^T, Q_3^T, Q_4^T)^T \in \Omega \subset \mathbb{R}^{12}$  and, analogously, generalized momenta  $p \in \mathbb{R}^{12}$ , we can define the energy functions

$$E_k(p) = \frac{1}{2} p^T M^{-1} p$$

$$E_p(q) = \sum_{i=1}^3 \sum_{j=i+1}^4 k (r_{ij}(q) - r)^2$$

with force constant  $k$  and bond length  $r$ . Note that this is an extremely simplified potential energy function, which in practice would be a sum of different energy terms describing the various intra- and intermolecular forces acting on the molecule.

Obviously, this artificial potential energy function is minimal for  $r_{ij} = r \forall i \neq j$ . Hence, two metastable conformations can be detected which only differ in the orientation of the apex of the tetrahedron  $Q_4$ . During a molecular simulation, the tetrahedron may not be able to change into another conformation because of the high energy barriers. Therefore, we divided state space  $\Omega$  into eight subsets  $B_k$ ,  $k = 1, \dots, 8$ , ran the separate samplings and applied the reweighting strategies. The results are listed in table 1.

In this case, Direct Free Energy Reweighting leads to a clearly better result than Estimation of Entropy Differences Reweighting. We have to keep in mind, however, that the abstract tetrahedral molecule example was modeled using a highly simplified potential energy function.

subset	exact	Direct Free Energy RW	Entropy Difference RW
$B_1$	0	0.00007	0
$B_2$	0.5	0.50001	0.45025
$B_3$	0	0	0
$B_4$	0	0	0
$B_5$	0	0	0
$B_6$	0	0	0
$B_7$	0.5	0.49987	0.54958
$B_8$	0	0.00005	0.00017

Table 1: Results of the statistical weighting of the conformations of the tetrahedral molecule model.

### 3.2 Pentane

To provide a more realistic example, we want to derive the conformational weights of the alkane pentane. As the conformational distribution of pentane at temperature of 300 K has been determined in a robust manner before, see [13], we choose to use this molecule for evaluation purposes. Pentane is able to adopt nine metastable conformations, defined by the two torsion angles  $\omega_1$  and  $\omega_2$  between carbon atoms  $C_1, C_2, C_3, C_4$  and  $C_2, C_3, C_4, C_5$ , respectively, see figure 2. At a torsion angle of  $180^\circ$  (*trans*), the geometry is energetically most favorable. Further-

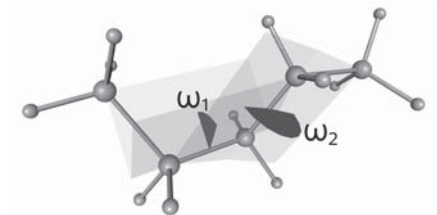


Figure 2: The torsion angles  $\omega_1$  and  $\omega_2$  define the metastable conformations of pentane.

more, local minima of the torsion potential can be found at  $60^\circ$  (*gauche*<sup>-</sup>) and  $300^\circ$  (*gauche*<sup>+</sup>). As pentane is a symmetric molecule, one would expect to find approximately equal weights for corresponding conformations, e.g.  $w(\text{trans}/\text{gauche}^-) \approx w(\text{gauche}^-/\text{trans})$ .

To obtain a dataset for reweighting, we first parameterized pentane according to the Merck Molecular Force Field (MMFF) [14], a force field optimized for small, drug-like molecules. Nine geometries corresponding to the conformations listed in table 2 were generated using the molecule editor from visualization software Amira [15], subsequently serving as starting points for nine separate Hybrid Monte Carlo (HMC) [2] samplings. HMC was set up to sample

five chains with 1000 HMC steps each per starting geometry. For each HMC step, a 60 fs molecular dynamics trajectory was calculated to generate a trial state. The pentane molecule was modeled in vacuo. The average acceptance probability of HMC runs was more than 90%. Convergence of the HMC runs was monitored using the Gelman-Rubin acceptance criterion [16] with a threshold value of 1.2. The potential function was modified so as to restrain sampling to the current local minimum, assuring thorough sampling of even the energetically unfavorable conformations.

conformation	exact [13]	Direct Free Energy RW	Entropy Difference RW
<i>trans/trans</i>	0.473	0.464	0.440
<i>gauche<sup>-</sup>/trans</i>	0.120	0.100	0.103
<i>gauche<sup>+</sup>/trans</i>	0.132	0.127	0.130
<i>trans/gauche<sup>-</sup></i>	0.117	0.102	0.103
<i>trans/gauche<sup>+</sup></i>	0.132	0.131	0.134
<i>gauche<sup>+</sup>/gauche<sup>+</sup></i>	0.013	0.040	0.044
<i>gauche<sup>-</sup>/gauche<sup>-</sup></i>	0.012	0.034	0.044
<i>gauche<sup>+</sup>/gauche<sup>-</sup></i>	< 0.005	0.001	0.001
<i>gauche<sup>-</sup>/gauche<sup>+</sup></i>	< 0.005	0.001	0.001

Table 2: Results of the statistical weighting of the conformations of pentane at 300 K.

The results of the reweighting are presented in table 2. In this case, it is not easy to determine the better strategy. Both methods approximate the exact weights rather well. When comparing the weights with the Euclidean and the infinity norm, however, the Direct Free Energy Reweighting provides a marginally better result.

## 4 Conclusion

After having described new direct reweighting strategies for use in conformation dynamics in section 2, we could prove their applicability in section 3. Both algorithms provide a good approximation of statistical weights in a direct way. While the Direct Free Energy Reweighting determines the weights only on the basis of the potential energy values of the separate samplings, and, furthermore, evaluates only one equation, namely (7), the Estimation of Entropy Differences Reweighting requires additional data to compute distances between conformations (e. g. dihedral angle values or intramolecular distances) and handles more than one equation. Therefore, the first method is faster and requires less memory.

In both algorithms we mentioned certain evaluation regions and representatives, respectively. By adapting the region’s size or the quantity of points to the present examples, one can

reach better results. In the process, one has to regard that both too few and too many reference points may lead to a bad statistics.

In contrast to existing methods we do not have to generate overlapping distributions in order to weight them against each other. Our new direct strategies offer a fast and efficient possibility for the computation of statistical weights without producing excess data or approximating functions.

## References

- [1] C. Chipot, M. Scott Shell, A. Pohorille, *Free Energy Calculations — Theory and Applications*, Springer–Verlag, Berlin, 2007.
- [2] S. Duane, A. D. Kennedy, B. J. Pendleton, D. Roweth, Hybrid Monte Carlo, *Phys. Lett. B* **195** (1987) 216–222.
- [3] W. G. Hoover, *Time Reversibility, Computer Simulation and Chaos*, World Scientific, River Edge, 1999.
- [4] H. J. C. Berendsen, J. P. M. Postma, W. F. Van Gunsteren, A. Dinola, J. R. Haak, Molecular dynamics with coupling to an external bath, *J. Chem. Phys.* **81** (1984) 3684–3690.
- [5] W. Huisinga, *Metastability of Markovian Systems: A Transfer Operator Based Approach in Application to Molecular Dynamics*, Doctoral thesis, Freie Universität Berlin, 2001.
- [6] T. Schlick, *Molecular Modelling and Simulation*, Springer–Verlag, Berlin, 2002.
- [7] W. G. Hoover, A. Kenichiro, C. G. Hoover, S. V. de Groot, Time–reversible deterministic thermostats, *Physica D* **187** (2004) 253–267.
- [8] K. L. Chung, *Markov Chains with Stationary Transition Probabilities*, Springer–Verlag, Berlin, 1967.
- [9] C. R. Johnson, R. A. Horn, *Matrix Analysis*, Cambridge Univ. Press, Cambridge, 1990.
- [10] M. Weber, S. Kube, L. Walter, P. Deuffhard, Stable computation of probability densities for metastable dynamical systems, *Multiscale Model. Simul.* **6** (2007) 396–416.
- [11] H. A. Simon, A. Ando, Aggregation of variables in dynamical systems, *Econometrica* **29** (1961) 111–138.

- [12] M. Weber, K. Andrae, A simple method for the estimation of entropy differences, *MATCH Commun. Math. Comput. Chem.* **63** (2010) 319–332.
- [13] C. Schütte, *Conformational Dynamics: Modelling, Theory, Algorithm and Application to Biomolecules*, Habilitation thesis, Freie Universität Berlin, 1999.
- [14] T. A. Halgren, Merck molecular force field. I-V, *J. Comp. Chem.* **17** (1996) 490–641.
- [15] J. Schmidt-Ehrenberg, D. Baum, H. C. Hege, Visualizing dynamic molecular conformations, in: *Proceedings of IEEE Visualization 2002*, IEEE Comp. Soc. Press, Washington, 2002, pp. 235–242.
- [16] A. Gelman, D. Rubin, Inference from iterative simulation using multiple sequences, *Statist. Sci.* **7** (1992) 457–511.