The Adaptively Biased Molecular Dynamics method revisited: New capabilities and an application

Journal of Physics: Conference Series **640** (2015) 012020 doi:10.1088/1742-6596/640/1/012020

The Adaptively Biased Molecular Dynamics method revisited: New capabilities and an application

Mahmoud Moradi[1], Volodymyr Babin[2], Christopher Roland[2,*], and Celeste Sagui[2,*]

1. Beckman Institute, University of Illinois, Urbana, IL 61801
2. Center for High Performance Simulations (CHiPS) and Department of Physics, North Carolina State University, Raleigh, NC 27695

E-mail: cmroland@ncsu.edu, sagui@ncsu.edu

Abstract. The free energy is perhaps one of the most important quantity required for describing biomolecular systems at equilibrium. Unfortunately, accurate and reliable free energies are notoriously difficult to calculate. To address this issue, we previously developed the Adaptively Biased Molecular Dynamics (ABMD) method for accurate calculation of rugged free energy surfaces (FES). Here, we briefly review the workings of the ABMD method with an emphasis on recent software additions, along with a short summary of a selected ABMD application based on the B-to-Z DNA transition. The ABMD method, along with current extensions, is currently implemented in the AMBER (ver.10-14) software package.

Introduction

The central role of atomistic free energy calculations for basic chemical and biological research is now well established. The free energy is one of the central quantities that determines the behavior of a system at or near equilibrium. For instance, the free energy determines the delicate three dimensional structure of biomolecules, their conformations, their binding, and the reactions they catalyze. Unfortunately, the FES of most biomolecular systems is multi-dimensional, often quite rough and complicated, so that very long time-scales are typically needed in order to explore a FES with any reasonable accuracy. This is generally precluded in a regular molecular dynamics (MD) simulation, even under present computer capabilities. Given the central importance of this problem, considerable effort has been invested in developing and refining new free energy methods, especially those tailored for complex FES when entropic contributions cannot be neglected.

Over the past decade, methods targeting the computation of free energy using non-equilibrium simulations have become popular. These methods all estimate the free energy as a function of a set of "collective variables" from an evolving ensemble of realizations [1,2], and use that estimate to bias the system dynamics and thereby flatten the effective free energy surface. Perhaps the first method to implement this idea was the Local Elevation Method [3]. More recent approaches include the adaptive force bias method [4], the Wang-Landau approach [5], the metadynamics method [6], and the Adaptively Biased Molecular Dynamics (ABMD) method [7] which was developed by our group. Collectively, these methods may be considered to be umbrella sampling methods. In ABMD it is the biasing potential that eventually reproduces the negative of the unknown FES in the long time limit.

This short paper briefly reviews the current state of the ABMD method. The focus will be on a review of the method, as well as on the recent additions based on the idea of coupled multiple walkers [8] and the multi-dimensional string method [9]. An application of the ABMD method to the B-Z DNA transition [10] will also be considered. Currently, the ABMD method is part of the AMBER software package [11], with the implementation of the newer methodological additions pending.

1. ABMD theory

To investigate the equilibrium properties of a complex multi-atomic system, it typically is customary to identify a suitable set of collective variables $\sigma(r_1, ..., r_N)$ which maps the atomic positions $r_1, ..., r_N$ onto a manifold, and then to study its equilibrium probability density:

$$p(\xi) = \langle \delta[\xi - \sigma(r_1, ..., r_N)] \rangle,$$

with the angular brackets indicating an ensemble average. This probability density $p(\xi)$ provides useful information as to the relative stability of the states corresponding to different values of ξ. In practice, the Landau free energy

$$f(\xi) = -k_B T log[p(\xi)],$$

(with k_B representing the Boltzmann constant and T the temperature) is typically preferred. Typically, one thinks of $p(\xi)$ or $f(\xi)$ as providing for a reduced dimensionality description of the system in terms of the chosen collective variables only: the other degrees of freedom have all been effectively integrated out. It is therefore important that the collective variables investigated are chosen in as judicious of a manner as possible, thereby bringing out the underlying physics in the most transparent manner possible. Often, the collective variable is associated with the slowest degrees of freedom, although this is not a formal requirement.

The idea behind the ABMD method [7] is the following. Consider an unknown FES of a system as a function of a suitable set of collective variables. Typically, free energy barriers for biomolecular systems are quite large, and so its difficult for the system to escape the different minima and explore other regions of phase space. To facilitate this, the system dynamics is determined by an *effective* free energy, which is given by the sum of the unknown free energy of the system and a time-dependent biasing potential. Initially, the effective free energy coincides with the unknown free energy of the system. However, with time a biasing potential is built up, which penalizes the locations that the system has visited. The net effect of this is to force the system out of any given minimum and to explore other regions. Ultimately, for very long times, the effective free energy is flattened and the biasing potential gives the negative of the unknown free energy.

The ABMD method is formulated in terms of the following equations:

$$m_a \frac{d^2 r_a}{dt^2} = F_a - \frac{\partial}{\partial r_a} U\big[t|\sigma\left(r_1, \ldots, r_N\right)\big],$$

$$\frac{\partial U(t|\xi)}{\partial t} = \frac{k_B T}{\tau_F} G\big[\xi - \sigma\left(r_1, \ldots, r_N\right)\big],$$

where the first one contains the Newton's equations that govern ordinary MD (temperature and pressure regulation terms not shown), augmented with an additional force coming from the time-dependent biasing potential $U(t|\xi)$ (with $U(t = 0|\xi) = 0$), whose time evolution is given by the second equation. We refer to τ_F as flooding timescale and to $G(\xi)$ as the kernel (in analogy to the kernel density estimator widely used in statistics. The kernel should be positive definite ($G(\xi) > 0$) and symmetric ($G(-\xi) = G(\xi)$). It may be viewed as a smoothed Dirac

delta function. For large enough τ_F and small enough kernel width, the biasing potential $U(t|\xi)$ converges towards $-f(\xi)$ as $t \to \infty$.

Our numerical implementation of the ABMD method [7] involves the use of cubic B-splines (or products of thereof) to discretize $U(t|\xi)$. We use the biweight kernel for $G(\xi)$, and an Euler-like discretization scheme for the time evolution of the biasing potential, a very simple scheme that satisfies the goal of flattening $U(t|\xi) + f(\xi)$ in the long time limit. Note that the numerical cost of evaluating the time-dependent potential is constant over time, and so ABMD scales trivially as $O(t)$. Storage requirements are also relatively mild, especially since sparse arrays may be used because only the nonzero elements of U_m need to be stored explicitly. In addition, ABMD is characterized by only *two* control parameters: the flooding timescale τ_F and the kernel width $4\Delta\xi$.

We have also implemented two important extensions to ABMD. The first is identical in spirit to the multiple walkers metadynamics [1,12]. It amounts to carrying out several different MD simulations biased by the same $U(t|\xi)$, which evolves via:

$$\frac{\partial U(t|\xi)}{\partial t} = \frac{k_B T}{\tau_F} \sum_{\alpha} G\big[\xi - \sigma\left(r_1^{\alpha}, \ldots, r_N^{\alpha}\right)\big] \,,$$

where α labels different MD trajectories. A second extension is to gather several different MD trajectories, each bearing its own biasing potential and, if desired, its own distinct collective variable, into a generalized ensemble for "replica exchange" with modified "exchange" rules [13,14,15]. Both extensions are advantageous and lead to a more uniform flattening of $U(t|\xi) + f(\xi)$. This enhanced convergence to $f(\xi)$ is due to the improved sampling of the "evolving" canonical distribution.

We have implemented the ABMD method in the AMBER simulation package (ver.10-14) [11], with support for both replica exchange and multiple-walkers. In pure "parallel tempering" replica exchange (same collective variable for all replicas), N_r replicas are simulated at different temperatures T_n, $n = 1, \ldots, N_r$. Each replica has its own biasing potential $U^n(t|\xi)$, $n = 1, \ldots, N_r$, that evolves according to its own dynamical equation. Exchanges between neighboring replicas are attempted at a prescribed rate, with an exchange probability given by [13]:

$$w(m|n) = \begin{cases} 1, & \Delta \leq 0, \\ \exp(-\Delta), & \Delta > 0, \end{cases}$$

$$\Delta = \left(\frac{1}{k_B T_n} - \frac{1}{k_B T_m}\right)\left(E_p^m - E_p^n\right) + \frac{1}{k_B T_m}\Big[U^m(\xi^n) - U^m(\xi^m)\Big] - \frac{1}{k_B T_n}\Big[U^n(\xi^n) - U^n(\xi^m)\Big],$$

where E_p denotes the atomic potential energy. The biasing potentials are temperature-bound and converge in the $t \to \infty$ limit to the free energies at their respective temperatures.

In a general replica exchange scheme, we allow different replicas to have different collective variables and/or temperatures. We also allow for either an evolving or a static biasing potential (the latter obviously includes the case of $U^n(t|\xi) = 0$) on a per-replica basis (when all biasing potentials are static, the method reduces to a peculiar variant of the "Hamiltonian Replica Exchange" described in Ref.[13]). Exchanges between random pairs of replicas are then attempted at a prescribed rate. To illustrate the point, consider the *in vacuo* folding of the hydrophobic peptide $Ace - (Gly)_2 - Pro - (Gly)_3 - Nme$, which has a stable β-hairpin in the 1999 version of the Cornell *et al* [16] force field. First, accurate ABMD free energy calculations at eight different temperatures as a function of the radius of gyration R_g as the collective variable were carried out (see the inset in Fig.1). The global minimum corresponding to the smaller value of R_g is associated with a coil structure, while the metastable minimum associated with

the larger value corresponds to a β-hairpin structure. The relative depth of these wells changes, as expected, in the presence of explicit solvent environments [7,17]. However, there is more information that can be gleaned from these simulations. For instance, if one wants additional information with regards to the number of hydrogen bonds formed along the peptide backbone N_{OH}, it is possible to carry out an ABMD simulation with the modified replica exchange in such a way that one reuses the previously computed free energies for R_g in order to obtain the *two-dimensional* free energy landscape in the (N_{OH}, R_g) space. Here, the initial simulations play a double role in serving as a "sampling enhancement device" for different replicas, thereby greatly speeding up the calculation. The resulting *two-dimensional* free energy landscape shown in Fig.1 now conveys additional information not contained in the *one-dimensional* plots. In particular, the "coil minimum" associated with the *one-dimensional* plot is degenerate, and actually corresponds to two different minima for which the globular structures have a different number of hydrogen bonds. This example illustrates not only the type of information that may be obtained with ABMD, but also shows, that as one learns more about a given biomolecular system, one can "dynamically" change the questions asked. Previous calculations may often be readily reused, in order to speed up the new simulations.

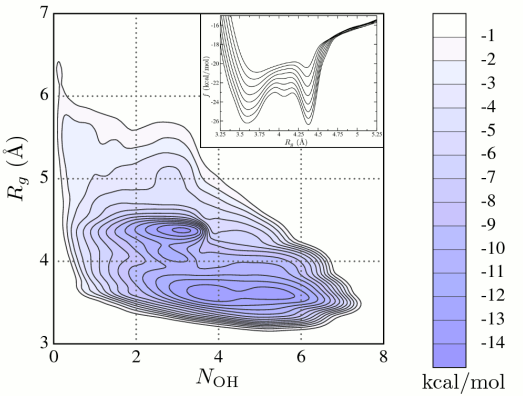

Figure 1. Two-dimensional free energy landscape for *in vacuo* Ace-$(Gly)_2$-Pro-$(Gly)_3$-Nme peptide at T=300 K as a function of collective variables R_g and N_{0H}. Inset shows the corresponding one-dimensional free energy as a function of only R_g for eight different temperatures ranging from T=300 (bottom) to T=600 K (top). See Ref.[7].

Finally, we note two important points. First, if one uses the ABMD method alone, one needs follow-up equilibrium umbrella sampling runs to improve the accuracy of the results, according to the recipe that we provided in the context of the metadynamics method [18]. Second, if in addition to the free energy map, one desires equilibrium, unbiased properties of the system, it is possible to set up a generalized REMD scheme. In this scheme, the choice of the Hamiltonians determines the performance of the method. If the exchange rate between the replicas is sufficiently high to guarantee a random walk between the replicas, then one needs only to consider the details of the so-called "hot" replica whose purpose is to facilitate barrier crossing (or, more formally, to decrease the ergodic time scale). One possibility for this is to run the hot replica at high temperature. Another possibility [19] is to construct the hot replica by adding a biasing potential to the original Hamiltonian that acts on some collective variable that describes one of the slow modes of the system that needs "acceleration". ABMD is therefore first used to generate potentials for these "hot" replicas. A combination of such Hamiltonian and Temperature based replica exchange molecular dynamics (HT-REMD) [18,19,20] provides for a practical way to reduce the computational costs associated with sampling, since it facilitates the sampling in the "hottest" replica by both means, and therefore also allows for a better "tuning" of the entire setup. These procedures ultimately can yield high accuracy free energy curves and equilibrium properties, and have already been applied to a variety of different biomolecular systems including small peptides [7,18,19], sugar puckering [19], polyproline systems [19,20,21], guest-host systems [22,23], polyglutamine systems [24], and the B-to Z DNA transition [10].

In addition, we have implemented Steered Molecular Dynamics (SMD) simulations to examine transition pathways and mechanisms, as well as to estimate free energy differences [20]. The SMD method is based on the following ideas. Consider a thermodynamic process that changes a system by means of a control parameter ξ, which runs from ξ_0 to ξ_t over a time interval t. The

second law of thermodynamics states that the work W done on the system cannot be smaller than the free energy difference: $\Delta F = f(\xi_t) - f(\xi_0) \leq W$. Thus, a non-equilibrium process can only provide us with an upper bound for the free energy difference. However, the remarkable Jarzynski [25] identity (JE) is an exact equality that holds even when the transformation is irreversible:

$$\exp(-\Delta F/k_B T) = \langle \exp(-W/k_B T) \rangle$$

(the angular brackets denote the average over all possible realizations of the process connecting the states ξ_0 and ξ_t). Thus, in principle ΔF can be estimated using a number of finite-time *non-equilibrium* simulations. Unfortunately, because the average of the exponential function is dominated by rare realizations, convergence is often poor unless a very large number of runs is used. It is, however, possible to improve the free energy estimates by using modifications based on the maximum likelihood estimator (MLE) and Bennet's acceptance ratio (BAR) method, which are considered to be equivalent [26,27,28]. These methods make use of both forward (F) and reverse (R) non-equilibrium pulling processes, in order to move the system between an initial and final state. Typically, the system is moved along a specific pathway in its configurational space, which is readily accomplished through the addition of a harmonic restraining potential to the Hamiltonian in the stiff-spring limit [20]. This approximation ensures that the calculated potential of mean force is well represented by the free energy of the system with the harmonic restraint. In our system, we carried out a number of non-equilibrium "pushing" and "pulling" simulations by changing the value of the collective variable ξ at a constant velocity over a prescribed pathway. The numerical values of the work for the forward (W^F) and reverse (W^R) pullings are then collected and ΔF is estimated from the transcendental equation [26,27,28]:

$$\sum_{i=1}^{n_{\mathrm{F}}} \frac{1}{1 + n_{\mathrm{F}}/n_{\mathrm{R}} \exp\left(W_i^{\mathrm{F}} - \Delta F\right)} - \sum_{i=1}^{n_{\mathrm{R}}} \frac{1}{1 + n_{\mathrm{R}}/n_{\mathrm{F}} \exp\left(W_i^{\mathrm{R}} + \Delta F\right)} = 0,$$

with $n_{\mathrm{F,R}}$ denoting the numbers of forward and reverse simulations, respectively. As may be expected, we find that ΔF calculated with this two-sided method displays an improved convergence over estimates obtained with the straightforward one-sided application of the Jarzynski equality [20]. To date, we have extensively tested SMD on proline and DNA-based systems. We have been able to use SMD to steer the system between different states over the ABMD FES, thereby being able to probe different transition mechanisms, their associated barriers, and to estimate free energy differences. It comes as no surprise that the modified formula based on the BAR method shows the fastest convergence and gives the best results [20].

2. New Methodological Developments

Recently, we have augmented our implementation of the ABMD method in a number of significant ways, including the addition of a multiple-walker selection mechanism [8], the well-tempered ABMD [29], driven ABMD [30], and a version of the string method [9]. These methods, which are currently being built into the latest release of AMBER, are briefly described in this section.

2.1. ABMD with Selection Algorithm

The multiple-walker ABMD method is based on a set of *non-interacting* walkers or replicas, all of which persist for the same duration. However, this can lead to problems because not all walkers are equally effective in sampling the configuration space. Often, these end up being clustered together, and kept near local metastable regions because of hidden barriers resulting from degrees of freedom orthogonal to the reaction coordinate. Thus, for maximum efficiency, one would like

to "encourage" walkers that are wandering in the under-explored regions, and penalize walkers in the oversampled regions, and achieve this on-the-fly. This requires a dynamic monitoring of the walkers, and their updating according to suitable, preselected fitness functions. A typical algorithm of this kind starts [8] by assigning the same weights to every walker, and lets them run for a preselected period of time. The "fitness" of each walker is then tested, which is a *de facto* measure of overlap of the explored phase space of each walker. Walkers who are wandering in under-sampled regions are then given a stronger weight and replicated a number of times, while walkers found to be exploring oversampled regions are correspondingly killed. This kind of procedure is then repeated periodically, thereby accelerating the convergence to a uniform distribution of walkers in the reaction coordinate space. This "selection" algorithm is thus a bootstrapping-based resampling of conformations in so far that the walkers are bootstrapped periodically to flatten the free energy landscape more efficiently. We note that a variation of selection algorithm similar to the one implemented here has previously been implemented in the NAMD [8] software package.

We have implemented an ABMD multiple-walker selection algorithm as follows. At fixed simulation time intervals τ, a weight w_i is assigned to walker i proportional to $\exp(\int_{(n-1)\tau}^{n\tau} S(\zeta_t^i)\,dt)$ in which ζ_t^i is the instantaneous value of reaction coordinate vector. Furthermore, we use $S(\zeta) = c\frac{\nabla^2\rho(\zeta)}{\rho(\zeta)}$, with $\rho(\zeta)$ representing the density of ζ, c is a constant, and index n denotes the time interval. This update is performed either for the duration of the simulation, or until some stopping criterion is satisfied. The stopping criterion is based on the entropy of weights $H = \sum_i w_i log(w_i)$. The selection algorithm will stop if $E_w = H - \log(\frac{1}{N})$ falls below $\epsilon \log(\frac{1}{N})$ in which N is the total number of walkers and $0 \leq \epsilon \leq 1$ is a constant. Note that $\log(\frac{1}{N})$ represents the entropy of the uniform weights.

Comparing to the original method [8], our implementation also incorporates the following technical features: (i) We use the biasing potential instead of a histogram to approximate the density; (ii) We use a generalized Laplacian instead of a simple second order derivative for multidimensional cases for added stability; and (iii) Since ABMD uses B-splines, we build up analytic functions for the biasing potential and therefore can calculate the needed derivatives analytically for additional stability and speedup. Tests indicate that our current implementation of the described selection mechanism for multiple walkers leads to a speed-up in the convergence by a factor ranging from 2 to 5 over a corresponding system with noninteracting multiple walkers.

2.2. The well-tempered ABMD

In its original implementation, ABMD uses a history-dependent biasing potential that is updated at a fixed rate:

$$U_a(\zeta, t) = U^0(\zeta) + \int_0^t dt'\, \omega\, K(\zeta - \zeta^{t'}), \tag{1}$$

in which K is a kernel function, U^0 is an arbitrary function (*e.g.*, this typically represents the initial guess for the biasing potential and is typically assumed to be flat; ω represents the contant, unbiased rate). It has been shown [45] that once converged, $\left\langle U_a(\zeta, t \to \infty)\right\rangle_a \approx U_a^s(\zeta) + u(t)$ (here $u(t)$ is an additive constant at each t) in which $U_a^s(\zeta)$ (*i.e.*, converged biasing potential) equals $-F(\zeta)$.

It turns out that using the same rate of updating the biasing potential throughout the simulation often leads to poor convergence to the free energy. Instead, rather than converging smoothly to the desired result, the biasing potential ends up fluctuating around $-F(\zeta)$ (with an amplitude which depends on ω). A solution to this problem has been proposed [29] via the use of a "well-tempered" ω:

$$U_a(\zeta, t) = U^0(\zeta) + \int_0^t dt'\, \omega(\zeta^{t'}, t')\, K(\zeta - \zeta^{t'}), \tag{2}$$

in which $\omega(\zeta,t)$ is a time-dependent, non-uniform rate $\omega_0 \mathrm{e}^{-\beta' U_a(\zeta,t)}$ ($1/\beta' = k_B T'$ and T' is a pseudo-temperature) that reduces to a constant ω_0 in the $\beta' \to 0$ limit (*i.e.*, conventional ABMD). It has been shown [29] that $\left\langle U_a(\zeta, t\to\infty)\right\rangle_a \approx U_a^s(\zeta) + u(t)$, and so $U_a^s(\zeta)$ and $F(\zeta)$ are related via $U_a^s(\zeta) = -(1+\frac{\beta'}{\beta})^{-1}F(\zeta)$ or $F(\zeta) = -(1+\frac{T}{T'})U_a^s(\zeta)$. This way of updating leads to smoother convergence of the biasing potential to the desired free energy.

2.3. **Driven ABMD**

Both ABMD and SMD schemes are both powerful nonequilibrium free energy methods. However, each method has its own practical limitations associated with it. For instance, SMD is often associated with a very slow convergence when used to calculate free energies. However, it is very useful for the qualitative exploration of transition paths. ABMD, on the other hand, is very useful in mapping out free energy landscapes, even though this may take a considerable computational effort depending on the complexity of the FES. It turns out, however, that ABMD and SMD may be combined and integrated into a new driven adaptively biased method [30]. Such a driven ABMD (D-ABMD) takes advantage of both methods for faster convergence. D-ABMD has an advantage over conventional (or well-tempered) ABMD in that it ensures the exploration of the transition pathway (from one end to the other) in the early stages of the simulation and gradually improves the estimate of the free energies almost uniformly along the reaction coordinate. D-ABMD has also an advantage over the conventional SMD in that the effective free energy surface gradually becomes smooth and flat such that the system can move along the reaction coordinate with progressively less amount of work. The D-ABMD method implemented is a straighforward generalization of the so-called driven metadynamics method [30].

To combine the two schemes, we have developed a driven adaptive-bias scheme that adds an adaptive ($U_a(\zeta,t)$) and a driving ($U_d(\zeta,t)$) potential to the Hamiltonian. We use an iterative approach in which an independent simulation is performed from time $t=0$ to $t=T$ in the n^{th} iteration ($n=1,2,\ldots$), biased by the potential $U_d(\zeta,t)+U_a^n(\zeta,t)$ in which $U_d(\zeta,t) = \frac{k}{2}(\zeta-\eta(t))^2$ for all n ($\eta(t)$ is the moving center of the SMD harmonic potential in ζ space), and:

$$U_a^n(\zeta,t) = U^{n-1}(\zeta) + \int_0^t dt'\, \omega(\zeta^{t'},t')\, K(\zeta-\zeta^{t'})\, \mathrm{e}^{-\beta w^{t'}}, \tag{3}$$

in which w^t represents either the accumulated work or the transferred work. The accumulated and transferred works are defined as $w_{ac}^t = \int_0^t dt' \frac{\partial}{\partial t'} U_d(\zeta^{t'},t')$ and $w_{tr}^t = w_{ac}^t - U_d(\zeta^t,t)$ (note that this is actually the transferred work shifted by $-U_d(\zeta^0,0)$). Theoretically the $\mathrm{e}^{-\beta w_{tr}^{t'}}$ factor or "constant weight" is more accurate but for practical reasons the $\mathrm{e}^{-\beta w_{ac}^{t'}}$ factor or "pulling weight" is preferred. For a discussion on these weight functions used for the free energy estimation from conventional pulling experiments, please see Ref.[31]. Particularly, in our algorithm, the constant weight $\mathrm{e}^{-\beta w_{tr}^{t'}} = \mathrm{e}^{-\beta w_{ac}^{t'}} \mathrm{e}^{\beta U_d(\zeta^t,t)}$ may become unstable for large biasing potentials. To avoid this instability a cutoff for w^t is used. This method also provides for faster convergence and greater stability to the desired free energy.

2.4. **The string method**

The swarms-of-trajectories string method (STSM) [9] is a path-finding algorithm that refines a putative transition pathway (or a string of conformations) iteratively until the pathway is converged. The string is defined by a number of nodes or images parameterized in a high-dimensional space of collective variables, whose position is updated in each iteration. The drift applied to each image is estimated by averaging over the drifts of a swarm of short unbiased trajectories all starting at the current position of the image. The parallel variation of the

method – which is considerably more efficient than the original variation – requires hundreds to thousands of replicas of the system to run simultaneously [32].

For a string of N nodes, each consisting of M copies, $N \times M$ replicas will be required. Each unbiased drifting iteration is followed by a biased equilibration stage to move the replicas to their updated position. Unlike the ABMD simulations, STSM simulations can easily employ many collective variables since the sampling always stays along a 1D curve. Therefore one may use all collective variables that are somewhat relevant to the transition of interest, particularly those associated with slow dynamics. The method therefore enables the mapping out of Least Free Energy Pathways (LFEP) in a multi dimensional phase space. The number of images used is important for the accuracy of the pathway and the number of copies used is important in determining the convergence behavior.

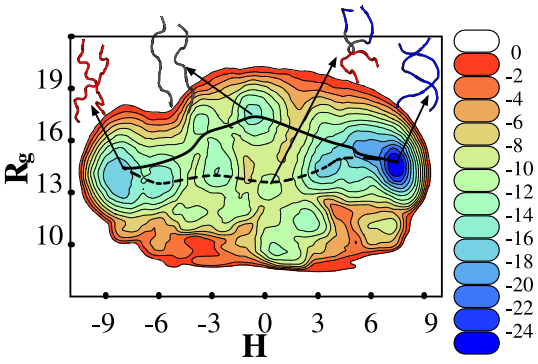

Figure 2. Free energy landscape of $(CG)_6$ DNA as a function of the radius of gyration R_g and handedness H. Ribbon diagrams associated with the four minima shown in blue (B-DNA), red (Z-DNA) and grey (transition structures). Two different Least Free Energy Paths (LFEPs) are identified with a solid line (stretch-collapse mechanism) and with a dashed line (zipper-like mechanism. See Ref.[10] for details.

Our parallel implementation of STSM method is based on iterative unbiased and restrained MD simulations followed with reparametrization of the restraint centers defined in the multidimensional collective variable space of ζ. At iteration s, M copies of image i restrained around old center η_i^{s-1} equilibrate for τ_E steps ($t = 0, \ldots, \tau_E$) such that $\zeta_{i,l}^t$ for each copy l of image i is expected to be close to η_i^{s-1} at $t = \tau_E$ assuming the restraint is stiff enough. The restraint is then released to allow drifting for τ_R steps. The drifted center η_i^s for image i is first determined by averaging over all drifted copies $\zeta_{i,l}^t$ at $t = \tau_E + \tau_R$. The resulting string of images is smoothed using the protocol proposed in Ref.[33]. The smoothing parameter $0 \leq \epsilon \leq 1$ determines the smoothness of the curve and is recommended to be on the order of $1/(N-1)$. The last step is the reparametrization which again follows the same protocol proposed [33] to generate a set of N sequentially equidistant centers along the same string formed by the smoothed centers.

3. An Application: The B-Z DNA transition

As already noted, the ABMD method has been extensively applied to a number of biomolecular systems. Here, we briefly discuss a single application, namely that of the B-Z DNA transition [10]. In 1979 the publication of the crystal structure of a d(CGCGCG)$_2$ duplex revealed a left-handed double-helix DNA with two antiparallel chains joined by Watson-Crick (WC) base pairs [33]. This conformation was called Z-DNA because the sugar-phosphate backbone had a characteristic zig-zag pattern. Left-handed Z-DNA is characterized by a dinucleotide repeat, and occurs in sequences that alternate a purine-pyrimidine repeat, mainly CG (or GC). The anti-syn alternation of base pairs that underlies the zig-zag pattern is due to the rotation of the guanine residue around its glycosidic bond, resulting in its syn conformation. The paired cysteine remains in its anti conformation, but it rotates 180° jointly with its attached sugar. As a result, the base pairs in Z-DNA, as compared to B-DNA where all the bases are in the anti conformation, are stacked on opposite sides of the line connecting the two phosphate groups across the helix [34,35], and there is a different stacking pattern of the GpC and CpG sequences involving a series of two base steps along each of the two strands [36].

Since its discovery, Z-DNA has evolved from an in vitro curiosity to a challenging entity with important biological roles. The first observation of a complete inversion of the ultraviolet

circular dichroism spectrum of poly(CG) DNA in the presence of a high concentration of salt took place in 1972 [36]. After the revelation of the Z-DNA structure in 1979, further evidence of the B-Z DNA transition was obtained [37]. Yet, the experimental evidence for the *in vivo* existence of Z-DNA has been slow to gather. It is now known that Z-DNA is highly immunogenic, and there are antibodies against it that have been further used to map regions favorable to Z-DNA conformations [38]. There is a high density of base sequences favoring Z-DNA, *i.e.,* bases with suitably alternating purines-pyrimidines, near transcription start sites. The formation of Z-DNA is also induced by Z-DNA binding proteins near the promoter region which boosts the transcription of the downstream genes. All in all, it is now thought that Z-DNA formation is closely related to gene expression, regulation, and recombination [39].

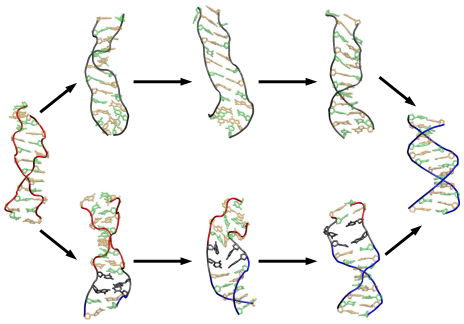

Figure 3. Alternative B-Z DNA transition pathways. Snapshots of a Z- to B-transition for the dodecamer $(CG)_6$ obtained from SMD simulations through the LFEP trajectories. Top: stretch collapse mechanism; Bottom: base extrusion mechanism. Colors in the ribbon representation are: red (Z-DNA); blue (B-DNA); grey (intermediate structures). In addition, the C and G nucleotides are colored in green and cream, respectively. See Ref.[10] for details.

Despite all the investigations about the biological role of left-handed Z-DNA, the microscopics behind the B-Z DNA transition have remained controversial, with several different mechanisms proposed in the literature [40]. Generally speaking, the models for the B-Z transition are classified into mechanisms that involve either base-pair opening or base-pair rotation without WC base-pair breaking. Each of these mechanisms may or may not have an intermediate structure. Among the models without intermediate structure, the most popular base-pair opening mechanism is the Wang model [33]. It proposes that base-pair opening occurs before base-pair plane and phosphate backbone angle rotation within the core of the helix. The Harvey model [41], on the other hand, proposes that the B-Z transition happens through the successive flipping of base-pair planes, without any disruption of the WC pairs. This process is supposed to be facilitated by breathing modes. Experimental evidence, however, seems to favor models with intermediate structures [42,43,44]. The Saenger-Heinemann model postulates that the transition takes place through two A-DNA-like intermediates (the second one with a dinucleotide unit), without breaking of the base pairs [45,46]. More recently, the extrusion of bases was observed in the crystal structure of a B-Z junction [47]. These extrusions and their propagation followed by the reformation of the base pairs in a new order also represents an alternative mechanism for the transition [47] (similarly to the Wang's model, except that the bases rotate outside the double-helix core).

We have recently reexamined the problem of the B-Z DNA transition [10] using the ABMD and SMD methods. This has the advantage of allowing for a unified description of the transition, incorporating a variety of mechanisms. A sample two-dimensional free energy for the B-Z transition is shown in Fig.2, as a function of the radius of gyration (R_g) and handedness (H). The most prominent minima are associated with B- and Z-DNA structures, respectively. Connecting these minima are two LFEPs associated with a stretch-collapse mechanism and a zipper mechanism (see Fig.3). The stretch collapse mechanism involves relatively little disruption of the base pairs. Instead, the DNA stretches and unwinds, forming two almost parallel strands that preserve their WC base pairs, and assuming an S-DNA type conformation. The conformation then rewinds to form a double helix of the opposite handedness. We note that such a mechanism has also been previously identified in other simulation studies [10]. Alternatively, the so-called zipper-like mechanism involves almost no change in R_g. However,

there are large changes in handedness within the structure, and the intermediates involve at least one B-Z junction with extruded base pairs. It is interesting that very similar free energy barriers are associated with both of these transition mechanisms [10], indicating that the transition is best described in terms of a reaction path ensemble.

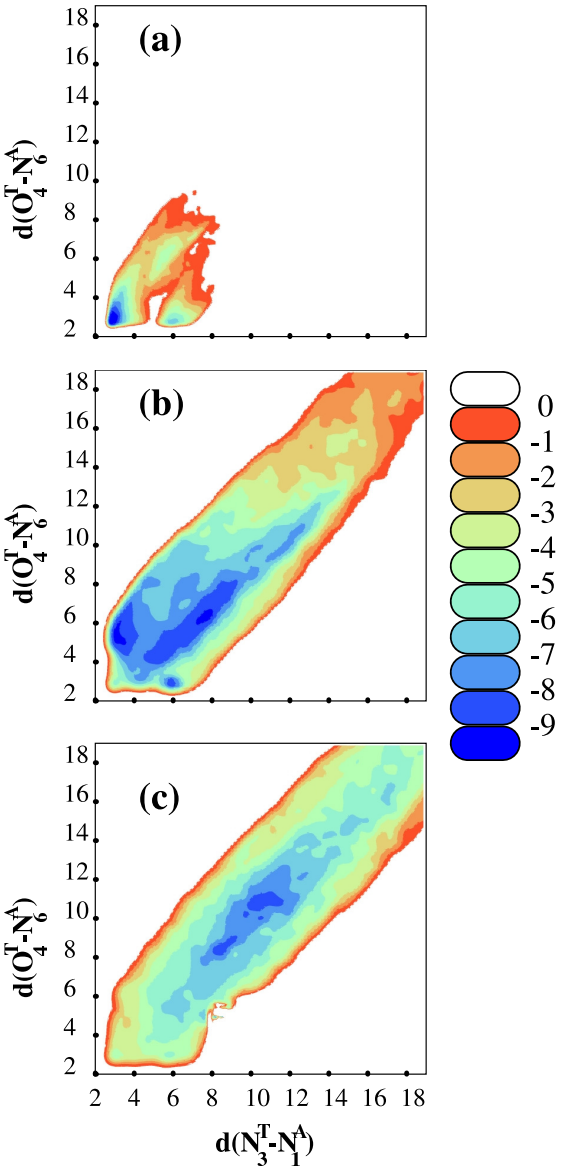

Figure 4. Free energy landscapes of an AT base in different conformation for $(CG)_3(AT)(CG)_3$ double helix DNA in terms of the collective variable $d(O_4^T - N_g^A)$ (*i.e.*, the distance between the O_4^T and N_6^A of the thymine and adenine bases) and $d(N_3^T - N_1^A)$ (*i.e.*, the N_3^T and N_1^A of the thymine and adenine bases) with the rest of the structure having a conformation corresponding to (a) B-NDA; (b) Z-Z junction; and (c) a B-Z junction. For (a) the free energy minima correspond to structures with the correct WC pairing, while in (b) the minima are associated with reverse and single WC pairing. See Ref.[10] for more details.

The crystal structure of a B-Z junction [47] has revealed the full extrusion from the helix of the two junctional bases. The formation of a B-Z junction requires little structural disruption, because it preserves the integrity of both the B- and Z-DNA helices as well as the base stacking between the two helices. While a B-Z junction is formed at the interface between B- and Z-DNA, a Z-Z junction is also commonly formed in sequences where the dinucleotide repeat is interrupted by single base-pair insertions, or deletions, that bring neighboring helices out of phase. The three-dimensional structure of a Z-Z junction has been recently described experimentally [48]. The Z-Z junction studied by de Rosa *et al* is stabilized by Zα, the Z-DNA binding domain of the RNA editing enzyme ADAR1, and consists of a single base pair AT in the middle of a CG sequence, with the resulting duplex sequence $5' - (CG)_3 - A - (CG)_3 - 3'$. The AT pair leads to a partial disruption of the helical stacking. While Z-DNA is resistant to intercalating agents, these can insert themselves into the junction region. In contrast to a B-Z junction, the basis of this structure are not fully extruded, and the stacking between the two left-handed helices is not continuous.

As part of our study of the B-Z transition, we have also investigated both the free energies of pure B-Z junctions as well as junctions with an A-T insertion. Sample free energy results for the latter are shown in Fig.4. Here the top panel (a) shows the results for a B-B junction. The free energy landscape is characterized by three minima, with the global minima corresponding – as expected – to a junction with the correct WC pairing. In (b), the free energy of a Z-Z junction and shows that the minima have shifted considerably. In this case the Z-Z junction may be conformations involving correct WC pairing, single WC pairing, and reverse WC pairing. The global minimum involves the latter, and characteristic structure is in good correspondence with the experimental results [48]. Finally, (c) shows results for a B-Z junction with minima corresponding to non-WC pairings. For these

and further details on the B-Z DNA transition, we refer the reader to Ref.[10].

4. Summary

In summary, we have briefly reviewed the workings of the ABMD method, along with recent extensions that are currently being worked into the AMBER simulation package. To illustrate the workings of the method, a single application involving the B-Z DNA transition is briefly summarized.

5. Acknowledgements

We thank NSF (grants MCB-1021883 and SI2/SEE-1148144) for funding, and the NCSU HPC for extensive computer support.

6. References

[1] Lelievre T, Rousset M, and Stoltz G 2007 *J. Chem. Phys.* **126**, 13411.
[2] Bussi G, Laio A, and Parrinello M 2006 *Phys. Rev. Lett.* 96, 090601.
[3] Huber T, Torda A and van Gunsteren W 1994 *J. Comput. Aided Des.* **8**, 695.
[4] Darve E, and Pohorille A 2001 *J. Chem. Phys.* **115** 9169.
[5] Wang F, and Landau D 2001 *Phys. Rev. Lett.* **115** 2050.
[6] Iannuzzi M, Laio A and Parrinello M 2003 *Proc. Natl. Acad. USA* **99** 12562.
[7] Babin V, Roland C, and Sagui C 2008 *J. Chem. Phys.* **128** 134101.
[8] Minoukadeh K, Chipot C, and Lelievre T 2010 *J. Chem. Theor. and Comput.* **6** 1008.
[9] Pan A, Sezer D, and Roux B 2008 *J. Phys. Chem.* **112** 3432.
[10] Moradi M, Babin V, Roland C, and Sagui C 2013 *Nucl. Acids Res.* **41** 33.
[11] AMBER 10, 2008 (http:ambermd.org), and higher versions.
[12] Raiteri P, Laio A, Gervasio F, Micheletti C, and Parrinello M 2006 *J. Phys. Chem.* **110** 3533.
[13] Sugita Y, Kitao A and Okamoto Y 2000 *J. Phys. Chem.* **113** 6042.
[14] Bussi G, Gervasio F, Laio A, and Parrinello M 2006 *J. Am. Chem. Soc. 128* 13435.
[15] Piana S and Laio A 2007 *J. Phys. Chem. B 111* 4553.
[16] Cornell W, *et al* 1995 *J. Am. Chem. Soc.* **117** 5179.
[17] Babin V, Karpusenka V, Moradi M, Roland C, and Sagui C 2009 *Int. J. Quant. Chem.* **109** 3666.
[18] Babin V, Roland C, Darden T, and Sagui C 2006 *J. Chem. Phys.* **125** 2049096.
[19] Babin V and Sagui C 2010 *J. Chem. Phys.* **132** 104108.
[20] Moradi M, Babin V, Roland C, and Sagui C 2010 *J. Chem. Phys.* **133** 125104.
[21] Moradi M, Babin V, Roland C, Darden T, and Sagui C 2009 *Proc. Natl. Acad. Sci. USA* **106** 20746.
[22] Moradi M, Babin V, Sagui C, and Roland C 2011 *Biophys. J.* **100** 1083.
[23] Moradi M, Babin V, Sagui C, and Roland C 2011 *J. Phys. Chem. B* **115** 8645.
[24] Moradi M, Babi V, Roland C, and Sagui C 2012 *PLoS Comput. Biol.* **8** e1002501.
[25] Jarzynski C 1997 *Phys. Rev. Lett.* **78** 2690.
[26] Crooks G 2000 *Phys. Rev. E* **61** 2361.
[27] Shirts M, Bair E, Hooker G and Pande VS 2003 *Phys. Rev. Lett.* **91** 140601.
[28] Kosztin I, Barz B, and Janosi L 2006 *J. Chem. Phys.* **124** 064106.
[29] Barducci A, Bussi G, and Parrinello M 2008 *Phys. Rev. Lett.* **100** 020603.
[30] Moradi M and Tajkhorshid E. 2013 *J. Phys. Chem. Lett.* **4** 1882.
[31] Oberhofer H and Dellago C, 2009 *J. Comp. Chem.* **30**1726.
[32] Jiang W, Phillips W, Huang L, Fajer M, Meng Y, Gumbart J, Luo Y, Schulten K, and Roux B 2014 *Comp. Phys. Comm.* **185** 908.
[33] Maragliano L, Fischer A, Vanden-Eijnden E and Ciccotti G 2006 *J. Chem. Phys.* 024106.
[34] Wang A, Quigley G, Kolpak F, Crawford J, van Boom J, van der Marel G, and Rich A, 1979 *Nature* **282** 680.
[35] Rich A, Nordheim A, Wang A 1984 *Annu. Rev. Biochem.* **53** 791.
[36] Nordheim A, and Rich A, 1983 *Proc Natl. Acad. Sci. USA* **80** 1821.
[37] Pohl F and Jovin T 1971 *J. Mol. Biol.* **72** 375.
[38] Thamann T, Lord R, Wang A, and Rich A, 1981 *Nucl. Acids Res.* **9** 5443.
[39] Lipps H, 1983 *Cell* **32** 435.
[40] Oh D, Kim Y, and Rich A, 2002 *Proc. Natl. Acad. Sci. USA* **99** 16666.

[41] Fuentes M, Cepeda V, Alonso C, and Perez J, 2006 *Chem. Rev.* **106** 2045.
[42] Harvey S, 1983 *Nucl. Acids Res.* **11** 4867.
[43] Goto S, 1984 *Biopolymers* **23** 2211.
[44] Ansevin A and Wang A, 1990 *Nucl. Acids Res.* **18** 6119.
[45] Lim W, and Feng Y, 2005 *Biophys. J.* **88** 1593.
[46] Saenger W, and Hienemann U 1989 *FEBS Lett.* **257** 223.
[47] Ha S, Lowenhaupt K, Rich A, Kim Y, and Kim K 2005 *Nature* **437** 1183.
[48] de Rosa M, de Sanctis D, Rosario A, Archer M, Rich A, Athanasiasdis A, and Carrondo M, 2010 *Proc. Natl. Acad. Sci. USA* **107** 9088.