J. Math. Biol. (2012) 64:917–931
DOI 10.1007/s00285-011-0432-6

Mathematical Biology

# Adaptive conformational sampling based on replicas

**Jeremy Curuksu**

Received: 2 November 2010 / Revised: 9 April 2011 / Published online: 8 June 2011


**Abstract** Computer simulations of biomolecules such as molecular dynamics simulations are limited by the time scale of conformational rearrangements. Several sampling techniques are available to search the multi-minima free energy landscape but most efficient, time-dependent methods do generally not produce a canonical ensemble. A sampling algorithm based on a self-regulating ladder of searching copies in the dihedral subspace is developped in this paper. The learning process using short- and long-term memory functions allows an efficient search in phase space while combining a deterministic dynamics and stochastic swaps with the searching copies conserves a canonical limit. The sampling efficiency and accuracy are indicated by comparing the ansatz with conventional molecular dynamics and replica exchange simulations.

**Keywords** Adaptive sampling · Convergence of molecular dynamics · Replica exchange · Dihedral angles

**Mathematics Subject Classification (2000)** 82B05 · 82C05

## 1 Introduction

Sampling the conformational variability of biomolecules permits to better understand their underlying functions. Molecular dynamics (MD) simulations are often preferred to stochastic sampling methods such as monte carlo because the connection in time of the sampled configurations allows us to ascertain time-dependent properties such as characteristic correlation functions derived from experiments.

J. Curuksu (✉)
Institute of Mathematics, Ecole Polytechnique Fédérale de Lausanne (EPFL), MATHGEOM, LCVMM, Station 8, MAC1 615, 1015 Lausanne, Switzerland
e-mail: jeremy.curuksu@epfl.ch

The time step in molecular dynamics simulations needs to be of the order of the highest frequency motion which for molecular systems correspond to $10^{-15}$ s (vibration of covalent bonds). The dynamics of biomolecules is characterized by diverse regimes of time scale going up to the second for the smooth bending of nucleic acids (Mesirov et al. 1996) or the folding of proteins (Thirumalai 1995). In addition, most degrees of freedom are highly discontinuous in relation to combined rotamer preferrences (Gauche+, Trans, Gauche−) and to hard-core repulsion between atoms. As a result, biomolecular free energy surfaces are rugged and highly anisotropic, with many maxima, minima and saddle points, implying a frequent kinetic trapping into its sub-minima (Beveridge et al. 2004). The small time step thus prevents the capture of many facets of biomolecular complexity from a continuous molecular dynamics trajectory, i.e. there is a time scale problem.

The lack of ergodicity due to kinetic trapping at room temperature can be reduced by the methods of replica (Hansmann 1997; Fukunishi et al. 2002), where stochastic exchanges of conformation with replicas of the system that run simultaneously but with a faster dynamics (increased temperature, Hansmann 1997 or artificially modified energy function, Fukunishi et al. 2002; Curuksu et al. 2009) provide new starting points on the energy surface.

An important question then is how to get some good starting points efficiently. On one hand the method becomes prohibitively time consuming when using a non specific increase of the dynamics (temperature) in the searching-replicas. On the other hand, using a specifically modified potential energy function may increase the dynamics only for some kinetically trapped conformations, and be inefficient for others.

In this paper, a method taking full advantage of the replica exchange formalism is derived using a sampling based on learning (Glover 1989), where the ladder of replicas (with a modified energy function) collects sufficient information from the target-replica to propose adapted solutions when this target replica is trapped.

This is made possible by tracking the current sampling of dihedral rotamers in the target replica and recording its cumulative history. The sampling scheme has been implemented and compared with different procedures for learning in the searching replicas on a dinucleotide nucleic acid system. The robustness of the method is indicated by an accurate sampling of dihedral clusters after short simulation time when compared to converged reference simulations. An increase in sampling efficiency more than 10-fold is obtained when compared to a conventional molecular dynamics simulation and non-specific (temperature) replica exchange simulations. In larger biomolecular systems the method can be applied to mutiple dihedral angles to promote global conformational rearrangments. This is illustrated by the study of a small polypeptide where all $\phi/\psi$ backbone dihedral angles in the central 10-residues segment are scaled and define the adaptive replica coordinate.

## 2 Background

Some early adaptive sampling techniques consisting in recognizing conformations sampled before were based on the iterative update of an umbrella potential, using the weighted histogram analysis method (Kumar et al. 1992) in the course of the

simulation. An adaptive umbrella sampling algorithm (Bartels et al. 1997) using the potential energy as the umbrella parameter (formally equivalent to multicanonical sampling, Berg et al. 1991; Wang et al. 2001) was developed to uniformly sample the potential energy surface (PES). Due to the non-specific restraint and the iterative procedure, a long convergence time is a natural drawback for complex systems. Even after convergence residual free energy barriers may impede fast transition between important conformational states (Bartels et al. 1997). In that case, selecting parameters in the potential energy function to explore a specific conformational subspace is a useful solution. Adaptive biasing force schemes go into this direction (Darve et al. 2001; Henin et al. 2010).

Learning processes during MD, consisting to deform the PES or FES (free energy surface), were also pioneered by the local elevation method (LE, Huber et al. 1994) where Gaussian potentials are cumulated to penalize conformations previously visited. The amplitude of the restraint is updated proportionally to the number of time a parameter set (dihedral values in Huber et al. 1994) has been sampled before. Metadynamics (Laio et al. 2002) raises energy wells based on the effective FES as the simulation proceeds and permits to estimate the (imaged) FES once the simulation converges –see also self-healing umbrella sampling (Marsili et al. 2006).

Self-modification of the energy surface during MD can also consist in introducing biasing potentials in the PES such that the surface near a minimum is raised (Hyperdynamics, Voter 1997) and the surface near a barrier or saddle point is left unaffected (or lower as in Gao et al. 2006). Hyperdynamics hence evolves as the simulation proceeds without any advanced knowledge of the hypersurface shape. However it requires either to calculate the Hessian matrix or to minimise its first derivatives numerical approximation at each step (Voter 1997). This can be avoided by relying on a threshold boost energy value and preset biasing potentials as in accelerated MD (Hamelberg et al. 2004). Still these two parameters have to be fine tuned using short pre-production runs and thus cannot guarantee an optimal sampling.

Multicopy search based techniques are a common way to improve sampling and a self-regulating conformational sampling on the PES was developed (Bitetti-Putzer et al. 2006) by propagating replicas of an identical system which compete against each other. To do so Bitetti-Putzer et al. made the replicas share a history dependent variable, i.e the PES deformation operator, and its sum over all replicas was subject to a holonomic constraint. Methods based on the *exchange* of conformations between some parallel replicas (Hansmann 1997), and hence exchange of the thermodynamic states, are also attractive compared to other generalized ensemble approaches because no pre-production runs are required for the determination of the replica weight factors. Moreover a Boltzmann-weighted distribution should be obtained in one (unbiased) replica. The method becomes prohibitively time consuming if temperature is used as the replica coordinate (as in Hansmann 1997). In contrast a non-ergodic sampling cannot be avoided if very specific parameters of the energy function are used as the replica coordinate (Fukunishi et al. 2002). For these reasons, the unbiased replica (target temperature and unmodified force field) may better be seen as a client in a scheme where the replica ladder responds effectively as a function of the current state and history recorded in the unbiased replica. The goal of the current development is to propose a self-regulatory scheme with enhanced transition in the dihedral angle distribution

function. This function is broadly defined, i.e. it can include all or just a subset of dihedral angles. The lost of efficiency on searching the entire conformational space, compared to competing replicas on the PES (Bitetti-Putzer et al. 2006), is replaced by a faster search on the dihedral phase space.

Dihedral angles have already been used to define a coarse grain phase space in molecular dynamics simulations related to peptide folding and refinement (Chen et al. 2005) and as simple constraints to successfully fold a protein (Ripoll et al. 2004). They were also used in recent replica exchange simulations to induce global conformational rearrangments such as protein folding (Kannan et al. 2009) and DNA bending (Curuksu 2009).

## 3 Theory and method

### 3.1 Deterministic dynamics

In biomolecular simulations, classical dynamics can be described by the Hamiltonian or Newtonian approach because the potential energy $U$ depends only on cartesian coordinates. As an alternative to stochastic sampling of all-atom systems such as monte carlo approaches, the molecular dynamics algorithm uses the classical Newton equations of motion to solve deterministically the dynamics of these systems.

For a large number of atoms a numerical integration of Newton equations has to be carried-out using a finite difference approximation, e.g. the Verlet algorithm

$$q(t+\delta t) \;=\; 2q(t) \;-\; q(t-\delta t) \;+\; \ddot{q}(t)\delta t^2 \;+\; O\left(\delta t^4\right) \tag{1}$$

where $\ddot{q}(t)$ is computed from the matrix of forces acting on each atom in three dimensions and thus depends on the chosen expression $U$ for the force field. From the definition of entropy and the equipartition theorem, it is straightforward to show that the state-probability of a molecular system with a given volume and coupled to an infinite heat reservoir at temperature T is proportional to the Boltzmann weight

$$w_B(q,\,p) = e^{-H(q\,,\,p)/k_B T} \tag{2}$$

where $H$ is the Hamiltonian (total energy) function and $k_B$ is Boltzmann constant. However, the trajectory generated by Newtonian dynamics implies the existence of a stability function $\mathbb{L}(t_0 \rightarrow t)$ that characterizes a time evolution of the probability density function $\rho$ in phase space, noted $\rho_t(q,\,p)$

$$\rho_t(q,\,p) = \mathbb{L}(t_0 \rightarrow t).\rho_0(q,\,p) \tag{3}$$

and the Liouville theorem asserts that the total time derivative of $\rho_t$ vanishes:

$$\frac{d\rho_t(q,\,p)}{dt} \;=\; \frac{\delta\rho_t(q,\,p)}{\delta t} + iL\rho_t \;=\; 0 \tag{4}$$

where iL is the Liouville operator. A formal solution for $\mathbb{L}(t_0 \rightarrow t)$ gives:

$$\mathbb{L}(t_0 \rightarrow t) \; = \; e^{-iLt} \tag{5}$$

As time diverges the probability density function becomes stationary

$$\frac{\delta \rho_t(q,\, p)}{\delta t} = 0 \tag{6}$$

so that

$$\mathbb{L}(t_0 \rightarrow t) = 1 \tag{7}$$

which represents an equilibrium ensemble, noted $\rho_{ens}(q\,,\; p)$. For a given stationary probability density function $\rho_{ens}$, the ergodic theorem (Birkhoff 1931) equates the expectation value over the ensemble with the average over time

$$\langle A \rangle_{ens} \; = \; \int A\, \rho_{ens}(q\,,\; p) \; = \; \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \int_0^{\tau} A(q,\, p,\, t)\, dt \; = \; \langle A \rangle_{time} \tag{8}$$

The validity of using Eq. 8 depends on whether the time evolution in (3) is Markovian and thus on the ability of the integrator to sample all important states in a given amount of time. However the time step $\delta t$ in equation (1) needs to be of the order of the highest frequency motion (vibration of covalent bonds) for numerical stability and model accuracy. Given the ruggedness of the energy landscape discussed in the introduction and that high energy conformations are unlikely (Eq. 2), this small time step prevents a continuous MD trajectory from visiting all relevant states in the phase space, meaning that Eq. 8 generally fails (broken ergodicity).

### 3.2 Stochastic sampling from replicas

To reduce the time spent within kinetic traps, one can extend equation (1) by introducing a stochastic term $\Gamma(t)\delta(t, n\tau_0)$ that essentially provides new starting points in configuration space:

$$q(t + \delta t) \; = \; [\, f\,(q(t), q(t - \delta t), \ddot{q}(t)) \; \sqcup \; \Gamma(t)\delta(t, r\,\tau_0)\,] \tag{9}$$

where $\sqcup$ is an exclusive 'or' operator, the function $f$ is the right hand side of (1), $\delta$ is the Kronecker delta, $\tau_0$ is a chosen relaxation time and $r$ is an integer.

The conformations $\Gamma(t)$ can be obtained by a stochastic swap of conformations coming from parallel simulations with faster dynamics according to the replica exchange ansatz. A faster dynamics can be achieved by increasing the temperature (Hansmann 1997) or be focused on some degrees of freedom such as dihedral angles by introducing a biasing potential in the force field (Kannan et al. 2009; Curuksu et al. 2009) that destabilizes one of the rotamer values Gauche+, Trans or Gauche−.

The replica exchange scheme consists in a periodic test for the exchange (every $\tau_0$ period) and an acceptance criterion satisfying the detailed balance equation of micro-reversibility. Given a set of simultaneous conformations for each replica

$$C = \{C_1, C_2, \ldots, C_n\} \tag{10}$$

where $n$ is the number of replicas, the weight of state $C$ in the generalized ensemble of all replicas is:

$$W_{GE}(C) = \exp\left(\sum_{i=1}^{n} \beta_i H(C_i)\right) \tag{11}$$

where $\beta_i = \frac{1}{k_B T_i}$. The detailed balance equation imposed on the swap of conformations, tested when $\delta(t, r\tau_0) = 1$, is:

$$W_{GE}(C)\,\omega(C \to C') = W_{GE}(C')\,\omega(C' \to C) \tag{12}$$

where C′ is a set of simultaneous conformations in which the swap has been carried-out and $\omega$ a transition probability, of going from set C to the new set C′ or back. This can be solved by the Metropolis criterion

$$\begin{aligned} \omega(C \to C') &= \min(1, e^{\Delta}) \\ \Delta &= \beta_i\,(U_i(q_j) - U_i(q_i)) - \beta_j(U_j(q_j) - U_j(q_i)) \end{aligned} \tag{13}$$

The term of kinetic energy $K(p)$ has cancelled-out since particle velocities are rescaled to correspond to their respective temperature in each replica after the exchange (e.g. by $\sqrt{\frac{T_{new}}{T_{old}}}$, Fukunishi et al. 2002).

In (13) the acceptance probability decreases exponentially with the difference in temperatures and potential energies, and will be significant only if the histograms of potential energy between replicas i and j overlap. Keeping a constant temperature and introducing in the force field a different biasing potential $v_i$ for the different replicas increases the acceptance probability since the expression for $\Delta$ simplifies into:

$$\begin{aligned} \Delta &= \beta\,(U_i(q_j) - U_i(q_i) + U_j(q_i) - U_j(q_j)) \\ &= \beta\,(v_i(q_j) - v_i(q_i) + v_j(q_i) - v_j(q_j)) \end{aligned} \tag{14}$$

i.e. $\Delta$ now depends only on the part of the Hamiltonian $v_i$ that differs between the replicas. The added term $v_i$ defined in (15) is chosen with increasing strength $\xi$ along the ladder of replicas and exchanges are attempted only between neighbor replicas in order to maximize the acceptance probability. One replica is simulated without added biasing potential and referred to as the target replica.

## 3.3 Adaptive replicas

### *3.3.1 Short term memory function*

Assuming we do not know the phase space in advance, the selection of new starting points in Eq. (9) has to rely on a self-regulating biasing scheme. The transition between the main rotamer substates Gauche+, Trans and Gauche− can be accelerated in the ladder of searching replicas by a biasing potential which destabilizes these substates (Curuksu et al. 2009):

$$v^k(dx) = \xi \times \begin{cases} ((dx - r)^2 - (R - r)^2)^2 & \text{if } r < dx < R \\ (R - r)^4 & \text{if } dx < r \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

$v^k(dx)$ is a one- or two-dimensional biasing potential which has the shape of a quasi-Gaussian with flat ceiling and destabilizes the rotamer values defined within radius $r$. $dx$ is the distance of the rotamer value at time $t$ to the center of the ceiling, $r$ is the radius of the ceiling and R the radius of the base (larger surface than the ceiling). Six different sets of the biasing potential parameters (0°, 30°, 60°, 90°); (60°, 90°, 120°, 150°); (120°, 150°, 180°, 210°); (180°, 210°, 240°, 270°); (240°, 270°, 300°,330°); (300°, 330°, 360°, 390°), where first and fourth values indicate the base width, cover the three rotamers Gauche+, Trans and Gauche− and the regions in-between. This base width of 90° is chosen to be much larger than the libration of dihedrals in a given substate in experimental structures of biomolecules.

Hence the subspace of each biased dihedral $k$ is discretized into an array of metavariables corresponding to the six regularly spaced intervals defined above. In two dimensions the subspace $(k_1, k_2)$ is discretized into a $(6 \times 6)$ matrix. A dynamical allocation of the biasing potential defined by (15) can be carried-out on the fly on these metavariables. An autocorrelation coefficient $a(k)$ computed at every time step is evaluated in the target replica (using a tolerance index $<10^{-2}$) to decide whether to switch the position of the biasing potential on the metavariable visited in the target replica:

$$a(k) = \frac{\sum_1^\tau \langle \sigma_{k,m}^2 \rangle}{\sum_1^{N_m} \sum_1^\tau \langle \sigma_{k,m}^2 \rangle} \tag{16}$$

where for each dihedral angle $k$, $\langle \sigma_{k,m}^2 \rangle$ is the variance over the relaxation time $\tau$ of the dihedral value with respect to the mean inside the metavariable $m$. $\langle \sigma_{k,m}^2 \rangle$ is computed independently for each metavariable whenever they are visited. $N_m$ is the total number of visits recorded inside the metavariable $m$. The convergence of $a(k) \to 1$ characterizes a convergence of the metavariable average value. When this happens the position of $v^k(dx)$ is switched to this metavariable $m$ in order to promote the sampling of other metavariables in the searching replicas.

The denominator in (16) is kept in memory for each metavariable independently of the other metavariables and thus referred to as short term.

*3.3.2 Long term memory function*

Several positions of metavariables which are accessible in the phase space are determined on the fly by the short term memory function, however some metavariables located far away from the ones visited can be ignored. Moreover inside a metavariable the shape of the energy well can be complex enough to prevent excursion to some sub-minima of these wells. A smoothing procedure of the energy surface by addition of Gaussian-like function centred close to the sampled rotamer value was initially proposed in (Huber et al. 1994).

$$g^k(l,t) = w_G \times \sum_{i=1}^{t} exp\left( - \frac{(k_l(i) - k_l)^2}{2\,\sigma_G} \right) \tag{17}$$

where, for each dihedral angle k, $l$ is a sub-interval defined by centre $k_l$ and width $\sigma_G$. The width $\sigma_G$ is chosen small enough to not mask any distinct sub-minima ($\sigma_G = 22.5°$ as recommended in Huber et al. 1994). The energy penalty $w_G$ added by each Gaussian unit in the sum of (17) is chosen to be many orders of magnitude lower than energy barriers for dihedral transitions ($w_G \sim 10^{-5}$ kcal/mol). The introduction of time-dependent Gaussian functions, in the spirit of Metadynamics (Laio et al. 2002), is possible because each added unit has an infinitesimal height $w_G$ and thus do not violate the detailed balance equation of microreversibility upon replica exchange (Bussi et al. 2006a). The functions $g^k$ are cumulated over the entire potential energy surface visited during the sampling and kept in memory. They flatten the most visited areas such that any kinetic trap will tend to disappear as the simulation proceeds (Bussi et al. 2006b). In light of the work of (Roitberg et al. 2007) on the replica exchange equations and because each added unit has an infinitesimal height, the following equation holds:

$$\begin{aligned} \omega(C_i \Leftrightarrow C_j) &= min(1,\, e^{\,\Delta}) \\ \Delta &= \beta\,(v_i(q_j) - v_i(q_i) + v_j(q_i) - v_j(q_j) \\ &\quad + g_i(q_j) - g_i(q_i) + g_j(q_i) - g_j(q_j)) \end{aligned} \tag{18}$$

3.4 Molecular dynamics simulation details

The ansatz has been implemented as a Fortran subroutine in Amber, a package for biomolecular simulations (Case et al. 2004), and then applied to a nucleic acid dinucleotide d(ApA) kept via distance restraints into a B-DNA helical arrangement (parmbsc0 parameters used for the potential energy, Perez et al. 2007). In the following the ansatz is referred to as the learning enhanced sampling scheme (LESS). The equations of motion were integrated using 1 femtosecond MD moves in explicit water and physiological KCl ions with periodic boundaries. The Velocity Verlet integrator (see Leach 1996) was used. The system was coupled to a heat reservoir with the Berendsen algorithm (Ryckaert et al. 1977) in order to simulate the canonical NVT ensemble of conformations. The same starting conformation obtained after a short stage of thermal equilibration was used for each replica simulation.

A conventional temperature replica exchange MD simulation was used for reference, with a temperature range spanning 300–535 K (exponential scaling) over 26 replicas. Based on published estimates for the transition energy between nucleic acid backbone substates (Curuksu et al. 2009), a LESS simulation with five replicas and Emax levels for the biasing potential $v_i$ of 0 (target replica), 1, 3, 6 and 10 kcal/mol was carried-out. The height of the Gaussian units $w_G$ was fixed at $10^{-5}$ kcal/mol.

For comparaison the simulation of a 5 replicas system containing only the short term memory function and a 2 replicas system containing only the long term memory function were carried out (two replicas were used in the latter because the long term memory function is identical in all replicas except the target replica).

Subsequently, the method was applied to a small protein of 20 residues (Trp-cage protein, PDB-ID: 1L2Y, Neidigh et al. 2002) by scaling all backbone dihedral angles in the 10 central residues and compared to conventional molecular dynamics simulations. Same Emax levels for the biasing potential and number of replicas were used as for the d(ApA) system. The MD protocol was also identical except that the NPT (constant pressure) ensemble of conformations was simulated. Also for the Trp-cage simulation, the SHAKE algorithm (Ryckaert et al. 1977) was used to constrain bond vibrations involving hydrogen atoms and allow a time step of 2 femtoseconds.

A computer code in Fortran that can be compiled as a subroutine in the Amber package as well as input files for the Trp-cage simulations can be downloaded freely from http://www.lcvmwww.epfl.ch/~curuksu/less.

## 4 Results

The pair of dihedral angles $\varepsilon$ $(C4' - C3' - O3' - P)$ and $\zeta$ $(C3' - O3' - P - O5')$ in d(ApA) was used to evaluate the accuracy of the sampling scheme. The coupled $\varepsilon/\zeta$ crankshaft backbone motion implies the presence of two important metastable substates in nucleic acids called BI: $\varepsilon/\zeta$ = t/g− and BII: $\varepsilon/\zeta$ = g−/t (Djuranovic et al. 2004).

In the conventional MD simulation only two substates closed to the BI starting state are sampled after 25 ns (Fig. 1). Both substates around the BI conformation plus additional states in the BII conformation are sampled in the lowest temperature replica of the temperature replica exchange (Fig. 1) with proportions given in Table 1. The energy barriers for dihedral transitions in the d(ApA) system used here could only be efficiently overcome in the temperature replica exchange MD simulation. This type of non-specific replica exchange using fewer replicas (6 and 16) and thus a lower range of temperature indexes also produced an $\varepsilon/\zeta$ sampling througth the BII state but were still not converged after 20 ns (not shown). The LESS ansatz with five replicas produces an $\varepsilon/\zeta$ sampling which is in agreement with the reference temperature replica exchange simulation already after 5 ns as shown in Fig. 1 and Table 1. This is however obtained at a fraction of the computational cost (25 vs. 520 ns in term of effective computation). A cluster analysis of the phase space coverage indicates that the positions of the centroids are in good agreement as well, Table 1.

Without the fine tuning effect of the long term memory function $g^k$, the sampling scheme shows important differences in the phase space coverage after 5 ns. In partic-

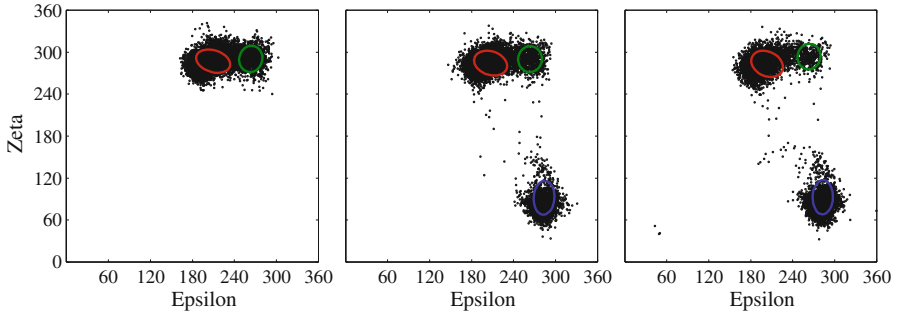


**Fig. 1** Comparison of the phase space distribution of $\varepsilon/\zeta$ dihedral angles in the dinucleotide model with conventional molecular dynamics during 25 ns at 300 K (*left*), the conventional temperature replica exchange ansatz during 20 ns in the target replica (*centre*) and the LESS ansatz during 5 ns in the target replica (*right*). The same number of sampled $\varepsilon/\zeta$ pairs is reported in each sample. Colored ellipses were obtained from a cluster analysis using the two dimensional direct least-square fitting of Fitzgibbon et al. 1999

**Table 1** Analysis of the $\varepsilon/\zeta$ phase space distribution in the dinucleotide model

| Ansatz | Trans/Gauche− | Gauche−/Gauche− | Gauche−/Gauche+ |
|---|---|---|---|
| TREMD (ref.) | 196/282 | 247/292 | 281/83 |
| CMD | 195/280 | 249/289 | –/– |
| LESS | 188/278 | 251/293 | 283/89 |
| S-memory | 189/281 | 223/289 | 283/90 |
| L-memory | 188/278 | 249/295 | 283/82 |
| TREMD (ref.) | 0.53 | 0.16 | 0.32 |
| CMD | 0.74 | 0.26 | 0.0 |
| LESS | 0.58 | 0.08 | 0.34 |
| S-memory | 0.29 | 0.06 | 0.65 |
| L-memory | 0.66 | 0.13 | 0.21 |

Position of the centroids (in degrees, up) and proportions (down) were computed by *k*-means partitioning in *R*. TREMD, temperature replica exchange (25 replicas) used as a reference; CMD, conventional molecular dynamics (25 ns); LESS, learning enhanced sampling scheme (5 replicas); S-memory, LESS ansatz without long term memory function (5 replicas); L-memory, optimized LESS ansatz (see text) without short term memory function (2 replicas)

ular the BII state is over-sampled and the $\varepsilon$-centroid value of the second BI cluster is too low, Table 1.

The second test, regarding the effect of the absence of short term memory function $v^k$, is equivalent to a 2-replicas system (scaled vs. unscaled) since the short term memory function $v^k$ is the only term that differs between the force fields of the searching replicas. Using the same unit height $w_G$ for $g^k$ as used for fine tuning in the complete LESS scheme ($10^{-5}$ kcal/mol) has no effect on the dynamics after 5 ns (not shown). The phase space coverage in the searching replica is almost identical to the target replica and nearly every (>95%) exchange attempt is accepted. Further tests were carried out by increasing the $w_G$ value from $10^{-5}$ to $10^{-3}$ kcal/mol. For $w_G = 10^{-3}$ kcal/mol, the $g^k$ function induces many transitions in the searching

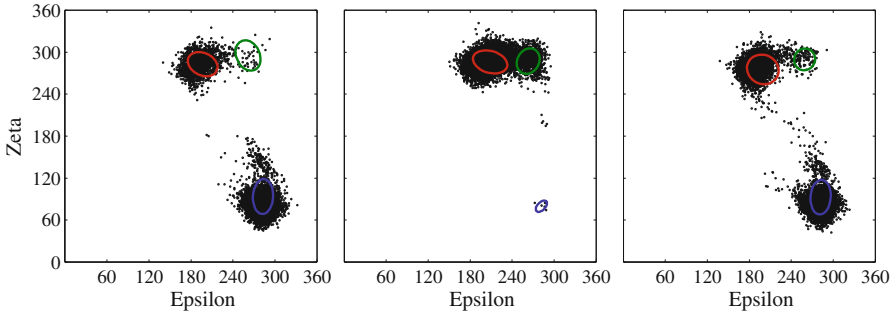


**Fig. 2** Comparison of the phase space distribution of $\varepsilon/\zeta$ dihedral angles in the dinucleotide model in the target replica of the LESS ansatz without the long term memory function (*left*), the 2-replicas LESS ansatz without the short term memory function and a unit height for the Gaussian functions of $10^{-3}$ kcal/mol (*centre*) and $5.10^{-4}$ kcal/mol (*right*). The same number of sampled $\varepsilon/\zeta$ pairs is reported in each sample. Colored ellipses were obtained from a cluster analysis using the two dimensional direct least-square fitting of Fitzgibbon et al. 1999

replica but almost no exchange with the target replica, Fig. 2. An optimal sampling could be found for $w_G = 0.5.10^{-3}$ which permits transitions in the searching replica but also sufficient exchanges with the target replica for providing new starting points in the phase space. The self-regulating 2-replicas approach could in principle be highly efficient in term of computational resources used but this approach would need many prior tests for determining the optimal $w_G$ value for each chosen set of rotamers. Using an infinitesimaly small value of $w_G$ as done in the complete LESS sampling scheme, only to fine tune the sampling of the main rotamers Gauche+, Trans and Gauche− promoted by the short term memory function, is in contrast a general approach for any set of rotamer variables.

To illustrate the potential applications of the method, the Trp-cage mini protein was chosen as a test system. Trp-cage mini protein is one of the smallest polypeptide that adopts a simple fold, consisting in a short $\alpha$ helix, a 3,10 helix and a stabilizing C-terminal poly-proline stretch (Neidigh et al. 2002). However, already for this system size molecular dynamics simulations are limited by the scaling of current softwares and computer architectures, e.g. an increased efficiency is not observed beyond 32 processors with SANDER or 64 processors with PMEMD (components of the AMBER suite, Case et al. 2004), during our benchmark tests on slightly larger systems (not shown). Hence we compare the sampling obtained after 5 ns in the reference replica of LESS using five replicas, with five conventional MD simulations of 5 ns each started with a different distribution of initial velocities. Indeed the production of 25 ns would take roughly five times longer on current computer architecture. For the Trp-cage simulation the pair of dihedral angles $\phi(\mathrm{C}_{i-1}-\mathrm{N}_i-\mathrm{C}\alpha_i-\mathrm{C}_i)$ and $\psi(\mathrm{N}_i-\mathrm{C}\alpha_i-\mathrm{C}_i-\mathrm{N}_{i+1})$ was scaled in LESS using Eq. 15 for each of the 10 central residues.

We found that already after 2 ns some conformations with a rmsd $<6$Å of the native fold (Neidigh et al. 2002) are sampled in the target replica of LESS (Fig. 3) and these become increasingly sampled afterward. In contrast the minimum rmsd value is still $>7$Å after 5 ns in five independent simulations using conventional molecular dynamics. In addition, the target replica of LESS continues to sample both near-native

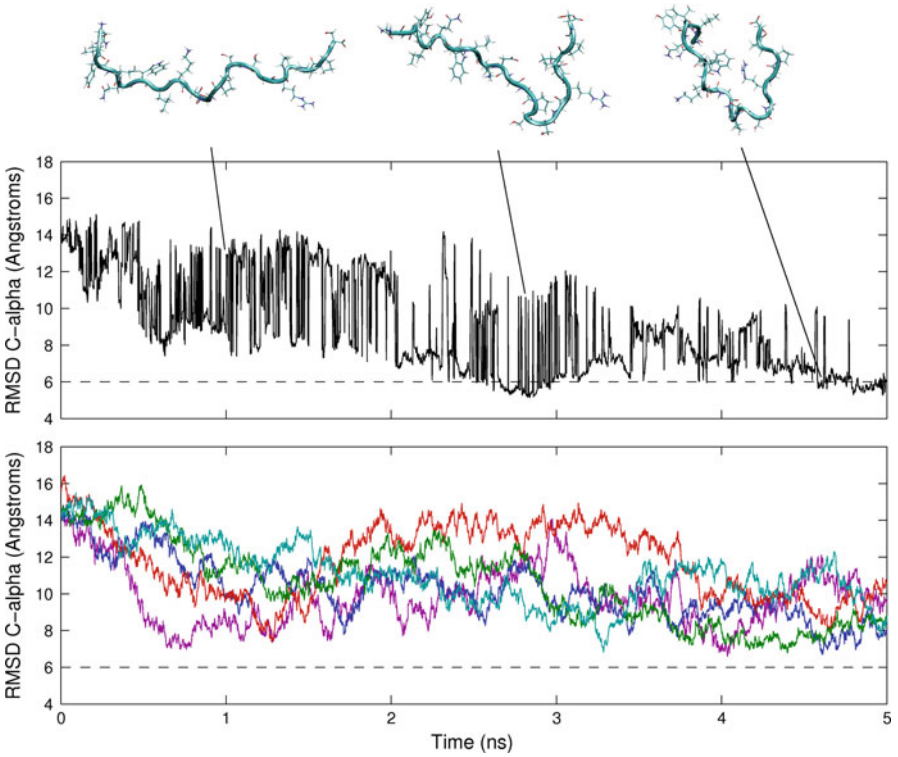


**Fig. 3** C$\alpha$ backbone root mean square deviation (rmsd) for the Trp-cage model with respect to the native conformation (Neidigh et al. 2002) in the target replica of the LESS ansatz (*black line*) and in five independent simulations of conventional molecular dynamics (*colored lines*). Representative simulation snapshots of the Trp-cage structure for three regimes of rmsd value (from *left* to *right*: 12, 9 and 6 Å) are shown as superposed stick and cartoon models. An horizontal hair dashed line indicates a rmsd value of 6 Å. Note a minimum rmsd value in LESS generally lower by 2 Å compared to conventional molecular dynamics

and more extended (rmsd values between 8 and 10 Å) conformations regularly during the simulation. This is in sharp contrast with conventional molecular dynamics where kinetic trapping impedes any fast transition between extended and partially folded conformations, indicated by steady rmsd values between 8 and 14 Å (Fig. 3) on the nanosecond time scale. As a result the current method can be used to estimate the relative probability of the sampled metastable conformations. Future work will go into this direction.

## 5 Conclusion

The proposed learning-enhanced sampling based on replicas is an extension of replica exchange approaches with a biasing potential by including a self-regulation of competing replicas. The results show that a trade off between a short term memory function that promotes transitions between large volumes in phase space (metavariables) and a long term memory function that corrects this sampling by smoothing the

local energy landscape is more efficient and generalizable than one of these techniques taken alone. The current approach is amenable to structure refinement of complex biomolecular systems containing non canonical dihedral substates. In particular the enhanced transition in backbone dihedral angles of DNA and proteins can induce more global conformational rearrangments, such as DNA bending (Curuksu 2009) and protein folding (Kannan et al. 2009). In this study we found that the enhanced transition in backbone dihedrals of the ten central residues in the Trp-cage mini protein leads to conformations close to the native fold in less than 5 ns. In the current method these conformations are sampled altogether with more extended structures, which allows to estimate their relative probability. Future work in this direction will consist in comparing the performance on different biomolecular systems and with respect to other approaches based on non-adaptive replica exchange (Kannan et al. 2009; Curuksu et al. 2009).

The theoretical development relies on several assumptions concerning convergence issues on which research by several laboratories is still on-going. The first (long term memory function) is that the inclusion of time dependent Gaussian functions preserves a canonical distribution in the unscaled, target replica. This is assumed because each Gaussian unit has an infinitesimal height (Bussi et al. 2006a) and the exchange equation (18) is derived from probability density functions including every biasing potential (Roitberg et al. 2007). The second (short term memory function) is that the alternative sampling of metavariables converges toward equilibrium after many switches between these substates (Maragakis et al. 2006). This situation is assumed in the target replica when the autocorrelation coefficient defined for each visited metavariable becomes stationary.

One drawback of the theory behind the method is its limitation to the dynamics of all-atom systems with explicit water, since in implicit solvent and coarse grain simulations the total number of degrees of freedom is likely to be of the same order as the number of dihedral angles. Still a consensus in the modeling community is that multi-scale modeling will dominate the scene as larger and larger biomolecular systems are investigated, and efficient algorithms such as the one proposed in this paper are needed to enhance the synergy of models at different scales.

**Acknowledgements** The author thanks the CLAMV (Computational Laboratories for Analysis, Modeling and Visualization) at Jacobs University, Germany for computational resources, the LCVMM (Laboratory for Computation and Visualization in Mathematics and Mechanics) at EPFL, Switzerland for computational resources and funding and M. Zacharias, J. Maddocks and M. Spichty for useful discussions.

## References

Bartels C, Karplus M (1997) Probability distribution for complex systems: adaptive umbrella sampling of the potential energy. J Comput Chem 18:80–865

Berg BA, Neuhaus T (1991) Multicanonical algorithms for first order phase transitions. Phys Lett B B267:53–249

Beveridge DL, Barreiro G, Byun KS, Case DA, Cheatham TE III, Dixit SB, Giudice E, Lankas F, Lavery R, Maddocks JH, Osman R, Seibert E, Sklenar H, Stoll G, Thayer KM, Varnai P, Young MA (2004) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides: I Research design, Informatics, and results on CpG steps. Biophys J 87:3799–3813

Birkhoff G (1931) Proof of the ergodic theorem. Proc Natl Acad Sci USA 17:656660

Bitetti-Putzer R, Dinner AR, Yang W, Karplus M (2006) Conformational sampling via a self-regulating effective energy surface. J Chem Phys 124:174901

Bussi G, Gervasio FL, Laio A, Parrinello M (2006) Free-energy landscape for $\beta$ hairpin foding from combined parallel tempering and metadynamics. J Am Chem Soc 128:13435–13441

Bussi G, Laio A, Parrinello M (2006) Equilibrium free energies from nonequilibrium metadynamics. Phys Rev Lett 96:090601

Case DA, Darden TA, Cheatham TE III, Simmerling CL, Wang J, Duke RE, Luo R, Maerz KM, Wang B, Pearlman DA, Crowley M, Brozell S, Tsui V, Gohlke H, Mongan J, Hornak V, Cui G, Beroza P, Schafmeister C, Caldwell JW, Ross WS, Kollman PA (2004) AMBER 8 University of California, San Francisco

Chen J, Im W, Brooks CL III (2005) Application of torsion angle molecular dynamics for efficient sampling of protein conformations. J Comput Chem 26:78–1565

Curuksu J, Zacharias M (2009) Enhanced conformational sampling of nucleic acids by an Hamiltonian replica exchange molecular dynamics approach. J Chem Phys 130:10411

Curuksu J (2009) Conformational sampling by molecular mechanics and dynamics simulations applied to the flexibility of nucleic acids. Dissertation, Jacobs University

Darve E, Pohorille A (2001) Calculating free energies using average force. J Chem Phys 115:83–9169

Djuranovic D, Hartmann B (2004) DNA fine structure and dynamics in crystals and in solution: the impact of BI/BII backbone conformations. Biopolymers 73:356–368

Fitzgibbon A, Pilu M, Fisher RB (1999) Direct least square fitting of ellipse. IEEE Trans Pattern Analy Mach Intell 21:446–480

Fukunishi H, Watanabe O, Takada S (2002) On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. J Chem Phys 116:9058–9067

Gao YQ, Zang L (2006) On the enhanced sampling over energy barriers in molecular dynamics simulations. J Chem Phys 125:114103

Glover F (1989) Tabu Search - Part I. ORSA J Comput 1:190–206

Hamelberg D, Mongan J, McCammon JA (2004) Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. J Chem Phys 120:29–11919

Hansmann UH (1997) Parallel tempering algorithm for conformational studies of biological molecules. Chem Phys Lett 281:140–150

Henin J, Fiorin G, Chipot C, Klein ML (2010) Exploring multidimensional free energy landscapes using time-dependent biases on collective variables. J Chem Theory Comput 6:35–47

Huber T, Torda AE, Van Gunsteren WF (1994) Local elevation: a method for improving the searching properties of molecular dynamics simulation. J CAMD 8:695–708

Kannan S, Zacharias M (2009) Folding simulations of Trp-cage mini protein in explicit solvent using biasing potential replica-exchange molecular dynamics simulations. Proteins 76:60–448

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. J Comput Chem 13:21–1011

Laio A, Parrinello M (2002) Escaping free-energy minima. Proc Natl Acad Sci USA 99:6–12562

Leach AR (1996) Molecular modelling. Principles and applications. Addison Wesley, Singapore, pp 356–358

Maragakis P, Spichty M, Karplus M (2006) Optimal estimates of free energies from multi-state nonequilibrium work data. Phys Rev Lett 96:100602

Marsili S, Barducci A, Chelli R, Procacci P, Schettino V (2006) Self-healing umbrella sampling: a non-equilibrium approach for quantitative free energy calculations. J Phys Chem 110:3–14011

Mesirov JP, Schulten K, Sumners DW (1996) Mathematical applications to biomolecular structure and dynamics, IMA volumes in mathematics and its applications. Springer, New York, pp 218–247

Neidigh J, Fesinmeyer R, Andersen N (2002) Designing a 20-residue protein. Nat Struct Biol 9:30–425

Perez A, Marchan I, Svozil D, Sponer J, Cheatham TE 3rd, Laughton CA, Orozco M (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. Biophys J 92:3817–3829

Ripoll D, Vila J, Scheraga H (2004) Folding of the villin headpiece subdomain from random structures. Analysis of the charge distribution as a function of pH. J Mol Biol 339:25–915

Roitberg AE, Okur A, Simmerling C (2007) Coupling of replica exchange simulations to a non-Boltzmann structure reservoir. J Phys Chem 111:2415–2418

Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys 23:327–341

Thirumalai D (1995) From minimal models to real proteins: time scales for protein folding kinetics. J Phys I France 5:1457–1467

Voter AF (1997) Hyperdynamics: accelerated molecular dynamics of infrequent events. Phys Rev Lett 78:3908

Wang F, Landau DP (2001) Efficient, multiple-range random walk algorithm to calculate the density of state. Phys Rev Lett 86:3–2050