Communication: Microsecond peptide dynamics from nanosecond trajectories: A Langevin approach

Andrzej J. Rzepiela,¹,a) Norbert Schaudinnus,¹,b) Sebastian Buchenberg,¹
Rainer Hegger,² and Gerhard Stock¹,c)

¹Biomolecular Dynamics, Institute of Physics, University of Freiburg, 79104 Freiburg, Germany
²Institute of Physical and Theoretical Chemistry, Goethe University, 60438 Frankfurt, Germany
³Freiburg Institute for Advanced Studies (FRIAS), University of Freiburg, 79104 Freiburg, Germany

(Received 14 November 2014; accepted 10 December 2014; published online 22 December 2014)

Based on a given time series, the data-driven Langevin equation (dLE) estimates the drift and the diffusion field of the dynamics, which are then employed to reproduce the essential statistical and dynamical features of the original time series. Because the propagation of the dLE requires only local information, the input data are neither required to be Boltzmann weighted nor to be a continuous trajectory. Similar to a Markov state model, the dLE approach therefore holds the promise of predicting the long-time dynamics of a biomolecular system from relatively short trajectories which can be run in parallel. The practical applicability of the approach is shown to be mainly limited by the initial sampling of the system’s conformational space obtained from the short trajectories. Adopting extensive molecular dynamics simulations of the unfolding and refolding of a short peptide helix, it is shown that the dLE approach is able to describe microsecond conformational dynamics from a few hundred nanosecond trajectories. In particular, the dLE quantitatively reproduces the free energy landscape and the associated conformational dynamics along the chosen five-dimensional reaction coordinate. © 2014 AIP Publishing LLC. [http://dx.doi.org/10.1063/1.4904894]

I. INTRODUCTION

While the structure and the function of proteins are closely interrelated, it is important to realize that biomolecules are often flexible and therefore may provide a large ensemble of thermally populated states.¹–² This conformational distribution can be described by the free energy landscape of the molecule, which defines the relative probabilities of the conformational states and the energy barriers between them.³–⁶ To explore the energy landscape, in particular, molecular dynamics (MD) simulations have been proven useful. However, there is a well-known gap between the integration time step of an all-atom MD trajectory (a few femtoseconds) and the timescale of biophysical processes (typically microseconds (µs) or longer).

To achieve sufficient statistical sampling of such dynamics (i.e., at least 10–100 events), long sampling times (or ≥10¹² time steps) are required. Since a single sufficiently long MD run still represents a considerable computational challenge,⁷ it would be advantageous to achieve the sampling via many short trajectories which can be run in parallel.⁵–¹¹ As a short trajectory can cover only a small part of the conformational space encountered by the system, one needs to invoke an interpolation scheme.

With this end in mind, several groups have adopted Markov state models which approximate the dynamics of the system in terms of a memoryless jump process.⁴–⁶ This is achieved by partitioning the continuous MD trajectory in discrete metastable states such that there is a timescale separation between fast intrastate and slow interstate transitions. In practice, however, various problems have been identified, e.g., concerning the unambiguous definition of Markov states and the sampling of the data. The data-driven Langevin equation (dLE) represents an alternative approach to construct a low-dimensional dynamical model from MD data.¹⁷–²⁴ Based on a given time series, the dLE estimates the drift and the diffusion field of the dynamics, which are then employed to reproduce the main features of the original time series. As the method requires only local information, the input data need not to be Boltzmann weighted in order to warrant that the Langevin model yields correct Boltzmann-distributed results. Similar to a Markov state model, the dLE fields can therefore be constructed using many short trajectories.

In this work, we study the practical applicability of the dLE approach for the interpolation of short trajectories. To this end, we perform extensive MD simulations of a short peptide helix, Aib₈, which was adopted to study energy transport and conformational dynamics previously.²⁵,²⁶ Being an achiral peptide, Aib₈ samples both left-handed and right-handed helical conformations (Fig. 1) and undergoes left-right transitions on a µs timescale at room temperature. Showing a number of transition pathways with various metastable conformational states, the system exhibits complex hierarchical structural dynamics,²⁷ which is nontrivial to reproduce by a low-dimensional dynamical model. To generate the starting conditions for the short trajectories, we first use statistically independent points of a previously performed 4 µs MD run of Aib₈. Alternatively, initial conditions may be generated by some enhanced sampling technique such as, for example,

---

¹Present address: Biozentrum, University of Basel and Swiss Institute of Bioinformatics, Basel, Switzerland.
²A. Rzepiela and N. Schaudinnus contributed equally to this work.
³Electronic mail: stock@physik.uni-freiburg.de
replica exchange methods or metadynamics. Here, we use the CONCOORD algorithm which exploits information from the experimentally derived structure in combination with a set of predefined geometrical constraints.

II. THEORY AND METHODS

A. MD details

Following previous work, all simulations were performed with the GROMACS program suite, using the GROMOS96 force field 43a1 (Ref. 33) to model the α-aminoisobutyric acid (Alb) and 700 SP water molecules. The simulations were performed at constant volume and the temperature of 300 K was controlled with the velocity rescaling algorithm. The equation of motion was integrated using a leap-frog algorithm with a time step of 2 fs. We employed the particle-mesh Ewald method to treat the long-range electrostatic interaction using a cutoff of 1.4 nm and updated the unbonded interaction pair-list every 10 fs. Bonds containing a hydrogen atom were constrained via the LINCS procedure. The output of the trajectories was saved every 10 ps. Following a standard equilibration protocol, first eight trajectories of 4 µs length were run. From one trajectory, 1650 statistically independent structures were obtained and used as initial conditions for the short trajectories of 7.5 ns length.

B. Dimensionality reduction

To reduce the dimensionality of MD data, we performed a dihedral angle principal component analysis (dPCA) which uses the sine/cosine-transformed φ and ψ dihedral angles of the peptide backbone. This avoids possible artifacts due to the mixing of overall rotation and internal motion, which may occur when a PCA using Cartesian coordinates is employed to study large amplitude processes such as folding. As the terminal ends exhibit largely uncorrelated fluctuations, we consider their autocorrelation function

where \( P \) represents the joint probability distribution along the first two PCs \( V_1 \) and \( V_2 \). To study the timescales of the PCs, we consider their autocorrelation function

with \( \delta V_n = V_n - \langle V_n \rangle \) and define its decay time \( \tau_n \) via \( C_n(\tau_n) = 1/e \).

Preceding the dPCA, a low pass filter was employed in order to reduce high-frequency noise of the MD data. We used a three-dimensional delay embedding \(^40\) on the time series of the dihedral angles, i.e., \( \mathbf{x}(t_n) = (x^1(t_n), x^2(t_{n-1}), x^2(t_{n-2}), x^2(t_{n-3}))^T \), which extends the dimensionality of the data by a factor three. As a consequence of the subsequent dPCA, the noise is shifted from the system coordinates to the bath coordinates.

C. dLE

The details of the derivation of the dLE have been discussed previously, hence we only introduce the main idea here. Starting point of the dLE is a given time series \( x(t) \) (the input data), which in our case is obtained from the dPCA of a MD simulation as described above. Under the assumption that the time evolution of \( x(t) \) can be described by a Langevin equation of the type

we aim to determine the fields \( h(x) \) and \( D(x) \) from the input data. Here, \( h(x) \) is the drift field which accounts for the deterministic part of the time evolution, and \( D(x) \) is the diffusion field which contains all spatial dependencies of the stochastic driving. Since the noise term \( \xi(t) \) is unknown, the vector fields \( h \) and \( D \) cannot be calculated directly but need to be estimated by exploiting the statistical properties of the signal. This is achieved by defining a local average \( \langle f(x) \rangle \) of the quantity \( f \) over the neighborhood of a given point \( x \) by

where the sum is taken over all points \( x_i \) of the time series. Due to the Heaviside step function \( \Theta \), the average is only performed in the neighborhood of size \( \epsilon \) around \( x \). In all calculations reported below, we adjusted \( \epsilon \) such that 50 neighbors contribute to the local average. Averaging over Eq. (3), we obtain for the drift field

where \( x_n = \langle x(t_n) \rangle / \delta t \) and \( \Delta x_n = x_{n+1} - x_n \). The diffusion field is obtained from the relation

followed by a Cholesky decomposition in order to get \( D \). The calculation of the drift and the diffusion field in Eqs. (5) and (6) is done locally and “on the fly,” that is, at every propagation step of the dLE. Employing a first-order Euler scheme to integrate Eq. (3), the dLE algorithm requires no continuous trajectory but only pairs of subsequent points. Further details on the practical implementation of the dLE method are given.
in Ref. 41. The dLE program can be downloaded from http://www.theochem.uni-frankfurt.de/~hegger/langevin.tar.gz.

D. CONCOORD

Based on a known molecular structure (usually obtained from experiment), the CONCOORD algorithm exploits the geometric restraints of a protein to generate an ensemble of new conformations. First all interatomic interactions of the starting structure are identified and specific geometric freedom is assigned to each interaction, yielding a set of upper and lower geometric bounds for all interacting pairs of atoms. In a second step, structures that fulfill all geometric bounds are generated by iteratively applying corrections to randomly generated coordinates. To obtain peptide structures for Aib9, we employed default options of CONCOORD. Starting from an equilibrium MD structure of Aib9, each following structure was generated from the previously accepted one in an iterative manner. Using a minimal root mean square difference to all previously accepted structures of 0.8 Å, we discarded conformations that were already sampled. The resulting structures were minimized in vacuum for 1000 steps, solvated, and subsequently minimized for 10 ps with the solvent molecules included. Starting with 5000 CONCOORD conformations, about 10% of the structures resulted in a stable MD trajectory, while the remaining structures were of high energy and therefore caused an abort of the MD integration.

III. RESULTS

A. Conformational dynamics of Aib peptide

As detailed in the Sec. II, we first performed eight 4 μs MD simulations of Aib9 in water at 300 K. These relatively long trajectories serve to reveal the conformational dynamics of the peptide and also represent a reference for the comparison to various dLE calculations discussed below. To obtain a first impression of the structural dynamics of Aib9, Fig. 1 shows a two-dimensional representation of the free energy landscape [Eq. (1)] of the peptide along the first two PCs V1 and V2 of a dPCA (see Methods). As expected from the achirality of Aib9, we find an overall symmetry with respect to V1, where the two main minima correspond to the all left-handed structure L at V1 ≈ −5 and the all right-handed structure R at V1 ≈ 5. Moreover, a number of metastable intermediate states exist that constitute pathways between L and R. Containing a mixture of right-handed and left-handed residues, the intermediate states can uniquely be described by a product state of these chiralities. The definition facilitates a network representation of the global energy landscape (see Fig. S1 in the supplementary material), which reveals all possible pathways of the L ↔ R transition of Aib9. Projections of the intermediate states are identified as local minima of the one-dimensional free energy curves along V1 and V2 displayed in Fig. 2. Also shown are the autocorrelation functions [Eq. (2)] of the first two PCs. Decaying within 500 and 50 ns, they reflect the typical timescale of the L ↔ R transition and of transitions between two intermediate states, respectively. A closer analysis of the structural dynamics of Aib9 shows that these relatively slow chiral transitions are regulated by the fast opening and closing of structure-stabilizing hydrogen bonds (within tens of picoseconds), thus providing a nice example of hierarchical peptide dynamics.

B. dLE modeling

To test the performance of the dLE approach to model the complex dynamics of Aib9, we now consider the above discussed MD runs as input data for subsequent dLE simulations. We employ the first five PCs of the dPCA as coordinate x(t) of the dLE (3), use 50 neighboring data points in the local averages [Eq. (4)], and perform 8 dLE runs of each 4 μs length (see Methods). Figure 2 shows the resulting free energy curves and autocorrelation functions for the first two PCs. Using the full MD data with an aggregate simulation time of 32 μs, we find that the dLE quantitatively reproduces the reference results from direct MD simulations. Similar agreement is found for the third, fourth, and fifth component, see Fig. S2 in the supplementary material. Hence, we have shown that the dLE accurately reproduces the original conformational dynamics of Aib9, given that the input trajectories are long (8 × 4 μs) compared to the slowest timescales of the dynamics.

We are now in a position to pursue our main objective, that is, to show that the dLE method works as well when we use an ensemble (say, hundreds) of short trajectories rather than a few long trajectories as input data. As the performance of such an approach depends crucially on the initial sampling of the system’s conformational space, we carefully need to choose the initial conformations of the short trajectories. To this end, we adopt a single long trajectory and sample from it = 1650 statistically independent conformations. This is done by requesting that the initial
positions are approximately uniformly distributed on the projection onto the first two principal components of the trajectory (see Fig. S3 in the supplementary material). We note that a single 4 µs trajectory still largely undersamples the conformational space of the system, due to the many existing pathways of the $L \leftrightarrow R$ transition (Fig. S1 in the supplementary material). Using these starting structures, we ran 1650 MD simulations of 7.5 ns length and employed the MD trajectories as input data for ten dLE simulations of 4 µs length. Figure 2 shows the resulting free energy curves and autocorrelation functions for the first two PCs. The dLE simulations based on the ns trajectories are indeed found to quantitatively match the MD reference results.

It is interesting to study how the quality of the dLE results depends on the number $N$ and the length $T$ of the short trajectories. To facilitate a simple representation, we characterize the free energy landscape by the ground state population $P_0 = [P(V_1 > 4) + P(V_1 < -4)]/2$ and the autocorrelation function by its exponential decay time $\tau$. Figure 3 shows the dependence of these quantities on $N$ and $T$. Roughly speaking, we find that we need more than 500 trajectories of at least 3 ns length in order to obtain converged values for $P_0$ and $\tau$. This corresponds to $N_m \geq 5 \times 10^5$ data points or 5 µs aggregate simulation time, which appears plausible considering we deal with a five-dimensional conformational space. The length of 3 ns roughly corresponds to the transition time between two metastable conformational states, which again represents a plausible requirement in order to connect different regions sampled by individual short trajectories. We note that in general, the population probability $P_0$ (reflecting the statistics of the problem) converges faster than the decay time $\tau$ which reflects the dynamics of the system. Interestingly, we find that the sampling is even improved by using many short trajectories, e.g., the symmetry $P(V_1 > 4) = P(V_1 < -4)$ is better recovered than in the case of the MD reference results.

So far, we have assumed that conventional MD runs are available which at least roughly capture the dynamics and thus enable us to generate appropriate initial conformations for the subsequent short MD runs. As this strategy may become cumbersome for larger molecules that involve long timescales, alternatively we may generate initial conditions by some enhanced sampling technique. In this work, we adopt the CONCOORD algorithm which represents a particularly simple and efficient method to generate plausible molecular structures. As described in the Sec. II, we generated 5000 CONCOORD structures of $\text{Aib}_{10}$, from which about 600 resulted in stable MD trajectories. Representing the initial structures with respect to the first two PCs of a dPCA (Fig. S3 in the supplementary material), we notice a higher density in the central region and around the two minima, meaning that the conformations are not as equally distributed as in the case of the MD initial conditions. Interestingly, though, we find that the PCs obtained for all MD runs and for short trajectories using MD and CONCOORD initial conditions are quite similar, see Fig. S4 in the supplementary material.

Using short trajectories with CONCOORD initial conditions as input data, we performed ten dLE simulations of 4 µs length. Figure 2 reveals that these calculations at least qualitatively reproduce the MD reference data, although the barriers are somewhat too low which results in a faster decay of the corresponding autocorrelation functions. Interestingly, though, Fig. 4 suggests that the deviations are not caused by a lack of data points. Rather we find that the dLE calculations converge to incorrect values of $P_0$ and $\tau$, which are a consequence of the unbalanced distribution of the CONCOORD initial conditions (Fig. S3 in the supplementary material). Typically we find that in lowly sampled regions, our CONCOORD structures have artificially high energy such that the subsequent MD simulation becomes unstable. We note that stable trajectories can be obtained by employing longer equilibration times; however, by then the system may already have left the conformational region of interest.

FIG. 3. Convergence of the dLE simulations with respect to the number $N$ (panels (a) and (c), with $T = 7.5$ ns), the length $T$ (panels (b) and (d), with $N = 1650$), and the total number of data points $N_m$ of the short trajectories, using MD initial conditions. Shown are the ground state population probability $P_0$ and the decay time $\tau$ of the autocorrelation function of $V_1$. The error bars reflect the standard errors of the corresponding ensemble. Horizontal lines indicate the reference values of $P_0$ and $\tau$, respectively.

FIG. 4. As Figure 3, but for dLE simulations using short trajectories with CONCOORD initial conditions.
IV. CONCLUSIONS

Adopting extensive MD simulations of the unfolding and refolding of a short peptide helix, we have shown that the dLE quantitatively reproduces the free energy landscape and its conformational dynamics along the chosen reaction coordinates. Because the dLE requires only local information [cf. Eq. (4)], it is able to describe microsecond peptide dynamics from nanosecond trajectories. Similar to the popular Markov state models, the dLE approach therefore holds the promise of predicting the long-time dynamics of a biomolecular system from relatively short trajectories which can be run in parallel. As may be expected, however, in practice the dLE approach suffers from similar problems as the Markov modeling, such as the definition of local averages and the sampling of the data.

The practical applicability of the dLE approach depends crucially on the initial sampling of the system’s conformational space, i.e., one needs to carefully choose the initial conformations of the short trajectories. Here, we have considered two strategies. First, we have assumed that conventional MD runs are available which at least roughly visit the mainly populated regions of the system’s free energy landscape. Since the sampled conformations are only used as initial conditions for the short trajectories, this can be achieved, e.g., by running a long MD simulation at high temperature. Although these input data are vastly undersampled, subsequent dLE simulations converge to the correct results in a straightforward and controlled manner. Alternatively, initial conditions may be generated by some enhanced sampling technique. As a particularly simple and highly efficient method, we have adopted the CONCOORD algorithm. While the resulting dLE calculations reproduce the MD reference data at least qualitatively, it has been found difficult to systematically converge the sampling, at least when we use the standard version and settings of CONCOORD. Further work to combine the dLE with some promising enhanced sampling strategy is in progress. The dLE program can be downloaded from http://www.theochem.uni-frankfurt.de/~hegger/langevin.tar.gz.

ACKNOWLEDGMENTS

We thank Bert de Groot, Björn Bastian, and Peter Hamm for numerous instructive and helpful discussions. This work has been supported by the Deutsche Forschungsgemeinschaft.