21 NOVEMBER 2012 Volume 137 Number 19

AIP | The Journal of Chemical Physics

jcp.aip.org

Accounting for the kinetics in order parameter analysis: Lessons from theoretical models and a disordered peptide

Ganna Berezovska, Diego Prada-Gracia, Stefano Mostarda, and Francesco Rao[a)]
Freiburg Institute for Advanced Studies, School of Soft Matter Research, Freiburg im Breisgau, Germany

(Received 17 August 2012; accepted 8 October 2012; published online 19 November 2012)

Molecular simulations as well as single molecule experiments have been widely analyzed in terms of order parameters, the latter representing candidate probes for the relevant degrees of freedom. Notwithstanding this approach is very intuitive, mounting evidence showed that such descriptions are inaccurate, leading to ambiguous definitions of states and wrong kinetics. To overcome these limitations a framework making use of order parameter fluctuations in conjunction with complex network analysis is investigated. Derived from recent advances in the analysis of single molecule time traces, this approach takes into account the fluctuations around each time point to distinguish between states that have similar values of the order parameter but different dynamics. Snapshots with similar fluctuations are used as nodes of a transition network, the clusterization of which into states provides accurate Markov-state-models of the system under study. Application of the methodology to theoretical models with a noisy order parameter as well as the dynamics of a disordered peptide illustrates the possibility to build accurate descriptions of molecular processes on the sole basis of order parameter time series without using any supplementary information. © *2012 American Institute of Physics*. [http://dx.doi.org/10.1063/1.4764868]

INTRODUCTION

Order parameters are conventionally used for the characterization of complex molecular processes.[1,2] Inter-atomic distances or a combination of them are common choices, providing an intuitive description in terms of free-energy projections.[3–7] Unfortunately, it has been repeatedly found that reduced descriptions based on order parameters are often inaccurate.[8–14] The origin of the failure is due to overlaps in the order parameter distribution, i.e., configurations with different properties corresponding to the same value of the coordinate, making the discrimination between states ambiguous.[10,12] From the point of view of the dynamics, spurious recrossings at the borders result in lower free-energy barriers and artificially faster kinetics.[15]

To improve on this situation, a new arsenal of tools emerged making use of complex networks and the theory of stochastic processes. Configuration-space-networks, referred to as Markov-state-models when the Markov property is satisfied, provide high resolution free-energy landscapes of complex molecular processes.[8,9,16–21] The main idea behind this approach is to map the molecular dynamics onto a transition network. Nodes and links represent sampled system configurations (*microstates*) and the transitions between them as observed in the molecular dynamics, respectively. The resulting transition network stores the entire kinetical information in the form of link weights and node connectivity, providing a compact representation of the molecular trajectory. Within this approach, free-energy representations are obtained in a more universal way without using arbitrary projections on order parameters. Both thermodynamics and kinetics come from the analysis of the transition network with methods like network clusterization algorithms,[17,20,22] network flow analysis,[9,21,23] and spectral methods.[18,19]

Besides advantages, a general strategy towards microstates building is still missing, making the initial mapping of the molecular trajectory onto a network non-trivial. Even for the well-studied case of structured peptides and the folding of small proteins, there is no consensus on the best practice.[21,24] Moreover, a broad set of problems including molecular association,[25,26] analysis of intrinsically disordered proteins[27,28] and liquids[29,30] are very hard to tackle with the current methodology. As shown for the case of water, *ad hoc* strategies are needed.[31–33] Ironically, many of these processes can be qualitatively described by the analysis of conventional order parameters.

In this work, an effort is made to reconcile the intuitive aspect of order parameters with the predictive power of transition networks, overcoming some of the limitations of both methodologies. The strategy couples a recently developed framework for the analysis of single molecule traces[34–36] with network clusterization techniques in order to obtain accurate Markov-state-models from conventional order parameter time series. Application to a two-state model in the presence of noise and molecular dynamics simulations of a disordered peptide are presented. Our results suggest a general approach to analyze molecular processes with high accuracy on the sole basis of conventional order parameter time series.

THEORY

Accurate kinetic models from conventional order parameters

Order parameters allow for intuitive descriptions of molecular processes. Unfortunately, such descriptions can

a)francesco.rao@frias.uni-freiburg.de.

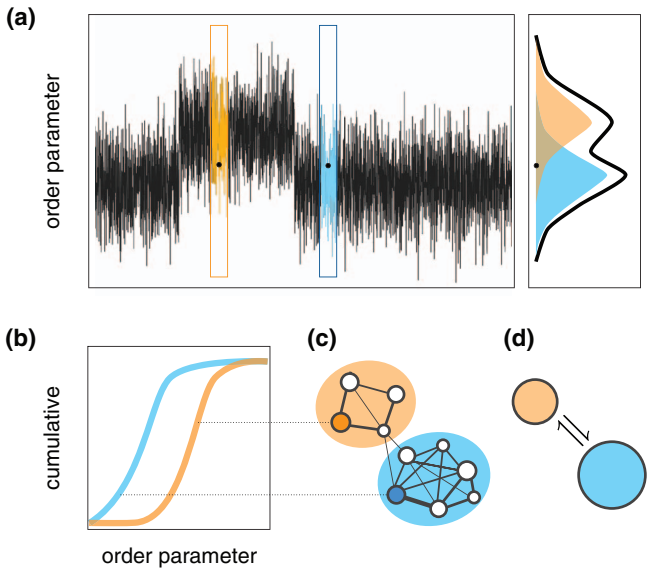

FIG. 1. Local-fluctuations order parameter analysis. (a) The time series of an order parameter and its distribution (black lines). Two snapshots with the same value of the order parameter but belonging to different states are characterized by distinct local distributions (orange and light blue, respectively). (b) A Kolmogorov-Smirnov test evaluates the similarity of the cumulative of the two distributions. Snapshots with similar distributions belong to the same *microstate*. (c) Microstates and the transitions between them represent nodes and links of a configuration-space-network, respectively. Network clusterization techniques allow the lumping of kinetically homogeneous regions of the network into states (orange and light blue areas). (d) States are used to build a reduced Markov-state-model of the original molecular process (see the Theory section for further details).

be highly inaccurate due to the presence of *overlaps*, i.e., configurations with different properties corresponding to the same value of the order parameter[10,12] (black dots in Fig. 1(a)). An important improvement in this respect was the introduction of configuration-space-networks, providing accurate and concise descriptions of the system kinetics and thermodynamics.[8,9,16–21] Apart from recent efforts to combine several order parameters to describe protein folding,[37] network approaches are not directly applicable for the analysis of conventional order parameter time series.

To overcome this impasse, a potential strategy makes use of a recently introduced framework for the analysis of single molecule experiments.[34–36] The idea consists in exploiting the information of the underlying free-energy surface by analyzing the fluctuations of a general coordinate that qualitatively describes the dynamical process. This is possible because two points with similar values of the coordinate but belonging to different states are characterized by distinct local distributions (see orange and blue areas in Fig. 1(a)). In order to characterize the molecular process, this information was processed in different ways, going from the concept of state "candidate" based on escape times[34,36] to cut-based free-energy profiles.[35] The latter approach presented a procedure that will be used below to build system microstates from local fluctuations of an arbitrary coordinate.

In this contribution a protocol to analyze common order parameter time series is presented to overcome the limitations due to overlaps. In this way, accurate kinetic models of the underlying dynamical process are obtained, providing a way to effectively use the information in order parameter time series. The proposed protocol works as follows: the first step of the analysis constructs a configuration-space-network where the nodes represent system *microstates* on the base of the order parameter time series. Depending on the complexity of the underlying dynamical process, the initial network usually contains hundreds to thousands of nodes (Fig. 1(c)). In a second step, this network is clusterized into kinetically homogeneous *states*. The clusterization process insures that no relevant barriers are lumped into the few (usually less than ten) states. As such, the complexity of the process is reduced through a kinetics-aware procedure by finding the slowest relaxations of the dynamical process. Being the inter-state transitions slow with respect to all other fluctuations, the detected states effectively represent free-energy basins.[17] Finally, a reduced kinetic model (Markov-state-model) reproducing the same relaxation dynamics of the original trajectory is built on the basis of the detected states (Fig. 1(d)). The accuracy of the predicted kinetics is monitored by first passage times analysis.

Most of these tools were developed elsewhere by us or others, here we combine them for a more accurate analysis of conventional order parameters. The technical details of all steps are described below in the Methods section. A code to reproduce the present analysis is freely distributed at the website www.raolab.com.

METHODS

Microstate building

The microstates accounting for the local fluctuations of the order parameter were built as suggested in Ref. 35. As such, each time point of the trajectory t_i was associated with a corresponding time window $[t_i - \tau/2, t_i + \tau/2]$. Two time points were considered to be *similar* if they had comparable distributions of the order parameter. Snapshot similarity D was estimated by comparing the cumulative of the two distributions via a Kolmogorov-Smirnov test[38] which checks whether two samples belong to the same distribution or not (Fig. 1(b)). (It is important to note that other metrics can be used to compare distributions; a nice discussion is found in Ref. 34.) D was defined as the maximum difference of the two cumulative distributions. Two samples belong to the same distribution and, thus, to the same microstate, if the condition $D \leq \zeta\sqrt{2/\tau}$ was fulfilled. The acceptance cutoff ζ corresponds to a certain confidence level. τ and ζ being related, we fixed the latter value to 0.5 and let τ vary. Comparisons were made along the trajectory using the leader algorithm in a way that each time point was associated to a microstate at the end of the procedure.[35,39]

The configuration-space-network

The resulting time series of microstates was mapped onto a configuration-space-network (Fig. 1(c)).[8,17,40] Microstates represent network nodes and a link between them exists if they are successively visited along the molecular trajectory. For each link detailed balance was imposed by making an

average of the number of transitions in both directions. This was only partially necessary because the original trajectories mostly satisfied detailed balance already.

Markovian clustering algorithm

The configuration-space-network is clusterized into kinetically homogeneous states (e.g., orange and blue areas in Fig. 1(c)). This is insured by the application of the Markov-clustering-algorithm (MCL).[41] We showed elsewhere that the resulting partition reflects the properties of the underlying free-energy surface where the detected states represent free-energy basins.[17]

MCL is based on the behavior of random walkers on the network. The algorithm works as follows: (i) normalize each column of the network in order to obtain a transition matrix T (i.e., sum over columns equal to one); (ii) compute $T^2 = T \times T$; (iii) take the pth power ($p > 1$) of every element of T^2 and normalize each column to one; and (iv) go back to step (ii). After several iterations MCL converges to a matrix $T_{\mathrm{MCL}}(p)$ invariant under transformations (ii) and (iii). Only a few lines of $T_{\mathrm{MCL}}(p)$ have some nonzero entries that give the clusters as separated basins (there is in general exactly one nonzero entry per column). Step (iii) reinforces the high-probability walks at a short time scale at the expense of the low-probability ones. The parameter p tunes the granularity of the clustering. If p is large, the effect of step (iii) becomes stronger and the random walks are likely to end up in small "basins of attraction" of the network, resulting in several small clusters. In free-energy language, the value of p determines the minimum barrier height detected by the algorithm. Small values of p split the network along the highest barriers.[17] As p increases lower barriers are detected.

Due to finite size sampling the MCL algorithm can return several states but only a few of them are statistically significant. For practical applications, the relevant interval of p may vary from 1.1 to 1.8:[17,20] best choice is to keep p as small as possible in a way that states are detected with minimal amount of noise. For dynamical processes with a clear separation of timescales (as the examples presented here), results are robust for different values of p.

Reduced kinetic model

The states detected by MCL were used to build a reduced kinetic model of the dynamical process (Markov-state-model, Fig. 1(d)). Transition probabilities were estimated from the original configuration-space-network by summing up all the links connecting any two states. It is worth noting that while the reduced kinetic model satisfies the Markov property because it reflects the slow relaxations only,[21] this is not generally true for the starting transition network.[16] (This property is in any case not needed when it comes to network clusterization.[17,20])

First passage time distributions

The kinetic similarity between the original trajectory and the different kinetic models was investigated by comparing the distribution of first-passage-times (fpt).[21] This is the distribution of times to reach a given target conformation from any other snapshot of the trajectory. For the kinetic models new trajectories were generated by running a random walk on the network. Arrival times depend on the definition of the target only and not on the detailed decomposition of the trajectory. For the original trajectory, the target state is defined as snapshots with a value of the order parameter smaller than a cutoff while for the random walk trajectories the target is represented by a network state.

Simulation details

Two-state model

A stochastic two-state model with transition probability $p_{ij} = 0.01$ ($i \neq j$) was built. The latter probability completely defines the kinetics of the model. The time evolution was monitored by an artificially defined order parameter Q in a way that Q cannot distinguish between the two states unambiguously (see dark blue line in Fig. 2(a) for a sample time series). Q was associated to energy functions $U_1 = \alpha Q^2$ and $U_2 = \alpha(Q-1)^2$ for the two states, respectively. As such, the first and second states preferentially visit different values of the order parameter. Within each state, the time evolution of Q was obtained by a conventional metropolis criterion $\min[1, \exp(-\beta \Delta U)]$ with β regulating the amount of fluctuations. Choosing $\alpha = 16$, values of β close to 1 suppress fluctuations while smaller values enhance them (most of the treatment below was done for the case of large fluctuations with $\beta = 0.3$). This model was used to generate order parameter time series of 10^5 steps.

Disordered peptide

Simulations of the $(\mathrm{Gly\text{-}Ser})_2$ flexible linker peptide were performed using the all-atom CHARMM force-field version 27[42,43] as implemented in the program ACEMD.[44] All calculations were done on NVIDIA GTX680 graphics cards. The system was solvated into a box containing 1560 TIP3P waters. After equilibration, the molecular dynamics was run for 1 μs in the NVT ensemble at 300K, using the Langevin algorithm. An integration time step of 4 fs was used by rescaling the hydrogen mass to 4 amu together with mass repartitioning.[45] The characteristic relaxation times of the $(\mathrm{Gly\text{-}Ser})_2$ peptide being in the ns regime (see the Results section), we concluded that the sampling time of 1 μs was enough to provide an equilibrium trajectory (see also Ref. 46 for a free energy comparison).

The radius of gyration was calculated with the program WORDOM,[39,47] neglecting all peptide hydrogens.

RESULTS

A two-state process with large fluctuations

The protocol described in the Theory and Methods sections was applied to the time series of a generic order parameter Q with an underlying two-state dynamics. This model

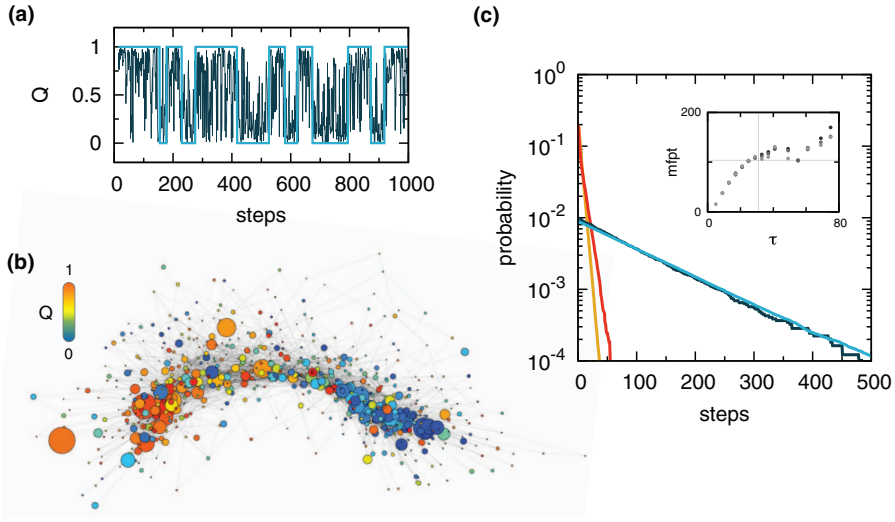

FIG. 2. A two-state model with large fluctuations. (a) The time series of the order parameter Q and the corresponding states obtained with the local-fluctuations analysis are shown in dark and light blue, respectively. (b) Configuration-space-network. Nodes and links represent local-fluctuations based microstates and the transitions between them as observed in the time series, respectively. Nodes are color coded according to the value of Q at the center of the time window while the size is proportional to the statistical weight. (c) First-passage time distributions obtained by different analysis techniques. The distribution corresponding to the original two-state model is shown in dark blue. Distributions obtained from the local-fluctuations reduced kinetic model, a *naive* two-state model, and along the original time series using $Q = 0.5$ as state separator are shown in light blue, yellow, and red lines, respectively. The dependence of the mean-first-passage-time (mfpt) with the time window size τ is shown in the inset (MCL p parameter of 1.3 and 1.4 for gray and dark gray points, respectively). The correct value of the mfpt and the time window value chosen for the analysis are shown as horizontal and vertical lines, respectively.

served as a benchmark for the proposed network approach since the underlying kinetics is known by construction. The amount of fluctuations of the order parameter Q is controlled by a parameter β (see the Methods section). To characterize the two-state behavior directly from the time series a *naive* strategy would take the distribution of Q, looking for the minimum separating the two states. The value of Q at the minimum, 0.5 in this case, would define the separator between the states. If the fluctuations around the separator are small, Q is a good order parameter in the sense that the number of crossings of the separator represents a good estimate of the barrier between the two states. This approach is valid for the case of small fluctuations but breaks down as soon as the overlap between the states increases (i.e., large fluctuations). In the latter case, most of the contributions to the separator come from re-crossings: transitions passing the separator come back very quickly to the initial state without reaching the end state. The origin of recrossings is mostly due to a sub-optimal choice of the order parameter rather than to a real physical property.

The time series of Q for the case of large fluctuations ($\beta = 0.3$) was characterized with different approaches. A particular stringent test consists in the calculation of the first-passage-time distribution to one of the two states. The correct distribution is shown as a dark blue line in Fig. 2(c) (mean-fpt of 103.85 steps). This distribution is greatly influenced by the definition of the target state. If the latter is correctly defined, the resulting distribution overlaps with the one calculated from the original two-state model. When the target state was chosen as $Q < 0.5$, a fpt distribution calculated along the trajectory resulted in a much faster kinetics (red line, Fig. 2(c)). With a mean-first-passage-time (mfpt) of 7.56 steps, the kinetics obtained by this analysis was one order of magnitude faster with respect to the input model.

A better description was obtained by analyzing the time series in terms of local-fluctuations and Markov-state-models. Application of the protocol presented in the Theory and Methods sections with a time window $\tau = 30$ found 636 microstates. The resulting configuration-space-network is depicted in Fig. 2(b). Nodes and links are microstates and transitions between them as observed along the time series, respectively. Colors reflect the value of Q at the center of the time window (average instantaneous values). It is expected that nodes with similar colors to be tightly connected together. While this is true in general, there are several nodes that do not follow this rule (e.g., orange nodes in the blue region). This reflects the presence of large fluctuations in the time series. That is, while the instantaneous value of Q would be compatible with the orange state ($Q \approx 1$), the distribution along the time window reflects the blue state ($Q \approx 0$, compare also with the time series depicted in Fig. 2(a)).

Network clusterization by the MCL algorithm with $p = 1.3$ detected exactly two states. Their evolution is shown as a step curve in Fig. 2(a) (light blue). These two states were used to build a Markov-state-model with transition probabilities directly calculated from the original transition network (see the Methods section). The fpt distribution was estimated by generating a new time series of 10^6 steps from this model. Strikingly, this distribution perfectly overlapped the one calculated on the original model as shown by the light blue line of Fig. 2(c). Using a time window $\tau = 30$, the resulting mfpt is of about 114.58 steps, very close to the correct value of 103.85. This is not the case when the Markov-state-model was built by estimating the transition probability from the number of times the separator

$Q = 0.5$ was crossed. As expected, the resulting fpt distribution provided very poor results, the inter-state separator being dominated by recrossings (mfpt = 5.2, yellow line, Fig. 2(c)).

It is important to note that the two kinetic models presented here (corresponding to the light blue and yellow lines) were built using the same starting information, i.e., the time series of the order parameter Q. Consequently, the improvement obtained by the local-fluctuations analysis is purely due to the different strategy applied rather than the use of supplementary information.

Some predictions of the network protocol may depend on the choice of the time window τ (see the Methods section). Caflisch and co-workers found a large range of validity for this parameter.[35] For the present two-state model, the time window range was evaluated against the mfpt to reach the target state (inset of Fig. 2(c)). For small time windows the value of the mfpt was smaller then the correct one (horizontal line). This is due to the inaccurate detection of states, with large fluctuations artificially considered as real transitions. As the value of τ increased, the mfpt first converged to values close to the correct one ($30 < \tau < 60$) and then increased again. It was found that when the time window was too large, some transitions were missed, resulting in an overall slower kinetics (i.e., large mfpt). Consequently, the behavior of the mfpt as a function of τ suggests a reasonable way to choose the time window as the location of the first slope change ($\tau = 30$, vertical line), just before the convergence region. Essentially identical results were found for MCL parameters $p = 1.3$–1.4 (gray and dark gray points in the inset of Fig. 2(c), respectively).

Summarizing, our procedure is characterized by two parameters: the time window τ and the MCL parameter p. The former has to do with the amplitude and frequency of the order parameter fluctuations. As such, it is critical to find an appropriate range of τ to avoid misleading descriptions of the kinetics. Although this is not trivial, we found that mfpt analysis gives useful information towards this aim. On the other hand, the MCL parameter p is related to the barrier height separating the states (see the Methods section and Ref. 17). Consequently, there is no optimal value for p because it depends on the minimal state-separating barrier height that one wants to consider. In the present model system, where only two states are present and sampling is almost ideal, $p > 1.2$ returned the expected two states. This is not the case for smaller values of p, i.e., the barrier separating the states is too small compared to the resolution power of that range of the parameter ($p = 1$ returns one cluster by construction). In general, multiple states separated by heterogeneous barrier heights are hierarchically resolved by MCL by incrementally increasing the value of p, as it was already shown elsewhere for the case of the alanine dipeptide.[17] Finally, more realistic problems (see the next section) are always affected by finite sampling. As a consequence, larger values of the parameter p tend to break apart the peripheries of the states, generating a number of clusters with very low populations. These artificial splittings can influence the description of the kinetics without affecting the thermodynamics.

Multi-state dynamics of a disordered peptide

Gly-Ser peptides have been used in experiments for quite some time as flexible linkers.[48,49] Short stretches of this peptide are interesting from a theoretical point of view because they are computationally tractable, presenting nontrivial conformational disorder.[22,46,50] In this section the local-fluctuations analysis is applied to the dynamics of a $(\text{Gly-Ser})_2$ peptide. To this aim, a long molecular dynamics simulation of 1 μs was performed. It has been shown[46] that the radius of gyration R_g qualitatively describes the conformational disorder of this peptide, suggesting the presence of multiple states. A time series stretch and the distribution of the R_g are shown as a dark blue line and a gray area in panel (a) and (b) of Fig. 3, respectively.

Application of the local-fluctuations analysis on the R_g time series revealed the presence of five major states ($\tau = 300$ ps and $p = 1.3$; these two values were chosen following the mfpt based strategy presented in the previous section; see also inset of Fig. 4) for a cumulative population of 98%. Figure 3(b) shows the R_g distribution of the five states (colored lines). Interestingly, the five distributions are largely overlapping making their detection impossible by simply looking at the total distribution. In fact, the total distribution suggested no more than three states as indicated by the number of peaks (gray are in Fig. 3(b)). Interestingly, the presence of the five states was already apparent in the raw R_g time series (dark blue line, Fig. 3(a)). Those states were hidden in the total distribution due to the large fluctuations, reiterating the idea that using free-energy projections for the characterization of molecular processes can be ambiguous.

From a structural point of view, all five states are well characterized. A molecular representation of the three most compact states is shown in Fig. 3(c). The most compact one, coded in light blue in Fig. 3(b), corresponds to a loop-like structure typically found in β-strands turns. This structure is stabilized by a hydrogen bond between the first backbone oxygen O_1 and nitrogen N_4. This conformation has a population of around 23% and $R_g \approx 3.1$ Å. The second state has a population of 22% and $R_g \approx 3.5$ Å (green curve in Fig. 3(b)). Its topology is very similar to the turn-like structure but it is more disordered due to the formation of a non optimal backbone hydrogen bond. The third state instead is stabilized by the interaction of the side chain oxygen of SER_2 with the backbone nitrogen N_4 (population of 20% and $R_g \approx 3.7$ Å, yellow curve in Fig. 3(b)). In this structure, the side chain substitutes the backbone oxygen as a partner in the hydrogen bond, acting as a trap towards further compaction. Finally, the last two states (orange and red curves in Fig. 3(b)) are rather unstructured, providing similar realizations of almost completely extended conformations.

To check whether the kinetics of the Markov-state-model built on top of the five detected states reflected the same dynamics of the original trajectory, a fpt analysis was performed as shown in Fig. 4. The fpt distribution to the compact state was calculated on a trajectory originated from the Markov-state-model (light blue line). To compare it with the molecular dynamics simulation, the fpt to conformations with

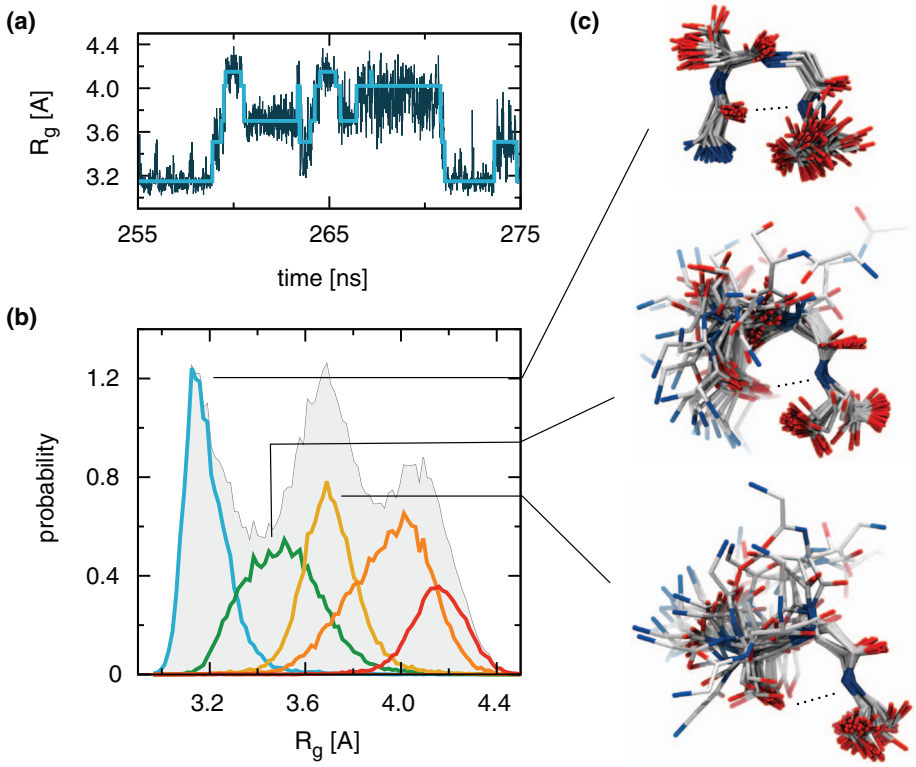

FIG. 3. The $(Gly\text{-}Ser)_2$ peptide. (a) A time series stretch of the radius of gyration (R_g, dark blue) and of the detected states after the local-fluctuations analysis (light blue). (b) Distribution of the radius of gyration. The distributions from the entire trajectory and for the five most populated states are shown as a gray area and colored lines, respectively. (c) Structural characterization of the three most compact states. For each of them, 50 random snapshots were overimposed.

$R_g = 3.16 \pm 0.01$ (the top of the first peak in the R_g distribution) was calculated along the original molecular dynamics trajectory (dark blue line in Fig. 4). This represents a good estimate of the fpt to the compact state at long times. Strikingly, the two curves nicely overlapped. An exponential fit of the data, i.e., $\sim\exp(-t/t_r)$, showed a relaxation time of 6.5 ns in both cases.

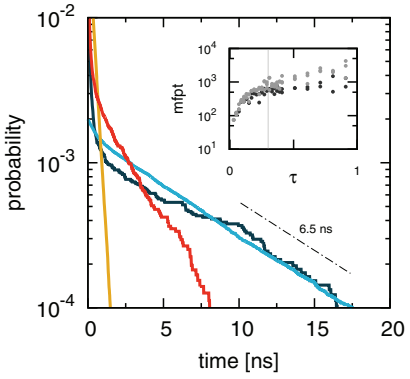

FIG. 4. The $(Gly\text{-}Ser)_2$ first-passage-time distribution. Distributions obtained from the local-fluctuations analysis, a *naive* two-state model and along the original time series are shown in light blue, yellow, and red, respectively. For the latter two cases the target of the relaxation was $R_g < 3.4$. The relaxation kinetics to the compact state defined as the conformations belonging to the first peak of the R_g ($R_g = 3.16 \pm 0.01$, see Fig. 3(b)) is shown in dark blue. The dependence of the mfpt with the time window size τ is shown in the inset (MCL p parameters 1.2 and 1.3 are shown in dark gray; in gray the values for 1.4, 1.5, and 1.6). The time window chosen for the analysis, $\tau = 300\ ps$, is indicated with a vertical line.

The fpt distribution calculated along the original trajectory using as target state all conformations with $R_g < 3.4$ (the position of the first minimum of the R_g distribution) provided a much faster kinetics (red line in Fig. 4). However, processes involving barrier crossings are expected to have the same relaxation kinetics when either the whole state is selected as a target or just the most probable conformation of the state (i.e., conformations at the peak of the order parameter distribution).[21] Consequently, the disagreement between the fpt distribution to $R_g < 3.4$ and to a more stringent target (i.e., $R_g = 3.16 \pm 0.01$), provides strong evidence that the correct estimation of the kinetics cannot be obtained directly from the R_g distribution due to the presence of recrossings. As expected, this becomes even worse when the Markov-state-model is built directly from the R_g distribution, i.e., by choosing as states the portions of the distribution separated by minima and as transition probabilities the number of times the minima were crossed (yellow line in Fig. 4).

DISCUSSION

Nowadays, molecular dynamics simulations easily produce tera scale of data. As such, an important bottleneck in the understanding of molecular processes is in the analysis rather than the data generation per se.

In this work, a strategy for the analysis of conventional order parameters time series that is kinetics compliant was investigated. Our results provided strong evidence that the coupling of order parameter fluctuations with complex network analysis represents a powerful approach to deconvolute

crowded order parameter distributions of molecular systems. This procedure allows the construction of kinetically accurate Markov-state-models in a natural and intuitive way, largely overcoming the problems raised by conventional order parameter analysis. Moreover, a wide range of experimentally generated time traces coming from FRET or optical tweezers can be readily tackled by this methodology. Taking into account the fluctuations within a time-window τ, the approach is able to distinguish between snapshots belonging to different states but having the same value of the order parameter. Towards an accurate characterization of the kinetics the value of τ needs to be chosen appropriately. We proposed to estimate it on the basis of a mean-first-passage-times analysis. Very similar in spirit to what Caflisch and collaborators proposed in their work,[35] this procedure is more suitable to our network clusterization approach.

Finally, it is worth mentioning that instead of looking at better ways to analyze conventional order parameters time series, some groups focused their attention on the development of optimal reaction coordinates.[51–53] These abstract coordinates aim to correctly characterize the molecular kinetics. Among them, one method based on cut-based free-energy profiles seems very promising.[53] In this approach, the coefficients of a linear combination of physical distances are optimized against the cut-based free-energy profile. At the end of the process, the combination which maximizes the barrier height with respect to a target state provides the optimal reaction coordinate. A fundamental difference between this method and the local-fluctuations analysis is that the latter requires the time evolution of a single (non-optimal) coordinate while the optimization procedure makes use of a very large number of physical distances (usually around thousands[15,53,54]) to perform properly, making the latter more computationally intensive.

[1] R. Du, V. Pande, A. Grosberg, T. Tanaka, and E. Shakhnovich, J. Chem. Phys. **108**, 334 (1998).
[2] S. Benkovic, G. Hammes, and S. Hammes-Schiffer, Biochemistry **47**, 3317 (2008).
[3] T. Lazaridis and M. Karplus, Science **278**, 1928 (1997).
[4] A. Dinner, A. Sali, L. Smith, C. Dobson, and M. Karplus, Trends Biochem. Sci. **25**, 331 (2000).
[5] R. Zhou, B. Berne, and R. Germain, Proc. Natl. Acad. Sci. U.S.A. **98**, 14931 (2001).
[6] F. Rao and A. Caflisch, J. Chem. Phys. **119**, 4035 (2003).
[7] M. Cecchini, F. Rao, M. Seeber, and A. Caflisch *et al.*, J. Chem. Phys. **121**, 10748 (2004).
[8] F. Rao and A. Caflisch, J. Mol. Biol. **342**, 299 (2004).
[9] S. V. Krivov and M. Karplus, Proc. Natl. Acad. Sci. U.S.A. **101**, 14766 (2004).
[10] F. Rao, G. Settanni, E. Guarnera, and A. Caflisch, J. Chem. Phys. **122**, 184901 (2005).
[11] R. Hegger, A. Altis, P. Nguyen, and G. Stock, Phys. Rev. Lett. **98**, 28102 (2007).
[12] S. Muff and A. Caflisch, Proteins **70**, 1185 (2008).
[13] G. Maisuradze, A. Liwo, and H. Scheraga, Phys Rev. Lett. **102**, 238102 (2009).
[14] V. Voelz and V. Pande, Proteins **80**, 342 (2012).
[15] S. Krivov, J. Phys. Chem. B **115**, 12315 (2011).
[16] W. C. Swope, J. W. Pitera, and F. Suits, J. Phys. Chem. B **108**, 6571 (2004).
[17] D. Gfeller, P. De Los Rios, A. Caflisch, and F. Rao, Proc. Natl. Acad. Sci. U.S.A. **104**, 1817 (2007).
[18] F. Noé, I. Horenko, C. Schütte, and J. C. Smith, J. Chem. Phys. **126**, 155102 (2007).
[19] J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope, J. Chem. Phys. **126**, 155101 (2007).
[20] D. Prada-Gracia, J. Gómez-Gardeñes, P. Echenique, and F. Falo, PLOS Comput. Biol. **5**, e1000415 (2009).
[21] F. Rao and M. Karplus, Proc. Natl. Acad. Sci. U.S.A. **107**, 9152 (2010).
[22] F. Rao, Phys. Chem. Lett. **1**, 1580 (2010).
[23] S. V. Krivov and M. Karplus, J. Phys. Chem. B **110**, 12689 (2006).
[24] E. Hua-Mei Kellogg, O. Lange, and D. Baker, J. Phys. Chem. B **116**, 11405 (2012).
[25] I. Buch, T. Giorgino, and G. De Fabritiis, Proc. Natl. Acad. Sci. U.S.A. **108**, 10184 (2011).
[26] D. Huang and A. Caflisch, PLOS Comput. Biol. **7**, e1002002 (2011).
[27] V. Uversky, C. Oldfield, and A. Dunker, Annu. Rev. Biophys. **37**, 215 (2008).
[28] M. Knott and R. Best, PLOS Comput. Biol. **8**, e1002605 (2012).
[29] J. Errington and P. Debenedetti, Nature (London) **409**, 318 (2001).
[30] T. Yan, C. Burnham, M. Del Pópolo, and G. Voth, J. Phys. Chem. B **108**, 11877 (2004).
[31] F. Rao, S. Garrett-Roe, and P. Hamm, J. Phys. Chem. B **114**, 15598 (2010).
[32] S. Garrett-Roe, F. Perakis, F. Rao, and P. Hamm, J. Phys. Chem. B **115**, 6976 (2011).
[33] D. Prada-Gracia, R. Shevchuk, P. Hamm, and F. Rao, J. Chem. Phys. **137**, 144504 (2012).
[34] A. Baba and T. Komatsuzaki, Proc. Natl. Acad. Sci. U.S.A. **104**, 19297 (2007).
[35] P. Schuetz, R. Wuttke, B. Schuler, and A. Caflisch, J. Phys. Chem. B **114**, 15227 (2010).
[36] A. Baba and T. Komatsuzaki, Phys. Chem. Chem. Phys. **13**, 1395 (2011).
[37] Y. Yin, G. G. Maisuradze, A. Liwo, and H. A. Scheraga, J. Chem. Theory Comput. **8**, 1176 (2012).
[38] N. V. Smirnov, Rec. Math. (Mat. Sbornik) (NR) **6**, 3 (1939).
[39] M. Seeber, M. Cecchini, F. Rao, G. Settanni, and A. Caflisch, Bioinformatics **23**, 2625 (2007).
[40] D. Gfeller, D. de Lachapelle, P. De Los Rios, G. Caldarelli, and F. Rao, Phys. Rev. E **76**, 026113 (2007).
[41] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, Nucleic Acids Res. **30**, 1575 (2002).
[42] B. Brooks, R. Bruccoleri, B. Olafson, S. Swaminathan, M. Karplus *et al.*, J. Comput. Chem. **4**, 187 (1983).
[43] B. Brooks, C. Brooks III, A. Mackerell, Jr., L. Nilsson, R. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch *et al.*, J. Comput. Chem. **30**, 1545 (2009).
[44] M. Harvey, G. Giupponi, and G. Fabritiis, J. Chem. Theory Comput. **5**, 1632 (2009).
[45] K. A. Feenstra, B. Hess, and H. J. C. Berendsen, J. Comput. Chem. **20**, 786 (1999).
[46] F. Rao and M. Spichty, J. Comput. Chem. **33**, 475 (2012).
[47] M. Seeber, A. Felline, F. Raimondi, S. Muff, R. Friedman, F. Rao, A. Caflisch, and F. Fanelli, J. Comput. Chem. **32**, 1183 (2011).
[48] O. Bieri, J. Wirz, B. Hellrung, M. Schutkowski, M. Drewello, and T. Kiefhaber, Proc. Natl. Acad. Sci. U.S.A. **96**, 9597 (1999).
[49] A. Möglich, K. Joder, and T. Kiefhaber, Proc. Natl. Acad. Sci. U.S.A. **103**, 12394 (2006).
[50] F. Rao, J. Comput. Chem. **32**, 1113 (2011).
[51] R. Best and G. Hummer, Proc. Natl. Acad. Sci. U.S.A. **102**, 6732 (2005).
[52] A. Ma and R. Aaron, J. Phys. Chem. B **109**, 6769 (2005).
[53] S. Krivov, J. Phys. Chem. B **115**, 11382 (2011).
[54] S. Steiner and A. Caflisch, Proteins **80**, 2562 (2012).