THE JOURNAL OF PHYSICAL CHEMISTRY B

ARTICLE

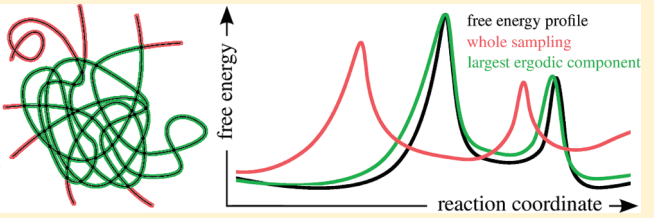
pubs.acs.org/JPCB

Equilibrium Distribution from Distributed Computing (Simulations of Protein Folding)

Riccardo Scalco and Amedeo Caflisch*

Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

ABSTRACT: Multiple independent molecular dynamics (MD) simulations are often carried out starting from a single protein structure or a set of conformations that do not correspond to a thermodynamic ensemble. Therefore, a significant statistical bias is usually present in the Markov state model generated by simply combining the whole MD sampling into a network whose nodes and links are clusters of snapshots and transitions between them, respectively. Here, we introduce a depth-first search algorithm to extract from the whole conformation space network the largest ergodic component, i.e., the subset of nodes of the network whose transition matrix corresponds to an ergodic Markov chain. For multiple short MD simulations of a globular protein (as in distributed computing), the steady state, i.e., stationary distribution determined using the largest ergodic component, yields more accurate free energy profiles and mean first passage times than the original network or the ergodic network obtained by imposing detailed balance by means of symmetrization of the transition counts.



I. INTRODUCTION

Atomistic molecular dynamics (MD) simulations are widely used for Boltzmann-weighted (i.e., equilibrium) sampling of the phase space of biological macromolecules.[1] In principle, MD simulations of length significantly longer than the process of interest should generate a molecular "movie" at very high spatial and temporal resolution. In practice, because of the many degrees of freedom in the (poly)peptide chain and the related complexity of the free energy surface it is very challenging to sample the conformational space of proteins and even peptides by standard MD techniques, which have an inherently "slow" time step of about 1−5 fs. At low temperatures, MD simulations can get trapped in the starting basin. At elevated temperatures, on the other hand, the accessible phase space increases enormously so that not all possible conformations are visited. A number of simulation techniques have been introduced to enhance the sampling of the conformational space.[2−8] At the same time, the availability of hundreds to thousands of processors has been exploited by intrinsically parallel jobs like distributed computing[9,10] and loosely coupled MD simulations.[11] Because of the significant time-scale gap between the actual protein folding process (microseconds to seconds) and simulation length (nanoseconds), it is not possible to extract folding kinetics *directly* from distributed computing simulations.[10,12]

Markov chain models have been used to determine transition probabilities between coarse-grained states. These states (or more precisely clusters of MD snapshots) usually range in number between 100 and 1000, and have been derived from multiple short MD runs[13−16] or from long trajectories with multiple folding and unfolding events.[17] One advantage of Markov state models is that they can be used to combine (short) independent MD simulations for extracting information on processes whose time-scale is longer than the one of the individual MD runs.[13,18−20] A potential disadvantage is that the sampling of phase space by multiple, independent short runs can be affected by a statistical bias.[21] Such bias is easily understood considering that the starting nodes are selected following a probability distribution which is different from the Boltzmann-weighted distribution. Under the assumption that the transition probabilities between coarse-grained states, which are conditional probabilities of local transitions, are sampled correctly, the bias can be removed by calculating the steady state of the Markov chain. For the steady state calculation, the Markov chain must be ergodic,[22] in other words it must be irreducible (from any state the system can reach every other state) and aperiodic (there are no states which show up at a fixed period of time). Markov chains derived from multiple MD trajectories are usually not ergodic. The nonergodicity is a consequence of the sampling by multiple (short) runs, e.g., most initial and final conformations act as sources and sinks, respectively, which make the Markov chain nonirreducible.

Here we show that an automatic procedure for the identification of the largest ergodic component from a nonirreducible directed network (shown schematically in Figure 1) is able to remove the statistical bias which is typical of MD sampling by multiple short trajectories. The procedure used here is based on a theorem[23] that expresses the possibility to subdivide every directed graph (the terms "network" and "graph" are used as synonyms) in its irreducible components, the largest of which is likely to be the most relevant. The method has several advantages: first, it does not require any parameter; second, it translates

Received: February 15, 2011
Revised: April 8, 2011
Published: April 25, 2011

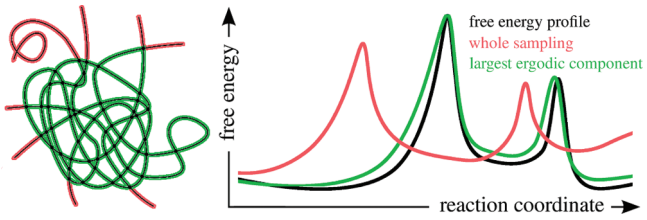


Figure 1. Schematic illustration of the procedure described in this paper. Left: The largest subset of nodes that yields an ergodic network (green) is extracted from the whole sampling (red and green). In this way, it is possible to remove the statistical bias, which is usually introduced by multiple independent trajectories. Right: The free energy profile of the largest ergodic component (green) is much closer to the actual profile (black) than the profile obtained considering the whole sampling (red).

the problem of obtaining an irreducible network in the search for the minimum number of links to be removed to obtain irreducibility; third, several computational implementations already exist. Here we use the algorithm presented by Tarjan,[23] which makes use of a *depth-first search* in the graph and thus is very efficient.

This paper is structured as follows. The Theory section presents the principles of Markov state models derived from MD sampling, the procedure for the extraction of the largest ergodic component, and an analytical formula of the dependence of the statistical bias on the number of simulations, their length, and the fundamental matrix associated with the transition matrix. In the Examples section, the kinetics obtained by extraction of the largest ergodic component are compared with the broadly used imposition of detailed balance by simple symmetrization of the matrix of transition counts. It is shown that for multiple short trajectories, the statistical bias cannot be removed by imposing detailed balance which results in wrong statistics. In contrast, the largest ergodic component yields free energy profiles and mean first passage times that better reflect the kinetics than the results obtained by detailed balance imposition. The Conclusion summarizes the main points of this work.

II. THEORY

A. Markov State Models from Multiple MD Simulations. We consider the frequent case of m independent molecular dynamics or Metropolis Monte Carlo runs for which it is convenient to introduce the abbreviation m-trj instead of the more generic term *trajectory*. More precisely, with m-trj we intend the m symbolic trajectories obtained usually by a clustering procedure of the whole sampling.[24]

We assume that the system being studied is ergodic and can be formalized as a finite homogeneous Markov chain. Thanks to the ergodic hypothesis, we can use Birkhoff's theorem[25] to extract from the m-trj the transition matrix P associated with the Markov chain. If the m-trj is long enough, the transition probabilities will converge. Here, we assume to obtain such convergence, or at least to obtain it in a certain subspace of interest of the system phase space. Indeed such transition probabilities are local, i.e., they are conditional probabilities and therefore not affected by an incomplete sampling of the phase space.

Given an m-trj, we use *naïve* definitions, namely the maximum likelihood estimates,[21] for the probability distribution p_i over the nodes set $\{i\}$, the transition probabilities P_{ij} between nodes, and the transition rates q_{ij}. Let start by defining q_{ij} as the number of one step transitions $i \rightarrow j$ observed in the m-trj, normalized by the total number of transitions. From q_{ij} we derive $p_i = \Sigma_j q_{ij}$, so that $\Sigma_i p_i = 1$. Finally, the transition probabilities are the conditional probabilities $P_{ij} = q_{ij}/p_i$. Note that, if $p_i > 0$ for every node i, P is by definition a right stochastic matrix, i.e., a square matrix whose rows consist each of non-negative real numbers that sum up to 1: $\Sigma_j P_{ij} = \Sigma_j (q_{ij})/(p_i) = 1$.

In general, given an m-trj, the Markov chain P obtained with the above definitions is not ergodic. Because of the finite sampling, one or more segments of the m-trj act as attractor(s) so that the directed graph associated with the chain is not irreducible. We stress that the non irreducibility of the chain could be a solvable problem which is easily fixed by considering that such attractors are contained in the statistically less informative part of the sampling, so that they can be discarded without any significant loss of statistics. In essence, we are concerned with the statistical bias of which every m-trj is affected. In other words, the choice of starting point(s) of the m 1-trj does not reflect the correct probability distribution over the nodes. To remove such bias, we need an equilibration procedure, namely we calculate the steady state π of the Markov chain, the state that satisfies the equation $\pi = \pi P$.[22] Only ergodic chains possess one and only one steady state, the entries of which are all positive. Thus, we need to retrieve an ergodic chain from the m-trj before calculating the steady state.

B. The Largest Ergodic Component. The problem of how to obtain an irreducible graph from a nonirreducible one does not have a straightforward solution, because the problem itself is not well-defined. In other words, given a directed graph, many different irreducible graphs can be generated from it. A common solution, due to its simplicity, is to impose detailed balance by symmetrizing the count matrix, i.e., defining $q_{ij}^{db} = q_{ji}^{db} = (q_{ij} + q_{ji})/2$, which corresponds to including the counts that would have been obtained by the time-reversed simulations.[21,26] The symmetrization of the count matrix introduces an error which is larger the larger the difference between the actual sampling and the equilibrium sampling is (see below). In other words, imposing detailed balance directly to the nonergodic graph renders impossible the removal of the statistical bias connected to the m-trj. We suggest to obtain ergodicity from the collected transitions without modifying their statistical nature, i.e., we advise against the insertion of spurious transitions as in the symmetrization of the count matrix. Instead, we suggest *to remove the minimum amount of links to obtain an irreducible directed graph*. We prefer to remove instead of insert links in order to affect the statistics as little as possible.

With the above task in mind, i.e., removing the minimum amount of links for generating an ergodic graph, we make use of following graph theory theorem.[27] Given a directed graph $G = (V,E)$, where V and E are the sets of vertices and edges, it is possible to define an equivalence relation on V such that two vertices v and w are equivalent if there is a path from v to w and a path from w to v. Let V_i, i:1, ..., n, the n distinct equivalence classes, defining $G_i = (V_i E_i)$, where $E_i = \{(v,w) \in E | v,w \in V_i\}$, one can prove that

- each G_i is strongly connected (irreducible)
- no G_i is a proper subgraph of a strongly connected subgraph of G (a proper subgraph of G is a subgraph which contains at least one and not all the edges of G)

The subgraphs G_i are called the strongly connected (or irreducible) components of G. We note that the subdivision in equivalence classes is unique, so we do not need any parameter to obtain the strongly connected components. Moreover, the condition of minimal removal of links is satisfied by the second point, namely the subgraphs G_i are the largest possible components.

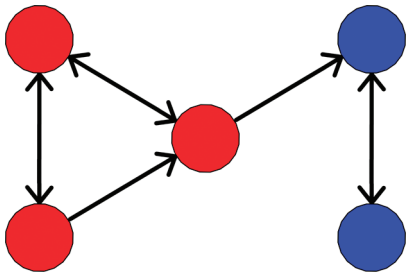


Figure 2. Example of a directed graph with two strongly connected components emphasized by different colors.

Figure 2 shows an example of a directed graph with two strongly connected components.

Hopefully, the largest strongly connected component covers the most relevant part of the original graph. It will be shown in the results that the larger is the sampling, the smaller is the number of links to be removed to obtain the largest strongly connected component. Assuming the largest strongly connected component is aperiodic, as is always the case for complex networks describing free energy surfaces of peptides and proteins,[24,28] it is also the largest ergodic component. Finally, we emphasize that such procedure is not a community detection algorithm,[29−31] it simply solves the well-defined mathematical problem regarding the subdivision in strongly connected components of a generic directed graph.

The theorem enunciated above has been employed in different computer algorithms. Here, the algorithm published by Tarjan is used.[23] It is based on a *depth-first search* procedure in the graph which is very efficient. For the largest network in the examples mentioned below (network of example A with 3652 nodes and 18948 links) the Tarjan algorithm requires less than one second on a 3 GHz commodity processor. In general, it requires $O(V,E)$ space and time, namely the algorithm needs space bounded by $k_1V + k_2E + k_3$, where k_1, k_2, and k_3 are constant.

C. The Statistical Bias. To illustrate the origin of the bias, it is useful to formulate an analytical formula of the deviation from the stationary distribution. In the following, P is the transition matrix associated with an ergodic Markov chain C. We remember that from any m-trj it is possible to generate an ergodic chain C by means of the Tarjan algorithm, which, as mentioned above, extracts the strongly connected components from the directed graph drawn following the m-trj. The matrix P we are looking for will be the one associated with the ergodic chain C generating the largest irreducible component. In general, the chain C consists of several trajectories of different length extracted from the original m-trj, which all together draw an irreducible directed graph.

Given the chain C, we calculate the transition rates q_{ij}, the probability distribution p_i, the transition matrix P and the steady state π. The latter is the solution of $\pi = \pi P$, and in this work it is calculated iteratively by means of $p_n = p_{n-1}P$ until convergence is reached. The relevant question is: *How does the difference $\pi - p$ depend on the sampled m-trj?* This question can be formalized for each node i using the transition rates q_{ij} as follows:

$$\sum_j q_{ij} = k_i + \sum_j q_{ji}$$

Here the index runs over the nodes. The quantity k_i is the difference between the *outgoing* and *ingoing* flow of the node i and is caused by the finite length of the m 1-trj. Of course the total sum must be zero:

$$\sum_i k_i = \sum_{ij} q_{ij} - \sum_{ij} q_{ji} = 1 - 1 = 0$$

Note that for equilibrated transition rates we expect $k_i = 0\ \forall i$, i.e., the flow conservation law $\Sigma_j q_{ij} = \Sigma_j q_{ji}$. This can be easily seen by defining q_{ij} by means of the steady state π, i.e., $q_{ij}^{eq} = \pi_i P_{ij}$, so that

$$\sum_j q_{ij}^{eq} = \sum_j \pi_i P_{ij} = \pi_i \sum_j P_{ij} = \pi_i$$

$$\sum_j q_{ji}^{eq} = \sum_j \pi_j P_{ji} = [\pi P]_i = \pi_i$$

In other words, the presence of $k_i \neq 0$ requires the determination of the steady state to have an unbiased statistics.

With the naïve definitions of the probability distribution over the nodes ($p_{0_i} = \Sigma_j q_{ij}$) and the entries of the transition matrix P ($P_{ij} = q_{ij}/p_{0_i}$) derived from the chain C, we are able to prove the following equality for $p_m = p_0 P^m$:

$$\lim_{m\to\infty} p_m = p_0 - \lim_{m\to\infty} \sum_{n=0}^{m} kP^n$$

Proof. From equation $\Sigma_j q_{ij} = k_i + \Sigma_j q_{ji}$, we have

$$p_{0_i} = k_i + \sum_j p_{0_j} P_{ji} = k_i + p_{1_i}$$

or, without the index

$$p_0 = k + p_0 P = k + p_1$$

So, we can write

$$\begin{aligned} p_1 &= p_0 - k \\ p_2 &= p_1 P = p_0 P - kP = p_1 - kP = p_0 - k - kP \\ p_3 &= p_2 P = p_0 - k - kP - kP^2 \\ &\vdots \\ p_m &= p_0 - \sum_{n=0}^{m-1} kP^n \end{aligned}$$

taking the limit, given that P is ergodic, we briefly write

$$\pi \equiv p_\infty = p_0 - \sum_{n=0}^{\infty} kP^n$$

Summing up, the difference we are looking for is $\pi - p_0 = -\Sigma_{n=0}^{\infty} kP^n$.

The equation indicates that the series $\Sigma_{n=0}^{\infty} kP^n$ converges. To appreciate this point we rewrite the result in terms of the fundamental matrix[32] associated with the transition matrix P:

$$Z = \mathbb{I} + (P - P^\infty) + (P^2 - P^\infty) + \cdots = (\mathbb{I} - P + P^\infty)^{-1}$$

Observing that

$$\lim_{n\to\infty} [kP^n]_i = \sum_j k_j [P^\infty]_{ji} = \pi_i \sum_j k_j = 0$$

we can now rewrite the series in the form

$$\pi = p_0 - \sum_{n=0}^{\infty} kP^n = p_0 - kZ$$

dx.doi.org/10.1021/jp2014918 |J. Phys. Chem. B 2011, 115, 6358–6365

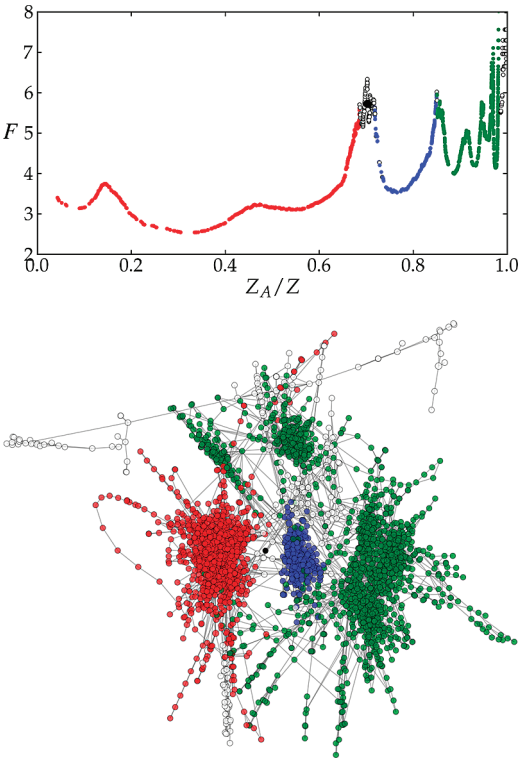


Figure 3. Free energy surface of the model system used as benchmark. The sampling was obtained by implicit solvent MD simulations of a simplified-sequence variant of protein G at 330 K[17] for which a C_α rmsd coarse-graining with 3.5 Å cutoff resulted in 3683 clusters, i.e., nodes of the network (see text). (Top) cFEP of the model. The free energy F is given in kcal/mol in all cFEPs in this work. (Bottom) The network representation was generated by the Fruchterman−Reingold force-directed algorithm.[39] The 27742 links between pairs of node represent MD transitions at 20 ps saving frequency. The coloring reflects the main basins, red and blue, which have been identified by plotting the cFEP separately from their representative node. The collection of green nodes could be furthermore divided in three smaller basins. The white nodes are unassigned, i.e., at free energy barriers. The black node at the barrier is the starting node of example D.

Moreover, we could rewrite the equality $\pi = p_0 - kZ$ separating the different factors involving the sampled m-trj. Let k_i be the difference in outgoing and ingoing transitions in the node i, m the number of 1-trj of which the chain C if formed and s the total number of transitions observed; then we note that $\Sigma_i|k_i| \leqslant 2m$ and $k_i = k_i/s$. Extracting m we define $\lambda_i \equiv k_i/m$, then we have $\Sigma_i|\lambda_i| \leq 2$ and the equality becomes $\pi = p_0 - (m/s)\lambda Z$. Thus, the difference $\pi - p_0$ depends on distinct factors:

1. The multiplicative term m/s shows that $\pi - p_0$ increases with the number m of 1-trj and decreases with the total length s of them, as expected. Note that s/m is the mean number of steps per 1-trj.
2. The vector $\lambda \equiv k/m$ depends on the initial and final nodes of the m 1-trj and it is influenced by the choice of starting nodes. For example, in the exotic case of multiple runs each of them starting and ending at the same node, λ is the null vector and there is no bias, whatever m is.
3. The shape of the visited free energy surface, which affects the fundamental matrix Z.

Some observations are needed. In the case of a m-trj consisting of few long simulations, the ratio m/s is small and, as expected, p_0 is a good approximation of π. Vice versa, for a m-trj of many short runs, the mean number of steps per simulation is small (m/s big) and the steady state calculation could be significant. Only for a choice of starting nodes that exactly follows the probability distribution at equilibrium π, the term λZ will have entries nearly equal to zero, so to make p_0 a good approximation of π. Summing up, for multiple trajectory analysis of many short runs, typical of parallel and distributed computing, that start always from the same conformation or from different conformations of unknown distribution, the steady state π offers the correct statistics.

Let us now compare π with the statistics resulting from the detailed balance imposition. The probability distribution over the nodes obtained by count symmetrization is stationary because

$$\sum_j p_j^{db} P_{ji}^{db} = \sum_j p_j^{db} \frac{q_{ji}^{db}}{p_j^{db}} = \sum_j q_{ij}^{db} = p_i^{db}$$

The stationary probability vector p_i^{db} differs from the stationary distribution π as follows:

$$p_i^{db} = \sum_j (q_{ij} + q_{ji})/2 = (p_{0_i} + p_{1_i})/2 = p_{0_i} - k_i/2$$

$$= \pi_i + [kZ]_i - k_i/2$$

It is crucial to note here that the difference between p_i^{db} and π strongly depends on the sampled m-trj. Thus, there are situations for which imposing detailed balance by simple count symmetrization is not appropriate, particularly when the m-trj consists of many short runs like in parallel and distributed computing as will be shown in the next section.

III. EXAMPLES

In this section, we illustrate the usefulness of the automatic procedure to extract the largest ergodic component, which is a subset of nodes whose transition matrix corresponds to an ergodic Markov chain. The protein used is a simplified-sequence variant of protein G[17] which is sampled by implicit solvent[33] MD at 330 K. First, the ~220000 snapshots saved every 20 ps along the MD simulations are clustered by C_α rmsd and a threshold of 3.5 Å using the leader-algorithm as implemented in WORDOM.[34,35] The clustering yields 3683 nodes, and there are 27742 links between them. The transition matrix associated with the 3683 clusters is ergodic as detailed balance condition was imposed. This transition matrix and associated stationary distribution are referred to as the "model" in the following. The cut-based free energy profile (cFEP)[36] and conformational space network[24] of the model are shown in Figure 3.

The transition matrix of the model is used to generate the m-trj sampling, i.e., to propagate m (short) trajectories of a random walker, which emulate m independent MD runs. Every step of the random walker represents a time interval of 20 ps because of the saving frequency of the MD simulations from which the network is extracted. Four examples of m-trj are discussed. They differ in the choice of the starting node(s), the number of random walker trajectories m, and/or the length $l = s/m$ of each trajectory (see Table 1 for details). Using the naïve definitions it is straightforward to determine $[P^{db}]_{ij} \equiv q_{ij}^{db}/p_i^{db}$ which is the transition matrix derived imposing detailed balance by symmetrization of the count matrix. The transition matrix derived from the chain C associated with the largest ergodic component is $[P^C]_{ij} \equiv q_{ij}^C/p_i^C$ where q_{ij}^C is the number of one step transitions $i \rightarrow j$ observed in

Table 1. Examples of *m*-trj Analysis[a]

	sampling				% visited phase space[b]		
example	*m*	*l*	(μs)	starting node	total	in largest erg. comp.	% discarded sampling[c]
A	10 000	10	2	random	99 (3652)	80 (2791)	18
B	1000	500	10	most pop.	86 (1868)	75 (1664)	1
C	1000	200	4	most pop.	73 (1356)	68 (1283)	0.5
D	1000	200	4	at the barrier	78 (1940)	76 (1817)	0.001

[a] The first five entries of each row list the name of the example, the number of random walkers *m* (i.e., number of emulated MD runs), the length of each run *l*, the total sampling, and the starting conformation(s), respectively. Note that the starting conformation of each of the 10000 runs of example A is drawn randomly from the 3683 nodes of the model, whereas it is a single node for examples B, C, and D. [b] The visited phase space is the sum of the state probabilities of the model over the states visited by the *m*-trj. The number of nodes visited is in parentheses. [c] The discarded sampling is the percentage of random walker steps that is not included in the largest ergodic component. For examples B, C, and D it is much smaller than the difference of the values in the two preceding columns because the discarded sampling concerns a part of phase space not sampled enough to reach its correct probability.

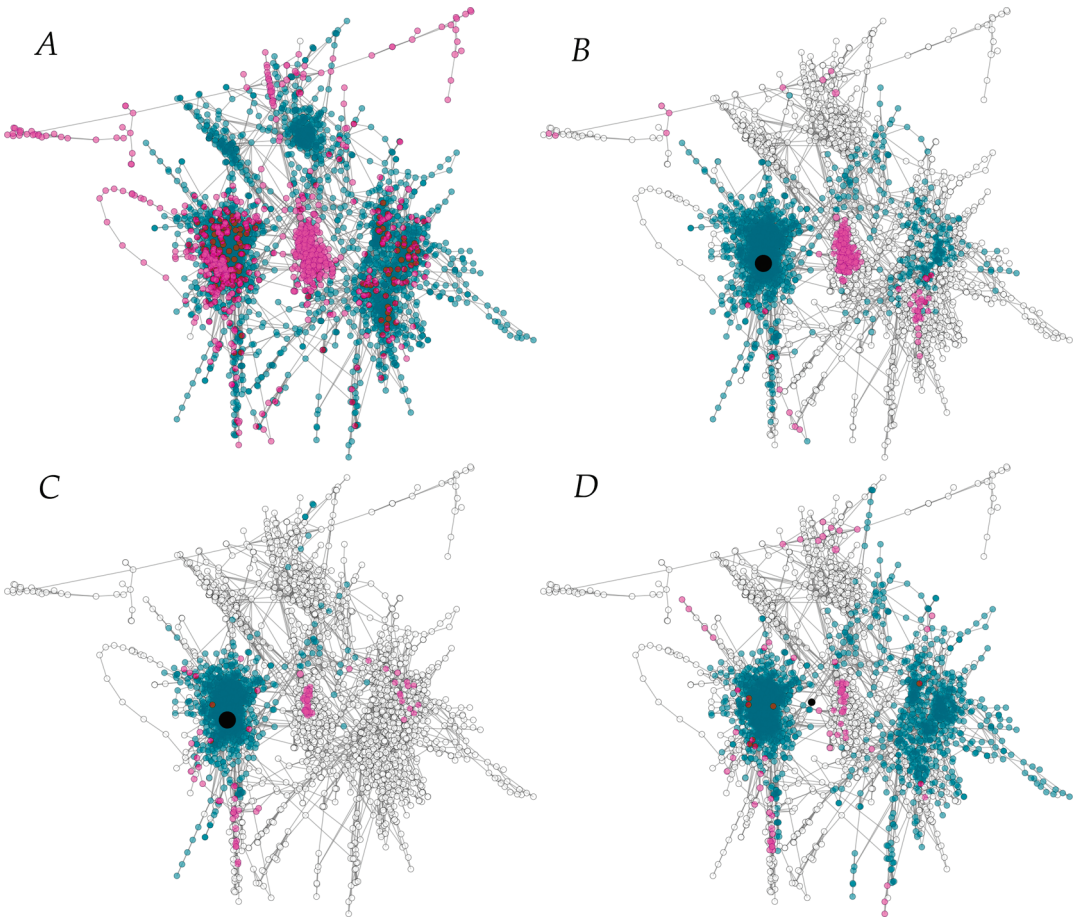


Figure 4. Network representation of the sampling in each of the four examples A to D. The nodes visited by the *m*-trj are in cyan or magenta if they are inside or outside the largest ergodic component, respectively. The white nodes were not visited by the *m*-trj and the black node is the starting node except for example A which used almost all nodes as starting nodes. The details of the four examples are given in Table 1.

the chain *C* and $p_i^C = \Sigma_j q_{ij}^C$. Therefore, $[P^{C,eq,db}]_{ij} \equiv q_{ij}^{C,eq,db}/p_i^{C,eq,db}$ is the transition matrix derived imposing detailed balance on the equilibrated transition rate probabilities $q_{ij}^{C,eq} = \pi_i P_{ij}^C$, where π is the steady state of the chain *C*, i.e., $\pi = \pi P^C$.

For each example, we calculate the cut-based free energy profile (cFEP)[36] using the most probable node as reference and the mean first passage time (mfpt) as progress coordinate.[37] The analysis focuses on the differences between the straightforward (but in most case inappropriate) count symmetrization (P^{db}) and the steady state of the largest ergodic component ($P^{C,eq,db}$).

A. Distributed Computing. Example A is an *m*-trj consisting of $m = 10000$ very short ($l = 10$) random walkers, which is the equivalent of 2 μs of sampling by implicit solvent MD, starting at nodes selected randomly (Table 1). Note that the 200 ps length of each walker corresponds to an explicit water MD time scale of about 2−20 ns (i.e., 10 to 100 longer[38]) because of the lack of

dx.doi.org/10.1021/jp2014918 |*J. Phys. Chem. B* 2011, 115, 6358–6365

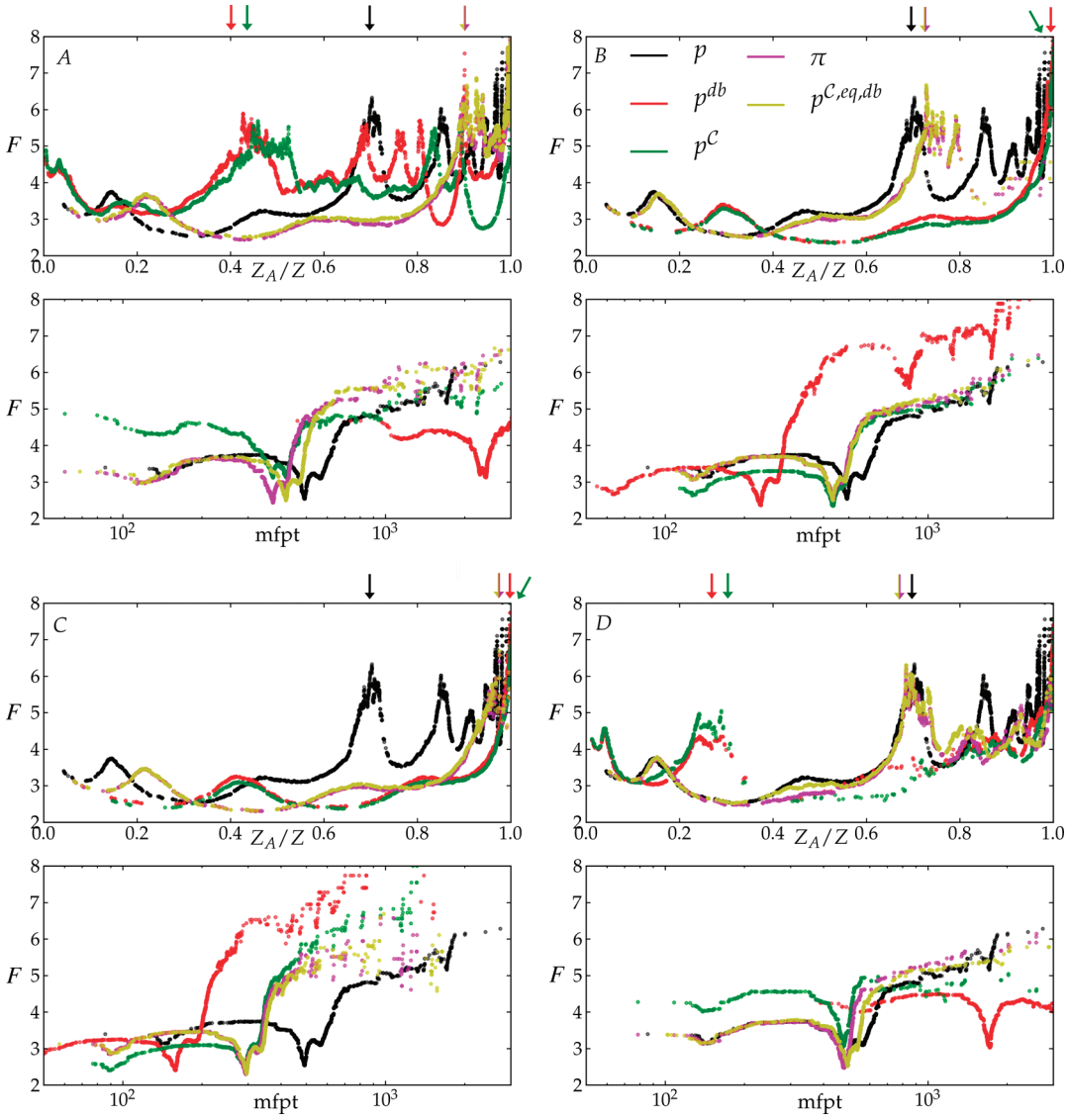


Figure 5. Free energy profiles and mfpt values of examples A, B, C, and D. The top and bottom parts of each of the four panels show the cFEP plotted using as reaction coordinate the relative partition function Z_A/Z (ref 36) and mfpt (ref 37), respectively. Note that to improve resolution the cFEPs plotted as a function of mfpt include only the range up to 3000 steps, i.e., 60 ns. The vertical arrows above the Z_A/Z cFEPs indicate the relative partition function value corresponding to mfpt =60 ns. The cFEP plotted as a function of Z_A/Z illustrates barrier heights and locations as obtained by different transition matrices, while the cFEP with mfpt as reaction coordinate allows the direct comparison of the mfpt values. As shown in the legend of panel B, individual cFEPs are colored as follows: Black for the original transition matrix P with probability distribution $p_i = \Sigma_j q_{ij}$ and transition rate probabilities q_{ij}; red for the naïve symmetrization of the transition counts, resulting in the transition matrix P^{db} with probability distribution $p_i^{db} = \Sigma_j q_{ij}^{db}$, where $q_{ij}^{db} = q_{ji}^{db} = (q_{ij} + q_{ji})/2$; green for the largest ergodic component P^C with naïve definitions of p_i^C and q_{ij}^C; magenta for P^C with the steady state π and $q_{ij}^{C,eq}$; yellow for $P^{C,eq,db}$ with $q_{ij}^{C,eq,db} = (q_{ij}^{C,eq} + q_{ji}^{C,eq})/2$ and $p_i^{C,eq,db} = \Sigma_j q_{ij}^{C,eq,db}$. The comparison of P^{db} (red) and $P^{C,eq,db}$ (yellow) is useful to analyze the statistical bias.

friction in the implicit solvent MD simulations.[17] Since the starting nodes are chosen uniformly and not according to the distribution of node size, the initial ensemble does not reflect the Boltzmann distribution, which is often the case in distributed computing.[10,12] A total of 18% of random walker steps (i.e., 18% of the *m*-trj sampling) are outside of the largest ergodic component identified by the Tarjan algorithm (in less than 1 s). A comparison of the networks colored according to the individual free energy basins (Figure 3) and according to the largest ergodic component (Figure 4A) indicates that most of the discarded sampling lies outside of the most populated basin and is located in the region of the free energy surface colored in blue in Figure 3. This part of sampling concerns the second largest irreducible component, which could be connected with the largest one by means of further sampling at the barrier between the red and the blue basins.

The symmetrization of the count matrix of the whole *m*-trj sampling yields a free energy profile very different from the model and too large mfpt values of the nodes within the most populated basin (Figure 5A, red curves). Because of the very short length of the random walker trajectories ($1/l \equiv m/s = 0.1$) and the choice of the starting nodes (shape of the λ vector), steady state calculation is expected to be necessary. Despite the 18% loss of *m*-trj sampling due to the extraction of the largest ergodic component, the transition matrices P^C with the steady state and $P^{C,eq,db}$ yield very good approximations of the main

dx.doi.org/10.1021/jp2014918 |*J. Phys. Chem. B* 2011, 115, 6358–6365

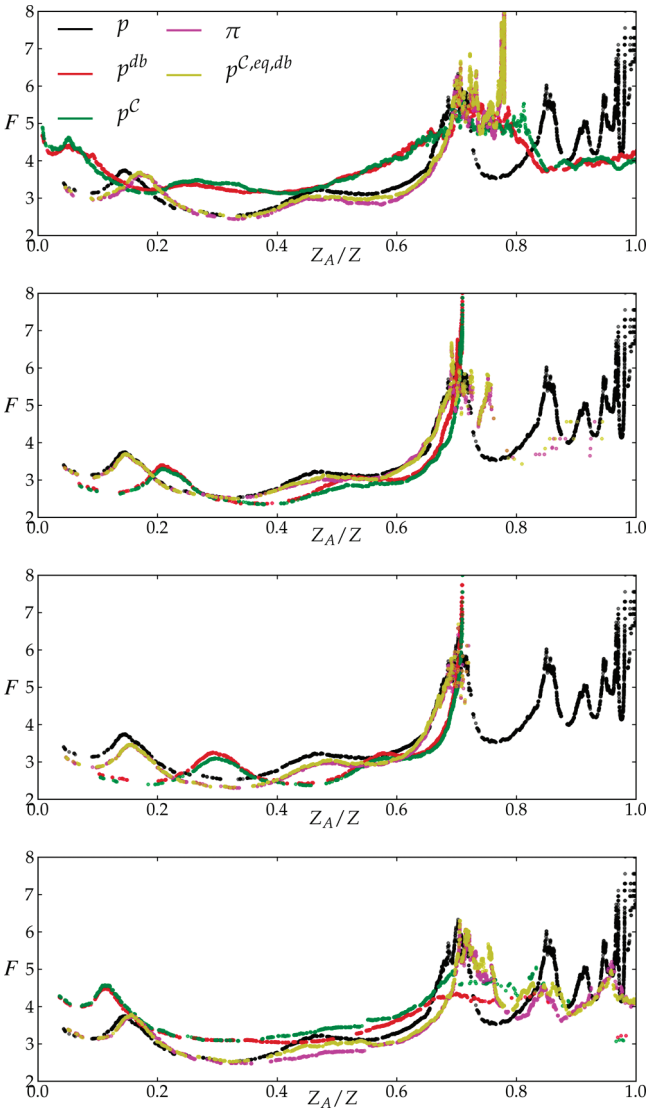


Figure 6. Stationary distribution on the largest ergodic component extracted from the *m*-trj of examples A, B, C, and D, yields the same cFEP of the main free energy basin as the original model. For a direct comparison, a rescaling of the Z_A/Z reaction coordinate is applied to the cFEP calculated using the $P^{C,eq,db}$ transition matrix. More precisely, given for the original model $p_A = \Sigma_i p_i$ for all nodes i in the main basin (whose range is $0 \leq Z_A/Z < 0.7$), and p'_A the same quantity for the other transition matrices, a coordinate transformation is applied to the x-axis by means of a rescaling factor $C = p_A/p'_A$ for the p_i values of the *m*-trj. Note that the rescaling is possible because the cFEP is invariant with respect to any continuous invertible transformation of the reaction coordinate.[40] The colors of the cFEPs are explained in the legend of Figure 5.

barrier on the cFEP and mfpt values. Moreover, their cFEPs are essentially identical. The position of the barrier on the x-axis corresponds to the statistical weight of the most populated basin. The shift of the barrier by about 18% in the cFEPs of P^C and $P^{C,eq,db}$ with respect to the original model is consistent with the result of the aforementioned network-coloring comparison, i.e., that the steps of the walkers not included in the largest ergodic component are mainly located outside of the most populated basin. Note however that the barrier height is preserved. Both of these findings are further illustrated by a transformation of the reaction coordinate (i.e., rescaling of the relative partition function) in the cFEP plot (Figure 6A).

Since a significant fraction of nodes is lost upon extraction of the largest ergodic component, it is somewhat surprising that mfpt values and cFEP profiles up to the first barrier are very accurate. There are two main reasons for these observations. First, both mfpt and cFEP are calculated using the transition probabilities and the equilibrium distribution, which are not affected by the bias introduced by multiple trajectory sampling. Second, the relaxation kinetics inside a free energy basin depends only on the profile of the basin up to the barrier to leave the basin.

B. Influence of Simulation Length. Examples B and C are *m*-trj composed of $m = 1000$ random walkers starting always at the most populated node of the model. They differ in the length of each walker trajectory which is $l = 500$ (resulting in a total of 10 μs sampling) and $l = 200$ (4 μs sampling) in examples B and C, respectively. As in example A, P^C with the steady state and $P^{C,eq,db}$ yield similar cFEPs and mfpt values. Importantly, P^C and $P^{C,eq,db}$ approximate correctly the original model in example B but not in example C (Figures 5B and 5C), which reflects that the simulation length plays an important role particularly when all runs start from the same structure. The network illustrations and cFEPs show that the $l = 500$ walkers have sufficient time to jump over the main barrier and visit other basins besides the most populated one (example B, i.e., Figures 4B and 6B) while the $l = 200$ walkers do not leave the main basin (example C, i.e., Figures 4C and 6C).

C. Influence of Starting Structure. Examples C and D are *m*-trj composed of $m = 1000$ random walkers each of $l = 200$ steps, which is the equivalent of 4 μs of sampling by implicit solvent MD. These two examples differ in the starting structure which, as mentioned above, is the most probable node of the model in example C (Figure 4C), and a very low-populated node at the top of the main barrier, i.e., the barrier to escape from the most probable basin, in example D (Figure 4D).

The stationary distribution of the largest ergodic component and the associated transition matrices P^C and $P^{C,eq,db}$ yield a much better approximation of the main barrier on the cFEP and mfpt values than P^{db} (Figure 5C,D). Moreover, the mfpt values are more accurate in example D than C which is consistent with the location of the starting node. Notably, using the stationary distribution, the sampling generated by starting at the main barrier yields the most accurate mfpt values of the four examples (compare magenta and yellow profiles with black profile in Figure 5D). In striking contrast, for the nodes within the most populated basin the simple symmetrization of the count matrix (P^{db}) in example D (Figure 5D) yields mfpt values that are significantly larger than the model, which is another indication of the error related to naïvely considering the transitions of the time-reversed simulations.

IV. CONCLUSIONS

Distributed computing and massively parallel computers have fostered the sampling of (small) protein conformational space by multiple, independent MD runs. The individual MD simulations are usually much shorter, particularly in distributed computing, than the time-scales associated with relevant conformational transitions, like protein folding and protein/protein association. As a consequence, the sampling obtained by independent MD runs is usually biased because of the short length of each run and/or the choice of the starting conformation(s).

In the present work, the statistical bias of multiple MD trajectories is formulated by an analytical expression that describes the dependence on the length of the trajectories, the choice of the starting conformation(s), and the underlying free energy surface. More precisely, an analytical formulation is given for the difference between the stationary distribution (or steady state) and the probability distribution obtained by simple symmetrization of the count matrix.

An automatic procedure is introduced for extracting the largest irreducible component from the whole conformational space network, or more precisely, the largest subset of nodes of the network whose associated transition matrix reflects an ergodic Markov chain. From the latter, the stationary distribution can be determined and used for calculating mfpt values and cFEP. The algorithm by Tarjan for the determination of the irreducible components is very efficient (linear on the number of nodes and links). Its application to four examples of MD sampling by multiple short trajectories shows that the stationary distribution on the largest ergodic component of the original network yields more accurate mfpt values and cFEPs than the naïve symmetrization of the count matrix. Thus, Tarjan's algorithm could be combined to network and cFEP analysis to search for weakly sampled regions of conformational space between two (or more) strongly connected components. This information could be very useful for improving an initial sampling by further MD simulations.

The automatic procedures for extracting the largest ergodic component and for determining the stationary distribution have been implemented in WORDOM.[35]

AUTHOR INFORMATION

Corresponding Author

*Telephone: +41 44 635 55 21. Fax: +41 44 635 68 62. E-mail: caflisch@bioc.uzh.ch.

ACKNOWLEDGMENT

We thank Dr. Andreas Vitalis for interesting discussions and comments to the manuscript. We thank Dr. Michele Seeber for the WORDOM implementation of the procedure for extraction of the largest ergodic component. This work was supported by a Swiss National Science Foundation grant to A.C.

REFERENCES

(1) McCammon, J. A.; Karplus, M. *Nat. Struct. Biol.* **2002**, *9*, 646–652.
(2) Krivov, S. V.; Chekmarev, S. F.; Karplus, M. *Phys. Rev. Lett.* **2002**, *88*, 038101.
(3) Dickson, A.; Dinner, A. R. *Annu. Rev. Phys. Chem.* **2010**, *61*, 441–459.
(4) Okamoto, Y. *J. Mol. Graph. Model* **2004**, *22*, 425–439.
(5) Melchionna, S. *Phys. Rev. E* **2000**, *62*, 8762–8767.
(6) Bonomi, M.; Parrinello, M. *Phys. Rev. Lett.* **2010**, *104*, 190601.
(7) Zhang, C.; Ma, J. *J. Chem. Phys.* **2010**, *132*, 244101.
(8) Muff, S.; Caflisch, A. *J. Phys. Chem. B* **2009**, *113*, 3218–3226.
(9) Snow, C. D.; Nguyen, H.; Pande, V. S.; Gruebele, M. *Nature* **2002**, *420*, 102–106.
(10) Paci, E.; Cavalli, A.; Vendruscolo, M.; Caflisch, A. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 8217–8222.
(11) Settanni, G.; Gsponer, J.; Caflisch, A. *Biophys. J.* **2004**, *86*, 1691–1701.
(12) Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 14122–14125.
(13) Pitera, J. W.; Swope, W. C.; Suits, F. *J. Phys. Chem. B* **2004**, *108*, 6571–6581.
(14) Singhal, N.; Snow, C. D.; Pande, V. S. *J. Chem. Phys.* **2004**, *121*, 415–425.
(15) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. *J. Chem. Phys.* **2007**, *126*, 155101.
(16) Noé, F. *J. Chem. Phys.* **2008**, *128*, 244103.
(17) Guarnera, E.; Pellarin, R.; Caflisch, A. *Biophys. J.* **2009**, *97*, 1737–1746.
(18) Pitera, J. W.; Chodera, J. D.; Swope, W. C.; Dill, K. A. *Multiscale Model. Simul.* **2006**, *5*, 1214–1226.
(19) Sriraman, S.; Kevrekidis, I. G.; Hummer, G. *J. Phys. Chem. B* **2005**, *109*, 6479–6484.
(20) Yang, S.; Banavali, N. K.; Roux, B. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 3776–3781.
(21) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. *J. Chem. Phys.* **2009**, *131*, 124101.
(22) Kemeny, J. G.; Snell, J. L. *Finite Markov Chains*; Undergraduate Texts in Mathematics; Springer-Verlag: New York, 1976.
(23) Tarjan, R. *SIAM J. Comput.* **1972**, *1*, 146–160.
(24) Rao, F.; Caflisch, A. *J. Mol. Biol.* **2004**, *342*, 299–306.
(25) Khinchin, A. I. *Mathematical Foundations of Information Theory*; Dover: New York, 1957.
(26) Muff, S.; Caflisch, A. *J. Chem. Phys.* **2009**, *130*, 125104.
(27) Bondy, A.; Murty, U. S. R. *Graph Theory*; Graduate Texts in Mathematics; Springer: New York, 2010.
(28) Caflisch, A. *Curr. Opin. Struct. Biol.* **2006**, *16*, 71–78.
(29) Fortunato, S.; Lancichinetti., A. *Phys. Rev. E* **2009**, *80*, 56117.
(30) Fortunato, S. *Phys. Rep.* **2010**, *486*, 75–174.
(31) Schuetz, P.; Caflisch, A. *Phys. Rev. E* **2008**, *78*, 26112.
(32) Grinstead, C, M.; Snell, J. L. *Introduction to Probability*, 2nd ed.; American Mathematical Society: Providence, RI, 1998.
(33) Ferrara, P.; Apostolakis, J.; Caflisch, A. *Proteins* **2002**, *46*, 24–33.
(34) Seeber, M.; Cecchini, M.; Rao, F.; Settanni, G.; Caflisch, A. *Bioinformatics* **2007**, *23*, 2625–2627.
(35) Seeber, M.; Felline, A.; Raimondi, G.; Muff, S.; Friedman, R.; Rao, F.; Caflisch, A.; Fanelli, F. *J. Comput. Chem.* **2011**, *32*, 1183–1194.
(36) Krivov, S. V.; Karplus, M. *J. Phys. Chem. B* **2006**, *110*, 12689–12698.
(37) Krivov, S. V.; Muff, S.; Caflisch, A.; Karplus, M. *J. Phys. Chem. B* **2008**, *112*, 8701–8714.
(38) Cavalli, A.; Ferrara, P.; Caflisch, A. *Proteins* **2002**, *47*, 305–314.
(39) Reingold, E. M.; Fruchterman, T. M. J. *Softw. Pract. Exper.* **1991**, *21*, 1129–1164.
(40) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 13841–13846.