Efficient exploration of discrete energy landscapes

Martin Mann

Bioinformatics Group, University of Freiburg, Georges-Köhler-Allee 106, D-79110 Freiburg, Germany

Konstantin Klemm

Bioinformatics Group, Institute for Computer Science, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany

Many physical and chemical processes, such as folding of biopolymers, are best described as dynamics on large combinatorial energy landscapes. A concise approximate description of the dynamics is obtained by partitioning the microstates of the landscape into macrostates. Since most landscapes of interest are not tractable analytically, the probabilities of transitions between macrostates need to be extracted numerically from the microscopic ones, typically by full enumeration of the state space or approximations using the Arrhenius law. Here, we propose to approximate transition probabilities by a Markov chain Monte Carlo method. For landscapes of the number partitioning problem and an RNA switch molecule, we show that the method allows for accurate probability estimates with significantly reduced computational cost.

I. INTRODUCTION

Energy landscapes [1–3] are a key concept for the description of complex physical and biological systems. In particular, the dynamics of structure formation (folding) of biopolymers, e.g., protein or ribonucleic acids, can be understood in terms of their energy landscapes [4,5]. Formally, a landscape is determined by a set X of microstates (or conformations), a neighborhood structure of X that encodes which conformations can be reached from which other ones, and an energy function $E:X \to \mathbb{R}$, which assigns an energy value to each state. In the case of ribonucleic acids (RNAs), it has been demonstrated that the dynamics of the folding process can be captured in good approximation by merging large contiguous sets of microstates into macrostates [6,7]. A typical mapping is in terms of gradient basins: Each macrostate contains the microstates from which a given local minimum is reached by steepest descent in energy, including the local minimum itself. The so-defined macrostates also are called inherent structures in the context of continuous disordered systems, see Ref. [8] for a recent review.

Given a partitioning of the landscape, the dynamics is approximately described as a Markov chain on the set of macrostates. In order to obtain this description, the transition probabilities between macrostates in this Markov chain need to be extracted from the original energy landscape.

As a first approximation, the *Arrhenius* equation predicts that the transition probability is exponentially suppressed by the ratio between barrier height and temperature. The barrier height (also called activation energy) from minimum a to minimum b measures the minimal amount by which the system's energy must increase along a path from a to b [6,9,10]. The accuracy of this approach is limited because it ignores the multiplicity of low-energy paths [11]. A more severe drawback is the complexity of computing barrier height itself. For landscapes of RNA secondary structure [12], the problem is nondeterministic polynomial hard [13,14].

Commonly used methods [6,15–18] for precise transition rate estimation are based on enumeration of all microstates. For landscapes of real combinatorial problems or long biopolymers with a large number microstates, however, enumeration is impractical with the given time resources. Typically, limited storage capacity puts even more severe restrictions on the size of tractable problems because a large fraction of the enumerated microstates needs to be kept in working memory. Some studies partially circumvent this problem by considering only the low-energy fraction of the landscape that is tractable with the available resources [9,19–21]. Other heuristic approaches [22–26] restrict the landscape to the subset of states likely to be traversed by certain trajectories, e.g., folding from the open chain to the ground state of a biopolymer.

Here, we make a contribution to the original challenge of capturing an arbitrary discrete landscape in terms of macrostates and transition probabilities. We suggest a Markov chain Monte Carlo sampling method for transition matrix estimation. Different from the earlier approaches, the memory requirement scales linearly with the number of nonzero transition probabilities to be determined. Other recent methods of stochastic landscape exploration [27,28] use trajectories of the original dynamics for counting transitions between macrostates. In contrast, the idea behind the present method is to explicitly explore boundaries between macrostates. To this end, we confine the dynamics into a single macrostate b and find and count possible transitions from b to all adjacent macrostates. This strategy allows us to select the regions of the landscape to be explored and the precision to be applied.

II. LANDSCAPE AND MICROSTATE DYNAMICS

A discrete energy landscape is a triple (X, E, M) where
(1) X is a finite set of states,
(2) $E:X \to \mathbb{R}$ is an energy function on X, and
(3) $M:X \to \mathcal{P}(X)$ is a neighborhood function or move set that assigns, to each state $x \in X$, the set of its directly accessible neighboring states. $\mathcal{P}(X)$ is the power set of X. Here, we assume that M is symmetric, i.e., $x \in M(y) \Rightarrow y \in M(x)$. By Δ, we denote the maximum number of neighbors, $\Delta = \max_{x \in X} |M(x)|$.

We consider a time-discrete stochastic dynamics on the state set X. Having the Markov property, the dynamics is defined by giving the transition probability $p_{x \to y}$ from each $x \in X$ to each $y \in M(x)$. Provided the system is in state x at time t, $p_{x \to y}$ is the probability that the system is in state y at time $t + 1$. With probability $p_{x \to x} = 1 - \sum_{y \in M(x)} p_{x \to y}$, the system remains at state x.

Specifically, the Metropolis probabilities at inverse temperature β,

$$p_{x \to y} = \Delta^{-1} \min\{\exp(\beta[E(x) - E(y)]), 1\} \quad (1)$$

are used throughout this contribution. This choice, however, is not compulsory. All that follows, and, in particular, the estimation by sampling, applies to arbitrary choices of transition probabilities leading to *ergodic* Markov chains. The ergodicity is important because we need a unique stationary distribution $P(x)$ on X.

III. PARTITIONING AND MACROSTATE DYNAMICS

A partitioning of the landscape is a mapping F from the set of microstates X into a set of macrostates B. Our goal here is to find a dynamics on B that does have the Markov property while following the original microstate dynamics as closely as possible. In general, however, a Markov chain is not obtained as the direct mapping $[F(x_t)]_{t=0}^{\infty}$ of a Markov chain $(x_t)_{t=0}^{\infty}$ generated by the dynamics on X. The reason can be sketched as follows. When the system is in a macrostate $b \in B$, the probability of exiting to a macrostate c depends on where exactly (in which microstate) the system is inside b. The microstate assumed inside b, however, depends on how the system entered b, which is again influenced by the macrostate a assumed before entering b.

Thus, the following simplified assumption is made [29]. Given that the system is found in macrostate $b \in B$, the microstate $x \in X$ is distributed as

$$P_b(x) = \begin{cases} P(x) / \sum_{y \in F^{-1}(b)} P(y), & \text{if } x \in F^{-1}(b), \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

This is the stationary distribution P of the whole system restricted to microstates in b and normalized appropriately. Under this assumption, the probability of a transition to macrostate c, when being in macrostate $b \neq c$ is

$$q_{b \to c} = \sum_{x \in F^{-1}(b)} \left(P_b(x) \sum_{y \in M(x) \cap F^{-1}(c)} p_{x \to y} \right). \quad (3)$$

The inner sum is the probability of going to a microstate y belonging to macrostate c and being a neighbor of x, given that the system is in state x. The outer sum represents the equilibrium weighting of the microstates x inside the given macrostate b. A straightforward method determines the exact transition probabilities by performing the sums in Eq. (3), i.e., exhaustive enumeration of all microstates and all neighbors [6,17].

Throughout this contribution, we consider the usual partitioning of X with respect to gradient basins, but the method is not restricted to this choice. Two microstates $x, y \in X$ lie in the same macrostate $F(x) = F(y)$ if and only if the steepest descent walks starting in x and y terminate in the same

local minimum. A state $u \in X$ is called local minimum, if $E(v) > E(u)$ for all $v \in M(u)$. For a given landscape and partitioning, the macrostate transition probabilities can be estimated by the sampling algorithm presented in Sec. IV.

IV. SAMPLING METHOD

The method we introduce computes an estimate of the transition probabilities q in Eq. (3) by a standard importance sampling restricted to a macrostate b using the microstate probabilities $P_b(x)$ defined in Eq. (2). Being in state $x_t \in F^{-1}(b)$ at time t, a neighbor $z \in M(x_t)$ is drawn at random with equal probabilities. The suggestion is accepted as the next state, $x_{t+1} = z$, with probability $\min\{1, P_b(z)/P_b(x_t)\}$. Otherwise, the state remains the same, $x_{t+1} = x_t$. This choice guarantees that the relative frequency of state x tends toward the relative frequency $P_b(x)$ for increasing chain length $t \to \infty$ [30]. For a realization of a Markov chain of length $t_{\max}$, transition probabilities are estimated as

$$q'_{b \to c} = \frac{1}{t_{\max}} \sum_{t=1}^{t_{\max}} \sum_{y \in M(x_t) \cap F^{-1}(c)} p_{x_t \to y}. \quad (4)$$

In practice, the inner summation is performed only once at each time t because each neighbor y of x_t contributes to the transition probability to exactly one macrostate $F(y)$.

Computation time is saved by storing visited microstates of basin b and their sets of neighbors with transition probabilities in a data structure with fast search access, e.g., in a hash table. This is particularly advantageous in cases with broadly distributed microstate probabilities, such as Boltzmann weights at low temperature. Here, the Markov chain will encounter the highly probable (low-energy) microstates many times, but neighbor sets and transition probabilities are computed only once per state. In the usual cases where macro-states are defined as basins of local minima, memory of visited states also saves time in evaluating the macrostate assignment function F: When the gradient walk starting at state x reaches a microstate known to be in basin b, x itself is known to belong to b. Thus, in many cases, the walk may be terminated before reaching the ground state. Keeping previously visited microstates in memory, however, is not necessary for the method to work. It may be handled according to the available resources. One may simply stop storing microstates when the designated memory has been filled.

So far, we have described how to estimate probabilities of transitions from *one* macrostate b to others. The result is the bth column vector $(q'_{b \to c})_{c \in B}$ of the estimated transition matrix q' as given in Eq. (4). By applying the procedure separately to each macrostate, the full matrix q' is obtained. This can be implemented as an iterative exploration of the energy landscape without initial knowledge of the set of macrostates. Whenever a neighbor y of a state x in the Markov chain belongs to a macrostate $F(y)$ not previously seen, we add the pair $[F(y), y]$ to a queue Q of macrostates yet to be worked on. Initially, Q may contain only one particular pair (b, x_0), e.g., the completely unfolded state x_0 of a polymer and the corresponding macrostate $b = F(x_0)$. The iterative exploration of the landscape is implemented in the following loop. (i) Extract a pair (b, x_0) from Q;

(ii) generate a Markov chain inside b, starting at x_0; (iii) obtain estimates according to Eq. (4) and add newly discovered macrostates to Q; (iv) if Q is not empty, resume at (i). Note, this method is directly parallelizable and will easily profit from distributed computing. Several independent realizations of Markov chains with respect to different macrostates can be run simultaneously, extracting from and feeding to the same queue. An implementation of the method is part of the Energy Landscape Library [31].

V. NUMBER PARTITIONING LANDSCAPE

The number partitioning problem (NPP) is a decision problem in the theory of computation and computational complexity [32–34]. It asks if a given set A of N real non-negative numbers can be partitioned into two subsets B,C such that numbers in B have the same sum as those in C. In an equivalent formulation, we label the numbers in A as $a_1,\dots,a_N$ and use spin variables $x_1,\dots,x_N$ to encode if a_i is in subset B ($x_i = +1$) or in subset C ($x_i = -1$). This system has the set of microstates $X = \{-1, +1\}^N$. We define the energy of state $x \in X$ as

$$E(x) = \left| \sum_{i=1}^{N} x_i a_i \right|. \tag{5}$$

Then, the NPP amounts to the question if the ground-state energy of this system is zero.

The number partitioning *landscape* is obtained by using the hypercube as the neighborhood structure. For each $x \in X$, we have

$$M(x) = \{y \in X | d(x,y) = 1\}, \tag{6}$$

as the set of neighbors. The usual Hamming distance d is used, so $d(x,y)$ is the number of entries i such that $x_i \neq y_i$. A local move on the landscape means flipping one of the N spin variables x_i.

Random instances are typically generated by drawing the a_i as statistically independent random variables uniformly distributed in the unit interval. Then, the expected number of local minima grows exponentially with N, more precisely $\langle |B| \rangle \sim 2^N N^{-3/2}$ [35]. Here, we use special instances of the NPP where

$$a_i = (i-1)^{-\alpha}, \tag{7}$$

with $\alpha = 0.55$. For these instances, we have found the number of local minima to grow exponentially with N for $N \leqslant 40$. However, the growth is much slower than for randomly generated instances. At $N = 40$, the instance of Eq. (7) has 318 local minima, to be compared with an expected number of $\approx 10^{15}$ local minima for randomly generated instances. Each landscape with $10 \leqslant N \leqslant 40$ has, at least, one basin with an energy barrier $\geqslant 0.1$. We explore the landscape at temperature $1/\beta = 0.1$.

Figure 1 shows the convergence of the probability estimates. For each system size N, the sampling error decreases inversely proportional to the number of sampling steps performed per basin. Larger systems need more computational effort to reach a certain precision. The inset of Fig. 1 indicates that the total computational effort required for the error to
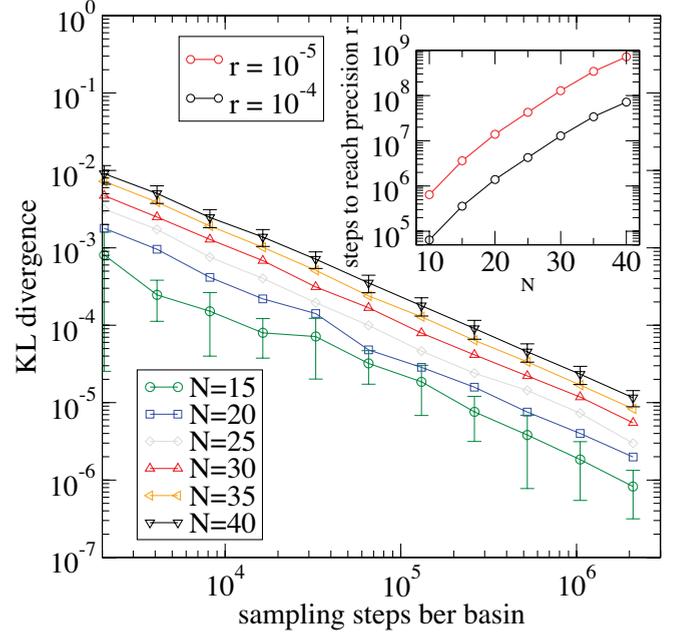
FIG. 1. (Color online) The deviation of estimated transition probabilities from the exact values is inversely proportional to the number of sampling steps (main panel). Shown are the analyzed special instances of number partitioning landscapes [see Eq. (7)] for various sizes N. Error bars ($N = 15$ and $N = 40$) indicate the standard deviation between errors for different basins. The inset shows the N dependence of the total number of sampling steps required for reaching a given precision, i.e., lowering the error below r. Given a macrostate a, we employ the Kullback-Leibler (KL) divergence $D(\cdot || \cdot \cdot)$ [36] to define the error as $\epsilon(s,a) := D[q'(s,a)||q'(2s,a)]$, making a comparison of the estimate of the outgoing transition probability vector $q'(s,a) = (q'_{a \to b})_{b \in B}(s)$ after s sampling steps with its estimate after $2s$ sampling steps. The plotted values are the equally weighted average of the errors $\epsilon(s,a)$ over all macrostates $a \in B$.

fall below a given value grows subexponentially with N, to be compared with a number of microstates increasing as 2^N. Thus, under growing N, sampling a strongly decreasing fraction of microstates is sufficient in order to reach a given precision.

VI. FOLDING LANDSCAPE OF AN RNA SWITCH

As a real-world example of folding landscapes of biopolymers, we consider RNA molecules [37]. The primary structure of an RNA molecule is a finite sequence (a string) over the alphabet of the four nuclear bases {A,C,G,U}. An RNA secondary structure is a list of pairs (i,j) of positions in the primary structure such that the following conditions hold. (1) Base combinations at pairing positions must be A-U or G-C (Watson-Crick pairs) or G-U (wobble pair); (2) each position i can pair with at most one other position j; (3) there are no two pairs (i,j) and (k,l) with $i < k < j < l$. The latter condition forbids so-called pseudoknots and makes the graph representation of a secondary structure outer planar (see Fig. 2).

In the folding landscape of an RNA sequence, the set of microstates X contains the valid secondary structures. The energy $E(x)$ of a secondary structure $x \in X$ is a sum over
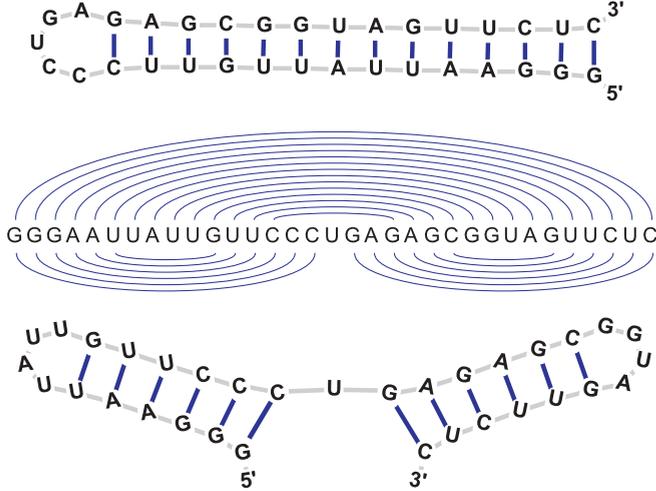
FIG. 2. (Color online) RNA secondary structures (top and bottom) with energies $-14.4\frac{\text{kcal}}{\text{mol}}$ and $-14.3\frac{\text{kcal}}{\text{mol}}$ of the tested bistable RNA d33 and their outer-planar linear Feynman diagrams (middle) (drawn using jViz.Rna v1.77 [38]). Energy evaluation and sequence design are based on the Vienna RNA package v1.8.2 [39] and the method from Ref. [40].

binding energies of stacks (contiguous regions of binding) and entropic contributions from open (unbound) sections of the RNA chain. For details of energy calculations, we refer to the literature [39,41,42]. Microstates $x,y \in X$ are adjacent, i.e., $y \in M(x)$ and $x \in M(y)$, if y can be generated from x by adding or removing a single base pair. Shift moves [37] are not considered in this paper. When the lowest-energy neighbor of a structure is not unique, the degeneracy is resolved by the lexicographic ordering on string representations of the structures [6,7,9,37].

Multistable RNAs, so-called RNA switches, are essential for the regulation of cellular processes. Thus, an understanding of the folding kinetics of such molecules is of great importance. For a detailed overview, see Ref. [40]. Specifically, we work with the bistable RNA d33 sequence shown in Fig. 2. It has 29 759 371 microstates, allowing for full enumeration and, thus, for a comparative analysis with our method. Out of the 3223 local minima, the two lowest are the secondary structures given in Fig. 2. These two ground states practically have the same energy. A walk between the ground states involves breaking all base pairs, resulting in an energy barrier of height $\Delta E = 1.18 \times 10^{-19}$ J. The temperature for both sampling and energy calculation is $T = (273.15 + 37.00)$ K. Therefore, $1/\beta = k_B T = 4.28 \times 10^{-21}$ J is more than 1 order of magnitude below the barrier height ΔE of the RNA switch.

A comparison between exact and sampled transition probabilities is made in terms of the average time $\tau(b)$ from macrostate b to one of the ground states. For a biopolymer as considered here, $\tau(b)$ is the folding time when starting in an initial state b, such as the open chain.

Given a set of target states $A \subset B$, the time to target is $\tau(a) = 0$ when starting in one of the target states $a \in A$ (boundary condition). For a starting state $b \in B \setminus A$, the average time $\tau(b)$, until first reaching one of the target states,
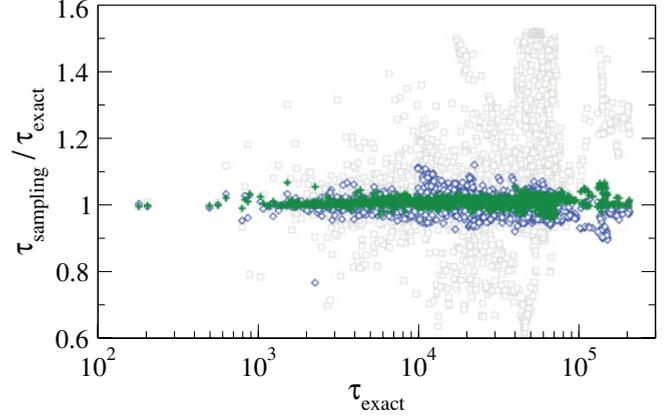
FIG. 3. (Color online) Sampling precision in terms of the predicted average time $\tau(b)$ to reach the ground state from a macrostate b for RNA d33. For each b, the corresponding data point gives the ratio between τ_{sampling} based on the sampled transitions q' and the value τ_{exact} from the exact ones q versus τ_{exact} itself. Symbols indicate the number of sampling steps per macrostate as 10^3 (squares), 10^4 (diamonds), and 10^5 (crosses). The target set contains both ground states.

obeys the recursion,

$$\tau(b) = 1 + \sum_{c \in B} q_{b \to c} \tau(c). \tag{8}$$

The average time to target from b is one time step plus the time to target from state c following b. The distribution of c is given by the transition probability $q_{b \to c}$. Time to target is also called exit time [43].

Figure 3 shows that $\tau(b)$ values based on the sampled transition probabilities have small relative error for all starting macrostates $b \in B$. With 10^4 sampling steps per basin, the ratios between exact and approximate times τ are in the range [0.75; 1.15]. They fall into [0.96; 1.07] when using 10^5 steps per basin.

To investigate the sampling error, we compare, separately for each macrostate b, the exact ones with the estimated transition probability vectors for leaving b. We quantify the discrepancy between the two vectors by the KL divergence [36],

$$D(q'_b || q_b) = \sum_{c \in B} q'_{b \to c} \ln \frac{q'_{b \to c}}{q_{b \to c}}. \tag{9}$$

Figure 4 summarizes the evolution of the sampling error for increasing sampling steps per basin. As in the number partitioning landscape, KL divergence decreases inversely proportional to the number of sampling steps. The mean of the distribution of KL values across basins is larger than the median by a factor of 3, indicating a broad distribution. This is due to a broad distribution of macrostate sizes. Probability vectors for macrostates comprising one or a few microstates reach a low error after fewer sampling steps than those for large macrostates. Still, the maximum of the error across all basins also decreases proportionally to the average. One of the extensions of the method outlined in Sec. VII chooses a number of sampling steps individually for each macrostate based on its estimated partition function.
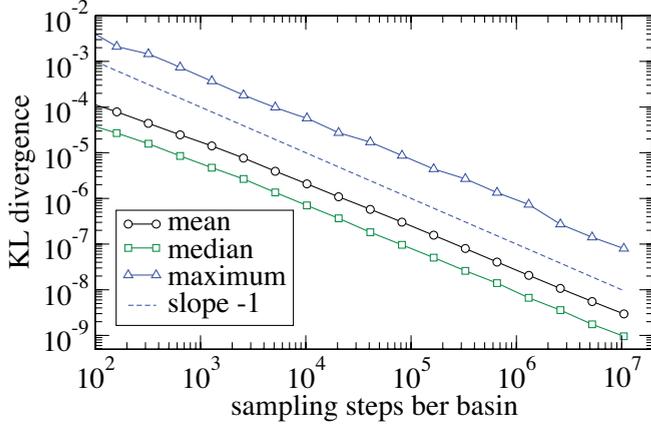
FIG. 4. (Color online) Time evolution of the sampling error [KL divergence Eq. (9)] for the folding landscape of the RNA switch molecule d33. Mean, median, and maximum are for the distribution of KL values over the $|B| = 3223$ macrostates (local minima).

In Fig. 5, we compare the kinetics of the molecule for the approximated transition probabilities via sampling and the exact ones obtained by enumeration of all microstates of the landscape. As an initial condition, we choose the whole ensemble to be in the macrostate of the open chain (structure without base pairs). As a qualitative description, the ensemble first populates the first and, somewhat later, the second ground state. On an intermediate time scale (10^6), an almost constant concentration vector is reached with the
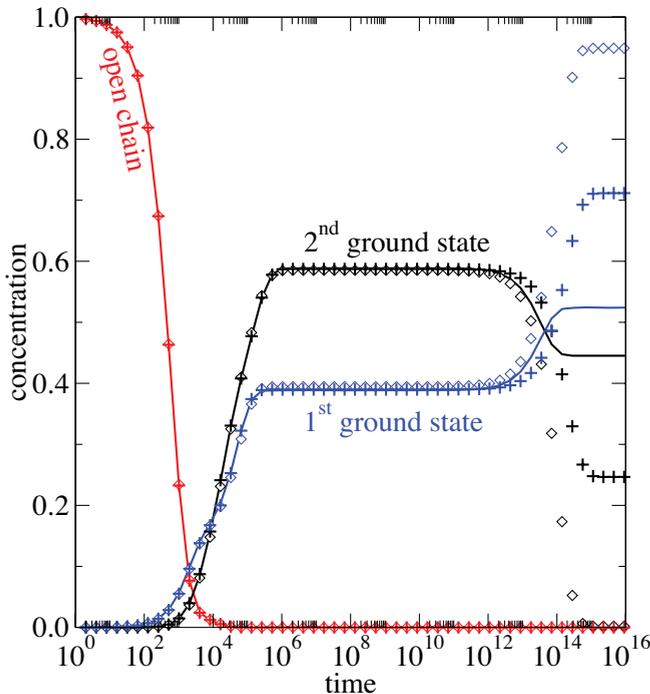
FIG. 5. (Color online) Time evolution of macrostate concentrations for RNA d33 from the exact transition probabilities q (curves) and from the estimates q' via sampling for 10^4 ($\diamond$) and 10^5 ($+$) steps per macrostate. The concentrations both for the exact and the sampled transition rates are at the stationary values for $t \geqslant 10^{15}$. See main text for a discussion of the discrepancy in these stationary concentrations.

second ground state dominating the first. However, this plateau concentration vector is transient. Probability mass flows from the second to the first ground state on a slow time scale (10^{15}) to reach the stationary concentrations.

With transition probabilities obtained by sampling for 10^4 steps per macrostate, the kinetics is reproduced with high precision both in the timing as well as in the absolute concentration in the plateau where the relative error is below 10^{-2}. Since we hash the probabilities for already visited structures, the computational effort per basin is dominated by the number of visited states instead of overall sampling steps. Thus, small basins are sampled faster than larger ones. When sampling 10^5 steps per macrostate, which renders the kinetics with even higher accuracy, computation time is still reduced by a factor of ≈ 9 compared with full enumeration.

The stationary concentrations found at time $t \geqslant 10^{15}$, however, do not agree with the exact solution. A much larger number of sampling steps is required to match these. Further tests with other RNA switch molecules yield the same qualitative result for a moderate number of sampling steps per basin. Both the concentration levels and the time scales are faithfully reproduced by the transition rates from sampling, except for the stationary concentrations.

A closer look at the particular structure of the landscape of RNA switches hints at an explanation for the discrepancy. Both ground states have large and deep basins. In RNA d33, barriers to neighboring basins are all at least 8.6×10^{-20} J above ground-state energy, which is more than $20 k_B T$. Exits from one ground-state basin toward the other lead through a small number of microstates. When sampling the large ground-state basins, these few salient microstates are likely to be missed. The subtle balance between incoming and outgoing probability flows in the equilibrium is distorted. Due to the symmetry of the move set, however, those missed nonzero transition probabilities can be identified to some extent. When the forward rate is nonzero, then the backward rate must be nonzero as well. For dynamics with detailed balance, as considered here, even quantitative correction of missed or undersampled rates is possible. This is suggested as one of the extensions in Sec. VII.

VII. EXTENSIONS AND MODIFICATIONS OF THE METHOD

We outline ideas for varying the method to potentially increase efficiency and applicability in various settings. These are not used in the applications in Secs. V and VI.

A. Guided sampling

The stopping criterion (iv) of the outer loop for full transition matrix estimation (Sec. IV) may be modified if we do not aim to explore the whole landscape but only a subset of the set of macrostates [18]. Then, Q may be handled as a priority queue. For instance, we may be interested only in transitions between macrostates below a certain energy threshold or those involved in typical trajectories. In the latter case, the next macrostate to be explored is the one that is reached from already explored macrostates with the largest probability.

B. Partition function estimation

In addition to transition probabilities, the canonical partition function,

$$Z_b = \sum_{x \in F^{-1}(b)} \exp[-\beta E(x)], \tag{10}$$

of the macrostate b may be estimated at any time during the sampling. Consider a subset $X \subseteq F^{-1}(b)$ of the microstates of macrostate b. When sampling in the macrostate b, the fraction of time the Markov chain spends in X is

$$r = \frac{1}{Z_b} \sum_{x \in X} \exp[-\beta E(x)]. \tag{11}$$

Therefore, Z_b can be calculated when knowing r and the energies of all states in X. Now X can be taken as the set of states visited in the first t^* steps of the Markov chain, $X = \{x_t | t = 1, \ldots, t^*\}$. An estimate r' of r is obtained by counting how often the Markov chain visits states in X during a sufficiently long time interval $[t_{\text{start}}, t_{\text{stop}}]$,

$$r' = \frac{|\{t | t_{\text{start}} \leqslant t < t_{\text{stop}} \wedge x_t \in X\}|}{t_{\text{stop}} - t_{\text{start}}}. \tag{12}$$

In order to obtain an unbiased estimate of r, this time interval must not overlap with the time steps during which X is recorded, thus, $t_{\text{start}} > t^*$. By solving Eq. (11) for Z_b and replacing r with the estimate r', we obtain

$$Z_b' = \frac{1}{r} \sum_{x \in X} \exp[-\beta E(x)], \tag{13}$$

as an unbiased estimate of the partition function Z_b.

C. Sampling time adjustment

The estimate $r(t)$ may also be used for adapting the length of the Markov chain exploring macrostate b to the size of b. The sampling will be run until the fraction of covered probability mass exceeds a certain threshold, e.g., stopping as soon as $r(t) > 0.5$.

D. Landscape coarse graining

The state space can be coarse grained beyond the initially chosen macrostate state partitioning by dynamically merging macrostates [44]. For merging macrostate b into macrostate a, the affected entries in the matrix q are replaced by

$$\hat{q}_{a \to c} = \frac{Z_a q_{a \to c} + Z_b q_{b \to c}}{Z_a + Z_b}, \tag{14}$$

$$\hat{q}_{c \to a} = q_{c \to a} + q_{c \to b} \tag{15}$$

for all macrostates $c \notin \{a, b\}$. The new diagonal element $\hat{q}_{a \to a}$ is obtained by normalization of probability. The row and column of macrostate b are set to zero (or are deleted). In a separate index, the mapping of macrostate b to macrostate a is stored.

Strategies for the choice of macrostates to be merged need to be explored yet. A reasonable starting point is to choose pairs of macrostates with high overlap in successor states,

e.g., choosing a and b such that

$$\sum_{c \in B} q_{a \to c} q_{b \to c} \tag{16}$$

is maximal.

E. Balancing the transition probability matrix

If the microstate dynamics in terms of the transition probabilities $p_{x \to y}$ fulfills detailed balance, then so does the macrostate dynamics with transition probabilities $q_{b \to c}$. Detailed balance means

$$Z_l q_{l \to s} = Z_s q_{s \to l} \tag{17}$$

for all macrostate pairs (l, s). The transition probabilities $q'_{l \to s}$ obtained by the sampling, however, need not fulfill the same condition. By the transformation,

$$q_{l \to s}^* = \frac{1}{2} q'_{l \to s} + \frac{1}{2} q'_{s \to l} \frac{Z_s}{Z_l}, \tag{18}$$

transition probabilities q^* with detailed balance are obtained. The transformation also serves to impose a known stationary distribution of concentrations on the transition probability matrix.

VIII. CONCLUSION AND DISCUSSION

When coarse graining the state space of an energy landscape into macrostates, transition probabilities between macrostates have to be obtained in order to capture the coarse-grained stochastic dynamics. Here, we have introduced a sampling method that allows for a fast yet accurate estimation of these transition probabilities. We have demonstrated the scalability of the approach with a system size for special instances of the number partitioning problem. As a real-world application, we have analyzed the folding landscape of the secondary structure of an RNA switch as an example of a biopolymer. Its rich dynamic behavior on separate fast and slow time scales is accurately rendered by transition probabilities obtained with low computational cost.

The general method introduced here may serve as a flexible framework for stochastic exploration of energy landscapes. As laid out in Sec. VII, several extensions and modifications may be made to obtain increased performance and wider applicability. In particular, the high variation of macrostate sizes may be exploited in a scheme for an automatic choice of sampling effort. Furthermore, the merging of small macrostates with larger neighbors during the sampling may lead to more manageable and potentially more meaningful partitions of the landscape akin to metabasins [8].

In ongoing and future work, the method will be applied to other energy landscapes including those of state-discrete protein folding dynamics [45–47]. Such landscapes have been shown to be amenable to sampling approaches [48]. Another field of application of our method is the clarification of concepts for dynamics on energy surfaces, such as the notion of a folding funnel [11,49,50].

[1] C. M. Reidys and P. F. Stadler, SIAM Rev. **44**, 3 (2002).

[2] D. Wales, *Energy Landscapes* (Cambridge University Press, Cambridge, U.K., 2003).

[3] Q. Zhou and W. H. Wong, Phys. Rev. E **79**, 051117 (2009).

[4] C. B. Anfinsen, Science **181**, 223 (1973).

[5] D. Ming, M. Anghel, and M. E. Wall, Phys. Rev. E **77**, 021902 (2008).

[6] C. Flamm, I. L. Hofacker, P. F. Stadler, and M. T. Wolfinger, Z. Phys. Chem. **216**, 155 (2002).

[7] M. T. Wolfinger, W. A. Svrcek-Seiler, C. Flamm, I. L. Hofacker, and P. F. Stadler, J. Phys. A **37**, 4731 (2004).

[8] A. Heuer, J. Phys. Condens. Matter **20**, 373101 (2008).

[9] M. T. Wolfinger, S. Will, I. L. Hofacker, R. Backofen, and P. F. Stadler, Europhys. Lett. **74**, 726 (2006).

[10] M. Baiesi, L. Bongini, L. Casetti, and L. Tattini, Phys. Rev. E **80**, 011905 (2009).

[11] P. Garstecki, T. X. Hoang, and M. Cieplak, Phys. Rev. E **60**, 3219 (1999).

[12] W. Fontana, P. F. Stadler, E. G. Bornberg-Bauer, T. Griesmacher, I. L. Hofacker, M. Tacker, P. Tarazona, E. D. Weinberger, and P. Schuster, Phys. Rev. E **47**, 2083 (1993).

[13] J. Manuch, C. Thachuk, L. Stacho, and A. Condon, in *Proceedings of DNA Computing and Molecular Programming*, Lecture Notes in Computer Science Vol. 5877 (Springer, 2009), p. 106 .

[14] C. Thachuk, J. Manuch, L.-A. M. Arash Rafiey, L. Stacho, and A. Condon, in *Proceedings of Pacific Symposium on Biocomputing, Kamuela, Hawaii, 2010* (World Scientific, Singapore, 2010), Vol. 15, p. 108 .

[15] H. S. Chan and K. A. Dill, J. Chem. Phys. **100**, 9238 (1994).

[16] S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster, Biopolymers **49**, 145 (1999).

[17] P. Sibani, R. van der Pas, and J. C. Schön, Comput. Phys. Commun. **116**, 17 (1999).

[18] Z. Burda, A. Krzywicki, O. C. Martin, and Z. Tabor, Phys. Rev. E **73**, 036110 (2006).

[19] P. Sibani, J. C. Schön, P. Salamon, and J.-O. Andersson, Europhys. Lett. **22**, 479 (1993).

[20] P. Sibani and P. Schriver, Phys. Rev. B **49**, 6667 (1994).

[21] J. C. Schön and P. Sibani, J. Phys. A **31**, 8165 (1998).

[22] A. Xayaphoummine, T. Bucher, F. Thalmann, and H. Isambert, Proc. Natl. Acad. Sci. U.S.A. **100**, 15310 (2003).

[23] X. Tang, B. Kirkpatrick, S. Thomas, G. Song, and N. M. Amato, J. Comput. Biol. **12**, 862 (2005).

[24] S. Thomas, G. Song, and N. M. Amato, Phys. Biol. **2**, S148 (2005).

[25] S. Thomas, X. Tang, L. Tapia, and N. M. Amato, J. Comput. Biol. **14**, 839 (2007).

[26] M. Geis, C. Flamm, M. T. Wolfinger, A. Tanzer, I. L. Hofacker, M. Middendorf, C. Mandl, P. F. Stadler, and C. Thurner, J. Mol. Biol. **379**, 160 (2008).

[27] D. Gfeller, P. De Los Rios, A. Caflisch, and F. Rao, Proc. Natl. Acad. Sci. **104**, 1817 (2007).

[28] D. Prada Gracia, J. Gómez-Gardeñes, P. Echenique, and F. Falo, PLoS Comput. Biol **5**, e1000415 (2009).

[29] H. A. Kramers, Physica **7**, 284 (1940).

[30] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, and A. H. Teller, J. Chem. Phys. **21**, 1087 (1953).

[31] M. Mann, S. Will, and R. Backofen, in *Proceedings of BIRD'07* (OCG, 2007), Vol. 217, p. 83 .

[32] M. R. Garey and D. S. Johnson, *Computers and Intractability* (Freeman, San Francisco, 1979).

[33] S. Mertens, Theor. Comput. Sci. **265**, 79 (2001).

[34] P. F. Stadler, W. Hordijk, and J. F. Fontanari, Phys. Rev. E **67**, 056701 (2003).

[35] F. F. Ferreira and J. F. Fontanari, J. Phys. A **31**, 3417 (1998).

[36] S. Kullback, American Statistician **4**, 340 (1987).

[37] C. Flamm, W. Fontana, I. Hofacker, and P. Schuster, RNA **6**, 325 (2000).

[38] K. C. Wiese and E. Glen, in *IEEE Symposium on Computer-Based Medical Systems* (IEEE, Los Alamitos, CA, 2006), Vol. 659 (2006).

[39] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster, Chem. Monthly **125**, 167 (1994).

[40] C. Flamm, I. L. Hofacker, S. Maurer-Stroh, P. F. Stadler, and M. Zehl, RNA **7**, 254 (2000).

[41] I. Tinoco, O. C. Uhlenbeck, and M. D. Levine, Nature (London) **230**, 362 (1971).

[42] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner, Proc. Nat. Acad. Sci. U.S.A. **83**, 9373 (1986).

[43] G. R. Grimmett, D. R. Stirzaker, *Probability and Random Processes* (Oxford University Press, New York, 2001).

[44] L. Bongini, L. Casetti, R. Livi, A. Politi, and A. Torcini, Phys. Rev. E **79**, 061925 (2009).

[45] K. A. Dill, Biochemistry **24**, 1501 (1985).

[46] M. Mann, S. Will, and R. Backofen, BMC Bioinformatics **9**, 230 (2008).

[47] M. Mann, D. Maticzka, R. Saunders, and R. Backofen, HFSP J. **2**, 396 (2008).

[48] T. Wüst and D. Landau, Comput. Phys. Commun. **179**, 124 (2008), ISSN 0010-4655, special issue based on the Conference on Computational Physics 2007—CCP 2007.

[49] P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **89**, 8721 (1992).

[50] K. Klemm, C. Flamm, and P. F. Stadler, Eur. Phys. J. B **63**, 387 (2008).