Particle Filter Method to Integrate High-Speed Atomic Force Microscopy Measurements with Biomolecular Simulations

Sotaro Fuchigami*, Toru Niina, and Shoji Takada*

Department of Biophysics, Graduate School of Science,

Kyoto University, Kyoto 606-8502, Japan

* Correspondence to:

Sotaro Fuchigami

sotaro@biophys.kyoto-u.ac.jp

Phone: +81-75-753-4220, Fax: +81-75-753-4222

&

Shoji Takada

takada@biophys.kyoto-u.ac.jp

Phone: +81-75-753-4220, Fax: +81-75-753-4222

ABSTRACT

  High-speed atomic force microscopy (HS-AFM) can be used to observe the structural dynamics of biomolecules at the single-molecule level in real time under near-physiological conditions; however, its spatiotemporal resolution is limited. Complementarily, molecular dynamics (MD) simulations have higher spatiotemporal resolutions, albeit with some artifacts. Here, to integrate HS-AFM data and coarse-grained molecular dynamics (CG-MD) simulations, we develop a particle filter method that implements a sequential Bayesian data assimilation approach. We test the method in a twin experiment. First, we generate a reference HS-AFM movie from the CG-MD trajectory of a test molecule, a nucleosome; this serves as the "experimental measurement." Then, we perform a particle filter simulation with 512 particles, which captures the large-scale nucleosome structural dynamics compatible with the AFM movie. Comparing particle filter simulations with 8–8192 particles, we find that using greater numbers of particles consistently increases the likelihood of the whole AFM movie. By comparing the likelihoods for different ionic concentrations and timescale mappings, we find that the "true" concentration and timescale mapping can be inferred as the largest likelihood of the whole AFM movie but not that of each AFM image. The particle filter method provides a general approach for integrating HS-AFM data with MD simulations.

INTRODUCTION

Biomolecular functions are generally realized via their structural dynamics. Albeit they are of

central importance, characterizing protein structures at both high spatial and temporal resolutions

remains a major unsolved challenge. On the one hand, X-ray crystallography, cryo-electron

microscopy, and other techniques can provide structure models at high spatial resolution, i.e.,

atomic resolution, though these are limited mainly to static snapshot information without time

resolution. On the other hand, time-resolved spectroscopies, fluorescent imaging, and other

techniques can directly observe the temporal dynamics of proteins, though their spatial resolution

is limited. Among the other possible techniques, high-speed atomic force microscopy (HS-AFM)

has the unique feature of being able to directly observe biomolecular structural dynamics under

near-physiological conditions and at the single-molecule level.[1–7] However, the resolutions of the

obtained AFM movies are typically limited in both time and space to ~sub-100 ms and ~1 nm (in

the lateral direction to an AFM stage), respectively. Moreover, the HS-AFM data of biomolecules

generally contain noise. Thus, HS-AFM data by themselves are not sufficient to model structural

dynamics at near-atomic resolutions.

On the other hand, molecular dynamics (MD) simulations provide information with a very high

spatiotemporal resolution: ~1 Å in space and ~1 fs in time. However, MD simulations are limited

to reachable timescales; typically, these are ~1 µs for the standard all-atom model under standard computational capacities.[8,9] More importantly, MD simulations are based on force fields that approximate to the "true" atomic forces. While recent progress in force fields has in many cases enabled MD simulations to estimate observables within experimental error margins, in some cases these MD simulation estimations deviate significantly from the experimental data. In such cases, it is crucial to systematically correct the MD simulation data using experimental data. In fact, many recent studies have attempted to develop methodologies to achieve this integration.[10–20]

In this study, we develop a particle filter method to integrate HS-AFM data with MD simulations. Using this combination, we aim to exploit the advantages of both types of data to obtain high-resolution structural dynamics insights. Our particle filter method uses a sequential Bayesian data assimilation approach.[21–24] Compared against other data assimilation approaches, such as the Kalman filter method for linear dynamical system and the ensemble Kalman filter method, the particle filter method offers one of the most general yet computer-intensive approaches. The method represents the probability distribution of structures at a time as a finite set of structure samples—referred to as "particles"—for data assimilation. For all particles (i.e., structure samples), MD simulations are used to propagate protein structures for a short time period corresponding to the time interval between two consecutive HS-AFM image acquisitions in a procedure referred to

as "prediction." This is followed by a likelihood estimation of the HS-AFM still-images at the time and a resampling of particles according to the estimated likelihoods in the so-called "filtering" procedure. Then, we proceed to the next round of prediction and filtering procedures. Alternate iterations of these two procedures results in an ensemble of structural dynamics that are compatible with the experimental time-series data. Considering the gap between the reachable timescales of all-atom MD simulations and the time resolutions of HS-AFM measurements, we chose a coarse-grained (CG) molecular model for the simulation; with this, we could accelerate the MD simulation by several orders of magnitude, achieving ~ms timescales.[25,26]

The particle filter method has been successfully applied to nonlinear deterministic dynamical systems with relatively few degrees of freedom. Compared to many of these cases, the application of the particle filter method to HS-AFM data poses major challenges. First, AFM provides surface height data of molecules bound on the stage plane, which is displayed as a two-dimensional image. Thus, filtering is performed in the measurement space of the pixel numbers, which contains unusually numerous degrees of freedom.[24,27,28] Second, biomolecular dynamics are Brownian and intrinsically stochastic. How this intrinsic stochasticity is reconciled with the particle filter method remains an open question. Recently, Matsunaga et al. successfully developed a particle filter method for the Förster resonance energy transfer data of a highly stochastic system—that of a

5

folding protein.[14] Third, an order-of-magnitude difference exists between the timescales of a single

MD step in CG-MD simulations (~1 ps per MD step) and the time interval between two consecutive

frames in HS-AFM measurements (~sub-100 ms per frame). Data assimilation in such cases poses

a major challenge.

THEORY AND METHODS

In this section, we first introduce the general theory of the particle filter method. Then, the specific

procedures for particle filter simulations using AFM measurement data and CG-MD simulations

are described. Next, details of the CG-MD simulation are provided. Finally, the protocol of the

twin experiment used to test the method is presented.

Particle Filter: General Theory

The particle filter method—also referred to as sequential Monte Carlo—is a general Bayesian data

assimilation approach that integrates experimental time-series data with a simulation model. We

begin this subsection with a brief introduction to the general theory (see, e.g., Refs. 22 and 23 for

a thorough description). For a simpler explanation of the particle filter theory, we use a discrete

time description, in which time takes the integer values $t = 0, 1, \cdots, \text{T}$. In the current context, one

integer corresponds to the time interval between two consecutive experimental data acquisitions, which is considerably longer than the time for a single MD step. We denote the molecular structure of interest at time t as a state-space vector $\boldsymbol{x}_t$ and model the system time propagation as a stochastic simulation model, as follows:

$$\boldsymbol{x}_t = \boldsymbol{f}_t(\boldsymbol{x}_{t-1}, \boldsymbol{v}_t). \tag{1}$$

Here, $\boldsymbol{f}_t$ defines the system time propagation, and $\boldsymbol{v}_t$ is a system noise vector. Next, we define the experimental measurement at time t as a vector $\boldsymbol{y}_t$; this vector depends on the state-space vector $\boldsymbol{x}_t$ and can be written as

$$\boldsymbol{y}_t = \boldsymbol{h}(\boldsymbol{x}_t, \boldsymbol{w}_t), \tag{2}$$

where $\boldsymbol{h}$ denotes the measurement model and $\boldsymbol{w}_t$ is a measurement noise vector.

Under the sequential approach, the molecular system $\boldsymbol{x}_t$ at time t is estimated from the experimental data $\boldsymbol{y}_{1:t}$ using the conditional probability distribution $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$; here, the time-series experimental data from time 1 to t are abbreviated as $\boldsymbol{y}_{1:t} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_t\}$. Using Bayes' theorem, we obtain

$$
\begin{aligned}
p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t}) &= p(\boldsymbol{x}_t|\boldsymbol{y}_t, \boldsymbol{y}_{1:t-1}) \\
&= \frac{p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{y}_{1:t-1})p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1})}{p(\boldsymbol{y}_t|\boldsymbol{y}_{1:t-1})} \\
&= \frac{p(\boldsymbol{y}_t|\boldsymbol{x}_t)p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1})}{\int p(\boldsymbol{y}_t|\boldsymbol{x}_t)p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1})d\boldsymbol{x}_t},
\end{aligned}
\tag{3}
$$

where we have used Eq. (2) to derive the third from the second line. Using this formula, the prior

distribution $p(x_t|y_{1:t-1})$ is updated to the posterior distribution $p(x_t|y_{1:t})$ by incorporating the

new experimental data $p(y_t|x_t, y_{1:t-1})$ as a likelihood function and the denominator

$p(y_t|y_{1:t-1})$ as a normalization factor. This procedure is referred to as "filtering," and $p(x_t|y_{1:t})$

is termed the filtering distribution. The prior $p(x_t|y_{1:t-1})$ in Eq. (3) can be calculated using the

time propagation of the system probability distribution from $p(x_{t-1}|y_{1:t-1})$, as

$$p(x_t|y_{1:t-1}) = \int p(x_t, x_{t-1}|y_{1:t-1})dx_{t-1}$$

$$= \int p(x_t|x_{t-1}, y_{1:t-1})p(x_{t-1}|y_{1:t-1})dx_{t-1}$$

$$= \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})dx_{t-1}, \tag{4}$$

where we have assumed x_t to have the Markovian property when deriving the third line. Eq. (4)

is referred to as "prediction," and $p(x_t|y_{1:t-1})$ is termed the predictive distribution. Thus, the

time evolution from $p(x_t|y_{1:t})$ to $p(x_{t+1}|y_{1:t+1})$ can be performed in one round of the two-step

procedure: prediction followed by filtering (Figure 1a). By repeating these rounds, we can

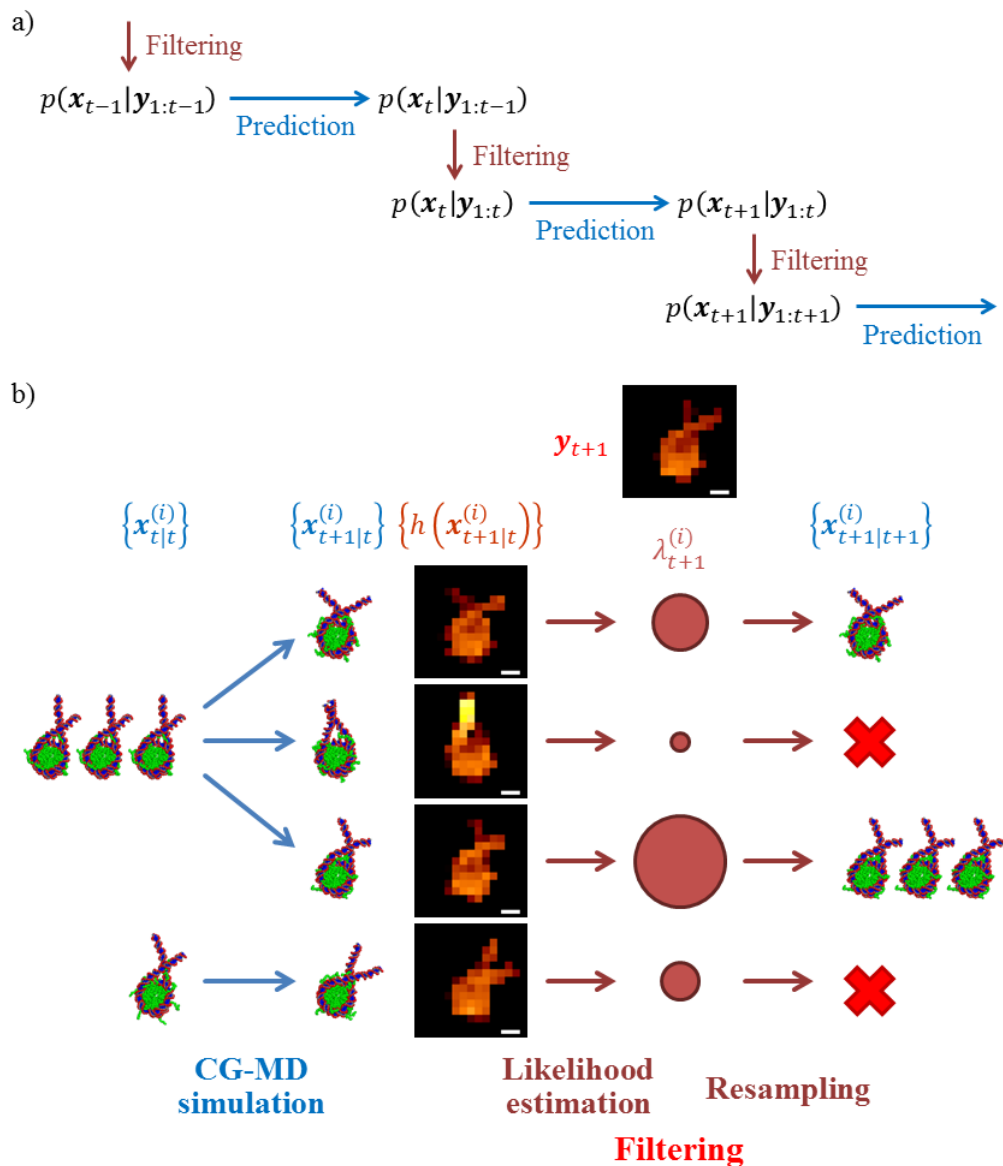sequentially estimate the time series of the distribution.

Figure 1. The particle filter method used to integrate the high-speed atomic force microscopy (HS-AFM) data with molecular simulations. a) Sequential updating via the two-step procedure, involving prediction (horizontal arrows) and filtering (vertical arrows) steps in a sequential Bayesian filtering framework. b) Procedure for one round of the particle filter simulation.

Under the particle filter method, $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$ is approximated by a set of N independent structure samples $\left\{\boldsymbol{x}_{t|t}^{(i)}\right\}$ drawn from the distribution $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$, where i is used to index the structure samples $(i = 1, \cdots, N)$. The approximation is expressed as follows:

$$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t}) \approx \frac{1}{N} \sum_{i=1}^{N} \delta\left(\boldsymbol{x}_t - \boldsymbol{x}_{t|t}^{(i)}\right). \tag{5}$$

In the field of data assimilation, structure samples are conventionally called "particles;" this is the origin of the particle filter method's name. In accordance with this convention, the term "particle" is used in this study. In the prediction step, the simulation for each of the N particles (i.e., structure samples) is performed independently for certain steps; the set of final structures obtained by the simulation are denoted as $\left\{\boldsymbol{x}_{t+1|t}^{(i)}\right\}$ and approximate the following distribution:

$$p(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t}) \approx \frac{1}{N} \sum_{i=1}^{N} \delta\left(\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t+1|t}^{(i)}\right). \tag{6}$$

The subsequent filtering step is realized by estimating the likelihood of each particle i as $\lambda_{t+1}^{(i)} \equiv p\left(\boldsymbol{y}_{t+1}\middle|\boldsymbol{x}_{t+1|t}^{(i)}\right)$ and drawing N particles from the approximate predictive distribution $\left\{\boldsymbol{x}_{t+1|t}^{(i)}\right\}$ according to the corresponding likelihood; this results in a new set of N particles $\left\{\boldsymbol{x}_{t+1|t+1}^{(i)}\right\}$, which estimate the filtering distribution $p(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t+1})$. This process is referred to as "resampling." Using Eq. (3), $p(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t+1})$ can be approximated as follows:

10

$$p(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t+1}) = \frac{p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1})p(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t})}{\int p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1})p(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t})d\boldsymbol{x}_{t+1}}$$

$$\approx \frac{p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1})\frac{1}{N}\sum_{i=1}^{N}\delta\left(\boldsymbol{x}_{t+1}-\boldsymbol{x}_{t+1|t}^{(i)}\right)}{\int p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1})\frac{1}{N}\sum_{i=1}^{N}\delta\left(\boldsymbol{x}_{t+1}-\boldsymbol{x}_{t+1|t}^{(i)}\right)d\boldsymbol{x}_{t+1}}$$

$$= \frac{1}{\sum_{i=1}^{N}p\left(\boldsymbol{y}_{t+1}\middle|\boldsymbol{x}_{t+1|t}^{(i)}\right)}\sum_{i=1}^{N}p\left(\boldsymbol{y}_{t+1}\middle|\boldsymbol{x}_{t+1|t}^{(i)}\right)\delta\left(\boldsymbol{x}_{t+1}-\boldsymbol{x}_{t+1|t}^{(i)}\right)$$

$$= \frac{1}{\sum_{i=1}^{N}\lambda_{t+1}^{(i)}}\sum_{i=1}^{N}\lambda_{t+1}^{(i)}\delta\left(\boldsymbol{x}_{t+1}-\boldsymbol{x}_{t+1|t}^{(i)}\right)$$

$$= \sum_{i=1}^{N}\beta_{t+1}^{(i)}\delta\left(\boldsymbol{x}_{t+1}-\boldsymbol{x}_{t+1|t}^{(i)}\right), \tag{7}$$

where $\beta_{t+1}^{(i)} \equiv \lambda_{t+1}^{(i)}/\sum_{i=1}^{N}\lambda_{t+1}^{(i)}$ represents the weight of the i-th particle. By approximating $\beta_{t+1}^{(i)}$

with non-negative integers $m_{t+1}^{(i)}$ that satisfy

$$\beta_{t+1}^{(i)} \approx \frac{m_{t+1}^{(i)}}{N} \qquad \left(\sum_{i=1}^{N}m_{t+1}^{(i)} = N\right), \tag{8}$$

$p(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t+1})$ can be approximated by a set of N particles, as follows:

$$p(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t+1}) \approx \frac{1}{N}\sum_{i=1}^{N}m_{t+1}^{(i)}\delta\left(\boldsymbol{x}_{t+1}-\boldsymbol{x}_{t+1|t}^{(i)}\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\delta\left(\boldsymbol{x}_{t+1}-\boldsymbol{x}_{t+1|t+1}^{(i)}\right), \tag{9}$$

where $\left\{\boldsymbol{x}_{t+1|t+1}^{(i)}\right\}$ is a set of $m_{t+1}^{(i)}$-replicated $\boldsymbol{x}_{t+1|t}^{(i)}$, such that

$$\left\{ \boldsymbol{x}_{t+1|t+1}^{(i)} \right\} \equiv \left\{ \underbrace{\boldsymbol{x}_{t+1|t}^{(i)}, \quad \cdots, \boldsymbol{x}_{t+1|t}^{(i)}}_{m_{t+1}^{(i)}} \right\}. \tag{10}$$

As a result, the estimated filtering distribution is likely to include many copies of particles with large weights, excluding particles with small weights from the estimate.

Particle Filter: Application to AFM Movie with CG-MD Simulation

Here, we apply the general theory to the AFM measurement data, which describe the time propagation of the surface shapes of biomolecular structures bound on the stage plane. The state vector $\boldsymbol{x}_t$ represents the three-dimensional Cartesian coordinates of all atoms involved in the biomolecular structure. Note: here, an "atom" refers to a "CG particle," which represents a group of atoms; because the term "particle" is used to represent a structure sample in the particle filter method, we avoid using the term "CG particle." With the CG-MD method described below, we can simulate the time propagation of $\boldsymbol{x}_t$. The measurement vector $\boldsymbol{y}_t$ represents a two-dimensional (xy-plane) AFM image, which corresponds to the surface envelope height (the z-coordinate) of the biomolecule, stacked in a one-dimensional vector. The dimensions of the vector correspond to the image pixel size N_p.

12

To simplify calculations for the HS-AFM data, we here assume a linear additive noise $\boldsymbol{w}_t$ in the measurement model, as follows:

$$\boldsymbol{y}_t = \boldsymbol{h}(\boldsymbol{x}_t) + \boldsymbol{w}_t. \tag{11}$$

We further assume that $\boldsymbol{w}_t$ is white, spatially uncorrelated Gaussian noise with a standard deviation σ. Thus, the likelihood function for filtering, $\lambda_t \equiv p(\boldsymbol{y}_t|\boldsymbol{x}_t)$, can be written as

$$\lambda_t = \prod_{j \,\in\, \text{pixels}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left((\boldsymbol{y}_t)_j - h_j(\boldsymbol{x}_t)\right)^2}{2\sigma^2} \right)$$

$$= \frac{1}{(2\pi\sigma^2)^{N_p/2}} \exp\left(-\frac{\sum_{j \,\in\, \text{pixels}} \left((\boldsymbol{y}_t)_j - h_j(\boldsymbol{x}_t)\right)^2}{2\sigma^2} \right). \tag{12}$$

In the particle filter method, the procedure for one round is applied, as shown in Figure 1b. A set of N particles $\left\{\boldsymbol{x}_{t|t}^{(i)}\right\}$ $(i = 1, \cdots, N)$ is used to approximate the distribution of molecular structures of the target. Using each molecular model as an initial structure, we perform CG-MD simulations for certain MD steps, corresponding to the time interval between two consecutive HS-AFM image acquisitions. The set of N final structures of the CG-MD simulation $\left\{\boldsymbol{x}_{t+1|t}^{(i)}\right\}$ provides an approximated predictive distribution. For the subsequent filtering step, an AFM image of each final structure $\boldsymbol{h}\left(\boldsymbol{x}_{t+1|t}^{(i)}\right)$ is generated as described below. Then, the likelihood of each particle i, expressed as

13

$$\lambda_{t+1}^{(i)} = \frac{1}{(2\pi\sigma^2)^{N_p/2}} \exp\left(-\frac{\sum_{j \in \text{pixels}}\left((\boldsymbol{y}_{t+1})_j - h_j\left(\boldsymbol{x}_{t+1|t}^{(i)}\right)\right)^2}{2\sigma^2}\right), \tag{13}$$

is estimated. Finally, we resample the particles by drawing N particles with probabilities

proportional to their likelihoods, as $\beta_{t+1}^{(i)} \equiv \lambda_{t+1}^{(i)}/\sum_{i=1}^{N}\lambda_{t+1}^{(i)}$; this results in the next filtering

distribution $\left\{\boldsymbol{x}_{t+1|t+1}^{(i)}\right\}$. By repeating the rounds, we can estimate the time series of the

approximated distribution of biomolecular structure dynamics, in a way that integrates the AFM

measurement data.

Coarse-Grained Molecular Dynamics Simulation

One prerequisite to using the particle filter method is an accurate simulation model for the system

of interest. For HS-AFM data, we require a simulation model for biomolecular structural dynamics;

in this study, we employ a CG-MD simulation. To test this method, we chose a nucleosome as the

test molecule; it is a protein-DNA complex that serves as a fundamental unit of chromatin folding

in eukaryotic cells. The nucleosome contains a histone octamer and a duplex DNA of 223 base

pairs (bps); the 601 strong positioning sequence of 145 bps is flanked by 39-bp linker DNA of the

poly CG sequence, as described in our previous work.[29] The nucleosome was placed on a plane

that represents the AFM stage (Figure S1).

14

We used the CG-MD method, which we had already extensively applied to nucleosomes.[29–31] Because it was described at length in our previous work, we only summarize it here. Each amino acid in histones is represented by a CG atom located at its Cα position, whereas each nucleotide in DNA is represented by three CG atoms, representing sugar, phosphate, and base. The histone octamer was modeled using the structure-based AICG2+ potential[32] and residue-type-specific excluded volume repulsion. The missing histone tails of histone octamer in the nucleosome were modeled using a statistical potential.[33] The DNA was modeled with the 3SPN.2C model, developed by the de Pablo Group.[34] The protein-DNA interactions include electrostatic (in the form of the Debye–Hückel potential), generic excluded volume, and sequence-nonspecific hydrogen bond interactions.[29] Charges in the protein-DNA interaction are defined as follows: -1.0e for phosphate and acidic residues (Asp, Glu) in the flexible tails and +1.0e for basic residues (Lys, Arg) in the flexible tails; Partial charges in the histone octamer core are estimated using the RESPAC method.[35] The all-atom reference structure of the system is prepared using the crystal structure of the nucleosome (PDB ID: 3LZ0) and an ideal 39-bp segment of DNA.

All CG-MD simulations were performed in CafeMol version 3.2[36], using Langevin dynamics at a temperature of 300 K with default parameters (except those of the excluded volume repulsion). Residue-type-specific radius parameters were uniformly rescaled by a factor of 1.1, to prevent the

histone tails from becoming inserted between two DNA strands.[29] A stage potential of the Lennard–

Jones type (with an energy coefficient $\epsilon = 1.0$ and length scale equal to the radius of the CG

particles) was applied to model the AFM stage and to prevent unrealistic diffusion.

We used an MD step of 0.3 in CafeMol time units, which corresponded to ~15 fs. However, it

should be noted that to speed up conformational dynamics, we utilized a small friction coefficient

in the Langevin equation (the default value in CafeMol is two orders of magnitude smaller than the

friction coefficient estimated from water viscosity). Moreover, we ignored the hydrodynamic effect

in Langevin dynamics. As a result, the apparent MD timescale mapping of ~15 fs per MD step was

inappropriate; we had previously found that a single MD step can instead be approximately mapped

to ~1 ps.[26]

The system was first equilibrated for 10^6 MD steps, to stabilize it on the AFM stage. Then, we

performed a CG-MD simulation for 10^7 MD steps with position restraints applied at the centers of

mass of the three histone multimers (two H2A/H2B dimers and one H3/H4 tetramer), to prevent

diffusion on the AFM stage.

To construct AFM images from simulated nucleosome structures, we used the collision detection

algorithm implemented in the afmize[37,38] software, for which the probe tip radius is 5 Å and the

half apex angle is 5°. We monitored structural changes of the nucleosome using the first principal

component (PC1, 45% contribution) obtained from principal component analysis (PCA), [39,40] using an in-house program. In the PCA, we used the DNA phosphate's Cartesian coordinates for the trajectory snapshots. Notably, the coordinates of histones were excluded because the histone core structure is always intact, whereas the structure of histone tails is nearly random.

Computer Experiment Protocol

We tested the described method—combining HS-AFM data with CG-MD simulations via the particle filter method—using a twin experiment.

First, we performed a CG-MD simulation for a test system (a nucleosome) at a 0.2 M monovalent ion concentration for 10^7 MD steps; this served as the ground truth. In the simulation, the nucleosome was flat on the stage; furthermore, its core was sufficiently stable and did not disassemble under the simulation conditions. On the other hand, the histone tails and linker DNAs were highly flexible and adopted a wide variety of conformations (Figure S2). The histone tails changed their conformations very rapidly; however, the linker DNA movements were so large and slow that we expected to be able to observe them using HS-AFM. We took snapshots for every 10^6 MD steps in the simulation; using these, we constructed synthetic HS-AFM "measurement" data. The resulting 11 time-point AFM images (including the initial one) constituted the reference AFM

17

movie. To test the effect of the AFM measurement's lateral resolution, we prepared two sets of reference AFM movies: one with a resolution of 2 nm $\times$ 2 nm and a pixel size of 15 $\times$ 15, and one with a resolution of 1 nm $\times$ 1 nm and a pixel size of 30 $\times$ 30.

Then, using an in-house script, we performed ten rounds of particle filter simulations with the reference AFM measurement data; we selected several different particle population sizes (8, 32, 128, 512, 2048, and 8192), ionic concentrations (0.1, 0.2, and 0.4 M), simulation timescale mappings (10^6 and 10^5 MD steps per round), and AFM image pixel resolutions (2 nm $\times$ 2 nm and 1 nm $\times$ 1 nm). The standard deviation of the measurement noise—denoted as σ—was set as 3.0 Å.

RESULTS

Particle Filter Simulation to Infer Structural Dynamics of a Biomolecule

To examine the particle filter method for HS-AFM, we performed a twin experiment. We first created an MD trajectory of nucleosomes with a 0.2 M ionic concentration, to serve as the ground truth; from this, we created a synthetic AFM movie that served as the putative measurement data. With this reference AFM movie, we performed a particle filter simulation with 512 particles, i.e., 512 structure samples drawn from a given probability distribution. The reference AFM movie

18

contained ten frames in addition to the initial one; thus, we alternatively repeated the prediction and filtering [Eqs. (6) and (7), respectively] ten times. Each prediction corresponded to 512 independent CG-MD simulations of 10^6 MD steps. Figure 2a shows the trajectories of the 512 particles along the first principal component (PC1) of the ground truth trajectory (blue). PC1 describes the large-amplitude motions of the target molecules, for which the dominant contribution is due to the linker DNA fluctuations lateral to the AFM stage. During each prediction period, the stochastic nature of the MD simulations resulted in a rapid increase in structural diversity. In the final structure ensemble of each prediction process, the linker DNAs were oriented in various directions owing to their flexibility, whereas the histone core remained intact. Notably, even though the final structure ensemble was rather broad, it was still far from the equilibrium ensemble; the initial and final structures in each round were significantly correlated (the correlation coefficient of PC1 was 0.439).

The filtering process began by calculating the likelihood of the reference AFM image for each of the structures. The resulting likelihood distribution was extremely broad owing to the conformational diversity in the linker DNAs (Figure 2b; the constant pre-factors in the log-likelihood were neglected throughout). With these 512 likelihoods expressed as relative probabilities, we drew a new set of 512 particles (i.e., we resampled). We found that the likelihood

19

of a certain particle was—in nearly all cases (eight out of ten rounds)—overwhelmingly larger than those of the other particles; as such, this single particle was resampled 512 times as the initial structures for the next round, as shown in Figure 2a. This is often termed degeneracy; we discuss the issue in more detail later.
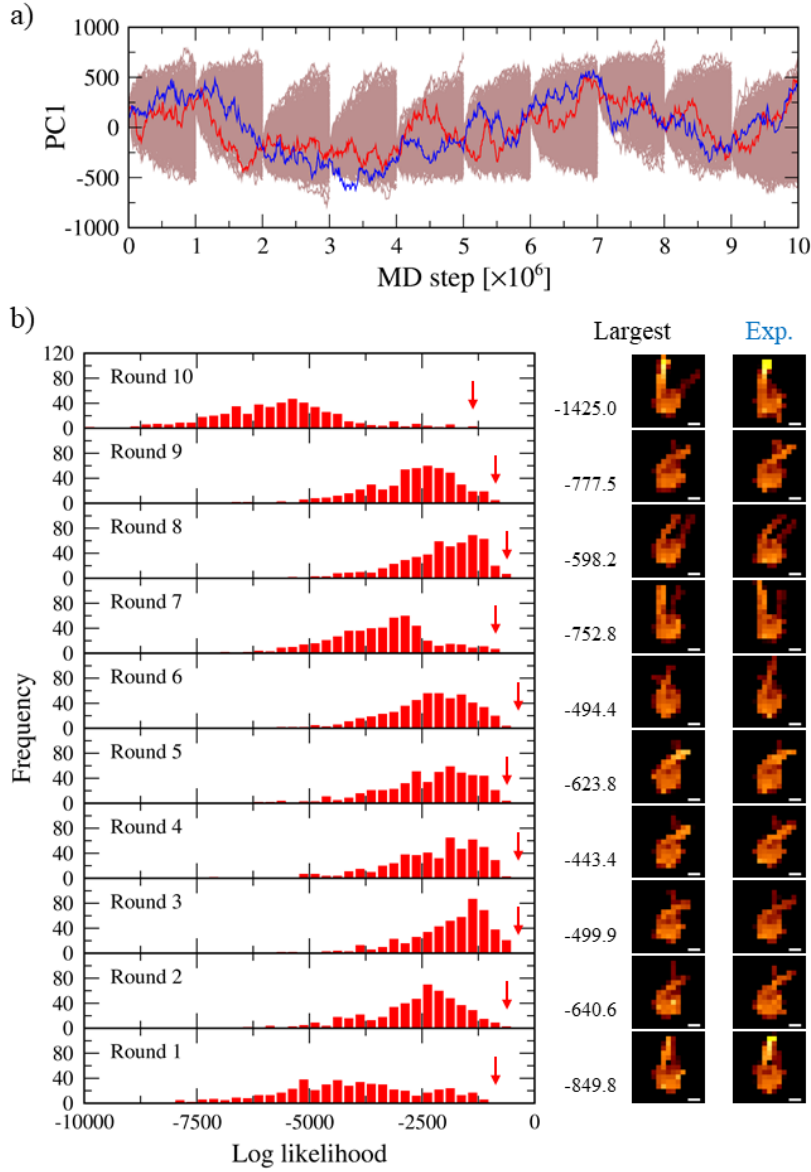
Figure 2. Results of particle filter simulations for a nucleosome with 512 particles. a) Trajectories of the first principal component (PC1) for all particles (brown), the particle that resulted in the largest likelihood (red), and the ground truth (blue). b) (left) Likelihood distribution of particles for each round; the largest log likelihood for each round is indicated by a red arrow and its value is

denoted on the right-hand side. (right) The AFM images with the largest likelihood for each round (Largest) and their corresponding synthetic experimental measurements (Exp.). Scale bar: 5 nm.

The widths of the likelihood distributions differ over the ten rounds (Figure 2b). The distributions of the first and the tenth rounds are much wider than those of the others. This difference seems to depend on the target conformations; the higher the position of the end of the linker DNA, the wider the likelihood distribution. This is most likely attributable to the fact that the linker DNA rarely moves away from the stage; thus, this conformation is difficult to sample. We confirmed the robustness of this likelihood distribution tendency by repeating the same particle filter simulations three times (Figure S3). From a diverse conformational distribution, it is possible to appropriately extract conformations that reproduce the pseudo-experimental results, by resampling of the particles based on the estimated likelihoods. In fact, we see that the computed AFM image of the particle with the largest likelihood for each round strongly resembles the corresponding pseudo-experimental one (Figure 2b, right-hand images). Thus, it was confirmed that the developed method works well, as expected. The values of largest likelihood for each round indicate a clear difference; the wider the likelihood distribution, the smaller the value of the largest likelihood. This tendency can be understood as follows: when the ground-truth molecule changes to a rare conformation, the

22

majority of the sample structures are far from the ground-truth, this results in a diverse range of likelihood values, and the largest likelihood tends to be smaller.

The total likelihood over ten rounds is simply a product of the likelihood in each round, and the most plausible movement of the target system can be estimated from the series of trajectories exhibiting the largest total likelihood (Movie S1). However, in general particle filter approaches, the entire trajectory exhibiting the largest total likelihood is not necessarily composed of the trajectories with the largest likelihood for each round; all three trials in the current work resulted in such a case. The results are reasonable, because only one or two particles with large likelihoods were resampled from the wide likelihood distributions. Generally, in the particle filter approach, it is preferable for a certain number of particles to be resampled, because this avoids particle degeneracy problems and can approximate well the conformational distribution. However, in our three trials with 512 particles, only one particle was resampled in almost all thirty rounds (except for two); this particle degeneracy might cause issues. Various methods have been proposed to solve the problem, although the simplest method is to increase the number of particles.

Dependency on Particle Numbers

The accuracy of the particle filter method depends on the number of particles; hence, we now investigate this dependency. In particular, the degeneracy problem can be resolved by increasing the particle number. We repeated the same particle filter simulations using 8, 32, 128, 2048, and 8192 particles, in addition to the 512-particle case described above.
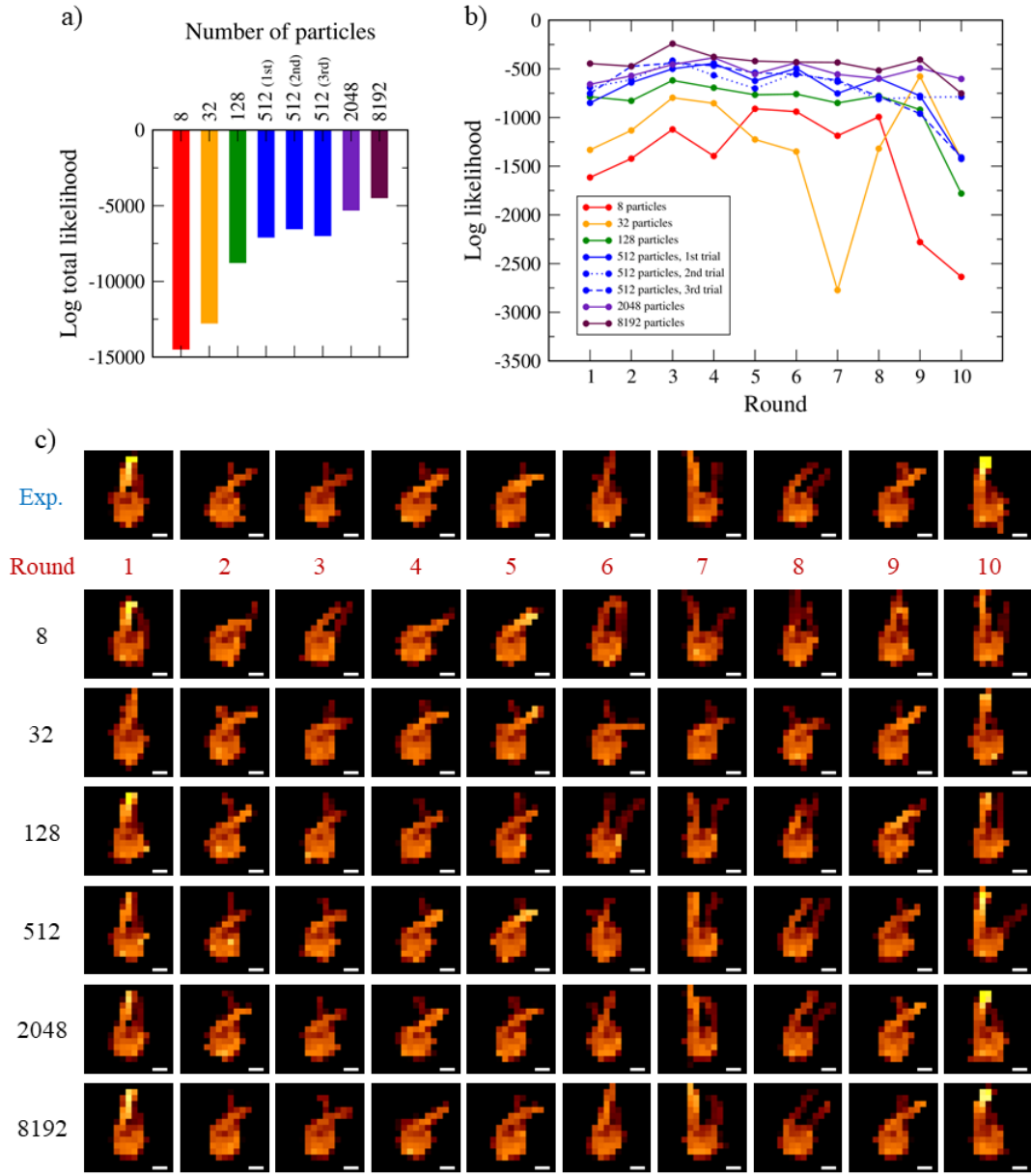
Figure 3. Results of the particle filter method with different numbers of particles: 8, 32, 128, 512, 2048, and 8192. a) The largest logarithm of the likelihood of the entire atomic force microscopy (AFM) movie. b) The corresponding likelihood for each round. c) The corresponding AFM image

for each round. For the case of 512 particles, the likelihoods obtained in the three trials are indicated

separately in a) and b); in c), the reference AFM images (Exp.) are also shown. Scale bar: 5 nm.

In Figure 3, we summarize the results of the most probable trajectory exhibiting the largest total

likelihood, obtained via particle filter simulations with different numbers of particles. Comparing

the AFM images obtained from the particle filter method (as the most probable trajectory) with

those of the reference movie (Figure 3c), we initially observe that these conformations resemble

each other in many filtering steps, regardless of the number of particles. Thus, the particle filter

method works reasonably well in all cases. However, by performing a more quantitative

comparison, we see that the maximum total likelihood has a clear dependence on the particle

number: it increases steadily as the particle number increases (Figure 3a). As expected, these results

suggest that a more accurate result can be obtained by simply increasing the number of particles.

Considering the likelihoods of each round for different numbers of particles, we identified subtler

cases (Figure 3b). In some cases, a smaller particle number led to a larger likelihood, which

opposed the overall trend. This reversal of the order was observed more frequently for smaller

particle numbers, suggesting that the conformational distribution was too wide to be properly

approximated by such small numbers of particles. The most prominent example is the comparison

26

between the 8- and 32-particle cases. Prior to the fourth round, the likelihoods for the 32-particle

case exceeded those for the 8-particle one, as expected; however, in the four subsequent rounds, a

reversal was observed. In the seventh round, the two linker DNAs in the reference AFM image did

not cross each other, which is a relatively rare conformation. For 8 particles, a parallel type of

structure was observed by chance, and it was chosen as the largest likelihood particle. However,

the structure selected in the 32-particle case exhibited a crossing of linker DNAs (Figure 3c). As

the number of particles increased, such a reversal was less likely to occur. However, reversal did

occur, albeit only rarely. Therefore, to assess the number of particles required in the particle filter

method, we needed to compare results from multiple rounds rather than those of just one.

Regarding the degeneracy issue: even for 2048 or 8192 particles, the resampling resulted in only

a single particle occupying all the samples for the next round (18 out of 20 rounds). Thus, increasing

the number of particles up to 8192 does not resolve the particle degeneracy problem.

Bayesian Inference of Physical Parameters: Ionic Concentration

One important purpose of data assimilation techniques (including the particle filter method) is to

infer the unobserved (hidden) properties/parameters. In the case of HS-AFM measurements, the

behaviors of biomolecules can be affected by environmental physical parameters that are difficult

27

to observe; these include the local ionic strength and interactions with the surface and probe tip.

Given measurement data, we can infer the physical parameters within the framework of the Bayesian approach. Here, we examine whether a physical parameter (i.e., the ionic concentration of the solution) can be inferred using the particle filter method for the HS-AFM data. To this end, we performed particle filter simulations of the nucleosome at 0.1, 0.2, and 0.4 M monovalent ion concentrations. We note that the ground-truth MD trajectory and subsequent reference AFM movie were obtained for the 0.2 M ionic concentration; furthermore, we considered the ion concentration by simply using the Debye screening length in the Debye–Hückel interaction model, although the counter-ion condensation around DNA is much more complex than a simple screening and calls for more refined treatment.[41–43] For each of the ionic concentrations, a ten-round particle filter simulation with 512 particles was repeated three times.

The MD simulation results showed that the structural dynamics of the nucleosome (particularly the linker DNAs) varied according to the ionic concentration. Compared to the case at a 0.2 M ionic concentration, DNAs in the terminal parts of the nucleosome at 0.4 M are more unwrapped, leading to larger fluctuations. On the other hand, at 0.1 M, the partial unwrapping of the DNAs is reduced, compared to the 0.2 M case.
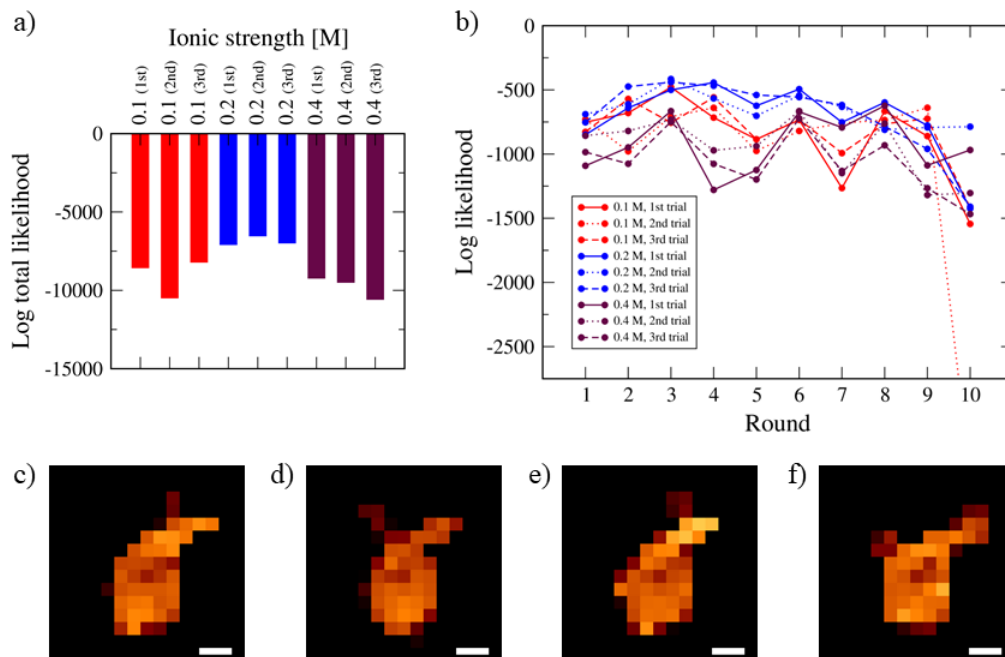
28

Figure 4. Inferences of ionic concentrations from particle filter simulations, for three trials at 0.1, 0.2, and 0.4 M ionic strengths. a) The largest logarithm of the likelihood of the entire atomic force microscopy (AFM) movie. b) The corresponding likelihood for each round. c-f) AFM images for the fifth round of c) the reference and the largest likelihood particle obtained for the first trial at d) 0.1 M, e) 0.2 M, and f) 0.4 M. Scale bar: 5 nm.

29

The largest total likelihoods obtained over ten rounds can be seen to depend on the ionic strength, as shown in Figure 4a. The largest total likelihood in the 0.2 M case was consistently the largest of the three ionic concentration cases. From these results, 0.2 M can be selected as the most probable value for the ionic strength.

Regarding the likelihood for each round: in some rounds (e.g., the fifth round), the difference in likelihood was seen to depend on the ionic concentrations (Figure 4b). All corresponding AFM images (Figure 4d-f) shared overall shapes with the reference image (Figure 4c) but showed significant differences in the details. Notably, the largest likelihood estimated for the 0.2 M ionic concentration in each round did not always exceed that at other ionic concentrations (Figure 4b). Thus, the ionic concentration must be inferred by combining the results of multiple rounds.

Bayesian Inference of Physical Parameters: Timescale Mapping

Although time is a physical parameter in which there is no apparent room for inference, we needed to infer the appropriate simulation timescale in the HS-AFM data-driven particle filter approach. First, we noted that an order-of-magnitude difference exists between the time intervals of one MD simulation step and two consecutive HS-AFM data acquisitions. Modern HS-AFM equipment can measure at a rate of 30–60 ms per frame,[44] whereas a recent all-atom MD simulation used ~1–5 fs

per MD step.[45] To speed up sampling, we employed the CG-MD simulation with a small friction coefficient and without hydrodynamic interactions. More generally, it has been argued that CG-MD simulations cannot accurately reproduce real molecular dynamics.[46–48] Thus, an apparent CG-MD time is not appropriate to use, but an effective timescale of one CG-MD step needs to be inferred posteriorly. On the other hand, AFM measurements assume interactions of modest strength between the molecule of interest and stage surface; otherwise, the molecules would move away from the surface (owing to the interactions with the probe tip) or would diffuse too rapidly on the stage to be observable. These surface interactions would slow down intrinsic molecular motions to some extent. However, in the CG-MD simulations, we did not explicitly include molecular structures of the AFM stage, which affect the motion of the specimen molecule (i.e., the nucleosome) in the lateral direction. In short, when we perform particle filter simulations for HS-AFM data, the effective simulation timescale (relative to the measurement) is not a straightforward parameter, and we may be required to infer an optimal timescale mapping between the MD step and HS-AFM measurement.

Motivated by these arguments, and using our twin-experiment framework, we examined whether a timescale mapping for pseudo-experimental data could be appropriately inferred. To this end, we performed a particle filter simulation for two different timescale mappings: 10^6 and 10^5 MD steps

31

per round (i.e., per the time interval between two consecutive HS-AFM data acquisitions). We

noted that the ground-truth MD trajectory consisted of 10^6 MD steps per round. Thus, the latter

case used an order-of-magnitude-faster timescale mapping than the ground-truth data. For each of

the timescale mappings, a ten-round particle filter simulation (with 512 particles) was repeated
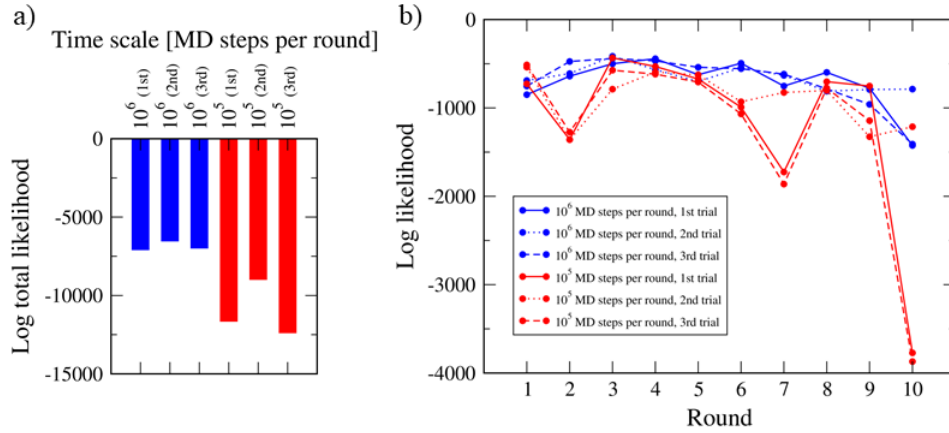
three times.

Figure 5. Timescale inferences obtained via particle filter simulations, using three trials for each

timescale mapping [i.e., 10^6 and 10^5 molecular dynamics (MD) simulation steps per one atomic

force microscopy (AFM) image]. a) The largest logarithm of the likelihood of the entire AFM

movie. b) The corresponding likelihood for each round.

32

The largest total likelihoods obtained over the ten rounds showed a clear difference depending on the timescale mappings (Figure 5a). When the timescale mapping matched (i.e., 10^6 MD steps per round), all three of the largest total likelihoods obtained by the particle filter simulation were significantly larger than those estimated for the timescale of 10^5 MD steps per round, as expected. These results suggest that using 10^6 MD steps per round as the timescale mapping is more suitable than using 10^5 MD steps per round.

On the other hand, the maximum likelihood for each round did not always distinguish the two timescale mappings; thus, the rounds could be classified into two types: "distinguishable" and "indistinguishable" (Figure 5b). The second, sixth, seventh, and tenth rounds were of type "distinguishable," in which there was a significant difference in likelihood between the two timescale mappings (except for the seventh and tenth rounds in the second trial for the faster timescale mapping). This suggests that it is difficult to obtain the structural changes observed in these rounds of pseudo-experimental measurements when using 10^5 MD steps. For example, the ground-truth trajectory exhibited marked structural DNA changes between the first and second rounds (Figure 3c, top line). Such large changes are not easily realized using 10^5 MD steps. Thus, the maximum likelihood in the second round was notably small for the case of 10^5 MD steps per

33

round. In the other rounds, the likelihood values for the two timescale mappings were similar and difficult to distinguish from each other (type "indistinguishable"). These results suggest that the target structures can be realized with a smaller number of MD steps than the actual value. Thus, the timescale mapping between the observation and simulation should not be estimated using a single round but several. To summarize, by using approximately ten rounds of measurement (though not within a single measurement), we can robustly infer the correct timescale mapping.

Effect of Pixel Resolution of AFM Images

The higher pixel resolutions of the AFM images could provide more detailed information regarding the structures of the observed biomolecules. Thus, the particle filter simulation using higher-resolution AFM images was expected to reproduce biomolecular motions more accurately. We performed two different types of particle filter simulation, using AFM images with low and high pixel resolutions (2 nm $\times$ 2 nm and 1 nm $\times$ 1 nm, respectively) to investigate the effects of differences in pixel resolution. For each resolution, a ten-round particle filter simulation with 512 particles was repeated three times.
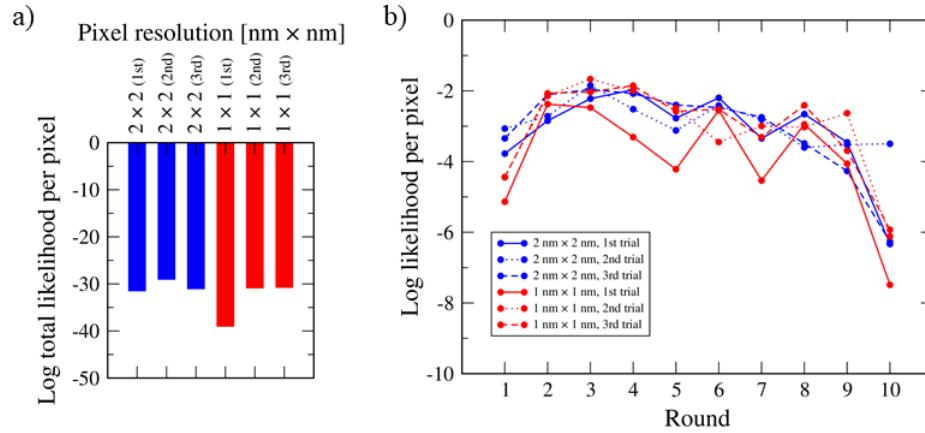
Figure 6. Dependencies of inferences on the pixel resolutions of atomic force microscopy (AFM) images, for three trials at each pixel resolution (2 nm × 2 nm and 1 nm × 1 nm). a) The largest logarithm of the likelihood of the entire AFM movie. b) The corresponding likelihood for each round.

Contrary to expectations, the largest total likelihoods "per pixel" obtained showed no significant difference between the two pixel resolutions (Figure 6a). We noted that the likelihood "per pixel" was used to compare results with different numbers of pixels, because the value of likelihood of an AFM image is defined as the product of the likelihood values for each pixel (as expressed in Eq. (12)); thus, it depends upon the number of pixels. Moreover, the likelihood per pixel for each round did not distinguish the two pixel resolutions (Figure 6b). These results suggest that using AFM images with higher pixel resolutions does not lead to the reproduction of more likely motions by

35

the current particle filter simulation. Thus, the original pixel resolution was sufficient to distinguish

biomolecular structures and resulted in the successful estimation of their plausible movements.

Because this resolution was readily attainable in the current HS-AFM experiment, we expect the

current particle filter method to be effective for actual HS-AFM data.

DISCUSSION

In the current particle filter method for HS-AFM movies, we found that during the filtering process, the likelihood of one particle overwhelmed those of all the other; thus, this particle dominated the resampling and—in nearly all cases—caused the ensemble to be occupied by copies of the single structure; this is termed degeneracy and is generally regarded as undesirable in the particle filter method. The severe degeneracy observed in the current study can be attributed to a combination of factors. First, it was partly due to the unusually many degrees of freedom in the measurement vector (i.e., the so-called "curse of dimensionality").[27] Second, it was partly due to the large disparity between the time interval of two consecutive HS-AFM image acquisitions and that of one MD step. Third, the intrinsic dynamics of biomolecules are an example of stochastic Brownian motion; that is, trajectories deviate over time. The second and third points combined cause the particles (i.e., structure samples) to deviate considerably from each other. Then, the many degrees of freedom and large deviations among particles together result in degeneracy. Although a straightforward method of avoiding degeneracy may be to increase the number of particles, we found that using large numbers of particles (as many as 8192) does not resolve the issue. In practical use, we cannot increase the number of particles beyond this.

We noted that, although undesirable, the degeneracy does not directly indicate the failure of the particle filter method. In fact, with the given number of particles, we successfully found the high-resolution spatiotemporal trajectory of the nucleosome, which was compatible with the reference HS-AFM movie to within a certain accuracy. More importantly, our examination of the parameter inference, ionic concentration, and timescale mapping suggested that we could infer the "true" physical parameter from the likelihood of the entire AFM movie. Thus, we conclude that, even with the severe degeneracy, the particle filter method for HS-AFM works well to a certain extent.

However, because of the degeneracy, we must proceed with caution when applying the particle filter method to HS-AFM. For instance, when inferring the ionic concentration, we found that within a single filtering step, the likelihood of the step derived from the "true" parameter simulation was smaller than that derived from the wrong one (see, e.g., the ninth round in Figure 4b). Thus, the parameter inference from a single-step likelihood may be erroneous. In other particle filter approaches, researchers often include particles featuring different sets of physical parameters and then infer these parameter values through repetitive filtering processes.[49] When sufficiently diverse (i.e., free from degeneracy) particles are picked up in each resampling process, this method of parameter inference works well, as demonstrated in previous studies. However, with the degeneracy observed in the current work, this protocol results in an erroneous inference. We must

38

apply multiple rounds of filtering to compare the likelihoods from different parameters. Resampling across different parameters must be performed after many rounds of filtering.

A popular alternative to the particle filter method is the ensemble Kalman filter method[50]; this retains the particle representation in the prediction step but uses Kalman gain in the filtering process. In general, the ensemble Kalman filter is a powerful method that can avoid degeneracy issues. However, in the assimilation to the HS-AFM data, we require a reverse mapping from the filtered image to the molecular structure. This in turn requires a flexible fitting method of the molecular structure to the given AFM image. Though possible[38], this flexible fitting demands a non-negligible computer time, which precludes the ensemble Kalman filter method's practical application to the current problem.

In this paper, we examined the particle filter method for a synthetic HS-AFM movie achieving promising results; thus, the next step is to apply it to real experimental HS-AFM data. Several additional factors should be considered. First, real HS-AFM data contain considerable measurement noise, the statistical nature of which is currently unclear. This noise should be spatially correlated and non-isotropic; that is, its correlation should be stronger in the x-direction (the direction of the probe scanning) than in the y-direction. This analysis is currently underway. Second, HS-AFM measurement is a complex process consisting of various unknown elements.

One such element is the unknown size and shape of the AFM probe tip. When analyzing

experimentally obtained data, we need to estimate the size and shape of the probe tip. This

estimation is possible using the particle filter simulation in the same way as the reported Bayesian

inference approach. As a preliminary twin experiment, we used the AFM images generated by

three different AFM probe tip radii (0.5, 5, and 50 Å) and performed resampling for the first round

of the first trial of the particle filter simulation (with 512 particles). In all three cases, only a single

particle was resampled; however, the three resampled particles differed from each other (Figure

S4). Among the three resampled particles, we found the largest likelihood values we obtained for

the 5 Å tip, which corresponded to the tip radius for the reference image. Thus, the radius of a

probe tip can be inferred by comparing likelihoods. Another element is the conformational

deformation of the sample molecule induced by tip-sample interaction, which results in HS-AFM

image distortion or disappearance. It is possible to perform explicit MD simulations that include

the tip-sample interaction. Third, the interactions between the biomolecules of interest and the stage

surface atoms are not homogeneous. Generally, accurate bottom-up modeling of these interactions

may be difficult owing to the lack of knowledge of the surface. We may need to infer effective

interactions via the Bayesian approach, as was performed for the ionic concentrations in this study.

Finally, related to the third point, mapping the timescales between the experiment and CG-MD

simulations is another factor to be considered (see the first paragraph in the subsection "Bayesian

Inference of Physical Parameters: Timescale Mapping"). Using the method described above, we

can apply Bayesian inference to calibrate the timescale mapping.

CONCLUSION

We developed a particle filter method to combine HS-AFM measurement data with CG-MD simulations. The particle filter method alternately propagates the structure ensemble using MD simulations over a short time period and then resamples the updated ensemble based on the likelihood of the HS-AFM image for that time. Through a twin experiment conducted upon a test molecule (i.e., a nucleosome), we confirmed that the developed method works well with ~512 particles. The particle filter simulations can infer the "true" ionic concentration and "true" timescale mapping, using the likelihood of the entire AFM movie.

Supporting Information

Figure S1. CG model of nucleosome (with linker DNAs) on the AFM stage. (PDF)

Figure S2. Conformations of nucleosomes generated via CG-MD simulations. a–j) Molecular model (upper) and corresponding synthetic AFM images (lower) for every 10^6 MD steps. Scale bar: 5 nm. (PDF)

Figure S3. Likelihood distribution of particles for each round, obtained using the particle filter simulation with 512 particles. a) First trial (same as Figure 2). b) Second trial. c) Third trial. The position of each bin containing a largest likelihood is indicated by a red arrow. (PDF)

Figure S4. AFM images for different AFM probe tip radii. a) The ground-truth AFM-like image produced with a 5 Å tip radius. b-d) AFM-like images for the largest likelihood particle obtained with an AFM probe tip radius of b) 0.5 Å (log-likelihood: -1386.40), c) 5 Å (log-likelihood: -849.8), and d) 50 Å (log-likelihood: -5504.6) for the first round of the particle filter simulation with 512 particles. Scale bar: 5 nm. (PDF)

Movie S1. A series of trajectories with the largest total likelihood obtained from the first trial of the particle filter simulation (512 particles). The reference AFM images are shown in the background. The number of MD steps (in units of 10^3) is displayed in the upper-left-hand corner. Scale bar: 5 nm. (AVI)

This information is available free of charge via the Internet at http://pubs.acs.org

REFERENCES

(1)   Ando, T.; Kodera, N.; Takai, E.; Maruyama, D.; Saito, K.; Toda, A. A High-Speed Atomic

Force Microscope for Studying Biological Macromolecules. *Proc. Natl. Acad. Sci.* **2001**,

98 (22), 12468–12472. https://doi.org/10.1073/PNAS.211400898.

(2)   Kodera, N.; Sakashita, M.; Ando, T. Dynamic Proportional-Integral-Differential Controller

for High-Speed Atomic Force Microscopy. *Rev. Sci. Instrum.* **2006**, *77* (8), 083704.

https://doi.org/10.1063/1.2336113.

(3)   Kodera, N.; Yamamoto, D.; Ishikawa, R.; Ando, T. Video Imaging of Walking Myosin v

by High-Speed Atomic Force Microscopy. *Nature* **2010**, *468* (7320), 72–76.

https://doi.org/10.1038/nature09450.

(4)   Uchihashi, T.; Iino, R.; Ando, T.; Noji, H. High-Speed Atomic Force Microscopy Reveals

Rotary Catalysis of Rotorless F 1-ATPase. *Science (80-.).* **2011**, *333* (6043), 755–758.

https://doi.org/10.1126/science.1205510.

(5)   Casuso, I.; Khao, J.; Chami, M.; Paul-Gilloteaux, P.; Husain, M.; Duneau, J. P.; Stahlberg,

H.; Sturgis, J. N.; Scheuring, S. Characterization of the Motion of Membrane Proteins

Using High-Speed Atomic Force Microscopy. *Nat. Nanotechnol.* **2012**.

https://doi.org/10.1038/nnano.2012.109.

(6)   Ando, T.; Uchihashi, T.; Kodera, N. High-Speed AFM and Applications to Biomolecular

        Systems. *Annu. Rev. Biophys.* **2013**, *42* (1), 393–414. https://doi.org/10.1146/annurev-

        biophys-083012-130324.

(7)   Ando, T.; Uchihashi, T.; Scheuring, S. Filming Biomolecular Processes by High-Speed

        Atomic Force Microscopy. *Chemical Reviews*. 2014. https://doi.org/10.1021/cr4003837.

(8)   Karplus, M.; McCammon, J. A. Molecular Dynamics Simulations of Biomolecules. *Nat.

        Struct. Biol.* **2002**, *9* (9), 646–652. https://doi.org/10.1038/nsb0902-646.

(9)   Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. Long-Timescale Molecular

        Dynamics Simulations of Protein Structure and Function. *Curr. Opin. Struct. Biol.* **2009**,

        19 (2), 120–127. https://doi.org/10.1016/j.sbi.2009.03.004.

(10)   Róycki, B.; Kim, Y. C.; Hummer, G. SAXS Ensemble Refinement of ESCRT-III CHMP3

        Conformational Transitions. *Structure* **2011**, *19* (1), 109–116.

        https://doi.org/10.1016/j.str.2010.10.006.

(11)   Pitera, J. W.; Chodera, J. D. On the Use of Experimental Observations to Bias Simulated

        Ensembles. *J. Chem. Theory Comput.* **2012**, *8* (10), 3445–3451.

        https://doi.org/10.1021/ct300112v.

(12)   Trendelkamp-Schroer, B.; Noé, F. Efficient Bayesian Estimation of Markov Model

Transition Matrices with given Stationary Distribution. *J. Chem. Phys.* **2013**, *138* (16),

164113. https://doi.org/10.1063/1.4801325.

(13)   Boomsma, W.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K. Combining Experiments and

Simulations Using the Maximum Entropy Principle. *PLoS Comput. Biol.* **2014**, *10* (2),

e1003406. https://doi.org/10.1371/journal.pcbi.1003406.

(14)   Matsunaga, Y.; Kidera, A.; Sugita, Y. Sequential Data Assimilation for Single-Molecule

FRET Photon-Counting Data. *J. Chem. Phys.* **2015**, *142* (21), 214115.

https://doi.org/10.1063/1.4921983.

(15)   Bonomi, M.; Heller, G. T.; Camilloni, C.; Vendruscolo, M. Principles of Protein Structural

Ensemble Determination. *Curr. Opin. Struct. Biol.* **2017**, *42*, 106–116.

https://doi.org/10.1016/j.sbi.2016.12.004.

(16)   Matsunaga, Y.; Sugita, Y. Linking Time-Series of Single-Molecule Experiments with

Molecular Dynamics Simulations by Machine Learning. *Elife* **2018**, *7*, e32668.

https://doi.org/10.7554/eLife.32668.

(17)   Capelli, R.; Tiana, G.; Camilloni, C. An Implementation of the Maximum-Caliber

Principle by Replica-Averaged Time-Resolved Restrained Simulations. *J. Chem. Phys.*

2018, *148* (18), 184114. https://doi.org/10.1063/1.5030339.

(18)   Bottaro, S.; Lindorff-Larsen, K. Biophysical Experiments and Biomolecular Simulations:

A Perfect Match? *Science (80-.).* **2018**, *361* (6400), 355–360.

https://doi.org/10.1126/science.aat4010.

(19)   Dixit, P. D.; Wagoner, J.; Weistuch, C.; Pressé, S.; Ghosh, K.; Dill, K. A. Perspective:

Maximum Caliber Is a General Variational Principle for Dynamical Systems. *J. Chem.

Phys.* **2018**, *148* (1), 010901. https://doi.org/10.1063/1.5012990.

(20)   Matsunaga, Y.; Sugita, Y. Use of Single-Molecule Time-Series Data for Refining

Conformational Dynamics in Molecular Simulations. *Curr. Opin. Struct. Biol.* **2020**, *61*,

153–159. https://doi.org/10.1016/j.sbi.2019.12.022.

(21)   Kitagawa, G. Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space

Models. *J. Comput. Graph. Stat.* **1996**, *5* (1), 1. https://doi.org/10.2307/1390750.

(22)   Chen, Z. H. E. Bayesian Filtering: From Kalman Filters to Particle Filters, and Beyond.

Statistics (Ber). **2003**, *182* (1), 1–69. https://doi.org/10.1.1.107.7415.

(23)   Higuchi, T.; Ueno, G.; Nakano, S.; Nakamura, K.; Yoshida, R. *Dēta Dōka Nyūmon:*

Jisedai No Shimyurēshon Gijutsu [An Introduction to Data Assimilation: The next

Generation Simulation Technology]; Asakura Publishing, 2011 (in Japanese).

(24)   Vetra-Carvalho, S.; van Leeuwen, P. J.; Nerger, L.; Barth, A.; Altaf, M. U.; Brasseur, P.;

Kirchgessner, P.; Beckers, J. M. State-of-the-Art Stochastic Data Assimilation Methods

for High-Dimensional Non-Gaussian Problems. *Tellus, Ser. A Dyn. Meteorol. Oceanogr.*

2018, *70* (1), 1445364. https://doi.org/10.1080/16000870.2018.1445364.

(25)   Takada, S. Coarse-Grained Molecular Simulations of Large Biomolecules. *Curr. Opin.*

Struct. Biol. **2012**, *22* (2), 130–137. https://doi.org/10.1016/j.sbi.2012.01.010.

(26)   Takada, S.; Kanada, R.; Tan, C.; Terakawa, T.; Li, W.; Kenzaki, H. Modeling Structural

Dynamics of Biomolecular Complexes by Coarse-Grained Molecular Simulations. *Acc.*

Chem. Res. **2015**, *48* (12), 3026–3035. https://doi.org/10.1021/acs.accounts.5b00338.

(27)   Bengtsson, T.; Bickel, P.; Li, B. Curse-of-Dimensionality Revisited: Collapse of the

Particle Filter in Very Large Scale Systems. In *Probability and Statistics: Essays in Honor*

of David A. Freedman; Institute of Mathematical Statistics, 2008; pp 316–334.

https://doi.org/10.1214/193940307000000518.

(28)   Li, T.; Sun, S.; Sattar, T. P.; Corchado, J. M. Fight Sample Degeneracy and
       Impoverishment in Particle Filters: A Review of Intelligent Approaches. *Expert Syst. Appl.*
       2014, *41* (8), 3944–3954. https://doi.org/10.1016/j.eswa.2013.12.031.

(29)   Niina, T.; Brandani, G. B.; Tan, C.; Takada, S. Sequence-Dependent Nucleosome Sliding
       in Rotation-Coupled and Uncoupled Modes Revealed by Molecular Simulations. *PLoS
       Comput. Biol.* **2017**, *13* (12), 1–22. https://doi.org/10.1371/journal.pcbi.1005880.

(30)   Brandani, G. B.; Niina, T.; Tan, C.; Takada, S. DNA Sliding in Nucleosomes via Twist
       Defect Propagation Revealed by Molecular Simulations. *Nucleic Acids Res.* **2018**, *46* (6),
       2788–2801. https://doi.org/10.1093/nar/gky158.

(31)   Brandani, G. B.; Takada, S. Chromatin Remodelers Couple Inchworm Motion with Twist-
       Defect Formation to Slide Nucleosomal DNA. *PLoS Comput. Biol.* **2018**, *14* (11), 1–25.
       https://doi.org/10.1371/journal.pcbi.1006512.

(32)   Li, W.; Wang, W.; Takada, S. Energy Landscape Views for Interplays among Folding,
       Binding, and Allostery of Calmodulin Domains. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*
       (29), 10550–10555. https://doi.org/10.1073/pnas.1402768111.

(33)   Terakawa, T.; Takada, S. Multiscale Ensemble Modeling of Intrinsically Disordered

Proteins: P53 N-Terminal Domain. *Biophys. J.* **2011**, *101* (6), 1450–1458.

https://doi.org/10.1016/j.bpj.2011.08.003.

(34)   Freeman, G. S.; Hinckley, D. M.; Lequieu, J. P.; Whitmer, J. K.; De Pablo, J. J. Coarse-

Grained Modeling of DNA Curvature. *J. Chem. Phys.* **2014**, *141* (16), 165103.

https://doi.org/10.1063/1.4897649.

(35)   Terakawa, T.; Takada, S. RESPAC: Method to Determine Partial Charges in Coarse-

Grained Protein Model and Its Application to DNA-Binding Proteins. *J. Chem. Theory

Comput.* **2014**, *10* (2), 711–721. https://doi.org/10.1021/ct4007162.

(36)   Kenzaki, H.; Koga, N.; Hori, N.; Kanada, R.; Li, W.; Okazaki, K. I.; Yao, X. Q.; Takada,

S. CafeMol: A Coarse-Grained Biomolecular Simulator for Simulating Proteins at Work.

J. Chem. Theory Comput. **2011**, *7* (6), 1979–1989. https://doi.org/10.1021/ct2001045.

(37)   Niina, T.; Kato, S. ToruNiina/Afmize: Version 1.1.0. **2019**.

https://doi.org/10.5281/ZENODO.3362044.

(38)   Niina, T.; Fuchigami, S.; Takada, S. Flexible Fitting of Biomolecular Structures to Atomic

Force Microscopy Images via Biased Molecular Simulations. *J. Chem. Theory Comput.*

2020, *16* (2). https://doi.org/10.1021/acs.jctc.9b00991.

(39)   Jolliffe, I. T. *Principal Component Analysis*, 2nd ed.; Springer series in statistics; Springer, 2002.

(40)   Fuchigami, S.; Fujisaki, H.; Matsunaga, Y.; Kidera, A. PROTEIN FUNCTIONAL MOTIONS: BASIC CONCEPTS AND COMPUTATIONAL. In *ADVANCING THEORY FOR KINETICS AND DYNAMICS OF COMPLEX, MANY-DIMENSIONALSYSTEMS: CLUSTERS AND PROTEINS: ADVANCES IN CHEMICAL PHYSICS, VOL 145*; Komatsuzaki, T., Berry, R. S., Leitner, D. M., Eds.; Advances in Chemical Physics; WILEY-BLACKWELL: COMMERCE PLACE, 350 MAIN STREET, MALDEN 02148, MA USA, 2011; Vol. 145, pp 35–82.

(41)   Oosawa, F. *Polyelectrolytes*; M. Dekker: New York, 1971.

(42)   Draper, D. E.; Grilley, D.; Soto, A. M. Ions and RNA Folding. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 221–243. https://doi.org/10.1146/annurev.biophys.34.040204.144511.

(43)   Hori, N.; Denesyuk, N. A.; Thirumalai, D. Ion Condensation onto Ribozyme Is Site Specific and Fold Dependent. *Biophys. J.* **2019**, *116* (12), 2400–2410. https://doi.org/https://doi.org/10.1016/j.bpj.2019.04.037.

(44)   Uchihashi, T.; Scheuring, S. Applications of High-Speed Atomic Force Microscopy to

Real-Time Visualization of Dynamic Biomolecular Processes. *Biochim. Biophys. Acta -*

Gen. Subj. **2018**, *1862* (2), 229–240.

https://doi.org/https://doi.org/10.1016/j.bbagen.2017.07.010.

(45)   Jung, J.; Kobayashi, C.; Sugita, Y. Optimal Temperature Evaluation in Molecular

Dynamics Simulations with a Large Time Step. *J. Chem. Theory Comput.* **2019**, *15* (1),

84–94. https://doi.org/10.1021/acs.jctc.8b00874.

(46)   Kouza, M.; Li, M. S.; O'brien, E. P. J.; Hu, C.-K.; Thirumalai, D. Effect of Finite Size on

Cooperativity and Rates of Protein Folding. *J. Phys. Chem. A* **2006**, *110* (2), 671–676.

https://doi.org/10.1021/jp053770b.

(47)   Yang, H.; Bandarkar, P.; Horne, R.; Leite, V. B. P.; Chahine, J.; Whitford, P. C. Diffusion

of TRNA inside the Ribosome Is Position-Dependent. *J. Chem. Phys.* **2019**, *151* (8),

85102. https://doi.org/10.1063/1.5113814.

(48)   Rudzinski, J. F. Recent Progress towards Chemically-Specific Coarse-Grained Simulation

Models with Consistent Dynamical Properties. *Computation* **2019**, *7* (3), 42.

https://doi.org/10.3390/computation7030042.

(49)   Nakamura, K.; Yoshida, R.; Nagasaki, M.; Miyano, S.; Higuchi, T. Parameter Estimation

of in Silico Biological Pathways with Particle Filtering towards a Petascale Computing. In

Pacific Symposium on Biocomputing 2009, PSB 2009; 2009; pp 227–238.

https://doi.org/10.1142/9789812836939_0022.

(50)   Evensen, G. The Ensemble Kalman Filter: Theoretical Formulation and Practical

Implementation. *Ocean Dyn.* **2003**, *53* (4), 343–367. https://doi.org/10.1007/s10236-003-

0036-9.

TOC graphic