PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Implementation of residue-level coarse-grained models in GENESIS for large-scale molecular dynamics simulations

Cheng Tan[1], **Jaewoon Jung**[1,2], **Chigusa Kobayashi**[1], **Diego Ugarte La Torre**[1], **Shoji Takada**[3], **Yuji Sugita**[1,2,4]*

1 Computational Biophysics Research Team, RIKEN Center for Computational Science, Kobe, Hyogo, Japan, **2** Theoretical Molecular Science Laboratory, RIKEN Cluster for Pioneering Research, Wako, Saitama, Japan, **3** Department of Biophysics, Graduate School of Science, Kyoto University, Kyoto, Japan, **4** Laboratory for Biomolecular Function Simulation, RIKEN Center for Biosystems Dynamics Research, Kobe, Hyogo, Japan

* sugita@riken.jp

OPEN ACCESS

Citation: Tan C, Jung J, Kobayashi C, Torre DUL, Takada S, Sugita Y (2022) Implementation of residue-level coarse-grained models in GENESIS for large-scale molecular dynamics simulations. PLoS Comput Biol 18(4): e1009578. https://doi.org/10.1371/journal.pcbi.1009578

Editor: Dina Schneidman-Duhovny, Hebrew University of Jerusalem, ISRAEL

Received: October 20, 2021

Accepted: March 26, 2022

Published: April 5, 2022

Data Availability Statement: The source code and user manual of GENESIS v1.7.1 is available at https://www.r-ccs.riken.jp/labs/cbrt/ website, and GENESIS-CG-tool is both included as a part of GENESIS v1.7.1 and deployed at https://github.com/noinil/genesis_cg_tool. Tutorials of performing CG simulations with GENESIS v1.7.1 are open on https://www.r-ccs.riken.jp/labs/cbrt/tutorials2019/ website. All the MD simulation files and data to produce the results are available from https://github.com/RikenSugitaLab/cg-development-GENESIS-1.7.0.

Abstract

Residue-level coarse-grained (CG) models have become one of the most popular tools in biomolecular simulations in the trade-off between modeling accuracy and computational efficiency. To investigate large-scale biological phenomena in molecular dynamics (MD) simulations with CG models, unified treatments of proteins and nucleic acids, as well as efficient parallel computations, are indispensable. In the GENESIS MD software, we implement several residue-level CG models, covering structure-based and context-based potentials for both well-folded biomolecules and intrinsically disordered regions. An amino acid residue in protein is represented as a single CG particle centered at the Cα atom position, while a nucleotide in RNA or DNA is modeled with three beads. Then, a single CG particle represents around ten heavy atoms in both proteins and nucleic acids. The input data in CG MD simulations are treated as GROMACS-style input files generated from a newly developed toolbox, GENESIS-CG-tool. To optimize the performance in CG MD simulations, we utilize multiple neighbor lists, each of which is attached to a different nonbonded interaction potential in the cell-linked list method. We found that random number generations for Gaussian distributions in the Langevin thermostat are one of the bottlenecks in CG MD simulations. Therefore, we parallelize the computations with message-passing-interface (MPI) to improve the performance on PC clusters or supercomputers. We simulate Herpes simplex virus (HSV) type 2 B-capsid and chromatin models containing more than 1,000 nucleosomes in GENESIS as examples of large-scale biomolecular simulations with residue-level CG models. This framework extends accessible spatial and temporal scales by multi-scale simulations to study biologically relevant phenomena, such as genome-scale chromatin folding or phase-separated membrane-less condensations.

Funding: The research was supported in part by MEXT as "Program for Promoting Researches on the Supercomputer Fugaku" (Biomolecular dynamics in a living cell (JPMXP1020200101) / MD-driven Precision Medicine (JPMXP1020200201)), Grant-in-aid for scientific research (Grant Numbers, 19H05645, 21H05249 (to YS), 21H05282 (to JJ and CT) and by RIKEN Pioneering research projects (Glycolipidologue Initiative, biology of intracellular environments) (to YS). The computer resources were provided by RIKEN (HOKUSAI BigWaterfall), RIKEN Center for Computational Science (Fugaku supercomputer), and HPCI (project IDs: hp200028, hp200129, hp200135, hp210172, hp210177). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Molecular dynamics (MD) simulations have been widely used to investigate biological phenomena that are difficult to study only with experiments. Since all-atom MD simulations of large biomolecular complexes are computationally expensive, coarse-grained (CG) models based on different approximations and interaction potentials have been developed so far. There are two practical issues in biological MD simulations with CG models. The first issue is the input file generations of highly heterogeneous systems. In contrast to well-established all-atom models, specific features are introduced in each CG model, making it difficult to generate input data for the systems containing different types of biomolecules. The second issue is how to improve the computational performance in CG MD simulations of heterogeneous biological systems. Here, we introduce a user-friendly toolbox to generate input files of residue-level CG models containing folded and disordered proteins, RNAs, and DNAs using a unified format and optimize the performance of CG MD simulations via efficient parallelization in GENESIS software. Our implementation will serve as a framework to develop novel CG models and investigate various biological phenomena in the cell.

This is a *PLOS Computational Biology* Software paper.

Introduction

Molecular dynamics (MD) simulation of a biomolecule in solution or membrane has been successfully applied to biophysical or biochemical problems at high spatial and temporal resolutions that are unattainable by experimental techniques. Recently, biological phenomena involving many biomolecules at larger scales have attracted much more attention than before in molecular and cellular biology, as well as medical and pharmaceutical studies. Genome-scale chromatin folding [1] and phase-separated membrane-less condensations [2] are well-known examples, which have already been targets of MD simulation studies [3–7]. To obtain biological insights on these phenomena, further methodological and computational challenges need to be met. An enormous number of atoms, for instance, more than 10–100 million atoms, are involved in such target systems, requiring us to improve the computational performance of the simulations. Alongside the size problem, there are also barriers in achieving sufficient conformational sampling in biologically meaningful time scales. Considering these issues, coarse-grained (CG) models of biomolecules are very useful in MD simulations, being a trade-off between modeling accuracy and computational efficiency. The CG models are able to decrease the number of degrees of freedom in a system and to smooth out the ruggedness of free energy landscapes, both of which efficiently accelerate MD simulations of biological systems [8].

Although there are various levels of coarse-graining [9–13], they can be categorized roughly into two classes of models in terms of their theories and parameterizations: physical-based or structure-based CG models. The MARTINI model [13] is one of the representatives in the former, while the Gō model [14–16], which describes the funnel-like energy landscape of protein folding, is probably the most well-known structure-based one. The resolution of each CG model, namely, how many heavy atoms are integrated as a single CG particle, is another crucial point in the choice of CG models. Besides, the solvent is treated

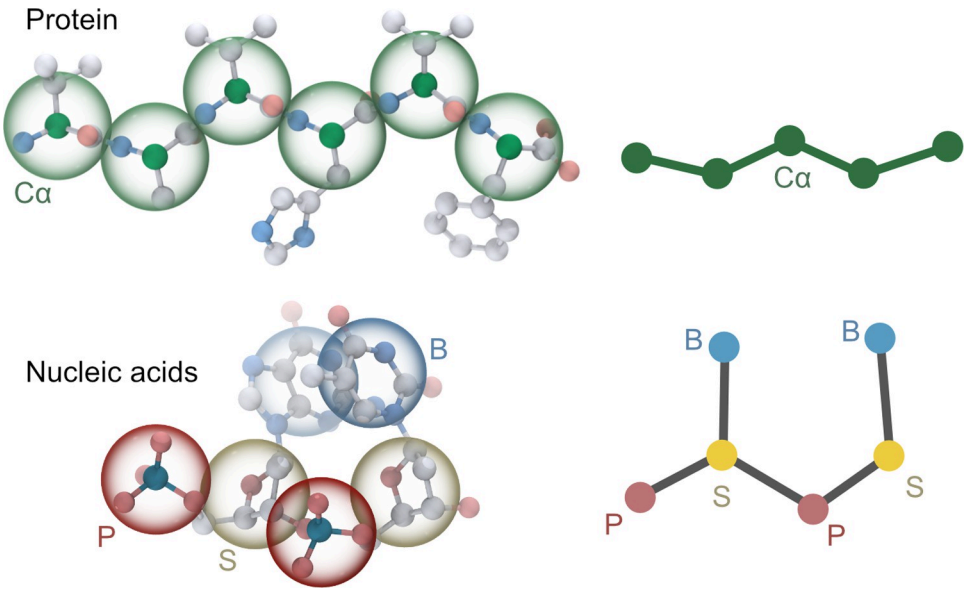

Fig 1. Mapping of a protein (top) and nucleic acids (bottom) from atomistic structures to the residue-level CG models. Each amino acid residue is represented with one CG particle (Cα), and each nucleotide is represented using three CG particles, phosphate (P), sugar (S), and base (B), respectively. Left: three-dimensional structure of protein and nucleic acids. Big and transparent spheres are the CG particles, with Cα in green, P in red, S in yellow, and B in blue. Small and solid spheres represent heavy atoms, with phosphorus in dark blue, nitrogen in pale blue, oxygen in red, Cα in green, and all the other carbon in white. Right: two-dimensional ball-and-stick cartoons of protein and nucleic acids abstracted from the left structures, using the same color scheme.

https://doi.org/10.1371/journal.pcbi.1009578.g001

differently, either represented by explicit but simplified water models [13] or modeled implicitly through effective interactions [12,17]. In the latter case, Brownian dynamics with hydrodynamic interactions [18,19] or the Lattice Boltzmann method [20,21] can be used to account for the effects of solvent-induced correlations. Here, we focus on residue-level CG models with non-hydrodynamic implicit solvation [22], whose interaction potentials are primarily described via structure-based ones. These models have been utilized in many simulation studies of folding dynamics of single proteins [14], conformational dynamics of nucleic acids [23,24], and complex molecular machinery [25]. In these models, an amino acid residue is usually simplified as one CG particle on the C_α atom, and a nucleotide is represented using three particles corresponding to the phosphate (P), sugar (S), and base (B), respectively (Fig 1). With this representation, proteins and nucleic acids are modeled with a similar resolution of around ten heavy atoms per single CG particle. Furthermore, as an extension from the original Gō model, the atomic interaction-based CG model version 2+ (AICG2+) [26] protein model introduces flexible local potentials and sequence-dependent contact parameters to gain a finer description of the free energy landscapes. The same strategy has been applied to RNA molecules [27]. On the other hand, for DNA, the 3SPN.2C model has been designed to describe the sequence-dependent geometric and mechanical properties of double-stranded DNA (dsDNA) [28].

The AICG2+ and 3SPN.2C models have been used jointly in simulations of protein-DNA complexes [29–31], although these two were developed by different research groups. In these simulations, electrostatic and excluded volume interactions between proteins or between proteins and nucleic acids are necessary for heterogeneous biological systems. In addition to

generic and physical interactions, the PWMcos method integrates high-throughput sequencing experimental results in the form of the position weight matrix (PWM) with complex structural information to describe the recognition of DNA bases by proteins [32]. Not only for the well-defined folded structures of proteins, residue-level CG models can also be used to study highly flexible and intrinsically disordered regions (IDR). In particular, two pairwise interaction models, Hydrophobicity scale (HPS) [33] and Kim-Hummer (KH) [34], have been tested and refined to reproduce the phase behaviors of protein IDRs [35,36] and RNAs [37] by Best, Mittal, and their coworkers.

The CG models mentioned above were implemented in general-purpose MD packages such as LAMMPS [38,39], GROMACS [40,41], NAMD [40,42], and OpenMM [43], or specialized CG software such as CafeMol [44]. Although these implementations are convenient tools for performing CG simulations, there are still some barriers for regular users or even developers. Each residue-based CG model is usually developed by different groups in different MD packages. If someone wants to use two of them simultaneously, it is not straightforward to combine two different CG models by treating their input or output data properly. Therefore, there is a necessity to incorporate the latest CG models into a unified working environment and provide a user-friendly tool for preparing MD input files. Furthermore, simultaneous usages of different CG models can be a powerful tool in the cutting-edge studies of protein-nucleic acid complexes in cellular environments. For example, the liquid-liquid phase-separated (LLPS) systems often involve complicated interactions among multi-domain proteins [45], IDRs [46], RNAs [47], and DNAs [48] and reach the length scales of hundreds to thousands of nanometers [49]. Such systems require both proper modeling of biomolecular interactions and efficient handling of computational resources.

In the current work, we present an implementation of several residue-level CG models of protein and nucleic acids in the GENESIS MD software [50,51]. In conjunction, we provide a collection of Julia [52] scripts, which we call GENESIS-CG-tool, to help users generate CG MD input files of complex biomolecular systems. The optimization and parallelization of CG MD simulations are the other focus in the current work. Although all-atom MD simulations in GENESIS have been highly optimized and parallelized, additional efforts are required to improve the performance of CG MD simulations because of the uniqueness of individual interaction potentials. We also noticed that some computations that take negligible fractions of time in all-atom MD simulations could be a bottleneck in CG MD simulations. After the implementation in GENESIS, memory and performance benchmark tests are carried out. CG MD simulations in GENESIS are applied to several cases of different sizes and molecular compositions, such as protein diffusion on DNA and LLPS consisting of protein IDRs and RNAs. The atomistic and CG MD simulations available in GENESIS can be a practical framework to investigate cellular-scale biological phenomena at multi-scale resolutions.

Design and implementation

Basic interaction terms

We first define basic interaction terms that are commonly used in various residue-level CG models. Each model, such as AICG2+, 3SPN.2C, HPS/KH, combines several basic interaction terms to represent intra- and inter-molecular potentials of proteins and/or nucleic acids.

Bond potential. The most typical bond potential has the following harmonic form:

$$E^{(1)}_{bond}(b_i) = k^{(1)}_{b,i}(b_i - b_{i,0})^2, \quad (1)$$

where b_i is the bond length between two neighboring CG particles, $b_{i,0}$ is the value of b_i in the reference structure, and $k_{b,i}^{(1)}$ is the force constant.

A higher-order term is also used in some CG models, such as the 3SPN.2C DNA model [39]:

$$E_{bond}^{(2)}(b_i) = k_{b,i}^{(2)}(b_i - b_{i,0})^2 + k_{b,i}^{(3)}(b_i - b_{i,0})^4, \tag{2}$$

where b_i and $b_{i,0}$ have the same meaning as in Eq (1); $k_{b,i}^{(2)}$ and $k_{b,i}^{(3)}$ are the force constants in the quadratic and quartic terms, respectively.

Angle potential. The harmonic angle potential is given by:

$$E_{angle}^{(1)}(\theta_i) = k_{a,i}(\theta_i - \theta_{i,0})^2, \tag{3}$$

where θ_i is the bond angle formed by the adjacent three CG particles, $\theta_{i,0}$ is its reference value, and $k_{a,i}$ is the force constant.

The AICG2+ protein model [26] uses a knowledge-based angle potential, which is based on the Boltzmann inversion method [53],

$$E_{angle}^{(2)}(\theta_i) = -k_B T \ln \frac{P_a(\theta_i)}{\sin\theta_i}, \tag{4}$$

where k_B is the Boltzmann constant, T is temperature in MD simulation, and $P_a(\theta_i)$ is an amino-acid type-dependent probability distribution of the angles analyzed from a set of PDB structures. $E_{angle}^{(2)}(\theta_i)$ is further refined via the iterative Boltzmann algorithm. Practically, in MD simulations, we calculate the energies and forces using spline functions fitted to values in a table containing the energy and its derivatives at several points.

Dihedral potential. The periodic proper dihedral potential is defined as follows:

$$E_{dihedral}^{(1)}(\phi_i) = \sum\nolimits_n k_{\phi,i,n}[1 + \cos(n(\phi_i - \phi_{i,0}))], \tag{5}$$

where ϕ_i and $\phi_{i,0}$ are the dihedral angle and its reference value, respectively, n is an integer number that controls the periodicity of the function, and $k_{\phi,i,n}$ is the force constant.

We also implement a Gaussian type dihedral potential, which is used in the AICG2+ [26] and the 3SPN.2C models [39]:

$$E_{dihedral}^{(2)}(\phi_i) = -\epsilon_{\phi,i} \exp\left(\frac{-(\phi_i - \phi_{i,0})^2}{2\sigma_{\phi,i}^2}\right), \tag{6}$$

where ϕ_i and $\phi_{i,0}$ have the same meaning as in Eq (5), $\sigma_{\phi,i}$ controls the Gaussian width and $\epsilon_{\phi,i}$ is the force constant.

Similar to Eq (4), AICG2+ [26] also uses a knowledge-based dihedral term, which is based on the Boltzmann inversion potential [53]:

$$E_{dihedral}^{(3)}(\phi_i) = -k_B T \ln P_d(\phi_i), \tag{7}$$

where $P_d(\phi_i)$ is the probability distribution of the dihedral angle. The $E_{dihedral}^{(3)}(\phi_i)$ is further refined via the iterative Boltzmann algorithm. This is also calculated as a tabulated function for better performance. Notably, in each of the dihedral potentials described above, we introduce a recently proposed algorithm [54] to improve the robustness and numerical stability in the force evaluations.

Nonbonded interactions. The structure-based Gō-like models for proteins usually use a 12–10 potential to model a native contact, defined as two residues containing heavy atoms at a

distance less than a specific cutoff in the native structure [14,26]. The potential [14] is defined as:

$$E_{G\tilde{o}}(r_i) = \epsilon_{G\tilde{o},i}\left[5\left(\frac{\sigma_i}{r_i}\right)^{12} - 6\left(\frac{\sigma_i}{r_i}\right)^{10}\right], \tag{8}$$

where r_i is the distance between two CG particles in a native-contact pair, σ_i is the native value of r_i, and $\epsilon_{G\tilde{o},i}$ is the force constant. Note that it is applicable not only to proteins but also to nucleic acids if their experimental structures are known.

The Lennard-Jones (LJ) potential (or 12–6 potential) is also widely used to describe pairwise nonlocal interactions and is commonly defined as:

$$E_{LJ}(r_i) = 4\epsilon_{LJ,i}\left[\left(\frac{\sigma_i}{r_i}\right)^{12} - \left(\frac{\sigma_i}{r_i}\right)^{6}\right], \tag{9}$$

where $\epsilon_{LJ,i}$ represents the minimum energy.

The LJ potential truncated at its energy minimum is a purely repulsive potential and can be used as the excluded volume interaction:

$$E^{(1)}_{exv}(r_i) = \begin{cases} \epsilon_{exv,i}\left[\left(\frac{\sigma_i}{r_i}\right)^{12} - 2\left(\frac{\sigma_i}{r_i}\right)^{6}\right] + \epsilon_{exv,i}, r < r_{C,exv} \\ 0, r \geq r_{C,exv} \end{cases} \tag{10}$$

where $r_{C,exv}$ is the cutoff distance and has the same value as σ_i. Note that the LJ forms in Eq (9) and Eq (10) are different in a coefficient of 2 for the 6th-power term. However, these two forms are mathematically equivalent when the parameters (σ and ϵ) are appropriately converted.

A simpler form of the excluded volume interaction is given by:

$$E^{(2)}_{exv}(r_i) = \begin{cases} \epsilon_{exv,i}\left(\frac{\sigma_i}{r_i}\right)^{12} + \epsilon'_{exv,i}, r < r_{C,exv} \\ 0, r \geq r_{C,exv} \end{cases} \tag{11}$$

where $r_{C,exv} = 2\sigma_i$ and $\epsilon'_{exv,i} = -2^{-12}\epsilon_{exv,i}$.

Other pairwise nonbonded interactions include the Gaussian potential:

$$E_{Gaussian}(r_i) = -\epsilon_{G,i}\exp\left(\frac{-(r_i - r_{i,0})^2}{2w_i^2}\right), \tag{12}$$

and the Morse potential:

$$E_{Morse}(r_i) = \epsilon_{M,i}(1 - e^{-\alpha_i(r_i - r_{i,0})})^2 - \epsilon_{M,i}, \tag{13}$$

where $\epsilon_{G,i}$ and $\epsilon_{M,i}$ are the "depth", and w_i and α_i are the "width" of the Gaussian and Morse potentials, respectively. In the 3SPN.2C model [39], the Morse potential is divided into two parts, the repulsive ($E^{(rep)}_{Morse}$) and the attractive ($E^{(attr)}_{Morse}$), as defined in the following:

$$E^{(rep)}_{Morse}(r_i) = \begin{cases} \epsilon_{M,i}(1 - e^{-\alpha_i(r_i - r_{i,0})})^2, r_i < r_{i,0} \\ 0, r_i \geq r_{i,0} \end{cases} \tag{13A}$$

and

$$E_{Morse}^{(attr)}(r_i) = \begin{cases} -\epsilon_{M,i}, r_i < r_{i,0} \\ \epsilon_{M,i}(1 - e^{-\alpha_i(r_i - r_{i,0})})^2 - \epsilon_{M,i}, r_i \geq r_{i,0} \end{cases} \quad (13B)$$

The electrostatic interactions are modeled using the Debye-Hückel theory:

$$E_{ele}(r_i) = \frac{q_{i_1} q_{i_2} e^{-r_i/\lambda_D}}{4\pi\varepsilon_0 \varepsilon_r r_i}, \quad (14)$$

where r_i is the distance between the charged particles i_1 and i_2, whose charges are q_{i_1} and q_{i_2}, respectively, ε_0 is the dielectric permittivity of vacuum, λ_D is the Debye screening length, ε_r is the relative permittivity of the solution, which is a function of the solution temperature T and salt molarity C: $\varepsilon_r = e(T)a(C)$, where $e(T) = 249.4 - 0.788T + 7.20 \times 10^{-4}T^2$ [55], and $a(C) = 1 - 0.2551C + 5.151 \times 10^{-2}C^2 - 6.889 \times 10^{-3}C^3$ [56]. The Debye length is defined as: $\lambda_D = \sqrt{\frac{k_B T \varepsilon_0 \varepsilon_r}{2N_A e_c^2 I}}$, where N_A is the Avogadro's number, e_c is the elementary charge, and I is the ionic strength of the solution.

Modulating functions. When residue-level CG models are used to describe specific non-bonded interactions, such as the hydrogen bonds (HBs) or the π-π stacking, it is necessary to introduce multi-body potentials. For instance, in the 3SPN.2C model, the base-stacking, base-pairing, and cross-base-stacking interactions are described with the angle and torsion-angle dependent multi-body potential functions [28,39]. Specifically, the angle-dependent modulating function in the 3SPN.2C model is defined as [39]:

$$f(\Delta\theta) = \begin{cases} 1, |\Delta\theta| < \gamma \\ 1 - \cos^2\left(\frac{\pi}{2\gamma}\Delta\theta\right), \gamma \leq |\Delta\theta| \leq 2\gamma \\ 0, |\Delta\theta| \geq 2\gamma \end{cases} \quad (15)$$

where $\Delta\theta$ is the difference between an angle (θ) and its reference value, and γ controls the tuning range.

Residue-level CG models

We next describe the CG models as the combinations of the basic interaction terms defined in the last section. We mainly focus on the energy function forms rather than the detailed parameters, which users can easily change in their MD simulations.

The AICG2+ model for folded proteins. The energy function of the AICG2+ model [26] is defined as:

$$V_{AICG2+}(\Gamma|\Gamma_0) = \sum_{b_i \in bonds} E_{bond}^{(1)}(b_i) + \sum_{\theta_i \in angles} E_{angle}^{(2)}(\theta_i) + \sum_{r_i \in 1-3pairs} E_{Gaussian}(r_i) + \sum_{\phi_i \in dihedrals} E_{dihedral}^{(2)}(\phi_i) + \sum_{\phi_i \in dihedrals} E_{dihedral}^{(3)}(\phi_i) + \sum_{r_i \in native\ contacts} E_{G\bar{o}}(r_i) + \sum_{r_i \in non-native\ contacts} E_{exv}^{(2)}(r_i), \quad (16)$$

where Γ and Γ_0 are the simulated and native structures, respectively. The reference values in each term are determined from the native values in Γ_0, except for the non-native contact term. In the non-native contact interactions, σ_i is dependent on the excluded volume radii of the particles in contact, using the combination rule of $\sigma_i = \frac{\varsigma_{i_1} + \varsigma_{i_2}}{2}$, where i_1 and i_2 are the indices of the particles in the i-th non-native contact, with ς_{i_1} and ς_{i_2} as their residue type-dependent radii.

The HPS and KH models for IDR. The two IDR models, HPS [33] and KH [34], share the same energy function [35]:

$$V_{HPS,KH}(\Gamma) = \sum_{b_i \in bonds} E^{(1)}_{bond}(b_i) + \sum_{r_i \in charged\ pairs} E_{ele}(r_i) + \sum_{r_i \in nonbonded\ pairs} E_{HPS,KH}(r_i), \tag{17}$$

where Γ is the conformation of an IDP. $E_{HPS,KH}(r_i)$ has the Ashbaugh-Hatch form [57]:

$$E_{HPS,KH}(r_i) = \begin{cases} E_{LJ}(r_i) + (1-\lambda_i)\epsilon_{LJ,i}, r_i \le 2^{1/6}\sigma_i \\ \lambda_i E_{LJ}(r_i), r_i > 2^{1/6}\sigma_i \end{cases} \tag{18}$$

where λ_i in the HPS model is the hydrophobicity scale, and in the KH model is a sign function to regulate the attractiveness or repulsiveness. All the other quantities, r_i, σ_i, and $\epsilon_{LJ,i}$, have the same meaning as defined in Eq (9). Specifically, $\epsilon_{LJ,i}$ and λ_i are based on the amino acid hydrophobicity in HPS [33] or the Miyazawa-Jerningan potential in KH [34,58], respectively.

The same potential functions are also used in the HPS RNA model [37], which has been developed to study the co-condensation of RNA and protein IDRs. In the HPS RNA model [37], the resolution is only one bead for one nucleotide, which is lower than the other nucleic acid CG models in GENESIS. Accordingly, the bond lengths and particle radii are also larger than the other models. Note that GENESIS also provides a structure-based three-bead-per-nucleotide RNA model, as described in a later section.

The 3SPN.2C model for DNA. The potential energy function of the 3SPN.2C model [28,39] is defined as follows:

$$\begin{aligned} V_{3SPN.2C}(\Gamma) = & \sum_{b_i \in bonds} E^{(2)}_{bond}(b_i) + \sum_{\theta_i \in angles} E^{(1)}_{angle}(\theta_i) + \sum_{\phi_i \in all\ dihedrals} E^{(1)}_{dihedral}(\phi_i) \\ & + \sum_{\phi_i \in backbone\ dihedrals} E^{(2)}_{dihedral}(\phi_i) + \sum_{r_i \in exv\ pairs} E^{(1)}_{exv}(r_i) + \sum_{r_i \in phosphate\ pairs} E_{ele}(r_i) \\ & + \sum_{base\ steps} E_{bstk} + \sum_{base\ pairs} E_{bp} + \sum_{cross-stacking\ pairs} E_{cstk}, \end{aligned} \tag{19}$$

where Γ is the conformation of the DNA molecule, and "exv pairs" are the nonbonded particle pairs that are not involved in either base-pairing or stacking interactions. The three terms for base-base interactions, E_{bstk}, E_{bp}, and E_{cstk}, are defined as multi-body energy functions, using the basic equations described in the previous section (Eq (13A), (13B) and (15)):

$$E_{bstk} = E^{(rep)}_{Morse}(r_i) + f(\Delta\theta_{BS,i}) E^{(attr)}_{Morse}(r_i), \tag{20}$$

$$E_{bp} = E^{(rep)}_{Morse}(r_i) + \frac{1}{2}\left(1 + \cos\Delta\phi_{BP,i}\right) f\left(\Delta\theta_{1,i}\right) f\left(\Delta\theta_{2,i}\right) E^{(attr)}_{Morse}(r_i), \tag{21}$$

$$E_{cstk} = f(\Delta\theta_{3,i}) f(\Delta\theta_{CS,i}) E^{(attr)}_{Morse}(r_i), \tag{22}$$

where r_i represents the distance between the two interacting bases, and the angles ($\theta_{BS,i}$, $\theta_{1,i}$, $\theta_{2,i}$, $\theta_{3,i}$, and $\theta_{CS,i}$) and dihedral angles ($\phi_{BP,i}$) are formed by the surrounding sugar and phosphate sites (for details of the definition, please refer to [39]).

The structure-based model for RNA. In addition to the one-bead (per-nucleotide) HPS model, we also implement a structure-based three-bead model for RNA [27], whose parameters are determined with the fluctuation matching method [59]. The energy function of this

model [27] is defined as:

$$V_{RNA}(\Gamma) = \sum_{b_i \in bonds} E^{(1)}_{bond}(b_i) + \sum_{\theta_i \in angles} E^{(1)}_{angle}(\theta_i) + \sum_{\phi_i \in dihedrals} E^{(1)}_{dihedral}(\phi_i)$$
$$+ \sum_{r_i \in native\ contacts} E_{G\tilde{o}}(r_i) + \sum_{r_i \in non-native\ contacts} E^{(2)}_{exv}(r_i)$$
$$+ \sum_{r_i \in phosphate\ pairs} E_{ele}(r_i), \quad (23)$$

where Γ is the conformation of the RNA molecule. The electrostatic term, E_{ele}, is optional in the intra-RNA interactions but required for protein-RNA interactions [27]. In the later case, r_i in $E_{ele}(r_i)$ represents the distance between a charged amino-acid residue and an RNA phosphate.

The PWMcos model for protein-DNA interaction. The protein-DNA binding is generally considered in two parts: the sequence-nonspecific interactions between amino acids with the DNA backbone groups (mainly the electrostatic interactions) and the sequence-specific interactions between amino acids and DNA bases. The PWMcos model can describe the latter, incorporating the PWM information into the structure-based interactions [32]. The model first defines a list of DNA-binding protein residues (DB-C_αs) forming contact with DNA in the native structure, and then the potential energy is then given by [32]:

$$\sum_{i \in bases} \sum_{j \in DB-C_\alpha} \sum_{m \in PWM\ columns} E_{PWMcos}(i, j, m, \vec{x})$$
$$= \sum_{i,j,m} (U_{m,j}(b_i, \vec{x}) + U_{m',j}(b_i, \vec{x})), \quad (24)$$

where m' is the base in the complementary strand that forms a base-pair with base m, b_i is the base type of any base i ($b_i \in [A, C, G, T]$), and $\vec{x}$ is the coordinates of particles in each conformation. The $U_{m,j}(b_i, \vec{x})$ function is defined as:

$$U_{m,j}(b_i, \vec{x}) = E_{Gaussian}(r_{ij}) f(\Delta\theta_1) f(\Delta\theta_2) f(\Delta\theta_3),$$

where r_{ij} is the distance between the i-th and the j-th C_α, and θ_1, θ_2, and θ_3 are three angles defined by the surrounding particles (for details, please refer to [32]). $E_{Gaussian}(r)$ and $f(\Delta\theta)$ functions are defined in Eq (12) and (15), respectively. The energy coefficient of $E_{Gaussian}(r)$ ($\epsilon_{G,i}$ in Eq (12)) is defined as a function of base type b_i based on the PWM:

$$\epsilon_{G,i} = \epsilon_{PWM,m}(b_i) = \gamma \left(\frac{e_{PWM,m}(b_i)}{N_m} + \epsilon' \right),$$

where $e_{PWM,m}(b)$ is the element in the mth column of the PWM and in the row corresponding to base type b, N_m is the total number of protein contacts formed with the mth base pair, ϵ' is a hyperparameter that shifts the absolute value of the function, and γ is a scaling factor to change the strength of the interaction.

Notably, a variation of the PWMcos has been used to model sequence-non-specific hydrogen bonds between protein and DNA backbone [30]:

$$E_{HB}(i, j, \vec{x}) = E_{Gaussian}(r_{ij}) f(\Delta\theta_1) f(\Delta\theta_2), \quad (25)$$

where r_{ij} represents the distance from the i-th phosphate (instead of base) to the j-th C_α, and θ_1 and θ_2 are the angles defined by the local particles around the target phosphate and C_α atom in the configuration with coordinates $\vec{x}$.

Table 1. Available CG models in heterogeneous biomolecular systems.

	Protein	DNA	RNA
Protein	V_{AICG2+}, $[V_{HPS,KH}]$		
DNA	V_{AICG2+}, $V_{3SPN.2C}$, $E^{(2)}_{exv}$, $[E_{ele}, E_{G\bar{o}}, E_{PWMcos}, E_{HB}]$	$V_{3SPN.2C}$	
RNA	V_{AICG2+}, V_{RNA}, $[V_{HPS,KH}]$, $E^{(2)}_{exv}$, $[E_{ele}, E_{G\bar{o}}, E_{HPS}]$	$V_{3SPN.2C}$, V_{RNA}, $E^{(2)}_{exv}$, $[E_{ele}, E_{G\bar{o}}]$	V_{RNA}, $[V_{HPS}]$

The interactions written in brackets are optional.

https://doi.org/10.1371/journal.pcbi.1009578.t001

Summary of the potential energy functions available in GENESIS. Although the models mentioned above are developed independently, some have been proved to work well with each other. For instance, the association of the AICG2+ and the 3SPN.2C models has shown the capability to study the protein-DNA binding systems [29,31]. Between proteins and DNA, electrostatic interactions contribute most to the sequence-nonspecific affinity. For the accuracy of electrostatics modeling, the RESPAC method was developed to calculate the surface charge distribution of proteins [60]. Furthermore, the angle-dependent HB potential enables modeling the rotation-coupled and uncoupled sliding of nucleosomal DNA [30,61]. The PWMcos method also facilitates the target search of transcription factors on both linear DNA and nucleosomes [32,62].

In addition to these well-developed methods, GENESIS also allows the combinations of different models. In Table 1, we list all the available CG models in GENESIS 1.7.0 in mixtures of proteins, DNAs, and RNAs.

Information flow

Fig 2 shows a flowchart for running CG simulations in GENESIS. We roughly divide the whole procedure into three steps: (i) preparation, (ii) simulation, and (iii) analysis. GENESIS provides tools to achieve the task in each step. In the preparation stage, we use a collection of scripts, "GENESIS-CG-tool", to translate experimental results into CG topology and coordinate files. Together with the MD control file, these files are then parsed by one of the GENESIS MD engines, atdyn, to perform energy/force calculations and time integrations. As the output data from MD simulations, GENESIS produces a log file of system properties including temperature, system size, and potential energy. The output coordinates from MD simulation are written in a trajectory file with the DCD format, which are processed by the GENESIS analysis tools and can be visualized by VMD [63], PyMOL [64], or other molecular graphics software. Here, we mainly focus on our developments in the first two stages, since the analysis using the DCD format files and analysis tools is common to those of all-atom MD simulations.

Preparation. We next explain the usage of the GENESIS-CG-tool, which can read structural information and generate force field parameters and coordinates for MD simulations. As input, the structural information is in the PDB format, which includes atom coordinates determined by experiments such as X-ray crystallography, NMR, or Cryo-EM. As output, we decided to use a unified GROMACS-like file format for all the CG models. Particularly, the output of GENESIS-CG-tool includes a major topology file (.top) and a bundle of molecule-specific files (.itp), as well as a coordinate file for all the particles in the system (.gro) (Fig 2). The molecule-specific topology (.itp) files contain the information of intra-molecular interactions, such as bond, angle, dihedral, and Gō-type native-contact interactions. Each term includes the indices of the involved particles, reference values of the variables, and the force parameters. Additionally, an integer variable (called the "function type") is used to distinguish the various potential functions of the same interaction type. For instance, we set the function types of 1 and 21 to $E^{(1)}_{bond}$ (Eq (1)) and $E^{(2)}_{bond}$ (Eq (2)), respectively.

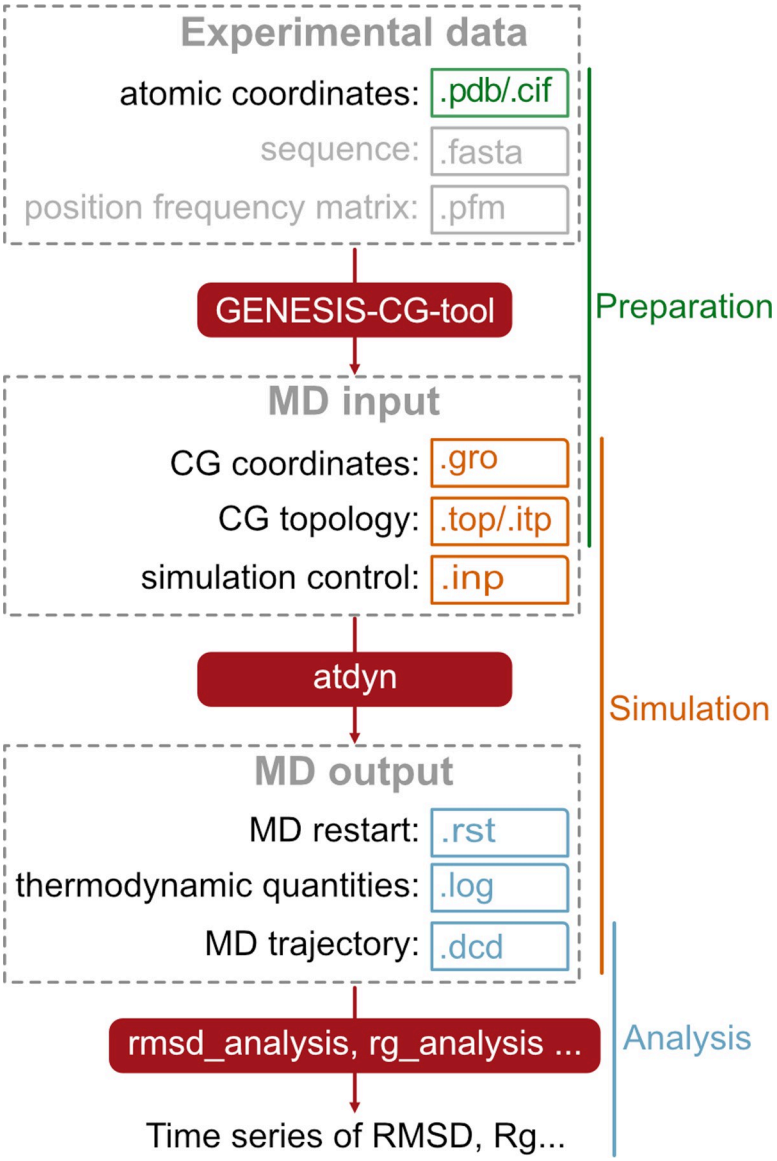

Fig 2. Information flow in GENESIS CG MD simulations. Dashed boxes represent collections of files used together as input or output at a particular step. Solid red boxes are used for the executable tools provided by GENESIS. Small boxes include typical filename extensions.

https://doi.org/10.1371/journal.pcbi.1009578.g002

GENESIS-CG-tool was developed with Julia [52], a fast, open-sourced, and reproducible programming language. The project's main entrance is a script called "aa_2_cg.jl", which works in the command-line environment of UNIX-like shells. The script takes a PDB file name as the positional argument and several model-related optional arguments. During the processing, the atomistic coordinates are read from PDB files. Then, local and nonlocal interactions, including their native values, are detected from the PDB structure. Other parameters, such as force constants, are determined by the models specified for each molecule.

As a simple example, assuming that one wants to prepare CG MD input files for a protein using a PDB file, "PRO1.pdb", the command "aa_2_cg.jl PRO1.pdb" will generate "PRO1_cg.top" and "PRO1_cg.itp" topology files and "PRO1_cg.gro" coordinate file (Figs 2 and S1). The

AICG2+ model is used for proteins by default, which can be changed with the option "--force-field-protein".

The 3SPN.2C model of DNA uses reference structures that include sequence-dependent geometric features. These structures can be generated by software such as 3DNA [65]. In GENESIS-CG-tool, with no requirement of external packages, we provide a simple "sequence-to-structure" script, which reads in DNA sequences and directly outputs both atomistic PDB and CG topology/coordinate files for dsDNA, including the necessary sequence-dependent information (see S1 Fig).

For heterogeneous systems that contain more than one type of biomolecules, the molecule-specific ".itp" files can be linked into the ".top" file, using the "#include" keyword.

Notably, recent improvements in experimental techniques such as cryo-EM have fostered more high-resolution structures of the huge biomolecule complexes, which are valuable resources in MD simulations. The spatial scales of these structures, however, outstrip the capacity of the traditional PDB file format. Another format, PDBx/mmCIF [66], has been used to store coordinate information of large structures. GENESIS-CG-tool also accepts files of the PDBx/mmCIF format as input. Besides, we also noticed that detecting the Gō-like native contacts is rather time-consuming for large protein complexes. We utilize the Julia thread parallelization to accelerate this procedure in the GENESIS-CG-tool.

We have packaged GENESIS-CG-tool as a part of GENESIS v1.7.0. We have also deployed the GENESIS-CG-tool in an individual Github repository: https://github.com/noinil/genesis_cg_tools. Users can refer to the online wiki page of this project for more details on the usage and available file formats.

MD algorithms

In GENESIS v1.7.0, there are two MD programs: spdyn and atdyn. Spdyn is parallelized using the mid-point cell method [67], one of the domain-decomposition schemes; while atdyn is parallelized with the atomic decomposition scheme [50,51]. We implement the residue-level CG models only in atdyn because of the feasibility of implementations and the number of CG particles required in most CG MD simulations. In atdyn, the Clementi Gō [14], Karanicolas-Brooks Gō [68], Domain-enhanced Model (DoME) [69], and dual/multi-basin Gō models [70] are also available and use the same topology and coordinate file formats as described in the previous section. Since atdyn is already a well-developed framework, the original source code is reused as much as possible to implement the residue-level CG models. In the following, we describe the newly introduced functions and optimization schemes only.

Cutoff scheme of the nonbonded interactions. The nonbonded interaction terms are the most time-consuming computations both in all-atom and CG MD simulations. Cutoff schemes are used for the nonbonded terms in CG MD simulations to reduce the computation time. For relatively short-range nonbonded interactions, such as $E_{exv}^{(1)}$ and $E_{exv}^{(2)}$, cutoff values are already included in the definition (see Eq (10) and (11)). In contrast, the other interactions do not have generally established cutoff values. We provide three parameters to control the cutoff lengths of the E_{bp}, E_{LJ}, and E_{ele} terms in our implementations. We set the default cutoff value of E_{bp} to 18Å, following the original 3SPN.2C model [28,39]. We use a value of 52Å for E_{ele}, which assures that the absolute value of the energy at the cutoff distance to be smaller than 10^{-4} kcal/mol (calculated at the temperature of 300K and with an ionic concentration of 150mM, for two charges of $\pm 1\ e^{-}$). This cutoff value of E_{ele} is slightly larger than the one used in the original 3SPN.2C model and guarantees accuracy [39]. As for the LJ potential used in the HPS/KH IDR models, the parameters $\epsilon_{LJ,i}$ and σ_i are dependent on residue-types [35]. Therefore, we determine the LJ cutoff using a strategy that, for the i-th pair, the absolute value

Table 2. Default values for the nonbonded cutoff and pair-list distances.

Nonbonded term	Cutoff (r_C, Å)	Neighbor list distance (r_P, Å)
E_{ele}	52	57
$E_{HPS,KH}$	39	44
E_{bp}	18	23
E_{PWMcos} and E_{HB}	–	23
$E_{exv}^{(1)}$ and $E_{exv}^{(2)}$	–	15

The cutoff of E_{PWMcos} and E_{HB} are given by $r_{i,0}$+5Å ($r_{i,0}$ is defined in Eq (12)). The cutoff of $E_{exv}^{(1)}$ and $E_{exv}^{(2)}$ are defined in Eq (10) and (11), respectively.

https://doi.org/10.1371/journal.pcbi.1009578.t002

of the energy at the cutoff distance is as small as 10^{-4} times of the energy minimum: $E_{LJ}(r_{C,LJ}) = 10^{-4}\epsilon_{LJ,i}$. This gives a value of $r_{C,LJ} \simeq 5.849\sigma_i$. We then choose the largest σ_i from the HPS/KH models to determine a unified value of 39Å as the cutoff for E_{LJ}. Note that these default values are considered as the "safest" ones. The users may be able to change them to smaller values by carefully evaluating the balance between computational efficiency and accuracy.

Several nonbonded interaction terms are considered only for pre-defined pairs of CG particles, such as the DNA base-stacking (E_{bstk}) and the Gō-type native contacts ($E_{Gō}$). For these terms, we don't apply the cutoffs but directly calculate all the interactions.

Neighbor lists. Neighbor lists are used to maintain a list of particles that can participate in nonbonded energy/force interactions within a few MD integration steps. We set the neighbor list distances (r_P, also called "pair-list distance" in GENESIS) to be roughly 5Å larger than the largest cutoff distance (r_C). A complete list of the default cutoff and neighbor list distances can be seen in Table 2.

We utilize the cell linked list strategy to construct the neighbor list. Fig 3 illustrates how we determine the neighbor lists of different nonbonded interaction terms. During every update of the neighbor lists, the system is first divided into small cells. In each cell C_i and for each interaction term, we construct a "particle list", $L_{nb-term}(C_i)$, which contains the particles located in C_i and that are involved in the interaction ("nb-term"). For example, $L_{ele}(C_i)$ is a list of all the charged particles in the cell C_i. We then determine the "neighboring cells" for each cell: $C_{neighbor}^{(nb-term)}(C_i) = \{C_k, r_{min}(C_i, C_k) < r_{P,nb-term}\}$, where $r_{min}(C_i, C_k)$ is the minimum distance between the cells C_i and C_k. As shown in Fig 3A, relatively short-range interaction terms such as $E_{exv}^{(1)}$ or $E_{exv}^{(2)}$ have a small number of the neighboring cells, whereas those with longer cutoff distances, such as E_{ele}, will have more neighboring cells.

After defining the "particle list" ($L_{nb-term}(C_i)$) and "neighboring cells" ($C_{neighbor}^{(nb-term)}(C_i)$) of each cell, the neighbor list for each particle can be determined by the algorithm in Fig 3B (for convenience, here we use electrostatic interaction as an example). The neighbor lists are updated every 20 steps in the CG simulations.

Using individual particle list and neighboring cell for each nonbonded interaction term, we can minimize the pairwise distance calculations. First, as described above, we can avoid do loops over unnecessary cells for short-ranged interactions by using the potential-dependent neighboring cells. Besides, in each cell, we just consider a pre-assigned subgroup of particles (the $L_{nb-term}(C_i)$ lists defined above) instead of all the particles. For example, in a 3SPN.2C DNA system, to construct the neighbor list for the electrostatic interactions, we only have to scan the charged particles (phosphates), which are only ~1/3 of all the particles. Whereas for the base-pairing interactions, we scan another group of particles, the bases, also ~1/3 of all the particles. Generally speaking, for the pairwise non-local interactions, considering 1/N of the

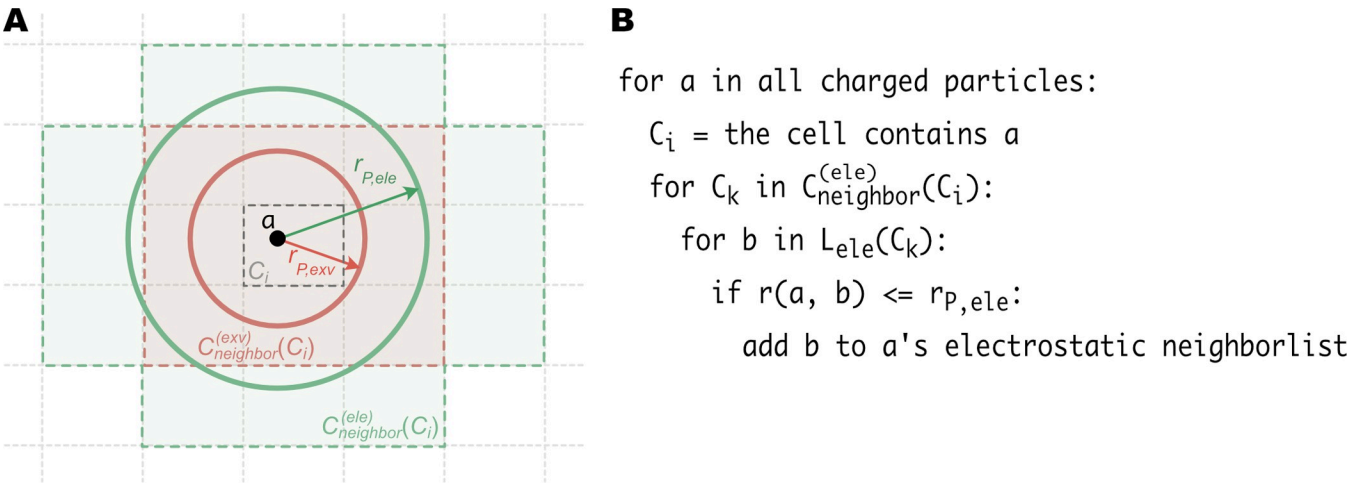

Fig 3. Using the cell linked list method to construct multiple neighbor lists. (A) An example of determining the neighbor list for E_{exv} (representing either $E^{(1)}_{exv}$ or $E^{(2)}_{exv}$) and E_{ele} on a two-dimensional space. The gray dotted lines show the edges of the cells. The center cell with dark dashed edges (C_i) contains the particle a (black dot) whose neighbor list is to be determined. $r_{P,exv}$ and $r_{P,ele}$ are the neighbor-list distances for E_{exv} and E_{ele}, respectively. The neighboring cells for the E_{exv} term ($C^{(exv)}_{neighbor}(C_i)$) are represented by the red cells and surrounded by the red dashed lines. Those for the $E_{ele}(C^{(ele)}_{neighbor}(C_i))$ are shown as the green area edged by green dashed lines. The red and green circles indicate the range of final neighbor lists for E_{exv} and E_{ele}, respectively. (B) Pseudo code of the algorithm to determine a neighbor list of the electrostatic interaction. $L_{ele}(C_k)$ is a list of all the charged particles in cell C_k.

https://doi.org/10.1371/journal.pcbi.1009578.g003

particles means only $1/N^2$ calculations compared to scanning over all the particles. Therefore, using cell linked list and grouping particles according to interaction terms can save a lot of computations in the neighbor list construction. On the other side, however, we also note that when periodic boundary conditions are used with the cell linked list method, the simulation box should have each dimension equal to or larger than three times the largest pair-list distance (57Å for the electrostatic interaction by default).

Time Integration. GENESIS atdyn has provided various integrators and thermostats/barostats. For the current CG implementation, we employ the velocity-Verlet integrator coupled with the Langevin thermostat. In the velocity-Verlet algorithm, coordinates ($\mathbf{r}_i$) and velocities ($\mathbf{v}_i$) are updated following:

$$\mathbf{v}_i(t) = \mathbf{v}_i(t-\Delta t) + \frac{\Delta t(\mathbf{F}_i(t) + \mathbf{F}_i(t-\Delta t))}{2m_i}, \tag{26A}$$

$$\mathbf{r}_i(t+\Delta t) = \mathbf{r}_i(t) + \Delta t\mathbf{v}_i(t) + \frac{\Delta t^2}{2m_i}\mathbf{F}_i(t), \tag{26B}$$

where t is the simulation time, Δt is the integration step size, m_i, $\mathbf{r}_i$, $\mathbf{F}_i$, and $\mathbf{v}_i$ are the mass, coordinate, force, and velocity of the i-th particle, respectively. In the Langevin thermostat, the force $\mathbf{F}_i$ is calculated by:

$$\mathbf{F}_i(t) = -\frac{\partial V_{total}(\mathbf{r}_i(t))}{\partial \mathbf{r}_i(t)} - m_i\gamma\mathbf{v}_i(t) + m_i\boldsymbol{\xi}_i(t), \tag{27}$$

where V_{total} is the total potential energy and γ is the friction constant. $\boldsymbol{\xi}_i(t)$ is the Gaussian noise vector satisfying:

$$\langle\boldsymbol{\xi}_i(t)\rangle = 0, \langle\boldsymbol{\xi}_i(t)\boldsymbol{\xi}_i(t')^T\rangle = \frac{2\gamma k_B T}{m_i}\frac{\delta_{t,t'}}{\Delta t}\mathbf{I}, \tag{28}$$

where k_B is the Boltzmann constant, T is temperature, $\delta_{t,t'} = 1$ when $t = t'$ and 0 otherwise, and $\mathbf{I}$ is an identity matrix of size 3×3. $\boldsymbol{\xi}_i(t')^T$ is the transpose of $\boldsymbol{\xi}_i(t')$ and the product follows matrix multiplication.

We use 10 femtosecond (fs) as the literal time step size (Δt in Eq 26), which is determined by the highest vibration frequency of the particles. On the other side, the time scale mapping of large-scale conformational dynamics of biomolecules is system-dependent and requires a careful comparison between simulated and experimental physical quantities [22].

Performance Optimization

Next, we optimize the performance of CG MD simulations implemented in GENESIS. At first, we optimize the nonbonded interactions, which are the most time-consuming parts in both residue-level CG and all-atom force-field models. There are two common strategies in the optimization. The first one is the vectorization of the most inner loops for utilizing as many SIMD (single instruction and multiple data) instructions as possible. The second one is to minimize the number of computations by using conditional statements. In contrast to all-atom models, using "conditional" statements is not avoidable due to the complex potential energy function forms. In our experience, the ratio r_P^3:r_C^3 gives us a good criteria whether the calculation would be vectorized or not using "conditional" statements. For instance, in the case of E_{ele}, the ratio is about 1.3, and it is better to apply vectorizations to the inner loop without any "conditional" statements. The ratio in $E_{exv}^{(1)}$ or $E_{exv}^{(2)}$ evaluations are greater than 3.5, and it is better to optimize by minimizing the calculation amount with conditional statements. The ratio in E_{LJ} or E_{bp} is less than 3, but we use conditional statements because the energy expression form itself is changed in HPS, KH, or 3SPN.2C models.

Another computational bottleneck in CG MD simulations is the generation of random numbers with Gaussian distributions when using the Langevin thermostat. This is mainly because of the replicated data MPI parallelization scheme used in GENESIS atdyn, in which all processes have a copy of the coordinate data. By keeping the replicated data scheme, we assign MPI parallelization in random number generation and apply "MPI_Allgatherv" to collect the information. We do not apply MPI parallelization to integrate coordinates or momenta because the communication time is comparable or even more time-consuming than the integration itself. Instead, OpenMP parallelization is applied in integrations.

Results

Benchmark of MD simulations with residue-level CG models

We first examine the usage of memory and the performance benchmark in MD simulations with residue-level CG models. The memory usage was examined with a single process and a single thread on a machine with 93GiB RAM (random access memory). The CPU benchmarks were executed with 5 OpenMP threads and various MPI processes on a computer server with Intel Xeon Gold-6148 (2.4GHz) CPU (20 cores per CPU and two CPUs per node) and InfiniBand EDR networking. GENESIS was compiled with the Intel compiler in couple with Intel MPI version 2019.5.281. Fig 4 shows the results of the memory benchmark (Fig 4A) and CPU benchmark using different biological systems (Fig 4B, 4C and 4D).

To test the memory consumption, we built systems consisting of different numbers of RNA Polymerase II (PolII). The smallest system has only one PolII and contains 3695 CG particles (3542 from protein, 124 from DNA, and 29 from RNA), based on the X-ray crystallography structure (PDB ID: 1R9T) [71]. We then duplicated the single PolII to create systems of different sizes, ranging from thousands to millions of particles. In all these tested systems, we used

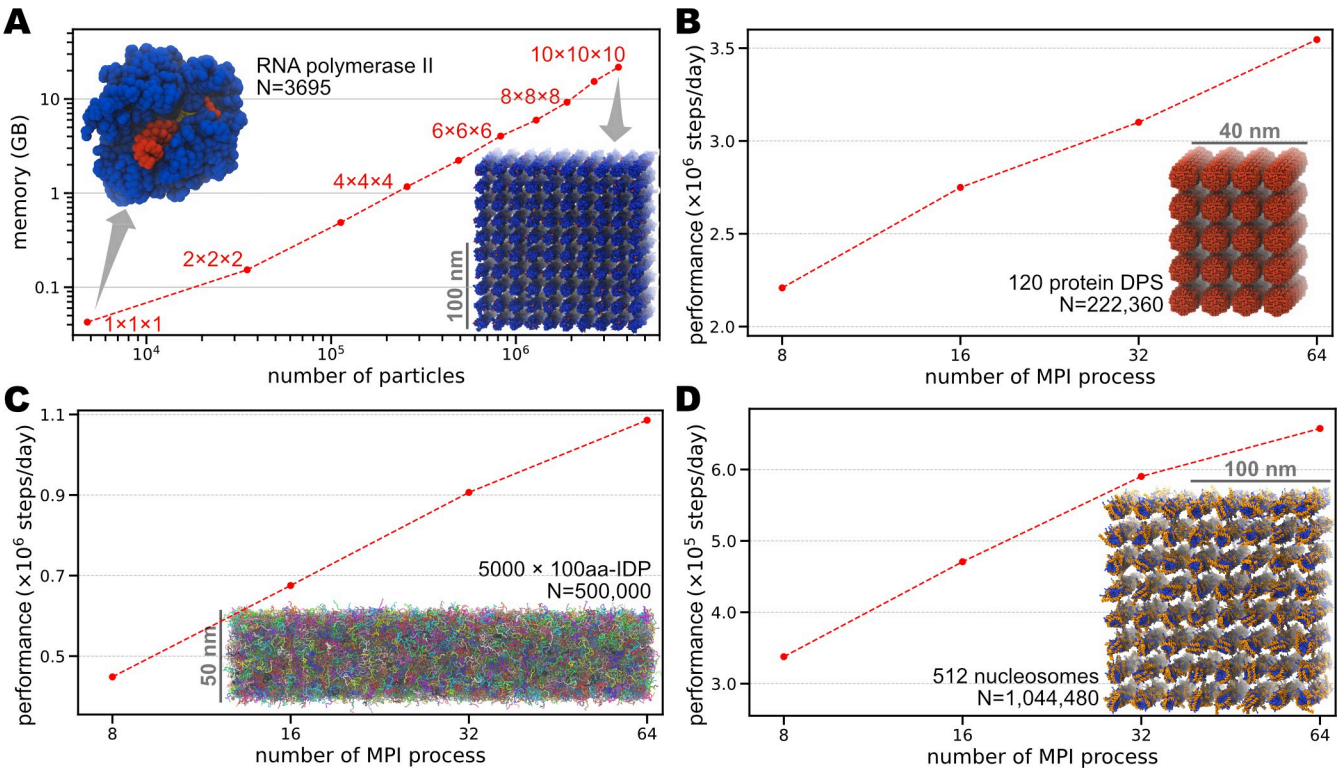

Fig 4. Benchmark of the GENESIS CG simulations. (A) Memory benchmark of the RNA Polymerase II (PolII) systems. A single PolII has 3,695 CG particles, which is then duplicated into a $n \times n \times n$ ($n = 1,\ldots,10$) grid space. The insets show the structure of a single PolII (left) and 1000 PolIIs (right). Protein is shown in blue, DNA in red, and RNA in yellow. (B) CPU benchmark of a system containing 120 DPS proteins. The total number of CG particles is 222,360. (C) CPU benchmark for a system of 5000 chains of 100aa IDPs. The total number of CG particles is 500,000. (D) CPU benchmark for a system including 512 nucleosomes. The total number of CG particles is 1,044,480. The insets of (B), (C), and (D) show the initial structure of each system, respectively. N represents the total number of CG particles in each system. For all the CPU benchmarks, we used 5 OpenMP threads.

https://doi.org/10.1371/journal.pcbi.1009578.g004

V_{AICG2+} for proteins, $V_{3SPN.2C}$ for DNAs, and V_{RNA} for RNAs. For proteins, we use integer charges distributed on the charged residues ($+e$ for Arg and Lys, $-e$ for Asp and Glu). For the interactions between different biomolecule types, we applied $E^{(2)}_{exv}$ and E_{ele} as general nonspecific interactions, and $E_{G\bar{o}}$ for native contacts between protein-RNA, protein-DNA, and RNA-DNA to maintain the whole structure. All the cutoff and pair-list distances are set to the default values (see Table 2). As shown in Fig 4A, we carried out simulations with n^3 ($n = 1,\ldots10$) duplicated PolIIs to track the memory usage. The consumed memory scales with the system sizes, from ~40 MB for a single PolII to ~22 GB for 1000 PolIIs. Notably, on an ordinary computer with 10GB memory, we can run GENESIS CG MD simulations for a system with about 2 million particles.

To examine the strong scaling of GENESIS CG simulations, we prepare three systems comprised of different types of biomolecules:

1. 120 copies of the protein DPSs [72]. In total, 222,360 CG particles modeled with V_{AICG2+};
2. 5000 chains of IDPs, each consisting of 100 amino acids. In total, 500,000 CG particles modeled with V_{HPS};
3. 512 copies of nucleosomes [73]. In total, 1,044,480 particles, using V_{AICG2+} for proteins, $V_{3SPN.2C}$ for DNAs, and $E^{(2)}_{exv}$, E_{ele} for protein-DNA interactions.

$E^{(2)}_{exv}$ and E_{ele} are used as general non-specific interactions for systems 1) and 3); and E_{HB} is additionally applied between protein and DNA in system 3). For all the cutoff and pair-list distances, we use the default values listed in Table 2. The benchmark results for these three systems are shown in Fig 4B, 4C and 4D, respectively. The results suggest that at least up to 64 MPI processes (8 nodes), the parallel performances are scalable in all three systems. Although this does not show perfect scalability, it is useful for simulating large biological systems. We discuss possible reasons for the insufficient scalabilities of our implementation and future plans to improve the performance in the Discussion section.

Application 1: protein target search on DNA

The AICG2+ protein model (V_{AICG2+}), and the 3SPN.2C DNA model ($V_{3SPN.2C}$), in combination with the sequence-specific (E_{PWMcos}) or nonspecific (E_{HB}) protein-DNA interactions, can be used to study the binding, diffusion, and conformational changes of protein-DNA complexes [29,30,61,62]. As a simple example, we performed MD simulations of the sex-determining region Y protein (sry) [74] and its target DNA. The modeling and simulation results are shown in Fig 5.

Sry uses a conserved High-Mobility Group (HMG) domain to recognize its consensus DNA sequence, "TAAACAAT" [75]. The HMG domain intercalates into the DNA minor groove and forms contacts with the bases. This type of minor groove binding of sry results in a sharp bending of the DNA [76], which plays a crucial role in regulating the transcription of its target gene [77]. Starting from the solution NMR structure of the sry-DNA complex (PDB entry 1J46 [76], see Fig 5A), we generated the CG model of sry using the GENESIS-CG-tool. We used the AICG2+ model (V_{AICG2+}) for the whole protein. Specifically, we applied Gō-type native contact interactions ($E_{Gō}$) to the folded HMG domain (residue 9–70) and assigned only local flexible potentials ($E^{(1)}_{bond}$, $E^{(2)}_{angle}$, and $E^{(3)}_{dihedral}$) to the N-terminal (residue 1–8) and C-terminal tails (residue 71–85). We also employed the RESPAC method [60] to calculate the partial charge distribution on the surface residues of sry (see Figs 5B and S2). As for the DNA, we designed a 50bp poly-CG sequence, with the 8-bp piece "TAAACAAT" inserted at the center (Fig 5C and 5D). The atomistic structure of the dsDNA was then constructed from this 50-bp sequence using the 3DNA package [65]. The 3SPN.2C CG topology and coordinate files were then generated from the atomistic structure, using the GENESIS-CG-tool. For sry to find its consensus sequence in the simulations, we utilize the PWMcos model (E_{PWMcos}) to incorporate information from the position frequency matrix (PFM, downloaded from JASPAR database [78] profile MA0084.1 [75]). For details of the model parameters, please see S1 Text.

After getting the input files, we carried out MD simulations of the sry-DNA complex for 10^7 steps. In the initial structure, sry was placed around 70Å away from the DNA. Soon after the simulation started, sry was bound onto DNA and began the sliding. We monitored the binding site of sry on DNA (Fig 5E) and the bending of DNA (Fig 5F). As can be seen, sry is preferentially bound to the consensus sequence (represented by the green region in Fig 5E) and occasionally left this target with a quick diffusion on the DNA. As expected, we observed coupling between the sequence-specific recognition and DNA bending caused by the intercalation of the HMG domain. These results show that the CG models (AICG2+, 3SPN.2C, and PWMcos) for protein-DNA interactions have been correctly implemented in GENESIS and can be used to study similar biological systems.

Application 2: phase behaviors of IDR and RNA

The HPS and KH IDR models have been used to study the phase behavior of IDRs [35]. Here, we used the IDR from protein Fused in Sarcoma (FUS) [79] to show that our implementation

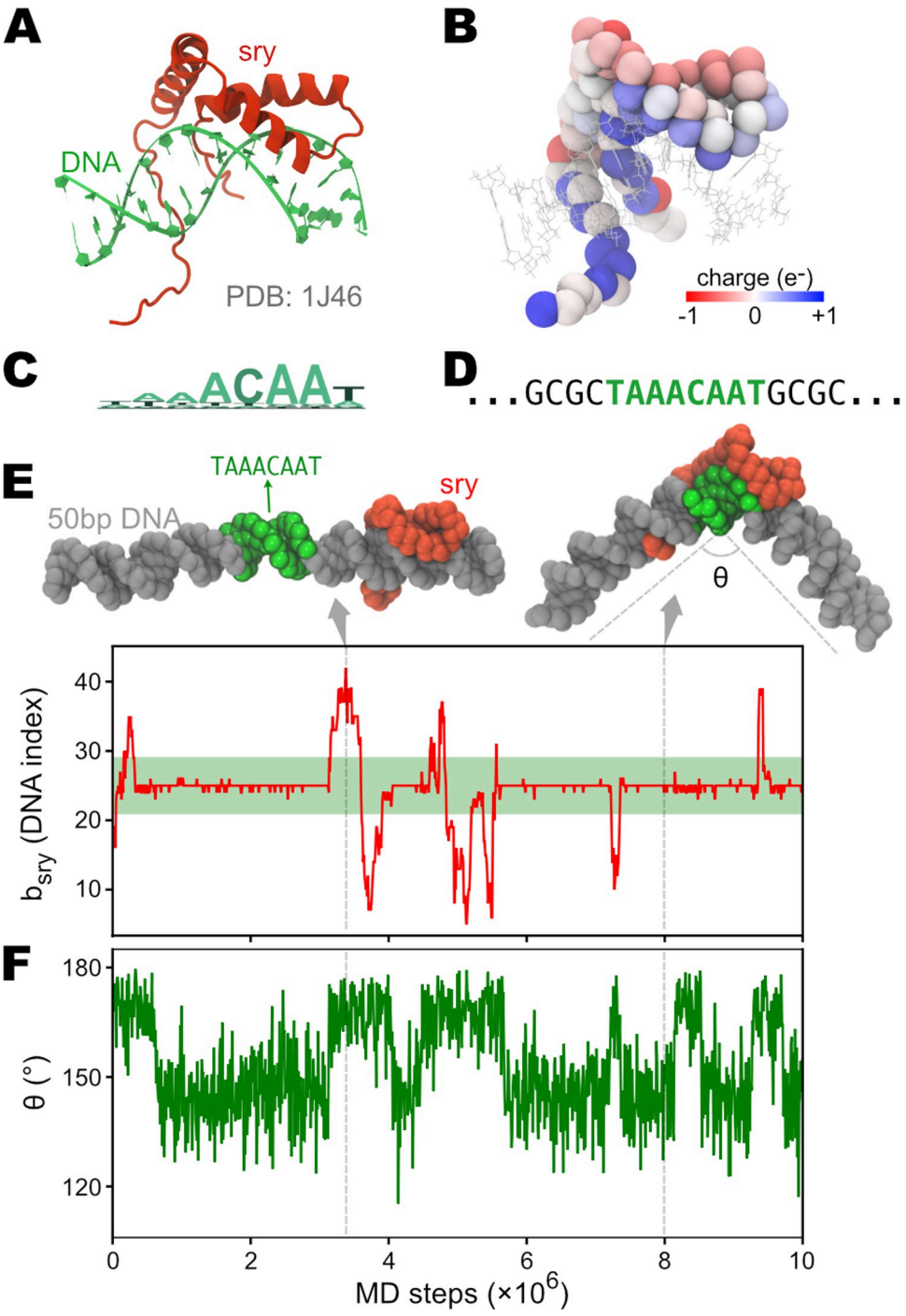

Fig 5. CG MD simulations of the DNA target search process by protein sry. (A) Solution NMR structure of sry binding on DNA (PDB entry: 1J46). (B) Coarse-grained sry with the Cα particles colored by the partial charges (determined using the RESPAC method). DNA structure from the PDB is also shown as lines for reference. (C) Sequence logo of sry's DNA target. (D) Sequence of DNA used in the CG simulations. The consensus sequence "TAAACAAT" is inserted at the center of a 50bp poly-CG DNA. (E) Time series of sry's binding position on DNA (*bsry*). The green region represents the consensus sequence. (F) Time series of the bending angle of DNA (θ). Two representative structures of sry binding at the poly-CG region or the consensus region are shown on top of (E). In these structures, DNA is colored in gray, except for the consensus sequence region in green. Sry is shown in red.

https://doi.org/10.1371/journal.pcbi.1009578.g005

of these models in GENESIS can simulate the condensation of proteins. We first simulated a single chain of the FUS IDR (163aa, see S1 Text for the sequence). The final structure of the single-chain simulation was then duplicated to create a 120-chains system. We carefully arranged the multiple chains so that the system had a boundary size of $18nm \times 18nm \times 200nm$, which is for the slab sampling method. We performed a 10^7-step simulation, as shown in Fig 6A and 6B. During the simulation, we monitored the density change of the CG particles along the z-axis (the longest dimension). As shown in the top panel of Fig 6A, at the beginning of the simulation, the FUS IDRs formed relatively small and low-density clusters. As simulation time increased, we observed several merge events of the condensates. After ~3.5×10^6 steps, all the FUS IDRs went into the same cluster, and the density finally reached around 5 particles per nm^3 (Fig 6A and 6B). This simple example shows the possibility of using the HPS model in GENESIS to explore the physical mechanisms of IDR condensation. Besides, we noticed that there are some recent updates of the parameters of the HPS model [36,80]. These changes of parameters can be easily used with GENESIS by modifying the parameter files, with no need to touch the code. In this sense, GENESIS can also serve as a handy framework for the users to re-calibrate their own set of parameters.

In conjunction with the HPS model for protein IDR, the HPS model for RNA has also been recently developed [37]. The HPS RNA model has a different resolution from the other CG models described above, with one CG particle per nucleotide. Nevertheless, we also implemented this model in GENESIS, preparing for possible applications in the study of biophysical condensations involving both protein and RNA. Here we built and simulated two systems of IDR from the protein LAF1 and 15-nt (nucleotide) poly-A RNA (hereafter called A15). The two systems consisted of 300 LAF1 together with 100 A15 (57,300 particles in total) and 750 LAF1 together with 250 A15 (143,250 particles), respectively. The slab method was used to simulate the first system for 5×10^6 steps, with the simulation box of $20nm \times 20nm \times 200nm$ (Fig 6C). The second system was simulated in a cubic box of size $(100nm)^3$ for 2×10^6 steps (Fig 6D). As can be seen in Fig 6C and 6D, in both situations, the A15 RNA entered the condensation of LAF1. We also quantitatively analyzed the co-condensation of LAF1 and RNA by plotting the radial distribution function (RDF) of protein and RNA particles as a function of the distance from the center of the LAF1 droplet (Fig 6E). There is a higher density of RNA inside the LAF1 condensation than in bulk (Fig 6E). These results are consistent with the previous simulation results [37].

Application 3: virus capsid

We also explored the ability of our GENESIS CG implementation to simulate large-scale biological systems, such as a virus capsid. Here we reported our modeling and simulation of the herpes simplex virus (HSV) capsid, whose high-resolution structure has been recently solved with the cryo-electron microscopy (cryo-EM) (PDB entry: 5ZAP) [81]. The capsid comprises about 4000 proteins assembled into 12 pentons (pentameric blocks) and 150 hexons (hexameric blocks), and 320 triplexes gluing all the subunits [81]. The number of atoms and the length scale exceed the PDB format's upper limit for such a vast structure. Therefore, the structural information of the HSV capsid was recorded in a file with the mmCIF format. We used the GENESIS-CG-tool to parse this file and generate the CG structure, which consists of 1,687,980 CG particles in total (Fig 7A). The AICG2+ potentials were used to model the whole structure. After generating the necessary input files, we then simulated the structure for 2×10^6 steps. We tracked the radius-of-gyration (R_g), the root-mean-square-deviation (RMSD), and the nativeness (Q, see S1 Text for definition) of the whole structure (shown in Fig 7B). As can be seen, during our simulation, the structure was stably maintained. Although here we only show a

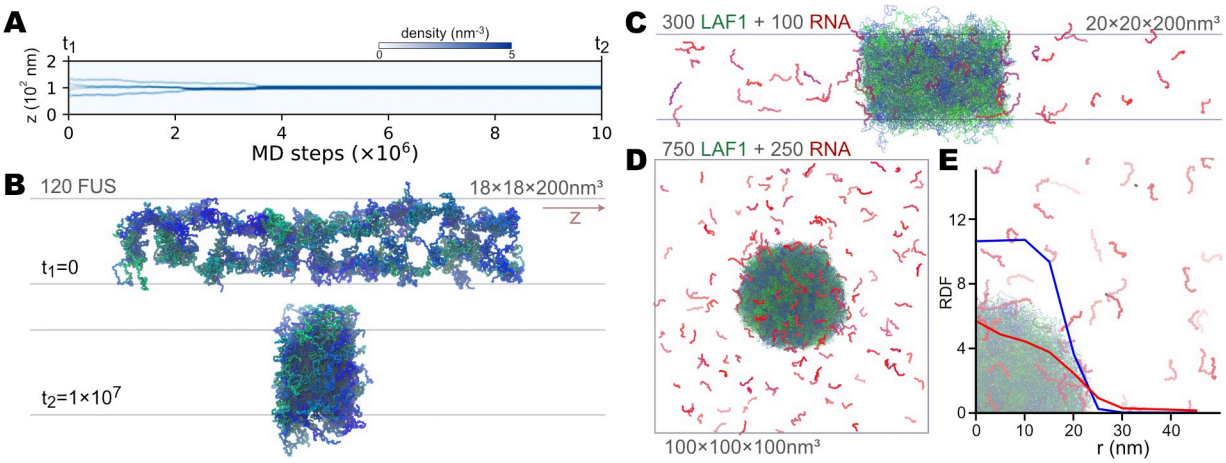

Fig 6. CG MD simulations of the condensation of IDRs and RNAs. (A) and (B) show the slab simulation results of 120 FUS IDRs. (A) Time series of FUS density along the long axis (z). The intensity of the blue color represents the density, as shown in the color bar above. (B) The first and last structure of the MD trajectory shown in (A). The FUS IDRs are shown as tubes. Different colors from green to blue are used to distinguish chains. (C), (D), and (E) show the simulation of the mixture of LAF1 IDRs and 15nt poly-A (A15) RNAs. (C) The last structure of a 5×10^6 steps slab simulation of 300 LAF1 IDRs and 100 A15 RNAs. LAF1 IDRs are in green or blue, whereas RNAs are in red or purple color. (D) The last structure of a 2×10^6 steps droplet simulation of 750 LAF1 IDRs and 250 A15 RNAs. The color scheme is the same as (C). (E) Radial distribution function of protein and RNA CG particles as a function of the distance from the center of mass of the LAF1 droplet (r). The structure of the condensation is shown as transparent background for reference.

https://doi.org/10.1371/journal.pcbi.1009578.g006

short equilibrium simulation of the HSV capsid, to our knowledge, it is one of the largest biological systems simulated with the residue-level CG models. We expect to run CG simulations of similar systems using GENESIS to probe the conformational changes and dynamic assembly or disassembly of the capsid.

Application 4: chromatin

Eukaryotic chromatin is the architecture that stores genetic information by packaging DNA into supercoiled structures around the histone octamer proteins. In addition to the function of genome organization, chromatin also hosts many biochemical reactions around the information flow between DNA and RNA. Due to its extraordinary importance, chromatin and related biological phenomena have been widely studied by MD simulations at different length scales and models at different resolutions [4,82–84]. In particular, residue-level CG models, including those we implemented here (V_{AICG2+}, $V_{3SPN.2C}$, E_{PWMcos}, and E_{HB}), have shown their abilities to decipher the dynamics of the nucleosome, which is the building block and basic unit of chromatin [30,31,61,62,85]. Here we tried to expand these residue-level CG MD simulations to a larger length scale. Instead of using a natural genomic sequence and constructing a Hi-C experiment-based structure, we chose poly-CG as the DNA sequence and built an artificial chromatin structure by connecting single nucleosome structures (based on PDB 1KX5) [73] with linear linker DNAs. The constructed structure contains 1024 nucleosomes linked by a 219,213-bp dsDNA (see Fig 8A and 8B). More details of the structure modeling can be found in S1 Text.

Similar to previous studies [30,61], we used V_{AICG2+} for histones, $V_{3SPN.2C}$ for DNA, and E_{HB} for hydrogen bonds between histone and DNA phosphates. In addition, we applied V_{HPS} to the flexible histone tails (see S1 Text for definition). We then performed 5×10^5-step MD simulations of this system at a temperature of 300K and 150mM ionic concentration. The time series of R_g of the four types of N-terminal histone tails and the number of wrapped DNA base-pairs are shown in Fig 8C. These results were based on the statistics over the 1024 nucleosomes. This example illustrates the possibility of using our GENESIS CG implementation to

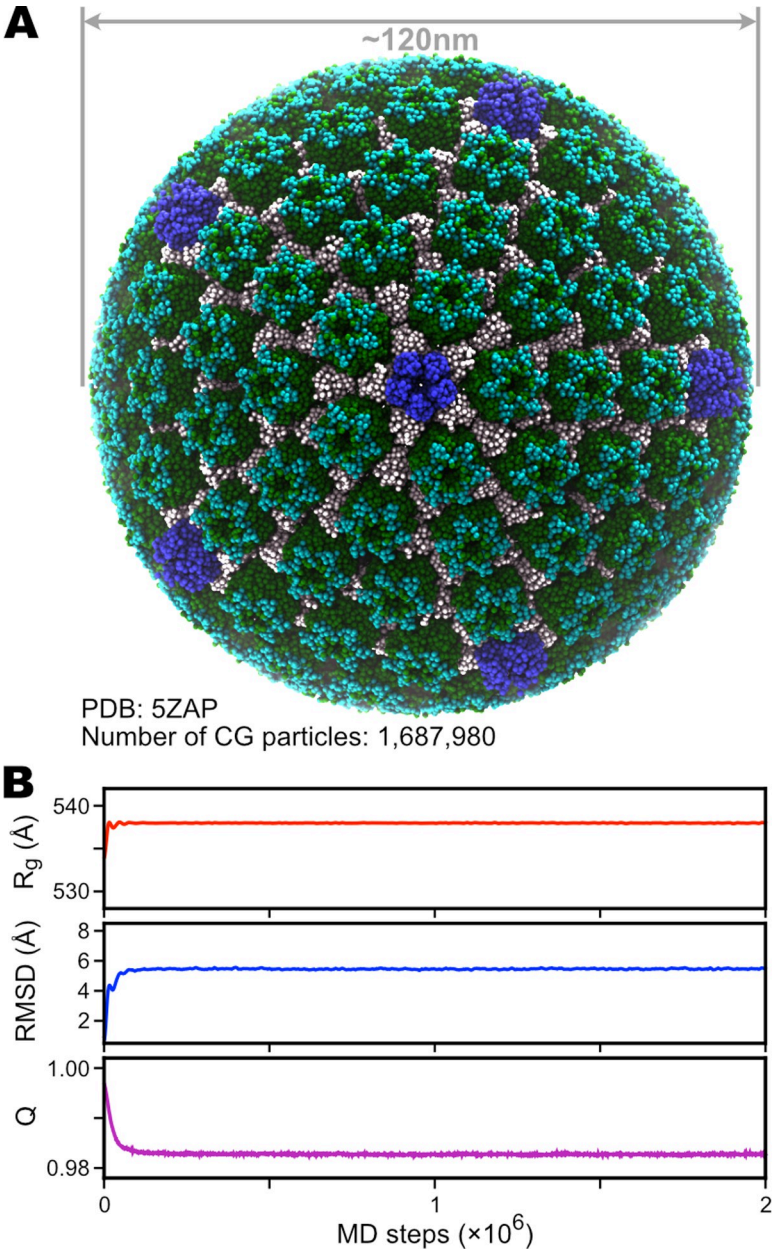

Fig 7. Simulation of the herpes simplex virus (HSV) type 2 B-capsid. (A) Structure of the coarse-grained HSV capsid, based on PDB entry 5ZAP. The pentameric blocks (pentons) are colored in blue, the hexameric blocks (hexons) are colored in green (VP5) and cyan (VP26). All the other proteins that glue together the pentons and hexons are colored in white. (B) Time series of the radius of gyration (*Rg*), RMSD, and nativeness (*Q*) of the whole structure during a 2×10^6-step simulation.

https://doi.org/10.1371/journal.pcbi.1009578.g007

carry out studies on the large-scale chromatin dynamics with accurate modeling of DNA and both well-folded and intrinsically disordered parts of proteins.

Discussion

In this paper, we have implemented residue-level CG models of biomolecules in GENESIS MD software and have applied them in CG MD simulations of various biomolecules by

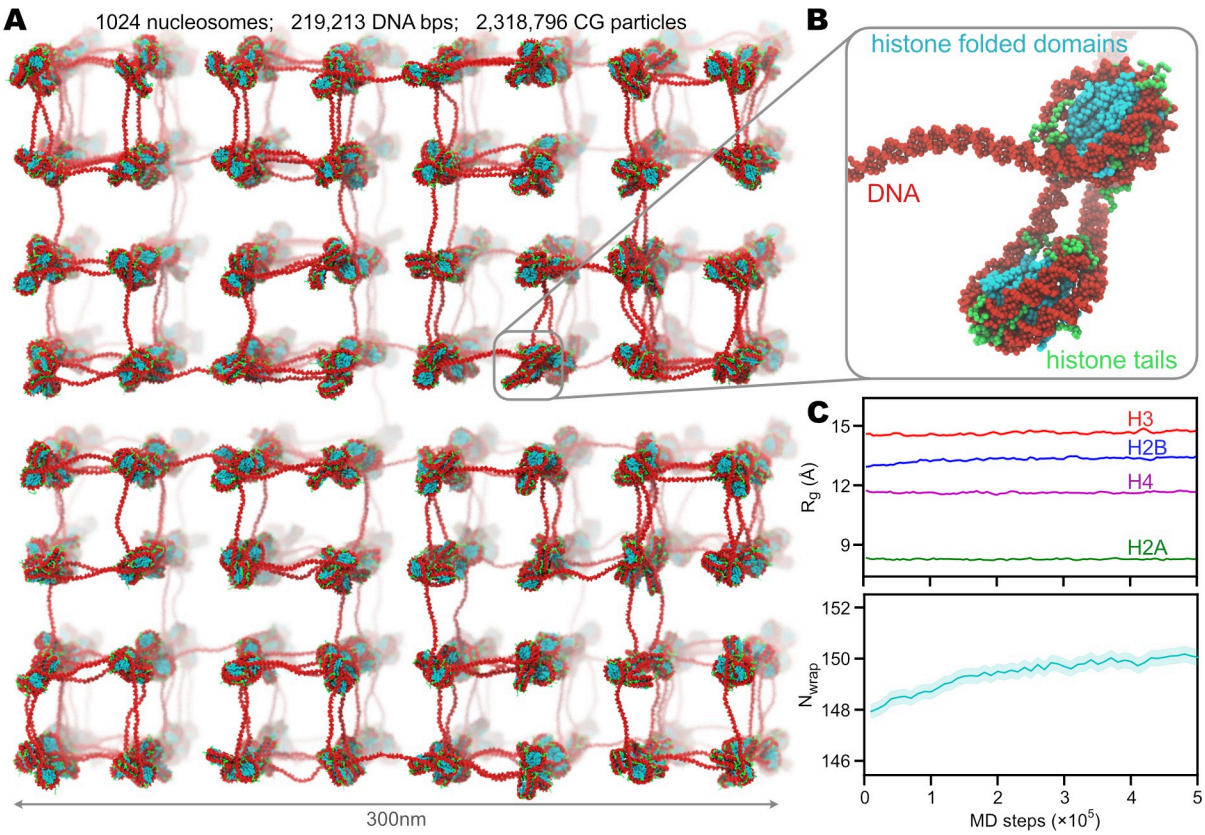

Fig 8. MD simulation of artificial chromatin composed of 1024 nucleosomes. (A) Structure of the whole artificial chromatin. DNA is colored in red, and histones are colored in cyan (for the folded domains) and green (tails). (B) A zoom-in structure of two adjacent nucleosomes as a part of the artificial chromatin. (C) Time series of the radius-of-gyration (*Rg*) of the N-terminal tails of the histones (top) and the number of wrapped DNA base-pairs on histone (bottom). The solid lines represent the average values, and the shadowed regions show the standard errors.

https://doi.org/10.1371/journal.pcbi.1009578.g008

combining different CG models. The combination of the AICG2+ model for proteins with the 3SPN.2C model for DNA and the PWMcos potential for protein-DNA sequence-specific interactions has been applied in the previous studies [62]. Whereas the combination of the HPS model for IDR and AICG2+ for folded protein is tested, for the first time, in this new framework. We also noticed several recently published CG models at an equivalent or similar resolution as ours, such as the TIS RNA model [23] and the iSoLF lipid model [86]. We will consider incorporating these models into GENESIS to cover a broader range of target biomolecules. In the applications presented here, we used the default interaction parameters in GENESIS. For more practical applications, such as genome-scale chromatin folding or phase-separated membrane-less condensations, one may need to more carefully optimize interaction terms between the IDR and the other components and their parameters. If necessary, the parameters should be calibrated carefully with experimental information as references. The GENESIS CG framework provides a convenient environment for these attempts.

Similar to the current work, other groups have also released combinatorial implementations of CG methods, such as the OpenAWSEM-Open3SPN2 within the OpenMM framework [43], the 3SPN.2C packages in the LAMMPS framework [38,39], and the CafeMol package [44]. Particularly, the OpenMM package utilized GPU calculations to gain better performances [43]. Compared with its implementation, our development aims to cover more models and provide a more flexible and feasible environment for the studies of larger-scale biomolecular

systems. CafeMol has been successfully applied for CG simulations in various biomolecular studies [26,29,30,62,87,88]. However, it is not designed for the high-performance computing of larger-scale systems. We have optimized our code in GENESIS to enable high-performance simulations of systems consisting of millions of particles on regular computers (Figs 4, 7 and 8). There is an upper limit of particle number with the atomic decomposition scheme in atdyn, which essentially limits our the applicability with the current code to huge biological systems. The domain-decomposition strategy and the efficient load balancer suitable to residue-level CG models are required to solve this problem. These two will be implemented in a new MD engine in GENESIS. However, it is worth noting that there are technical difficulties in developing the domain-decomposition scheme and load balancer for CG simulations. The most widely appreciated problem is the inhomogeneous particle distribution in the residue-level CG systems due to the implicit treatment of the solvent. In a conventional uniform spatial decomposition, the non-uniform particle density causes severe load imbalance among different domains. Therefore, special decomposition algorithms should be considered [89]. Another direction is to consider explicit solvent models [13,90], which intrinsically solve the load-balance problem, but introduce significantly more computations on the solvent particles. Besides, the large cutoff value of the non-bonded interactions also sets a lower limit of the target system size. We assume at least two domains in each dimension to apply domain decomposition. Each domain should have at least the size equal to or larger than the longest cutoff (in our implementation, 52Å of the electrostatic term). This requirement essentially limits the applicability of the program to relatively small systems. In developing the new CG MD engine in GENESIS, we will have to solve these problems.

Although the CG models have boosted the computational efficiency to be orders of magnitudes faster than the all-atom force fields [22], the speed-up still looks insufficient when studying systems of the organellar or cellular scales. It is necessary to employ enhanced sampling methods, many of which [91–94] have also been implemented in GENESIS. Taking this advantage, we will combine CG MD simulations with some of the enhanced sampling methods to obtain more statistically reliable data with reasonable computational times and to enable free-energy analysis in the organellar or cellular scales.

In parallel to the advances in computational biophysics, mesoscopic pictures of biomolecules in the organellar or cellular environments have been better described using the latest experimental techniques, such as cryo-electron microscopy (cryo-EM) [95], cryo-electron tomography (cryo-ET) [96], high-speed atomic force microscopy (AFM) [97], and in-cell NMR [98]. Integrating structural information by these experiments with computational biomolecular dynamics is one of the most important tasks in this research field. Recently, there have been several attempts [99–101] to make more realistic structural models of sub-cellular systems by utilizing available experimental results as much as possible. We hope that the implementation of residue-level CG models can provide more realistic structural dynamics of heterogeneous biological systems together with recent experimental information and modeling strategies. For this purpose, multi-scale and multi-resolution simulations are helpful in the trade-off between modeling accuracy and computational efficiency, if we can switch the resolutions of target biological systems easily from this residue-level CG models to other CG models with different approximations (MARTINI [13], SPICA [102], PACE [103], etc.) or all-atom force field models (AMBER [104], CHARMM [105], GROMOS [106], etc.). In the GENESIS software, most of the above models are already available, and some of them are being implemented by us. Mapping from the atomistic structure to CG is deterministic and easy, although information such as amino-acid backbone dihedral angles (φ and ψ) and hydrogen bonds are largely lost. The other direction, namely, reconstructing atomistic structure from CG coordinates, is usually much more difficult [107]. In our current implementation, the structure-

based models, such as AICG2+ [26], can uphold the secondary, tertiary, or even higher-order structures in the native structure. Additionally, by using the angle-dependent non-local potentials, the 3SPN.2C [28] and the PWMcos [32] models can partly preserve the base-stacking or hydrogen-bonding interactions in the DNA and protein-DNA interface, respectively. Structures sampled with these CG models can be remapped back to reasonable atomistic structures, provided the reconstruction methods [107,108] are good enough. In the near future, we may be able to investigate the inside of the cell at various resolutions, as Google Earth can do visualizations of both countries and local towns.

In summary, we have implemented several residue-level CG models in the GENESIS MD software, including the AICG2+ model for protein, the 3SPN.2C model for DNA, the PWMcos model for protein-DNA interactions, and the HPS/KH model for IDR and RNA. The MD input data for these CG models are given in a GROMACS-style file format by an easy-to-use toolbox, GENESIS-CG-tool. The computational performance of CG MD simulations is optimized to regular computers by preparing the nonbonded interaction lists suitable for the CG potential functions, vectorizing do-loop structures in the time-consuming subroutines, parallelizing the generation of random numbers in Langevin dynamics, and so on. With these attempts, GENESIS CG MD simulation is applied to large biological systems containing millions of CG particles efficiently on regular computers. The models and programs implemented here would be valuable for investigating heterogeneous biological phenomena involving folded/disordered proteins, RNAs, and DNAs at high concentrations. Implementations for larger systems containing billion CG particles and the use of enhanced sampling in the systems will follow the current implementation in the GENESIS software.

Availability and future directions

The source code and user manual of GENESIS v1.7.1 is available at https://www.r-ccs.riken.jp/labs/cbrt/ website, and GENESIS-CG-tool is both included as a part of GENESIS v1.7.1 and deployed at https://github.com/noinil/genesis_cg_tool. Tutorials of performing CG simulations with GENESIS v1.7.1 are open on https://www.r-ccs.riken.jp/labs/cbrt/tutorials2019/ website. All the MD simulation files and data to produce the results are available from https://github.com/RikenSugitaLab/cg-development-GENESIS-1.7.0. We plan to use GENESIS v1.7.1 to study large-scale biological phenomina such as LLPS of protein and nucleic acids. We also plan to implement more CG models of biomolecules into GENESIS.

Supporting information

S1 Text. Supplementary methods of molecular dynamics simulations.
(DOCX)

S1 Fig. Examples of using GENESIS-CG-tool to prepare CG topology and coordinate files. (A) Generate CG files (PRO1_cg.gro for coordinates; PRO1_cg.top and PRO1_cg.itp for topology) for a protein, based on its atomistic PDB file (PRO1.pdb). (B) Generate CG files (DNA1_cg.gro for coordinates; DNA1_cg.top and DNA1_cg.itp for topology) as well as an atomistic coordinate file (DNA1.pdb) for a 20bp DNA, from its DNA sequence (DNA1.fasta).
(TIF)

S2 Fig. Charge distribution on the surface residues of protein sry. The charges of the surface residues of the HMG domain (residue 9–70, dark blue) of sry was determined with the RESPAC method. Whereas the charges on the N-tail and C-tail were the original integer charges (green). The same data are used to plot the CG structure of sry (Fig 5B).
(TIF)

Acknowledgments

The authors thank Ai Shinobu and Giovanni B Brandani for testing the developments of CG models in GENESIS v1.7.0.

Author Contributions

Conceptualization: Cheng Tan, Jaewoon Jung, Chigusa Kobayashi, Yuji Sugita.

Data curation: Cheng Tan.

Formal analysis: Cheng Tan, Jaewoon Jung, Chigusa Kobayashi.

Funding acquisition: Cheng Tan, Jaewoon Jung, Shoji Takada, Yuji Sugita.

Investigation: Cheng Tan, Jaewoon Jung, Chigusa Kobayashi, Diego Ugarte La Torre.

Methodology: Cheng Tan, Jaewoon Jung, Chigusa Kobayashi.

Project administration: Cheng Tan, Jaewoon Jung, Chigusa Kobayashi, Shoji Takada, Yuji Sugita.

Resources: Cheng Tan, Jaewoon Jung, Chigusa Kobayashi.

Software: Cheng Tan, Jaewoon Jung, Chigusa Kobayashi.

Supervision: Shoji Takada, Yuji Sugita.

Validation: Cheng Tan, Jaewoon Jung, Chigusa Kobayashi, Diego Ugarte La Torre.

Visualization: Cheng Tan.

Writing – original draft: Cheng Tan, Jaewoon Jung, Chigusa Kobayashi.

Writing – review & editing: Cheng Tan, Jaewoon Jung, Chigusa Kobayashi, Diego Ugarte La Torre, Shoji Takada, Yuji Sugita.

References

1. Jerkovic´ I, Cavalli G. Understanding 3D genome organization by multidisciplinary methods. Nat Rev Mol Cell Biol. 2021; 22: 511–528. https://doi.org/10.1038/s41580-021-00362-w PMID: 33953379
2. Yoshizawa T, Nozawa R-S, Jia TZ, Saio T, Mori E. Biological phase separation: cell biology meets biophysics. Biophys Rev. 2020; 12: 519–539. https://doi.org/10.1007/s12551-020-00680-x PMID: 32189162
3. Di Pierro M, Zhang B, Aiden EL, Wolynes PG, Onuchic JN. Transferable model for chromosome architecture. Proc Natl Acad Sci. 2016; 113: 12168–12173. https://doi.org/10.1073/pnas.1613607113 PMID: 27688758
4. Nuebler J, Fudenberg G, Imakaev M, Abdennur N, Mirny LA. Chromatin organization by an interplay of loop extrusion and compartmental segregation. Proc Natl Acad Sci. 2018; 115: E6697–E6706. https://doi.org/10.1073/pnas.1717730115 PMID: 29967174
5. Borgia A, Borgia MB, Bugge K, Kissling VM, Heidarsson PO, Fernandes CB, et al. Extreme disorder in an ultrahigh-affinity protein complex. Nature. 2018; 555: 61–66. https://doi.org/10.1038/nature25762 PMID: 29466338
6. Martin EW, Holehouse AS, Peran I, Farag M, Incicco JJ, Bremer A, et al. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. Science (80-). 2020; 367: 694–699. https://doi.org/10.1126/science.aaw8653 PMID: 32029630
7. Conicella AE, Dignon GL, Zerze GH, Schmidt HB, D'Ordine AM, Kim YC, et al. TDP-43 α-helical structure tunes liquid–liquid phase separation and function. Proc Natl Acad Sci. 2020; 117: 5883–5894. https://doi.org/10.1073/pnas.1912055117 PMID: 32132204
8. Saunders MG, Voth GA. Coarse-Graining Methods for Computational Biology. Annu Rev Biophys. 2013; 42: 73–93. https://doi.org/10.1146/annurev-biophys-083012-130348 PMID: 23451897

9. Kmiecik S, Gront D, Kolinski M, Wieteska L, Dawid AE, Kolinski A. Coarse-Grained Protein Models and Their Applications. Chem Rev. 2016; 116: 7898–7936. https://doi.org/10.1021/acs.chemrev.6b00163 PMID: 27333362
10. Gopal SM, Mukherjee S, Cheng Y-M, Feig M. PRIMO/PRIMONA: A coarse-grained model for proteins and nucleic acids that preserves near-atomistic accuracy. Proteins Struct Funct Bioinforma. 2010; 78: 1266–1281. https://doi.org/10.1002/prot.22645 PMID: 19967787
11. Sterpone F, Melchionna S, Tuffery P, Pasquali S, Mousseau N, Cragnolini T, et al. The OPEP protein model: from single molecules, amyloid formation, crowding and hydrodynamics to DNA/RNA systems. Chem Soc Rev. 2014; 43: 4871–4893. https://doi.org/10.1039/c4cs00048j PMID: 24759934
12. Davtyan A, Schafer NP, Zheng W, Clementi C, Wolynes PG, Papoian GA. AWSEM-MD: Protein Structure Prediction Using Coarse-Grained Physical Potentials and Bioinformatically Based Local Structure Biasing. J Phys Chem B. 2012; 116: 8494–8503. https://doi.org/10.1021/jp212541y PMID: 22545654
13. Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, de Vries AH. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. J Phys Chem B. 2007; 111: 7812–7824. https://doi.org/10.1021/jp071097f PMID: 17569554
14. Clementi C, Nymeyer H, Onuchic JN. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? an investigation for small globular proteins. J Mol Biol. 2000; 298: 937–953. https://doi.org/10.1006/jmbi.2000.3693 PMID: 10801360
15. Go N. Theoretical Studies of Protein Folding. Annu Rev Biophys Bioeng. 1983; 12: 183–210. https://doi.org/10.1146/annurev.bb.12.060183.001151 PMID: 6347038
16. Takada S. Gō model revisited. Biophys Physicobiology. 2019; 16: 248–255. https://doi.org/10.2142/biophysico.16.0_248 PMID: 31984178
17. Arnarez C, Uusitalo JJ, Masman MF, Ingólfsson HI, de Jong DH, Melo MN, et al. Dry Martini, a Coarse-Grained Force Field for Lipid Membrane Simulations with Implicit Solvent. J Chem Theory Comput. 2015; 11: 260–275. https://doi.org/10.1021/ct500477k PMID: 26574224
18. Ermak DL, McCammon JA. Brownian dynamics with hydrodynamic interactions. J Chem Phys. 1978; 69: 1352–1360. https://doi.org/10.1063/1.436761
19. Ando T, Skolnick J. Sliding of Proteins Non-specifically Bound to DNA: Brownian Dynamics Studies with Coarse-Grained Protein and DNA Models. Clore GM, editor. PLoS Comput Biol. 2014; 10: e1003990. https://doi.org/10.1371/journal.pcbi.1003990 PMID: 25504215
20. Sterpone F, Derreumaux P, Melchionna S. Protein Simulations in Fluids: Coupling the OPEP Coarse-Grained Force Field with Hydrodynamics. J Chem Theory Comput. 2015; 11: 1843–1853. https://doi.org/10.1021/ct501015h PMID: 26574390
21. Brandner A F., Timr S, Melchionna S, Derreumaux P, Baaden M, Sterpone F. Modelling lipid systems in fluid with Lattice Boltzmann Molecular Dynamics simulations and hydrodynamics. Sci Rep. 2019; 9: 16450. https://doi.org/10.1038/s41598-019-52760-y PMID: 31712588
22. 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: 3026–3035. https://doi.org/10.1021/acs.accounts.5b00338 PMID: 26575522
23. Nguyen HT, Hori N, Thirumalai D. Theory and simulations for RNA folding in mixtures of monovalent and divalent cations. Proc Natl Acad Sci. 2019; 116: 21022–21030. https://doi.org/10.1073/pnas.1911632116 PMID: 31570624
24. Hinckley DM, Lequieu JP, de Pablo JJ. Coarse-grained modeling of DNA oligomer hybridization: Length, sequence, and salt effects. J Chem Phys. 2014; 141: 035102. https://doi.org/10.1063/1.4886336 PMID: 25053341
25. Kubo S, Niina T, Takada S. Molecular dynamics simulation of proton-transfer coupled rotations in ATP synthase FO motor. Sci Rep. 2020; 10: 8225. https://doi.org/10.1038/s41598-020-65004-1 PMID: 32427921
26. Li W, Wang W, Takada S. Energy landscape views for interplays among folding, binding, and allostery of calmodulin domains. Proc Natl Acad Sci. 2014; 111: 10550–10555. https://doi.org/10.1073/pnas.1402768111 PMID: 25002491
27. Hori N, Takada S. Coarse-Grained Structure-Based Model for RNA-Protein Complexes Developed by Fluctuation Matching. J Chem Theory Comput. 2012; 8: 3384–3394. https://doi.org/10.1021/ct300361j PMID: 26605744
28. Freeman GS, Hinckley DM, Lequieu JP, Whitmer JK, de Pablo JJ. Coarse-grained modeling of DNA curvature. J Chem Phys. 2014; 141: 165103. https://doi.org/10.1063/1.4897649 PMID: 25362344

29. Tan C, Terakawa T, Takada S. Dynamic Coupling among Protein Binding, Sliding, and DNA Bending Revealed by Molecular Dynamics. J Am Chem Soc. 2016; 138: 8512–8522. https://doi.org/10.1021/jacs.6b03729 PMID: 27309278

30. Brandani GB, Niina T, Tan C, Takada S. DNA sliding in nucleosomes via twist defect propagation revealed by molecular simulations. Nucleic Acids Res. 2018; 46: 2788–2801. https://doi.org/10.1093/nar/gky158 PMID: 29506273

31. Lequieu J, Schwartz DC, de Pablo JJ. In silico evidence for sequence-dependent nucleosome sliding. Proc Natl Acad Sci. 2017; 114: E9197–E9205. https://doi.org/10.1073/pnas.1705685114 PMID: 29078285

32. Tan C, Takada S. Dynamic and Structural Modeling of the Specificity in Protein–DNA Interactions Guided by Binding Assay and Structure Data. J Chem Theory Comput. 2018; 14: 3877–3889. https://doi.org/10.1021/acs.jctc.8b00299 PMID: 29806939

33. Kapcha LH, Rossky PJ. A Simple Atomic-Level Hydrophobicity Scale Reveals Protein Interfacial Structure. J Mol Biol. 2014; 426: 484–498. https://doi.org/10.1016/j.jmb.2013.09.039 PMID: 24120937

34. Kim YC, Hummer G. Coarse-grained Models for Simulations of Multiprotein Complexes: Application to Ubiquitin Binding. J Mol Biol. 2008; 375: 1416–1433. https://doi.org/10.1016/j.jmb.2007.11.063 PMID: 18083189

35. Dignon GL, Zheng W, Kim YC, Best RB, Mittal J. Sequence determinants of protein phase behavior from a coarse-grained model. Ofran Y, editor. PLOS Comput Biol. 2018; 14: e1005941. https://doi.org/10.1371/journal.pcbi.1005941 PMID: 29364893

36. Dannenhoffer-Lafage T, Best RB. A Data-Driven Hydrophobicity Scale for Predicting Liquid–Liquid Phase Separation of Proteins. J Phys Chem B. 2021; 125: 4046–4056. https://doi.org/10.1021/acs.jpcb.0c11479 PMID: 33876938

37. Regy RM, Dignon GL, Zheng W, Kim YC, Mittal J. Sequence dependent phase separation of protein-polynucleotide mixtures elucidated using molecular simulations. Nucleic Acids Res. 2020; 48: 12593–12603. https://doi.org/10.1093/nar/gkaa1099 PMID: 33264400

38. Plimpton S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. J Comput Phys. 1995; 117: 1–19. https://doi.org/10.1006/jcph.1995.1039

39. Hinckley DM, Freeman GS, Whitmer JK, de Pablo JJ. An experimentally-informed coarse-grained 3-site-per-nucleotide model of DNA: Structure, thermodynamics, and dynamics of hybridization. J Chem Phys. 2013; 139: 144903. https://doi.org/10.1063/1.4822042 PMID: 24116642

40. Noel JK, Levi M, Raghunathan M, Lammert H, Hayes RL, Onuchic JN, et al. SMOG 2: A Versatile Software Package for Generating Structure-Based Models. Prlic A, editor. PLOS Comput Biol. 2016; 12: e1004794. https://doi.org/10.1371/journal.pcbi.1004794 PMID: 26963394

41. Pronk S, Páll S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics. 2013; 29: 845–854. https://doi.org/10.1093/bioinformatics/btt055 PMID: 23407358

42. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, et al. Scalable molecular dynamics with NAMD. J Comput Chem. 2005; 26: 1781–1802. https://doi.org/10.1002/jcc.20289 PMID: 16222654

43. Lu W, Bueno C, Schafer NP, Moller J, Jin S, Chen X, et al. OpenAWSEM with Open3SPN2: A fast, flexible, and accessible framework for large-scale coarse-grained biomolecular simulations. Schneidman-Duhovny D, editor. PLOS Comput Biol. 2021; 17: e1008308. https://doi.org/10.1371/journal.pcbi.1008308 PMID: 33577557

44. Kenzaki H, Koga N, Hori N, Kanada R, Li W, Okazaki K, et al. CafeMol: A Coarse-Grained Biomolecular Simulator for Simulating Proteins at Work. J Chem Theory Comput. 2011; 7: 1979–1989. https://doi.org/10.1021/ct2001045 PMID: 26596457

45. Li P, Banjade S, Cheng H-C, Kim S, Chen B, Guo L, et al. Phase transitions in the assembly of multivalent signalling proteins. Nature. 2012; 483: 336–340. https://doi.org/10.1038/nature10879 PMID: 22398450

46. Borcherds W, Bremer A, Borgia MB, Mittag T. How do intrinsically disordered protein regions encode a driving force for liquid–liquid phase separation? Curr Opin Struct Biol. 2021; 67: 41–50. https://doi.org/10.1016/j.sbi.2020.09.004 PMID: 33069007

47. Roden C, Gladfelter AS. RNA contributions to the form and function of biomolecular condensates. Nat Rev Mol Cell Biol. 2021; 22: 183–195. https://doi.org/10.1038/s41580-020-0264-6 PMID: 32632317

48. Turner AL, Watson M, Wilkins OG, Cato L, Travers A, Thomas JO, et al. Highly disordered histone H1−DNA model complexes and their condensates. Proc Natl Acad Sci. 2018; 115: 11964–11969. https://doi.org/10.1073/pnas.1805943115 PMID: 30301810

49. Banani SF, Lee HO, Hyman AA, Rosen MK. Biomolecular condensates: organizers of cellular biochemistry. Nat Rev Mol Cell Biol. 2017; 18: 285–298. https://doi.org/10.1038/nrm.2017.7 PMID: 28225081

50. Jung J, Mori T, Kobayashi C, Matsunaga Y, Yoda T, Feig M, et al. GENESIS: a hybrid-parallel and multi-scale molecular dynamics simulator with enhanced sampling algorithms for biomolecular and cellular simulations. Wiley Interdiscip Rev Comput Mol Sci. 2015; 5: 310–323. https://doi.org/10.1002/wcms.1220 PMID: 26753008

51. Kobayashi C, Jung J, Matsunaga Y, Mori T, Ando T, Tamura K, et al. GENESIS 1.1: A hybrid-parallel molecular dynamics simulator with enhanced sampling algorithms on multiple computational platforms. J Comput Chem. 2017; 38: 2193–2206. https://doi.org/10.1002/jcc.24874 PMID: 28718930

52. Bezanson J, Edelman A, Karpinski S, Shah VB. Julia: A Fresh Approach to Numerical Computing. SIAM Rev. 2017; 59: 65–98. https://doi.org/10.1137/141000671

53. Terakawa T, Takada S. Multiscale Ensemble Modeling of Intrinsically Disordered Proteins: p53 N-Terminal Domain. Biophys J. 2011; 101: 1450–1458. https://doi.org/10.1016/j.bpj.2011.08.003 PMID: 21943426

54. Tan C, Jung J, Kobayashi C, Sugita Y. A singularity-free torsion angle potential for coarse-grained molecular dynamics simulations. J Chem Phys. 2020; 153: 044110. https://doi.org/10.1063/5.0013089 PMID: 32752657

55. Catenaccio A, Daruich Y, Magallanes C. Temperature dependence of the permittivity of water. Chem Phys Lett. 2003; 367: 669–671. https://doi.org/10.1016/S0009-2614(02)01735-9

56. Stogryn A. Equations for Calculating the Dielectric Constant of Saline Water. IEEE Trans Microw Theory Tech. 1971; 19: 733–736. https://doi.org/10.1109/TMTT.1971.1127617

57. Ashbaugh HS, Hatch HW. Natively Unfolded Protein Stability as a Coil-to-Globule Transition in Charge/Hydropathy Space. J Am Chem Soc. 2008; 130: 9536–9542. https://doi.org/10.1021/ja802124e PMID: 18576630

58. Miyazawa S, Jernigan RL. Residue–Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading. J Mol Biol. 1996; 256: 623–644. https://doi.org/10.1006/jmbi.1996.0114 PMID: 8604144

59. Chu J-W, Voth GA. Coarse-Grained Modeling of the Actin Filament Derived from Atomistic-Scale Simulations. Biophys J. 2006; 90: 1572–1582. https://doi.org/10.1529/biophysj.105.073924 PMID: 16361345

60. 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: 711–721. https://doi.org/10.1021/ct4007162 PMID: 26580048

61. Niina T, Brandani GB, Tan C, Takada S. Sequence-dependent nucleosome sliding in rotation-coupled and uncoupled modes revealed by molecular simulations. Panchenko ARR, editor. PLOS Comput Biol. 2017; 13: e1005880. https://doi.org/10.1371/journal.pcbi.1005880 PMID: 29194442

62. Tan C, Takada S. Nucleosome allostery in pioneer transcription factor binding. Proc Natl Acad Sci. 2020; 117: 20586–20596. https://doi.org/10.1073/pnas.2005500117 PMID: 32778600

63. Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. J Mol Graph. 1996; 14: 33–38. https://doi.org/10.1016/0263-7855(96)00018-5 PMID: 8744570

64. Schrödinger L. The PyMOL Molecular Graphics System, Version 1.8. 2015 Nov.

65. Lu X-J, Olson WK. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. Nat Protoc. 2008; 3: 1213–1227. https://doi.org/10.1038/nprot.2008.104 PMID: 18600227

66. Adams PD, Afonine P V., Baskaran K, Berman HM, Berrisford J, Bricogne G, et al. Announcing mandatory submission of PDBx/mmCIF format files for crystallographic depositions to the Protein Data Bank (PDB). Acta Crystallogr Sect D Struct Biol. 2019; 75: 451–454. https://doi.org/10.1107/S2059798319004522 PMID: 30988261

67. Jung J, Mori T, Sugita Y. Midpoint cell method for hybrid (MPI+OpenMP) parallelization of molecular dynamics simulations. J Comput Chem. 2014; 35: 1064–1072. https://doi.org/10.1002/jcc.23591 PMID: 24659253

68. Karanicolas J, Brooks CL. The origins of asymmetry in the folding transition states of protein L and protein G. Protein Sci. 2009; 11: 2351–2361. https://doi.org/10.1110/ps.0205402 PMID: 12237457

69. Kobayashi C, Matsunaga Y, Koike R, Ota M, Sugita Y. Domain Motion Enhanced (DoME) Model for Efficient Conformational Sampling of Multidomain Proteins. J Phys Chem B. 2015; 119: 14584–14593. https://doi.org/10.1021/acs.jpcb.5b07668 PMID: 26536148

70. Best RB, Chen Y-G, Hummer G. Slow Protein Conformational Dynamics from Multiple Experimental Structures: The Helix/Sheet Transition of Arc Repressor. Structure. 2005; 13: 1755–1763. https://doi.org/10.1016/j.str.2005.08.009 PMID: 16338404

71. Westover KD, Bushnell DA, Kornberg RD. Structural Basis of Transcription. Cell. 2004; 119: 481–489. https://doi.org/10.1016/j.cell.2004.10.016 PMID: 15537538

72. Grant RA, Filman DJ, Finkel SE, Kolter R, Hogle JM. The crystal structure of Dps, a ferritin homolog that binds and protects DNA. Nat Struct Biol. 1998; 5: 294–303. https://doi.org/10.1038/nsb0498-294 PMID: 9546221

73. Davey CA, Sargent DF, Luger K, Maeder AW, Richmond TJ. Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9Å Resolution. J Mol Biol. 2002; 319: 1097–1113. https://doi.org/10.1016/S0022-2836(02)00386-8 PMID: 12079350

74. Sekido R, Lovell-Badge R. Sex determination involves synergistic action of SRY and SF1 on a specific Sox9 enhancer. Nature. 2008; 453: 930–934. https://doi.org/10.1038/nature06944 PMID: 18454134

75. Harley VR, Lovell-Badge R, Goodfellow PN. Definition of a consensus DNA binding site for SRY. Nucleic Acids Res. 1994; 22: 1500–1501. https://doi.org/10.1093/nar/22.8.1500 PMID: 8190643

76. Murphy EC, Zhurkin VB, Louis JM, Cornilescu G, Clore GM. Structural Basis for SRY-dependent 46-X,Y Sex Reversal: Modulation of DNA Bending by a Naturally Occurring Point Mutation. J Mol Biol. 2001; 312: 481–499. https://doi.org/10.1006/jmbi.2001.4977 PMID: 11563911

77. Pontiggia A, Rimini R, Harley VR, Goodfellow PN, Lovell-Badge R, Bianchi ME. Sex-reversing mutations affect the architecture of SRY-DNA complexes. EMBO J. 1994; 13: 6115–6124. https://doi.org/10.1002/j.1460-2075.1994.tb06958.x PMID: 7813448

78. Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2019; 48: D87–D92.. https://doi.org/10.1093/nar/gkz1001 PMID: 31701148

79. Patel A, Lee HO, Jawerth L, Maharana S, Jahnel M, Hein MY, et al. A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. Cell. 2015; 162: 1066–1077. https://doi.org/10.1016/j.cell.2015.07.047 PMID: 26317470

80. Regy RM, Thompson J, Kim YC, Mittal J. Improved coarse-grained model for studying sequence dependent phase separation of disordered proteins. Protein Sci. 2021; 30: 1371–1379. https://doi.org/10.1002/pro.4094 PMID: 33934416

81. Yuan S, Wang J, Zhu D, Wang N, Gao Q, Chen W, et al. Cryo-EM structure of a herpesvirus capsid at 3.1 Å. Science (80-). 2018; 360: eaao7283. https://doi.org/10.1126/science.aao7283 PMID: 29622627

82. Öztürk MA, De M, Cojocaru V, Wade RC. Chromatosome Structure and Dynamics from Molecular Simulations. Annu Rev Phys Chem. 2020; 71: 101–119. https://doi.org/10.1146/annurev-physchem-071119-040043 PMID: 32017651

83. Wedemann G, Langowski J. Computer Simulation of the 30-Nanometer Chromatin Fiber. Biophys J. 2002; 82: 2847–2859. https://doi.org/10.1016/S0006-3495(02)75627-0 PMID: 12023209

84. Buckle A, Brackley CA, Boyle S, Marenduzzo D, Gilbert N. Polymer Simulations of Heteromorphic Chromatin Predict the 3D Folding of Complex Genomic Loci. Mol Cell. 2018; 72: 786–797.e11. https://doi.org/10.1016/j.molcel.2018.09.016 PMID: 30344096

85. Takada S, Brandani GB, Tan C. Nucleosomes as allosteric scaffolds for genetic regulation. Curr Opin Struct Biol. 2020; 62: 93–101. https://doi.org/10.1016/j.sbi.2019.11.013 PMID: 31901887

86. Ugarte La Torre D, Takada S. Coarse-grained implicit solvent lipid force field with a compatible resolution to the Cα protein representation. J Chem Phys. 2020; 153: 205101. https://doi.org/10.1063/5.0026342 PMID: 33261497

87. Terakawa T, Kenzaki H, Takada S. p53 Searches on DNA by Rotation-Uncoupled Sliding at C-Terminal Tails and Restricted Hopping of Core Domains. J Am Chem Soc. 2012; 134: 14555–14562. https://doi.org/10.1021/ja305369u PMID: 22880817

88. Dai L, Xu Y, Du Z, Su X, Yu J. Revealing atomic-scale molecular diffusion of a plant-transcription factor WRKY domain protein along DNA. Proc Natl Acad Sci. 2021; 118: e2102621118. https://doi.org/10.1073/pnas.2102621118 PMID: 34074787

89. Grime JMA, Voth GA. Highly Scalable and Memory Efficient Ultra-Coarse-Grained Molecular Dynamics Simulations. J Chem Theory Comput. 2014; 10: 423–431. https://doi.org/10.1021/ct400727q PMID: 26579921

90. Garaizar A, Espinosa JR. Salt dependent phase behavior of intrinsically disordered proteins from a coarse-grained model with explicit water and ions. J Chem Phys. 2021; 155: 125103. https://doi.org/10.1063/5.0062687 PMID: 34598583

91. Sugita Y, Okamoto Y. Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett. 1999; 314: 141–151. https://doi.org/10.1016/S0009-2614(99)01123-9

92. Sugita Y, Kitao A, Okamoto Y. Multidimensional replica-exchange method for free-energy calculations. J Chem Phys. 2000; 113: 6042–6051. https://doi.org/10.1063/1.1308516

93. Miao Y, Feher VA, McCammon JA. Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. J Chem Theory Comput. 2015; 11: 3584–3595. https://doi.org/10.1021/acs.jctc.5b00436 PMID: 26300708

94. Kamiya M, Sugita Y. Flexible selection of the solute region in replica exchange with solute tempering: Application to protein-folding simulations. J Chem Phys. 2018; 149: 072304. https://doi.org/10.1063/1.5016222 PMID: 30134668

95. Fernandez-Leiro R, Scheres SHW. Unravelling biological macromolecules with cryo-electron microscopy. Nature. 2016; 537: 339–346. https://doi.org/10.1038/nature19948 PMID: 27629640

96. Hylton RK, Swulius MT. Challenges and triumphs in cryo-electron tomography. iScience. 2021; 24: 102959. https://doi.org/10.1016/j.isci.2021.102959 PMID: 34466785

97. Uchihashi T, Ganser C. Recent advances in bioimaging with high-speed atomic force microscopy. Biophys Rev. 2020; 12: 363–369. https://doi.org/10.1007/s12551-020-00670-z PMID: 32172451

98. Luchinat E, Banci L. In-Cell NMR in Human Cells: Direct Protein Expression Allows Structural Studies of Protein Folding and Maturation. Acc Chem Res. 2018; 51: 1550–1557. https://doi.org/10.1021/acs.accounts.8b00147 PMID: 29869502

99. Johnson GT, Autin L, Al-Alusi M, Goodsell DS, Sanner MF, Olson AJ. cellPACK: a virtual mesoscope to model and visualize structural systems biology. Nat Methods. 2015; 12: 85–91. https://doi.org/10.1038/nmeth.3204 PMID: 25437435

100. Jewett AI, Stelter D, Lambert J, Saladi SM, Roscioni OM, Ricci M, et al. Moltemplate: A Tool for Coarse-Grained Modeling of Complex Biological Matter and Soft Condensed Matter Physics. J Mol Biol. 2021; 433: 166841. https://doi.org/10.1016/j.jmb.2021.166841 PMID: 33539886

101. Martínez L, Andrade R, Birgin EG, Martínez JM. PACKMOL: A package for building initial configurations for molecular dynamics simulations. J Comput Chem. 2009; 30: 2157–2164. https://doi.org/10.1002/jcc.21224 PMID: 19229944

102. Seo S, Shinoda W. SPICA Force Field for Lipid Membranes: Domain Formation Induced by Cholesterol. J Chem Theory Comput. 2019; 15: 762–774. https://doi.org/10.1021/acs.jctc.8b00987 PMID: 30514078

103. Han W, Wan C-K, Jiang F, Wu Y-D. PACE Force Field for Protein Simulations. 1. Full Parameterization of Version 1 and Verification. J Chem Theory Comput. 2010; 6: 3373–3389. https://doi.org/10.1021/ct1003127 PMID: 26617092

104. Tian C, Kasavajhala K, Belfon KAA, Raguette L, Huang H, Migues AN, et al. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. J Chem Theory Comput. 2020; 16: 528–552. https://doi.org/10.1021/acs.jctc.9b00591 PMID: 31714766

105. Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot BL, et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. Nat Methods. 2017; 14: 71–73. https://doi.org/10.1038/nmeth.4067 PMID: 27819658

106. Oostenbrink C, Villa A, Mark AE, Van Gunsteren WF. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. J Comput Chem. 2004; 25: 1656–1676. https://doi.org/10.1002/jcc.20090 PMID: 15264259

107. Badaczewska-Dawid AE, Kolinski A, Kmiecik S. Computational reconstruction of atomistic protein structures from coarse-grained models. Comput Struct Biotechnol J. 2020; 18: 162–176. https://doi.org/10.1016/j.csbj.2019.12.007 PMID: 31969975

108. Shimizu M, Takada S. Reconstruction of Atomistic Structures from Coarse-Grained Models for Protein–DNA Complexes. J Chem Theory Comput. 2018; 14: 1682–1694. https://doi.org/10.1021/acs.jctc.7b00954 PMID: 29397721