Discrete Molecular Dynamics Simulation of Biomolecules

Feng Ding and Nikolay V. Dokholyan

1 Introduction

Biological molecules are highly dynamic and coexist in multiple conformations in solution [1]. Molecular motions are observed on a broad range of time and length scales using spectroscopy and hydrogen–deuterium exchange experiments [2–5]. The internal motions and resulting conformational changes of these molecules play an essential role in their function. Sampling the structural and dynamic properties of biomolecules remains a challenge due to the large range of time and length scales associated with molecular life. Molecular modeling, especially molecular dynamics simulations of biomolecules and molecular complexes, has played a crucial role in bridging time and length scale gaps and has been pivotal to our understanding of the dynamic aspect of biomolecules [6].

Molecular dynamics (MD) is a computational simulation algorithm, where atoms move according to the laws of classical mechanics. Energetic interactions between atoms are modeled with empirical functions (a "force field") of varying complexities, usually composed of bonded terms representing chain connectivity (bonds, angles, and dihedrals) and nonbonded terms representing van der Waals (VDW) and electrostatic interactions. The dynamic trajectory of the molecular system can be obtained by integrating the equations of motions over a small time step ($\sim$1–2 fs). Analysis of the trajectories from MD simulations can provide great detail concerning the motions of individual particles as a function of time. Thus, these trajectories can be used to address specific questions about properties of a model system that are often inaccessible to experiments. For many aspects of biomolecular

F. Ding (✉) • N.V. Dokholyan
Department of Biochemistry and Biophysics, School of Medicine, University of North Carolina, Chapel Hill, NC, USA
e-mail: fding@unc.edu; dokh@unc.edu

N. Dokholyan (ed.), *Computational Modeling of Biological Systems: From Molecules to Pathways*, Biological and Medical Physics, Biomedical Engineering,
DOI 10.1007/978-1-4614-2146-7_3, © Springer Science+Business Media, LLC 2012

function, it is exactly these details that are of the highest interest and utility. MD simulations allow for the generation of experimentally testable hypotheses, and experiments play an essential role in validating simulation methodology.

The first MD simulation of a fluid system was reported by Alder and Wainwright in 1957 [7]. In a hard sphere fluid system, the authors found evidence of a solid–fluid phase transition that had not been observed in previous Monte Carlo simulations. The subject of hard sphere simulations falls in the general category of discrete potential MD (DMD), which is also called event-driven molecular dynamics, discontinuous molecular dynamics, or discrete molecular dynamics. The DMD methodology is continuously under development for hard-sphere and polymer systems [8–15], and has recently seen an increase in applications for studying biomolecules [16–22]. The development of continuous potentials for MD simulations has facilitated the inclusion of detailed aspects of atomic interactions [23, 24], which is the most common form of MD in current practice. Since the publication of the first MD simulation of bovine pancreatic trypsin inhibitor (BPTI) in 1977 [25], the application of MD simulations to study the structure, dynamics, and function of biomolecules has been increasing steadily. However, the time scales currently accessible in MD simulations are typically 10–100 ns, which restrict their application to many biological processes with large time and length scales (e.g., protein folding occurs in milliseconds to seconds). Even utilizing worldwide computing resources [26] or specialized high-performance computers dedicated to MD simulations (such as Anton [27, 28]), the time scale reached by MD is still in the range of microseconds. Conversely, with the recent development of DMD for biological systems, including the DMD force field [21], all-atom protein models [29–31], and hydrogen bond modeling [18], DMD simulations of realistic biomolecular systems can reach microsecond time scales on personal computers. All-atom DMD simulations have been applied to study protein folding [21, 30], protein design [32, 33], protein structure optimization [34], and post-translational modification of proteins [35]. In this chapter, we focus on DMD simulations of biomolecules. We briefly discuss the DMD algorithm and recent optimization approaches, important developments of DMD methodology for biomolecules, and several applications of all-atom DMD for biomolecules.

2 Discrete Molecular Dynamics

2.1 Algorithm

DMD simulations are based on pairwise interaction potentials that are discontinuous functions of the interatomic distance, r (Fig. 1). We assign for each atom a specific type—A, B, C, . . .—that determines its interaction with other atoms. The interaction potential between two atoms i (type A) and j (type B) is characterized by distances $r^{\mathrm{AB}}{}_{\mathrm{min}} < r^{\mathrm{AB}}{}_{1} < r^{\mathrm{AB}}{}_{2\ldots} < r^{\mathrm{AB}}{}_{k\ldots} < r^{\mathrm{AB}}{}_{\mathrm{max}}$, where $r^{\mathrm{AB}}{}_{\mathrm{min}}$ corresponds to the

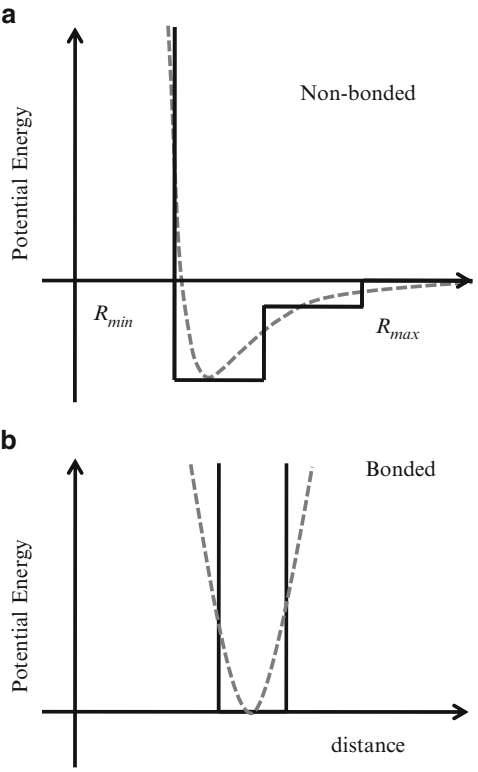

Fig. 1 A schematic diagram of DMD potentials. (**a**) Interaction between nonbonded atom pairs. $R_{\min}$ corresponds to the hard-core distance, $R_{\max}$ corresponds to the interaction range. (**b**) Interaction between bonded atoms. In both cases, *gray dashed lines* correspond to the continuous potential in traditional MD

hardcore collision distance and ${r^{\mathrm{AB}}}_{\max}$ corresponds to the maximal interaction range between the two atoms. If ${r^{\mathrm{AB}}}_k < r_{ij} < {r^{\mathrm{AB}}}_{k+1}$, the pairwise potential energy is assigned as $U_{ij} = {U^{\mathrm{AB}}}_k$. If $r_{ij} < {r^{\mathrm{AB}}}_{\min}$, $U_{ij} = \infty$ so that the two atoms do not come closer than the hard core distance; and if $r_{ij} > {r^{\mathrm{AB}}}_{\max}$, $U_{ij} = 0$ such that two atoms will not interact with each other. If atoms i and j are linked by a bond, the potential energy $U_{ij} = \infty$ when $r_{ij} > {r^{\mathrm{AB}}}_{\max}$. As the result, the two atoms will not escape from each other beyond ${r^{\mathrm{AB}}}_{\max}$, mimicking the bond (Fig. 1b). In DMD simulations, each atom moves with a constant velocity until its distance to another neighboring atom becomes equal to a potential step ${r^{\mathrm{AB}}}_k$, where the potential energy is not continuous. At this moment in time their velocities change instantaneously in accordance with the laws of energy, momentum, and angular momentum conservation. When the kinetic energy of the particles is not sufficient to overcome the potential barrier ${\epsilon^{\mathrm{AB}}}_k = {U^{\mathrm{AB}}}_{k-1} - {U^{\mathrm{AB}}}_k$ (only when the potential change is positive), the atoms undergo a hardcore reflection with no change in potential energy. Each of these events is termed as a collision. At each collision, positions and velocities are updated only for the two colliding atoms, and potential collisions with their neighboring atoms are recomputed. By iterating these calculations, the trajectory of the system is computed as a set of consecutive collision events.

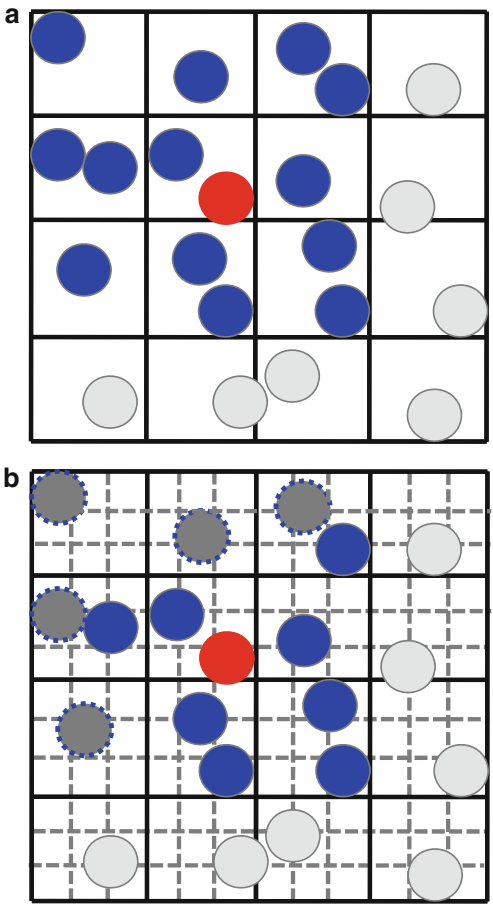

Fig. 2 Grid approach to facilitate the search of neighboring atoms. (**a**) The traditional approach to divide the simulation box into smallest cells, with cell dimension larger than the maximum interaction range. Only the atoms in the neighboring 27 cells (*in blue*) are counted as the neighboring atoms of the atom in *red*. (**b**) The new approach to further divide each cell into a finer grid. By dividing each dimension of the cell by three, the number of neighboring atoms can be greatly reduced (*dark gray spheres*)

In order to efficiently simulate collisions, Rapaport [8] proposed to divide the simulation box into subcells, with the dimension of the cell assigned as the largest interaction range of all the atom pairs and wall-crossing events treated as collisions. As the result, for each atom i, only the collisions between atom i and the atoms in the neighboring $3^3 = 27$ cells are required to be computed for predicting the next collisions of atom i (Fig. 2a). Assuming the average number of atoms in each cell is N_g, the average number of possible collisions to be evaluated for each atom is $27N_g$. To facilitate the evaluation of all possible collisions and prediction of the next collisions, Rapaport [9] proposed a priority tree containing all possible collisions between neighboring atoms ($\sim 27N_gN$), where N is the total number of atoms. The priority tree is sorted according to the collision time with computational complexity $O(\ln(27N_gN))$. As an alternative to this multievent scheduling, Allen and Tildesley [36] proposed a single-event scheduling approach, where only the soonest collision for each atom is stored in a fixed-length binary tree ($\sim N$) with

sorting time $O(\ln(N))$. Smith et al. [15] compared these two scheduling methods and found that in simulations of a polymeric system, the single-event scheduling approach is more efficient than multievent scheduling due to avoiding the insertion and deletion of superfluous potential collisions in the priority tree. Next, we discuss several additional optimization approaches.

2.2 Fine Grid

In DMD, the majority of calculation is the re-evaluation of collision times between a colliding atom and its neighbors. When the dimension of the cell ($l_c \sim \mathrm{IR}_{\max}$, the maximum interaction range) is large compared to the hardcore diameter, as in soft sphere systems (Fig. 1a), the number of atoms in each cell is often more than one. As discussed above, the number of atoms in the neighboring 27 cells is approximately $27\, l_c^3 \rho$, where ρ is the number density. However, assuming l_c approximately equal to the interaction range, the number of atoms inside the interaction range is $\sim(4\pi/3) l_c^3 \rho$ which is much less than $27\, {l_c}^3 \rho$. Therefore, many unnecessary atom pairs are included in the current scheme. We propose to divide each cell into a finer grid with each dimension divided evenly by a number, N_f (e.g., $N_f = 3$ in Fig. 2b). For each cell, we assign an integer address $(C_x,\ C_y,\ C_z)$. If the two cells have the address difference $(\Delta C^x,\ \Delta C^y,\ \Delta C^z)$ and

$$\sum_{d=x,y,z} (\max\{\Delta C^d - 1, 0\} \times l_c^d / N_f)^2 < l_c^2, \tag{1}$$

we consider the two cells as neighbors, and hence the atoms inside the cells are neighbors. Here, ${l_c}^d$ are the cell dimensions. As N_f increases, the number of atoms inside the neighboring cells asymptotically approaches $(4\pi/3){l_c}^3 \rho$, approximately 16% of the original number of neighboring atoms. As the result, the computational efficiency under the new scheme can be increased by as much as 6.4 times. On the other hand, the frequency of cell crossing and the corresponding CPU time spent are correspondingly increasing with this increase in N_f. Therefore, it is possible to find an optimal number of N_f for each type of DMD simulation system. In our all-atom protein model for DMD simulations, we use $N_f = 6$. *We find that in dense-packing cases such as folded proteins, we can improve the simulation efficiency by three to four times by using a finer grid.*

2.3 Reduce the Unnecessary Square Root Calculation

The most expensive calculation in the DMD algorithm is performed after each collision, when the DMD algorithm re-evaluates the collision times between the colliding atoms and their neighboring atoms. Because of the costly square root

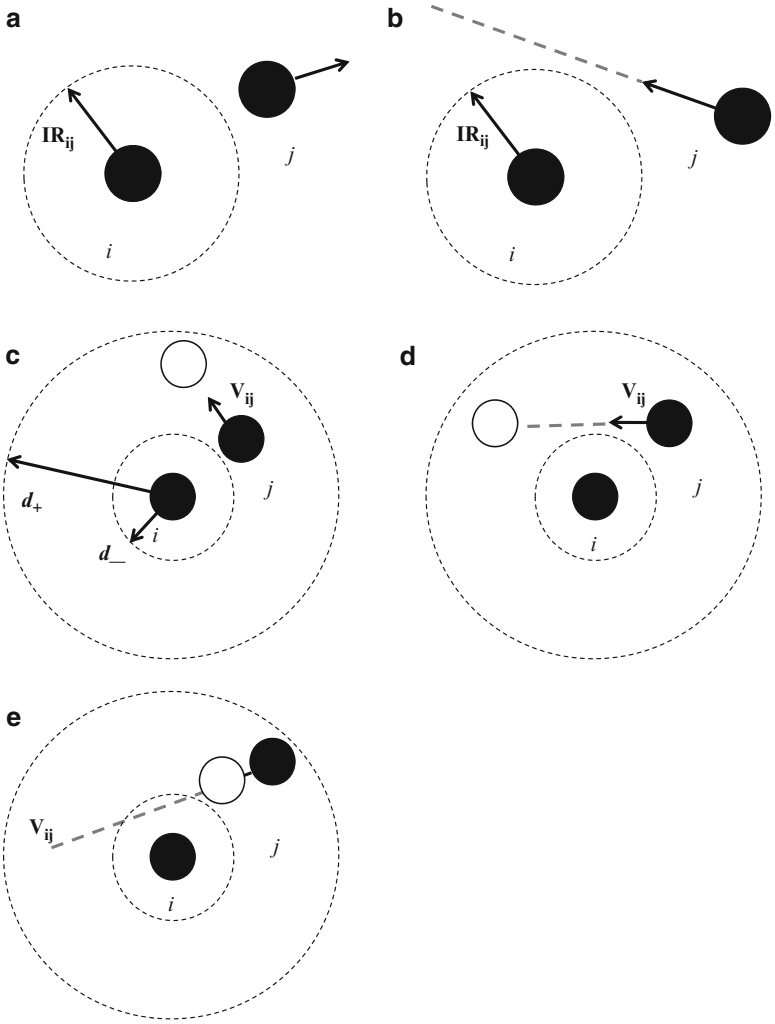

Fig. 3 Cases where the square root calculation to predict the next collisions is not necessary. If (**a**) two noninteracting (beyond the interaction range IR_{ij}) atoms are moving away from each other, and (**b**) two approaching, noninteracting atoms with the minimal distance larger than IR_{ij} (see the *dashed line* in **b**), the square root calculation is *not* necessary, since the operand is negative (collision will not happen). However, even if a collision can happen as in C, D, and E, the collision will not take place if some other event with respect to either i or j happens first. The *open sphere* along the direction of the relative velocity V_{ij} indicates the new position of atom j with respect to atom i

calculation involved in these calculations, it is important to devise a method to reduce the number of unnecessary collision time evaluations. For example, usually under two conditions (Fig. 3a, b), the predicted collision will not happen: (1) when the two atoms are moving away from each other ($\mathbf{R}_{ij} \cdot \mathbf{V}_{ij} > 0$) and the pairwise

distance is larger than the interaction range, IR_{ij} (Fig. 3a) and (2) when the two atoms are approaching the interaction range but the minimum distance is still larger than IR_{ij} $\left(\mathbf{R}_{ij}^2 - (\mathbf{R}_{ij} \cdot \mathbf{V}_{ij})^2/\mathbf{V}_{ij}^2 > IR_{ij}^2\right)$ (Fig. 3b). Here, $\mathbf{V}_{ij}$ is the relative velocity and $\mathbf{R}_{ij}$ is the relative displacement.

We developed a new approach to reduce further unnecessary square root calculations. During the recalculation of potential collisions (see Sect. 2.1), we assume a cutoff time Δt for each atom D. Within such a cutoff time, a collision will always happen to the atom of interest. Therefore, we may simply evaluate the pairwise displacement $\mathbf{R}_{ij}+\mathbf{V}_{ij}\Delta t$, with the pairwise distance R_{ij} within the potential steps $(d_-,\ d_+)$. In the following cases, collision will not occur:

1. Atoms moving away from each other $(\mathbf{R}_{ij} \cdot \mathbf{V}_{ij} > 0)$, but the two atoms do not collide within Δt at d_+, $(\mathbf{R}_{ij} + \mathbf{V}_{ij}\Delta t)^2 < {d_+}^2$ (Fig. 3c)
2. Atoms approaching each other $(\mathbf{R}_{ij} \cdot \mathbf{V}_{ij} < 0)$ with a minimum distance larger than d_- $\left(\mathbf{R}_{ij}^2 - (\mathbf{R}_{ij} \cdot \mathbf{V}_{ij})^2/\mathbf{V}_{ij}^2 > {d_-}^2\right)$, but the two atoms do not collide within Δt at d_+, $\mathbf{R}_{ij} + \mathbf{V}_{ij}\Delta t)^2 < {d_+}^2$ (Fig. 3d)
3. Atoms approaching each other $(\mathbf{R}_{ij} \cdot \mathbf{V}_{ij} < 0)$ with a minimum distance smaller than d_- $\left(\mathbf{R}_{ij}^2 - (\mathbf{R}_{ij} \cdot \mathbf{V}_{ij})^2/\mathbf{V}_{ij}^2 < {d_-}^2\right)$, but the two atoms do not collide within Δt, $[(\mathbf{R}_{ij} + \mathbf{V}_{ij}\Delta t) \cdot \mathbf{V}_{ij}]^2/\mathbf{V}_{ij}^2 > {d_-}^2 - \left[\mathbf{R}_{ij}^2 - (\mathbf{R}_{ij} \cdot \mathbf{V}_{ij})^2/\mathbf{V}_{ij}^2\right]$ (Fig. 3e)

and the collision time can be safely assumed to be infinity. The remaining question is how to define the cutoff time Δt. There are two types of events, the cell crossing and the random collision for the Anderson's thermostat [37], which can be used as the reference events since one of them will always happen if no pairwise collision takes place before these two events. We use the shorter time of these two events to define the cutoff time for each atom. Alternatively, one can dynamically define the cutoff time Δt for a given atom based on the atom's average collision time, $< t_{\text{col}} >$, which can be updated periodically. We set $\Delta t = 4 < t_{\text{col}} >$. *We find that such an optimization can improve the efficiency of simulation by 20–30%.*

2.4 Paul's O(1) Sorting Approach

In DMD, the next collision is obtained by sorting, using either the priority tree in the Rapaport approach (multievent scheduling [9]) or the binary tree in the Allen and Tildesley approach (single-event scheduling [36]). In both cases, the computational complexity is in the order of $O(\ln N)$. Recently, Paul [38] proposed a new sorting approach for DMD with a computational complexity of $O(1)$. In Paul's approach, a fixed length array (N_p) is used to hold the collision times, and the array is head–tail connected for repeated use (Fig. 4). The total time of the array is δt and the time step is $\delta t / N_p$. The pointer (index P_t) corresponds to the "current time" (t_C) in units of $\delta t / N_p$. Each collision at time t is added to the array with respect to the "current time": $[P_t + (t - t_C)/(\delta t / N_p)]\ \%\ N_p$. Each element in the array can hold more than one event since each element corresponds to a time window of $\delta t / N_p$. All the events

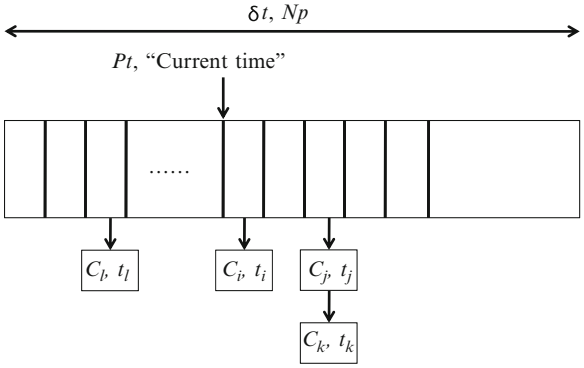

Fig. 4 Schematic for the O(1) sorting approach of collision events by Paul. The linear array of length N_g corresponds to the time interval δt. The array is head–tail connected for repeated usage. A pointer indicates the current time t_c in units of $\delta t/N_g$. Each collision time is inserted into the array with respect the current time: $[P_t + (t - t_c)/(\delta t/N_g)]\%\ N_g$. An element can hold more than one event connected by a simple linked list. The next collision is obtained by advancing the pointer until an occupied array is encountered, and choosing the event with the shortest collision time. By carefully select δt and N_g, the number of events in each element is small and the next collision can be found by a simple bubble sort

within this time window are linked by a simple "linked list." The next collision is obtained by moving the current time pointer forward to the first nonempty element, within which the soonest collision can be found by a simple "bubble sort" approach if the number of events within each element is small. One can define δt and N_p in such a way that the number of events within each element is small. *We find that when the system is large* ($\sim 10^5$ to 10^6 atoms), *sorting takes a significant amount of CPU time (~20% of total computation time). In this case, Paul's sorting approach greatly reduces the percentage of CPU time for sorting from ~20% to only 1–2%.*

Therefore, by carefully selecting the size of the fine grid, reducing the number of unnecessary square-root calculations, and adopting an $O(1)$ sorting algorithm, DMD simulation efficiency can be greatly improved over the traditional approach [9, 36], allowing for the simulation of biomolecular systems with realistic models and force fields. Next, we describe recent developments in the DMD force field and high-resolution molecular models

3 Development of DMD Force Field for Biomolecules

3.1 Hydrogen Bonds

The hydrogen bond interaction is the driving force for secondary structure formation in proteins and nucleotides. In contrast to the model used in continuous MD simulations, hydrogen bond interactions cannot be modeled as dipole–dipole interactions in DMD simulations. Liu and Elliot [39, 40] first proposed a hydrogen

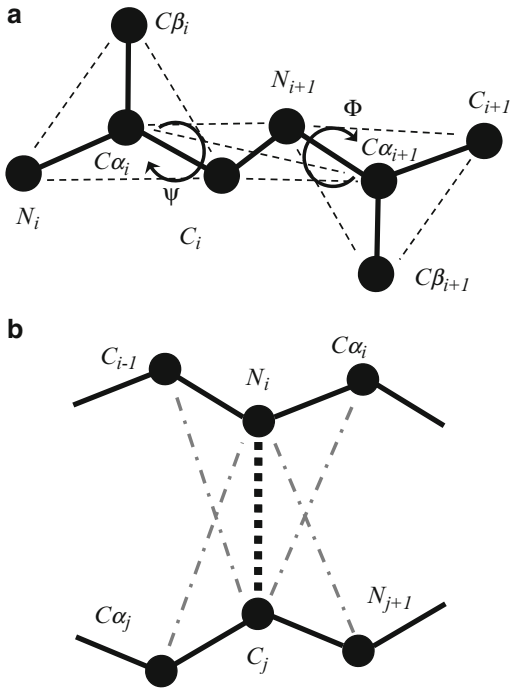

Fig. 5 Model of a hydrogen bond in a simple protein backbone model. (**a**) The four-bead model of a polypeptide. Backbone carbonyl oxygen and amide hydrogen are not explicitly modeled. (**b**) The schematic of a hydrogen bond between carbonyl carbon and amide nitrogen. The *gray dot-dashed lines* correspond to the auxiliary bonds

bond interaction model for DMD, where a hydrogen bond donor (proton) and acceptor (lone electron pair) are explicitly modeled as small attracting atoms positioned inside the hard spheres of the bonding atoms. As the result, the orientation dependence of the hydrogen bond is effectively modeled [39, 40]. However, the explicit modeling of hydrogens and lone electron pairs significantly reduces the computational efficiency of the simulations. Smith and Hall proposed [13] a different approach to model hydrogen bonds in a coarse-grained protein backbone model (alpha carbon C_a, backbone carbonyl carbon C, and nitrogen N; Fig. 5a). Although the backbone carbonyl oxygen O and amide H forming the hydrogen bond are not explicitly modeled, their coordinates can be computed based on the coordinates of existing backbone heavy atoms. A hydrogen bond is formed between N and C when they approach within a certain distance of the hypothetical O and H and are aligned collinearly based on angles of $N-H-O$ and $H-O-C$. When this linear alignment is changed, the hydrogen bond is allowed to dissociate, ignoring the impact of the dissociation energy on the dynamics and thus violating the energy conservation law. To overcome the energy conservation violation problem, we proposed an alternative approach to model the hydrogen bond [18]. The approach is based on a "*reaction*" algorithm in DMD: Two reactant atoms A and B can change their types to A' and B' upon collision at a given reaction interaction range. The total potential energy change ΔE associated with the atom type change is evaluated by summing over all interacting atoms. If the kinetic energy

is sufficient to overcome the potential energy change in the case of $\Delta E > 0$, the reaction takes place. Similarly, the reverse reaction can occur when the two atoms dissociate at the reaction interaction range. The reaction is intrinsically a multibody interaction model.

We explicitly model the hydrogen bond interaction using the reaction algorithm. For example, using the same coarse-grained backbone model as Smith and Hall (Fig. 5a), we assign auxiliary atoms for each hydrogen-bonding atom N and C that correspond to the nearest neighboring atoms along the backbone (Fig. 5b). If two atoms N_i and C_j form a hydrogen bond, we will explicitly assign a hydrogen bond between these two atoms and also assign auxiliary bonds between the auxiliary atoms of the donor and acceptor (gray lines in Fig. 5b). The two atoms then change their type to N_i' and C_j'. The auxiliary bonds will retain the alignment of the hydrogen bond during the simulation. The hydrogen bond and the corresponding auxiliary bonds will dissociate when the two hydrogen-bonded atoms move away from the reaction interaction range with a kinetic energy able to overcome the potential energy change. Upon dissociation, the atoms will revert to their original types. During both hydrogen bond formation and dissociation, the total energy change associated with type change and bond formation and breaking is evaluated. If the two approaching atoms cannot form a hydrogen bond, they will proceed with their regular predicted collision. The DMD potential function for hydrogen and auxiliary bonds can be derived from statistical analysis of the hydrogen bonds in high-resolution protein structures. Using this method, we were able to directly observe in silico a secondary structure transition between alpha helix and beta sheet, in which transition plays a crucial role in disease-associated protein misfolding and aggregation [18].

3.2 All-Atom Protein Model

In previous years, DMD has mainly been associated with coarse-grained modeling. Recently, we have developed an all-atom protein model for use in DMD simulations [21], where all heavy atoms and polar hydrogen atoms are explicitly represented, which is often referred to as the united-atom model. The all-atom model allows for the study of high-resolution conformational dynamics on the atomic level.

In the all-atom protein model (Fig. 6a), bonded interactions are modeled using distance constraints for the covalent bond length, bond angles, and dihedral angles (Fig. 6b). For covalently bonded atom pairs and also the bond angles, the interactions are modeled by a square-well potential (Fig. 1b). Dihedral interactions between atoms i and $i + 3$ are modeled by multistep potential functions [19] of pairwise distance. The set of distance parameters ($d_{\min}$, d_0, d_1, d_2, $d_{\max}$) for these potentials are experimentally determined from distance distributions in a nonredundant database of high-resolution protein structures (Fig. 6b).

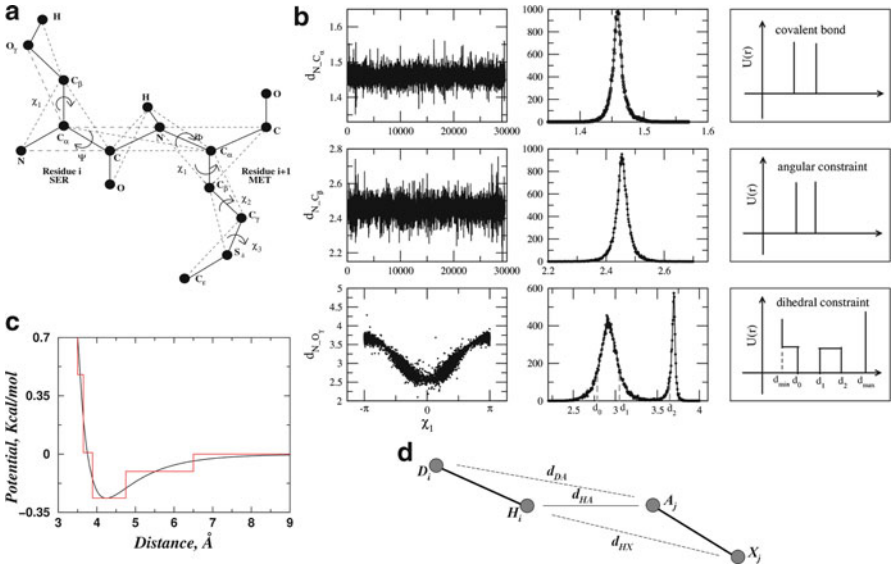

Fig. 6 All-atom protein model. (**a**) Schematic diagram for the all-atom protein model. Only two consecutive residues are shown. The *solid thick lines* represent the covalent and the peptide bonds. The *thin dashed lines* denote the effective bonds that are needed either to fix the bond angles, model the side chain dihedral angles, or to maintain the planarity of the peptide bonds. (**b**) Parameterization of the bonded interactions for representative atom pairs. The first column shows the distribution of the distances in serine between $N-C_\alpha$, $N-C_\beta$, and $N-O_\gamma$, respectively. The second column shows the corresponding histogram for the distribution of each atom pair. The third column shows the resulting constraint potentials schematically. For bonds (e.g., $N-C_\alpha$) and bond angles (e.g., $N-C_\beta$), the left and right boundaries of the constraint potential correspond to $d-\sigma$ and $d+\sigma$, respectively. Here, d is the average length and σ is the standard deviation of the distance distribution. (**c**) Parameterization of nonbonded interactions in all-atom DMD. The continuous *red line* corresponds to the van der Waals and solvation interaction between two carbon atoms. The *black step* function is the discretized potential for DMD. (**d**) A schematic for the hydrogen bonding interaction between hydrogen H_i and acceptor A_j. Atom D_i is the donor and X_j is the heavy atom directly bonded to A_j. Besides the distance between the hydrogen and the acceptor d_{HA}, we also assess the auxiliary distances d_{DA} (distance between atoms D_i and A_j) and d_{HX} (distance between atoms H_i and X_j)

In order to accurately represent nonbonded interactions, we discretized the continuous Medusa force field [34], in which the VDW and solvation interactions are included. VDW interactions use the standard Lennard-Jones potential, and solvation interactions are modeled by the Lazaridis–Karplus (LK) solvation model [41], which is expressed as the sum of pairwise distance-dependent effective solvation energies (EEF1). The discrete potential functions mimic the continuous potential $E_{ij}(d) = E_{ij}^{\mathrm{VDW}}(d) + E_{ij}^{\mathrm{LK}}(d)$ by capturing the attractions and repulsions while using a minimal number of steps (Fig. 6c). By trial and error in test simulations, we adopted the following discretization protocol: (1) we choose an interaction range of

6.5 Å, where the interaction potential attenuates in all atom pairs; (2) we assign a potential step between the distances corresponding to the energy minimum (force is zero) and the interaction action range (force approaching zero), where the force is maximum; (3) we choose the hard sphere distance with VDW–EEF1 energy equal to the minimum energy plus $2k_BT \sim 1.2$ kcal/mol, since thermodynamically the probability to find two atoms within this distance is very low. We choose the next repulsion step with VDW–EEF1 energy equals to the minimum energy plus $k_BT \sim 0.6$ kcal/mol, and the third repulsive step before the energy minimum with the repulsive force ~20 pN, a relative strong force in biology. The energy at each step of the potential is computed as the average of the continuous VDW–EEF1 function, except for the region corresponding to the energetic minimum.

We model the hydrogen bonding interaction using the reaction algorithm, which has been adapted to the all-atom representation (Fig. 6d). All possible interactions between backbone–backbone, backbone–side chain, and side chain–side chain atoms are included. Long-range electrostatic interactions were not included in the previous work [21]. Recently, we have included the electrostatic interaction between formal charges using the Debye–Hückel approximation, which results in better prediction of protein–peptide and protein–ligand interactions (unpublished work).

Other efforts in methods development of all-atom DMD model include those by Borreguero et al. [29], Emperador et al. [31], and Luo et al. [30]. However, these models are either nontransferable with structure-based interaction models [30] and constraints for specific secondary structure [31], or not systematically benchmarked [29].

3.3 Extension of the Force Field for Small Molecules

Recently, we have extended the Medusa force field in order to model small molecule ligands [42] by introducing new atom types and parameterizing the pairwise VDW and EEF1 interactions. We performed a benchmark of the new force field by predicting the binding affinities of a large set of protein–ligand complexes. The correlation coefficient between the computational and experimental affinities is approximately 0.6, which is comparable to other existing computational approaches. Additionally, we developed a flexible ligand docking method using the new force field for both ligand and pose selection [43]. The results of the docking benchmark are comparable to or better than those of other flexible docking programs on the market [43]. Therefore, the extended Medusa force field is useful in modeling small molecules.

We discretized the small molecule Medusa force field extension in order to model small molecules in DMD simulations. Using a similar discretization protocol to that described above for VDW–EEF1, we can readily obtain the nonbonded interactions for small molecules. Since there are an insufficient number of high-resolution small molecule structures to determine the parameters for the bonded terms, we

simply use the accepted average length R_0 and a fixed ratio $\sigma = 0.02$ to model the covalent bond and bond angles, $[R_0(1-\sigma),\ R_0(1+\sigma)]$. For the dihedral angles, we first determine the hybridization of the two central atoms, which determines the symmetry of the dihedral angle: threefold symmetry for sp^3–sp^3, twofold for sp^2–sp^2, and continuous for sp^2–sp^3. For simplicity, we assume a variation of 36° for each ideal angle and compute the multistep potential accordingly, with the energy barrier (ΔE) set as $2k_BT \sim 1.2\,\text{kcal/mol}$ to ensure enough transition between different rotamers ($p \sim \exp(-\Delta E/k_BT)$). Using the extended DMD force field, we are able to perform simulations of the interactions between proteins and small molecules. Since the extended force field also includes nucleotides, we are also able to model both DNA and RNA in DMD.

4 DMD Simulations of Biomolecules

4.1 Folding of Small, Fast-Folding Proteins

Given the vast conformational space available to proteins, the ability to capture protein native states provides an important, milestone benchmark test for all-atom DMD simulations. We performed ab initio folding simulations of six structurally diverse proteins using all-atom DMD with implicit solvation: Trp-cage (20 residues; a mini α/β protein); WW domain (26 residues; the central three strand β-sheet [Gly5-Glu30] of the all-β protein), villin head-piece (35 residues; an all-α protein); GB1 domain (56 residues; an α/β protein); bacterial ribosomal protein L20 (60 residues; an all-α protein); and the engrailed homeodomain (54 residues; an all-α protein). We demonstrate that, using our method, proteins can achieve the native or near-native states in all cases. For three small proteins—Trp-cage, WW domain, and villin headpiece—multiple folding transitions are observed, and the computationally characterized thermodynamics are in qualitative agreement with experiments. For example, our simulation reproduces the apparent two-state folding thermodynamics of WW domain (Fig. 7a), as observed in previous experiments [44, 45]. Additionally, following the folding trajectory in DMD simulations allows us to examine the folding pathway in detail. For the typical folding trajectory of WW (http://dokhlab.unc.edu/research/Abinitio/), we find that the initial folding event features the formation of the first two β-strands. This finding is consistent with experimentally observed kinetics, where the first two strands are more ordered in the folding transition state than the rest of the protein [46]. Such a kinetic folding intermediate was observed only recently in microsecond-long MD simulations with explicit solvent using the state-of-the-art Anton supercomputer, which is optimized specifically for MD simulations [28]. In contrast, our simulations were performed on personal computers, highlighting the computational efficiency of DMD simulations.

Due to the complex nature of protein folding and the fact that the tested proteins are small in size with relatively simple topologies, we do not expect our method to fully resolve the protein folding problem. We do posit that our all-atom DMD

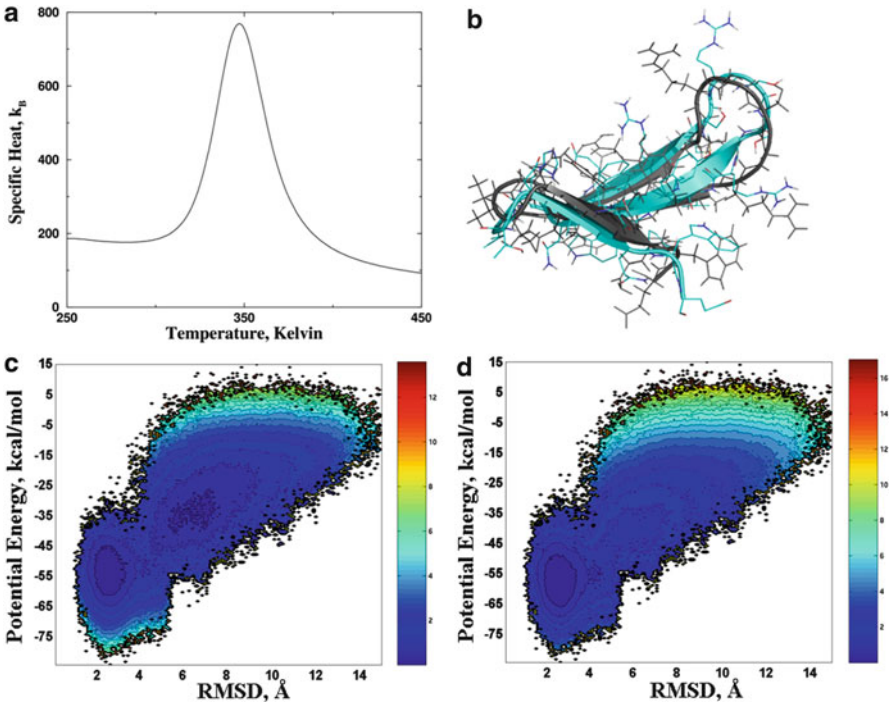

Fig. 7 All-atom DMD simulation of the WW domain. (**a**) Specific heat computed from simulations exhibits a sharp peak at $T \approx 350$ K. (**b**) The alignment between the native state and the representative folded structure in simulations. The contour plot of the 2D-PMF is plotted as the function of potential energy and RMSD at $T = 348$ K (**c**) and $T = 320$ K (**d**)

method can be used for the accurate sampling of conformational space for proteins and protein–protein complexes, which is crucial for protein engineering and the design of protein–protein and protein–ligand interactions.

4.2 Protein–Protein Design

Yin et al. used all-atom DMD simulations in *de novo* protein–protein interface design, where the amino acid sequences of a scaffold protein (human hyperplastic discs protein) were designed to bind a target protein (p21-activated kinase, PAK1). In the design protocol, DMD simulations were utilized for fast conformational sampling, and the RosettaDesign[93] software was used for sequence sampling. The DMD and RosettaDesign steps were performed iteratively in order to attain optimal protein designs that are at global energetic minima in both conformational and sequence spaces. We found that introducing DMD simulations allows for the effective sampling of the protein backbone conformation, which in turn remarkably enriched

the sequence space compatible with the target complex structure. Compared to the initial design obtained without using DMD, the final design had significant backbone (RMSD = 0.82 Å) and rigid-body (RMSD = 3.8 Å) movement. As a result of the backbone movement, 19 out of the 21 interface sites had different amino acids in the final design as compared to the initial design. The final design was experimentally verified to have a binding affinity of ~100 μM to the target protein, and significantly improved solubility as compared to the wild-type human hyperplastic discs protein [32] (Fig. 8).

4.3 Protein Dynamic Coupling and Allosteric Engineering of Kinases

The ability to modulate protein activity in a living cell with temporal control is crucial for our understanding of biological function. We hypothesize that protein dynamics is highly heterogeneous with long range dynamic coupling, and that perturbing distal regions dynamically coupled to the functional site can regulate a protein's function. Such an allosteric regulation is commonly utilized by cell, where the binding of a ligand on one site of the protein can turn the protein's function on or off. We performed DMD simulations of the catalytic domain of focal adhesion kinases (FAK). Based on the simulation trajectory, we found that the catalytically important loop, the G-loop, is strongly coupled to a loop (the insertion loop) that is connected by a β-hairpin (Fig. 9a, b) [33]. We reengineered the insertion loop by inserting a rationally designed unstable FK506-binding protein (iFKBP) domain. This intrinsically metastable domain is stabilized upon the addition of the drug rapamycin (or its analogs) in the presence of FRB. Using the DMD force field extended to include small molecules, we performed DMD simulations in order to study the impact of ligand binding on the conformational dynamics of the catalytic domain of FAK. We showed that the allosteric coupling of FKBP and the catalytic loop allows FAK to be activated via stabilization of FKBP by drug binding (Fig. 9c). *In vivo* experiments using the engineered FAK kinases showed that the protein's kinase function can indeed be regulated by the addition of the ligand. We have demonstrated the transferability of this design approach with other kinases, such as Src and p38 [33]. Therefore, using the allosteric interactions uncovered by DMD, we created a transferrable toolkit for creating regulatable kinases.

5 Conclusion

DMD was originally developed for simple hard sphere systems. In the past, DMD simulations were often associated with coarse-grained molecular systems. With the recent development of a high-resolution DMD force field as well as advances in DMD efficiency, DMD simulations have been applied to study the dynamics of

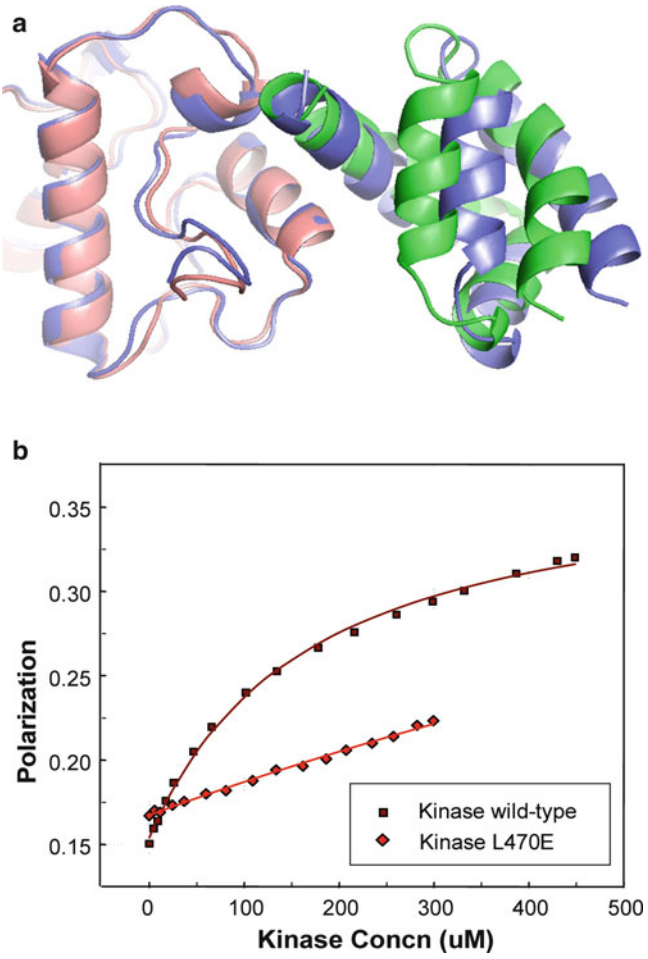

Fig. 8 De novo protein–protein interface design using DMD and Rosetta. (**a**) Starting from the initial structure (*blue*), the DMD assisted design has significant backbone movement in both the scaffold (*green*) and target (*magenta*) proteins. (**b**) The experimental binding assay of the protein–protein complex redesigned using DMD-Rosetta. The redesigned scaffold protein has a binding affinity of $\sim$100 μM with the wild-type target protein. No binding is found in the control experiment with PAK1 mutant L470E, indicating that the actual binding interface is the same as predicted

biological macromolecules. With the continuous development of the methodology, including the parallelization of simulation approaches [47, 48], in the future the DMD engine will be extended to sample the dynamics of ever larger molecules and molecular complexes with even longer time scales. With its ability to efficiently

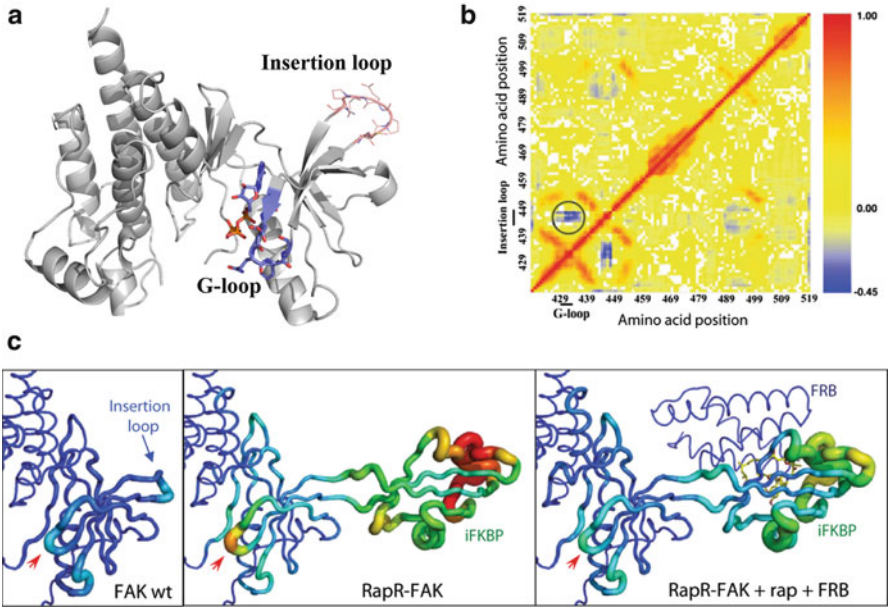

Fig. 9 Mechanism of regulation by iFKBP; Src regulation. (**a**) The portion of the FAK catalytic domain targeted for insertion of iFKBP (*blue*) and the G-loop (*red*). (**b**) Dynamic correlation analysis of the wild-type FAK catalytic domain (*red*, positive correlation; *blue*, negative correlation). The *circled region* indicates strong negative correlation between the movement of the insertion loop and the G-loop. (**c**) Tube representation depicting changes in the dynamics of the N-terminal lobe of the FAK catalytic domain, based on DMD simulations. Warmer colors and thicker backbone correspond to higher root mean squared fluctuation (RMSF) values, reflecting the degree of free movement within the structure. The *red arrows* point to the G-loop

sample the conformational dynamics of complicated systems, DMD simulations will play an important role in our understanding of biology and the effort to combat human diseases.

References

1. Bernado, P., Blackledge, M.: Structural biology: proteins in dynamic equilibrium. Nature **468**, 1046–1048 (2010)
2. Hvidt, A., Nielsen, S.O.: Hydrogen exchange in proteins. Adv. Protein. Chem. **21**, 287–386 (1966)
3. Linderstrom-Lang, K.U.: Deuterium exchange and protein structure. Methuen, London 1958
4. Englander, S.W., Kallenbach, N.R.: Hydrogen exchange and structural dynamics of proteins and nucleic acids. Q Rev. Biophys. **16**, 521–655 (1983)
5. Ishima, R., Torchia, D.A.: Protein dynamics from NMR. Nat. Struct. Biol. **7**, 740–743 (2000)
6. Karplus, M., McCammon, J.A.: Molecular dynamics simulations of biomolecules. Nat. Struct. Biol. **9**, 646–652 (2002)

7. Alder, B.J., Wainwright, T.E.: Phase transition for a hard sphere system. J. Chem. Phys. **27**, 2 (1957)
8. Rapaport, D.C.: Molecular-dynamics simulation of polymer-chains with excluded volume. J. Phys. A Math. Gen. **11**, L213–L217 (1978)
9. Rapaport, D.C.: Event scheduling problem in molecular dynamic simulation. J. Comput. Phys. **34**, 184–201 (1980)
10. Denlinger, M.A., Hall, C.K.: Molecular dynamics simulation results for the pressure of hard-chain fluids. Mol. Phys. **71**, 541–559 (1990)
11. Alejandre, J., Chapela, G.A.: Molecular-dynamics for discontinuous potentials.3. compressibility factors and structure of hard polyatomic fluids. Mol. Phys. **61**, 1119–1130 (1987)
12. Chapela, G.A., Martinezcasas, S.E., Alejandre, J.: Molecular-dynamics for discontinuous potentials.1. General-method and simulation of hard polyatomic-molecules. Mol. Phys. **53**, 139–159 (1984)
13. Smith, A.V., Hall, C.K.: alpha-helix formation: discontinuous molecular dynamics on an intermediate-resolution protein model. Proteins-Struct. Func. Genet. **44**, 344–360 (2001)
14. Smith, S.W., Hall, C.K., Freeman, B.D.: Large-scale molecular-dynamics study of entangled hard-chain fluids. Phys. Rev. Lett. **75**, 1316–1319 (1995)
15. Smith, S.W., Hall, C.K., Freeman, B.D.: Molecular dynamics for polymeric fluids using discontinuous potentials. J. Comput. Phys. **134**, 16–30 (1997)
16. Zhou, Y., Karplus, M.: Folding thermodynamics of a model three-helix-bundle protein. Proc. Natl. Acad. Sci. U. S. A. **94**, 14429–14432 (1997)
17. Zhou, Y., Karplus, M.: Interpreting the folding kinetics of helical proteins. Nature **401**, 400–403 (1999)
18. Ding, F., Borreguero, J.M., Buldyrey, S.V., Stanley, H.E., Dokholyan, N.V.: Mechanism for the alpha-helix to beta-hairpin transition. Proteins **53**, 220–228 (2003)
19. Ding, F., Buldyrev, S.V., Dokholyan, N.V.: Folding Trp-cage to NMR resolution native structure using a coarse-grained protein model. Biophys. J. **88**, 147–155 (2005)
20. Ding, F., Dokholyan, N.V., Buldyrev, S.V., Stanley, H.E., Shakhnovich, E.I.: Molecular dynamics simulation of the SH3 domain aggregation suggests a generic amyloidogenesis mechanism. J. Mol. Biol. **324**, 851–857 (2002)
21. Ding, F., Tsao, D., Nie, H., Dokholyan, N.V.: Ab initio folding of proteins with all-atom discrete molecular dynamics. Structure **16**, 1010–1018 (2008)
22. Peng, S., Ding, F., Urbanc, B., Buldyrev, S.V., Cruz, L., Stanley, H.E., Dokholyan, N.V.: Discrete molecular dynamics simulations of peptide aggregation. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. **69**, 041908 (2004)
23. Rahman, A.: Correlations in motion of atoms in liquid argon. Phys. Rev. A Gen. Phys. **136**, A405-A411 (1964)
24. Stilling, F.h., Rahman, A.: Improved simulation of liquid water by molecular-dynamics. J. Chem. Phys. **60**, 1545–1557 (1974)
25. McCammon, J.A., Gelin, B.R., Karplus, M.: Dynamics of folded proteins. Nature **267**, 585–590 (1977)
26. Shirts, M., Pande, V.S.: COMPUTING: screen savers of the world unite! Science **290**, 1903–1904 (2000)
27. Piana, S., Sarkar, K., Lindorff-Larsen, K., Guo, M., Gruebele, M., Shaw, D.E.: Computational design and experimental testing of the fastest-folding beta-sheet protein. J. Mol. Biol. **405**, 43–48 (2010)
28. Shaw, D.E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R.O., Eastwood, M.P., Bank, J.A., Jumper, J.M., Salmon, J.K., Shan, Y., Wriggers, W.: Atomic-level characterization of the structural dynamics of proteins. Science **330**, 341–346 (2010)
29. Borreguero, J.M., Urbanc, B., Lazo, N.D., Buldyrev, S.V., Teplow, D.B., Stanley, H.E.: Folding events in the 21–30 region of amyloid beta-protein (Abeta) studied in silico. Proc. Natl. Acad. Sci. U. S. A. **102**, 6015–6020 (2005)

30. Luo, Z., Ding, J., Zhou, Y.: Folding mechanisms of individual beta-hairpins in a Go model of Pin1 WW domain by all-atom molecular dynamics simulations. J. Chem. Phys. **128**, **225103** (2008)
31. Emperador, A., Meyer, T., Orozco, M. Protein flexibility from discrete molecular dynamics simulations using quasi-physical potentials. Proteins **78**, 83–94 (2009)
32. Jha, R.K., Leaver-Fay, A., Yin, S., Wu, Y., Butterfoss, G.L., Szyperski, T., Dokholyan, N.V., Kuhlman, B.: Computational design of a PAK1 binding protein. J. Mol. Biol. **400**, 257–270 (2010)
33. Karginov, A.V., Ding, F., Kota, P., Dokholyan, N.V., Hahn, K.M.: Engineered allosteric activation of kinases in living cells. Nat. Biotechnol. **28**, 743–747 (2010)
34. Ding, F., Dokholyan, N.V.: Emergence of protein fold families through rational design. PLoS Comput. Biol. **2**, e85 (2006)
35. Proctor, E.A., Ding, F., Dokholyan, N.V.: Structural and thermodynamic effects of post-translational modifications in mutant and wild type Cu, Zn Superoxide Dismutase. J. Mol. Biol. **408**, 555–567 (2011)
36. Allen, M.P., Tildersley, D.J.:. Computer simulation of liquids. Clarendon Press, New York, (1989)
37. Andersen, H.C.: Molecular-dynamics simulations at constant pressure and-or temperature. J. Chem. Phys. **72**, 2384–2393 (1980)
38. Paul, G.: A complexity O(1) priority queue for event driven molecular dynamics simulations. J. Comput. Phys. **221**, 615–625 (2007)
39. Liu, J.X., Bowman, T.L., Elliott, J.R.: Discontinuous molecular-dynamics simulation of hydrogen-bonding systems. Ind. Eng. Chem. Res. **33**, 957–964 (1994)
40. Liu, J.X., Elliott, J.R.: Screening effects on hydrogen bonding in chain molecular fluids: thermodynamics and kinetics. Ind. Eng. Chem. Res. **35**, 2369–2377 (1996)
41. Lazaridis, T., Karplus, M.: Effective energy function for proteins in solution. Proteins **35**, 133–152 (1999)
42. Yin, S., Biedermannova, L., Vondrasek, J., Dokholyan, N.V.: MedusaScore: an accurate force field-based scoring function for virtual drug screening. J. Chem. Inf. Model **48**, 1656–1662 (2008)
43. Ding, F., Yin, S., Dokholyan, N.V.: Rapid flexible docking using a stochastic rotamer library of ligands. J. Chem. Inf. Model **50**, 1623–1632 (2010)
44. Ferguson, N., Berriman, J., Petrovich, M., Sharpe, T.D., Finch, J.T., Fersht, A.R.: Rapid amyloid fiber formation from the fast-folding WW domain FBP28. Proc. Natl. Acad. Sci. U. S. A. **100**, 9814–9819 (2003)
45. Ferguson, N., Johnson, C.M., Macias, M., Oschkinat, H., Fersht, A.: Ultrafast folding of WW domains without structured aromatic clusters in the denatured state. Proc. Natl. Acad. Sci. U. S. A. **98**, 13002–13007 (2001)
46. Deechongkit, S., Nguyen, H., Powers, E.T., Dawson, P.E., Gruebele, M., Kelly, J.W.: Context-dependent contributions of backbone hydrogen bonding to beta-sheet folding energetics. Nature **430**, 101–105 (2004)
47. Miller, S., Luding, S.: Event-driven molecular dynamics in parallel. J. Comput. Phys. **193**, 10 (2004)
48. Herbordt, M.C., Khan, M.A., Dean, T.: Parallel discrete event simulation of molecular dynamics through event-based decomposition. In Application-specific Systems, Architectures and Processors, 2009. ASAP 2009. 20th IEEE International Conference, Boston, MA (2009)