Mini Review

Applications of Molecular Dynamics Simulation in Structure Prediction of Peptides and Proteins

Hao Geng a,1, Fangfang Chen b,1, Jing Ye b, Fan Jiang a,c,⁎

a Lab of Computational Chemistry and Drug Design, State Key Laboratory of Chemical Oncogenomics, Peking University Shenzhen Graduate School, Shenzhen 518055, China
b Guangdong and Shenzhen Key Laboratory of Male Reproductive Medicine and Genetics, Peking University Shenzhen Hospital, Shenzhen PKU-HKUST Medical Center, Shenzhen 518036, China
c NanoAI Biotech Co., Ltd., Silicon Valley Compound, Longhua District, Shenzhen 518109, China

Article info

Article history:
Received 2 April 2019
Received in revised form 7 July 2019
Accepted 23 July 2019
Available online 26 July 2019

Abstract

Compared with rapid accumulation of protein sequences from high-throughput DNA sequencing, obtaining experimental 3D structures of proteins is still much more difficult, making protein structure prediction (PSP) potentially very useful. Currently, a vast majority of PSP efforts are based on data mining of known sequences, structures and their relationships (informatics-based). However, if closely related template is not available, these methods are usually much less reliable than experiments. They may also be problematic in predicting the structures of naturally occurring or designed peptides. On the other hand, physics-based methods including molecular dynamics (MD) can utilize our understanding of detailed atomic interactions determining biomolecular structures. In this mini-review, we show that all-atom MD can predict structures of cyclic peptides and other peptide foldamers with accuracy similar to experiments. Then, some notable successes in reproducing experimental 3D structures of small proteins through MD simulations (some with replica-exchange) of the folding were summarized. We also describe advancements of MD-based refinement of structure models, and the integration of limited experimental or bioinformatics data into MD-based structure modeling.

© 2019 Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords:
Molecular dynamics
Protein folding
Cyclic peptides
Model refinement
Data-assisted modeling

1. Introduction

Three-dimensional (3D) structures of proteins and their complexes provide invaluable information, not only for understanding the molecular basis of the machinery of life, but also for screening and designing of new drugs for medical applications [1]. Since the first protein 3D structure (of myoglobin) was solved by X-ray crystallography sixty years ago [2,3], enormous efforts have been paid for protein structure determinations [4–9]. However, to obtain high-resolution structure of a protein experimentally is still quite expensive and time-consuming. On the other hand, the cost of obtaining new protein sequences has dramatically decreased due to significant progresses in high-throughput DNA sequencing technology [10,11]. Therefore, there is a huge and increasing gap between the numbers of known structures and sequences. Thus, protein structure prediction (PSP) has become a cost-effective and...
high-throughput way to provide structure information for biological and pharmaceutical researches.

Various methods for PSP have been developed, which can generally be classified into two types: template-based modeling (TBM) and de novo structure prediction. TBM predicts the native structure of a protein target by identifying known protein structure(s) as template(s) [12], based on sequence–sequence (homology modeling) or sequence–structure (threading) alignments. Without suitable template, de novo PSP (template-free modeling) can be used to predict novel protein fold. The most popular and successful ones are usually based on assembly of known structure fragments with potential energy (scoring) functions from mining know protein structures (such as Rosetta [13,14] and QUARK [15]). Also, inter-residue contacts inferred from co-evolutionary signals in sequence homologs can significantly facilitate the de novo PSP [16,17]. The predicted protein models can span a broad range of accuracies and are potentially suitable for different applications [18].

All above popular PSP methods require certain database(s) of sequences and structures, which are thus called knowledge-based. However, the fundamental theory supporting PSP is that the 3D structure of a native-state protein in physiological condition is encoded in its amino acid sequence, as its lowest-free-energy conformation [19,20]. Thus, in theory, PSP can be achieved with only an accurate energy function with a proper conformational search method. Most popular conformational search methods include Monte Carlo (MC) and molecular dynamics (MD). MC approach has been successfully used to study the folding of peptides and proteins, using either atomistic models [21–23] or more coarse-grained models [24–28]. It has also been used to predict protein loop structures [29]. In theory, MC methods can be as accurate as MD methods, but MC may suffer from lower efficiency when using fine-grained representation of the system, especially with large number of explicit solvent molecules. There is also a danger of biasing the sampling when using MC.

MD simulation methods have a long history. The method was originally developed about sixty years ago, for theoretical physicists to study systems of many interacting particles (such as atoms or atom groups) under classical mechanics [30,31]. The now-dominant version of MD was also soon established [32], which numerically solves classical equations of motion according to physical force on each particle. Because small integration time steps of femtoseconds (10^{-15} s) are usually necessary for all-atom simulations, to simulate biologically-relevant event (such as protein folding on > microsecond, 10^{-6} s) by MD requires a huge number of numeric calculations. Over the past 40 years, the time-scales that can be reached by atomistic MD simulation are increasing rapidly, even faster than the Moore’s law [33,34]. Now, MD simulations have become an important and pervasive physics-based method to explore the conformational space of peptides and proteins, which can even fold small proteins (<_80 amino acids) to their native structures [35].

Here we review applications of MD simulation in ab initio structure prediction of peptides and small proteins, refinement of protein structure models, as well as structure modeling assisted by experimental or bioinformatics data. However, we will not review following research areas traditionally not regarded as a PSP problem, although they are somehow related to structure prediction and heavily relying on MD simulations:

1. MD simulation has been extensively used in studying the conformational dynamics of proteins, for a long time [36]. This can be regarded as the prediction of an ensemble of structures for a protein in the native state [37]. However, this type of MD studies rely on known (representative or average) structure of a protein.

2. Some proteins or protein segments do not fold to ordered structures in the native condition [38]. The prediction of structure ensembles of intrinsically disordered proteins is very important, and MD simulation also plays a very important role [39].

Of course, we cannot cover all related works in this mini-review, but intended to give some representative examples about the topic of MD-based structure prediction of peptides and proteins.

2. Methodology Developments for MD Simulation

Since the first protein MD simulation in 1977 [36], with rapidly developing computing hardware and software, significantly longer simulation times and larger simulated systems can be achieved. Supercomputers have been built for biomolecular MD simulations such as protein folding [40]. Software packages for highly efficient MD simulations on parallel computing architectures have also been developed [41,42]. Using special-purpose supercomputer Anton [43], all-atom MD simulation reaching millisecond time-scale has been reported in 2010 [44]. Recently, the most exciting progress is the development of software for routine use of graphic processing unit (GPU) for MD simulations [45–47]. Now, MD simulation is becoming a powerful tool extensively used in studying biomolecular systems. Currently, a few hundred nanoseconds (ns) per day can be routinely achieved on a small protein system in explicit solvent.

In an MD simulation, forces are often calculated using a potential energy function (force field) of the system, which is crucial to the reliability of the MD simulation. A protein force field contains terms for both bonded (for bond lengths, bond angles, and dihedral angles) and non-bonded interactions (van der Waals and electrostatics). It has relatively simple mathematical formula, but usually contain many empirical parameters that determine its accuracy. With increasing computational performance and development of enhanced sampling methods (will be described below), more and more inaccuracies in protein force fields have been revealed [48,49].

These findings have been continuously spurring improvements of classical protein force fields, including AMBER [50–53], CHARMM [54–57], and OPLS-AA [58–60]. Most of the recent updates have been focused on the parameters of backbone and/or side-chain dihedral-angle (torsion) terms to fit ab initio quantum mechanics calculations or experimental (especially NMR) observables, because these parameters are closely related to the conformational behavior of simulated peptide or protein.

Water molecules play a crucial role in driving protein folding [61], and determining the structure and dynamics of protein molecules [62]. Water models like TIP3P [63] and TIP4P [64] developed in the early 1980s have been able to accurately describe the various physicochemical properties of water at room temperature, and are still widely used in MD. In these explicit–solvent simulations, a vast majority of the computational resources are consumed in calculating forces on water molecules. To increase efficiency, methods to treat water solvent implicitly have been developed, mostly based on the Generalized Born (GB) solvation model [64–66]. Sometimes, an energy term related to solvent-accessible surface areas (SA) was used for approximating the non-polar contributions to solvation [66].

Besides the accuracy and reliability of force field (including solvent model), another important factor severely limits the applications of MD: the time scale that can be easily achieved by a conventional MD simulation is usually much shorter than those related to real problems. Thus, enhanced sampling methods have been developed. Some use biased potential to force barrier crossing on pre-defined reaction coordinates (collective variables), such as umbrella sampling [67] and metadynamics [68]. However, these methods can hardly be used in actual structure prediction because the end point of folding should be unknown. Thus, enhanced sampling methods using energy as a reaction coordinate were developed, including biased potential methods such as accelerated molecular dynamics [69], generalized ensemble methods such as replica-exchange molecular dynamics (REMD) [70], and methods combining the two aspects [71]. Currently, the REMD method becomes the most popular one for ab initio folding, partly because it
can utilize current main-stream multi-node parallel computing architectures very efficiently. In addition, replica-exchange method can also be used with MC simulation, which has been used in I-TASSER [72] structure prediction pipeline by Zhang’s group to facilitate the fragment assembly and conformational search.

3. Peptide Structure Prediction

Peptides fill the gap between small-molecule drugs and proteins, and potentially can have both their advantages [73]. Numerous peptides have been found in natural products, and some of them have become drugs with great biological functional diversity. Still, millions of peptides sequences are estimated occurring in prokaryotic genomes [74], plants [75], venom in eukaryotes [76,77] and there are even larger sequence space for designed peptides as chemical tools and drug leads [78–81]. Because the function of a peptides is always related to its unique conformational behavior [82,83], accurate peptide structure prediction (PepSP) would contribute significantly to the peptide-based drug design.

Many attempts have been made for developing PepSP, including evaluations of some common PSP methods (Rosetta, I-TASSER) and specific development of PepSP methods (PepLook, Pep-Fold). However, these methods often cannot consistently predict the experimental structures of peptides. For example, Rosetta cannot reproduce the experimental structure of an α-conotoxin [84]. Peplook, Pep-Fold and I-Tasser predicted a set of 38 cyclic peptides consisting of 5–30 residues with average backbone RMSD (root-mean-square deviation from corresponding experimental structure) values of 3.8 Å, 4.1Å and 2.5 Å, respectively [85]. At the same time, solvent effect has been found very important for conformations of peptides [83,86], which can be treated explicitly in MD simulation, and the configuration entropy can be considered. In addition, a peptide in solution can have distinct conformations using REMD simulations with implicit solvent [91,92]. For 133 mutant were performed [100], and the resulted structural ensembles were validated against experimental NMR chemical shifts and backbone-RMSD to its crystal structure bound to integrin. A peptide structure prediction study of peptoid [98] shows, REMD simulation using generalized AMBER force field (GAFF) [99] with implicit solvation, combing a quantum mechanical refinement, correctly predicted the X-ray crystallographic structures of a N-aryl peptides in a cationic peptide nonamer to the accuracy of 0.2 Å and 1.0 Å RMSDs, respectively.

Recently, MD simulations using explicit solvent models have become more and more widely used. REMD simulation of Vasopressin and Oxytocin (peptide hormones cyclized by disulfide bond) and their mutants were performed [100], and the resulted structural ensembles were validated against experimental NMR chemical shifts and backbone-RMSD to its crystal structure bound to integrin. A peptide structure prediction study of peptoid [98] shows, REMD simulation using generalized AMBER force field (GAFF) [99] with implicit solvation, combing a quantum mechanical refinement, correctly predicted the X-ray crystallographic structures of a N-aryl peptides in a cationic peptide nonamer to the accuracy of 0.2 Å and 1.0 Å RMSDs, respectively.

Recently, MD simulations using explicit solvent models have become more and more widely used. REMD simulation of Vasopressin and Oxytocin (peptide hormones cyclized by disulfide bond) and their mutants were performed [100], and the resulted structural ensembles were validated against experimental NMR chemical shifts and backbone-RMSD to its crystal structure bound to integrin. A peptide structure prediction study of peptoid [98] shows, REMD simulation using generalized AMBER force field (GAFF) [99] with implicit solvation, combing a quantum mechanical refinement, correctly predicted the X-ray crystallographic structures of a N-aryl peptides in a cationic peptide nonamer to the accuracy of 0.2 Å and 1.0 Å RMSDs, respectively.

Recently, MD simulations using explicit solvent models have become more and more widely used. REMD simulation of Vasopressin and Oxytocin (peptide hormones cyclized by disulfide bond) and their mutants were performed [100], and the resulted structural ensembles were validated against experimental NMR chemical shifts and backbone-RMSD to its crystal structure bound to integrin. A peptide structure prediction study of peptoid [98] shows, REMD simulation using generalized AMBER force field (GAFF) [99] with implicit solvation, combing a quantum mechanical refinement, correctly predicted the X-ray crystallographic structures of a N-aryl peptides in a cationic peptide nonamer to the accuracy of 0.2 Å and 1.0 Å RMSDs, respectively.

Recently, MD simulations using explicit solvent models have become more and more widely used. REMD simulation of Vasopressin and Oxytocin (peptide hormones cyclized by disulfide bond) and their mutants were performed [100], and the resulted structural ensembles were validated against experimental NMR chemical shifts and backbone-RMSD to its crystal structure bound to integrin. A peptide structure prediction study of peptoid [98] shows, REMD simulation using generalized AMBER force field (GAFF) [99] with implicit solvation, combing a quantum mechanical refinement, correctly predicted the X-ray crystallographic structures of a N-aryl peptides in a cationic peptide nonamer to the accuracy of 0.2 Å and 1.0 Å RMSDs, respectively.

Recently, MD simulations using explicit solvent models have become more and more widely used. REMD simulation of Vasopressin and Oxytocin (peptide hormones cyclized by disulfide bond) and their mutants were performed [100], and the resulted structural ensembles were validated against experimental NMR chemical shifts and backbone-RMSD to its crystal structure bound to integrin. A peptide structure prediction study of peptoid [98] shows, REMD simulation using generalized AMBER force field (GAFF) [99] with implicit solvation, combing a quantum mechanical refinement, correctly predicted the X-ray crystallographic structures of a N-aryl peptides in a cationic peptide nonamer to the accuracy of 0.2 Å and 1.0 Å RMSDs, respectively.

Recently, MD simulations using explicit solvent models have become more and more widely used. REMD simulation of Vasopressin and Oxytocin (peptide hormones cyclized by disulfide bond) and their mutants were performed [100], and the resulted structural ensembles were validated against experimental NMR chemical shifts and backbone-RMSD to its crystal structure bound to integrin. A peptide structure prediction study of peptoid [98] shows, REMD simulation using generalized AMBER force field (GAFF) [99] with implicit solvation, combing a quantum mechanical refinement, correctly predicted the X-ray crystallographic structures of a N-aryl peptides in a cationic peptide nonamer to the accuracy of 0.2 Å and 1.0 Å RMSDs, respectively.
However, in spite of good conformational sampling and advanced force fields, MD simulation still cannot accurately predict peptide structures in all cases. Very recently, it was shown that the cis/trans isomerization of N-methylated cyclic hexapeptides cannot be reliably predicted [109]. Recently, simulation of guanylin (a 15-residue peptide with four cysteines) showed that the distribution of three disulfide-bond isomers is in qualitative agreement with experiment, but the most stable conformation of the isomer 2 is significantly different from the poorly ordered structure of the truncated peptide [110]. In another study, BE-META simulations were performed on five cyclic isoDGR-containing α/β-peptides using eight widely used force fields with explicit water to reproduce 79 NMR observables [111]. Most of the force fields display good agreement with experimental J(1H,15N,15Cα), but poor agreement was observed for NMR observables directly related to β-amino acids. In addition, accelerated MD simulations of three CPs were performed to predict their conformations [112], and benchmarked against inter-proton distances from NMR experiments and X-ray structures, revealing discrepancy between calculations and experimental observations. These studies are showing that MD-based PepSP can be a very good approach to evaluate force fields.

4. Ab Initio Protein Folding

As early as in 1975, Levitt & Warshel reported the folding simulation of a small protein (bovine pancreatic trypsin inhibitor) from fully extended conformation to RMSD of 5.3 Å, using a very simple model (two particles per residue) [113]. Limited by computational resource, energy minimization and normal-mode thermalisation were used instead of MD. Since then, there were numerous computational studies of folding and tertiary structure prediction of peptides and proteins, mostly using simplified protein models with Monte Carlo methods [114]. At early time, the direct atomistic MD simulation of folding has been impractical.

With increasing time resolution of experiments, people began to realize that some ultrafast folding proteins (with 20 ~ 106 residues) can fold on the time scale of a few μs. In 1999, Takada et al. reported successful folding of an artificial three-helix-bundle protein (54 residues) using overdamped Langevin dynamics (an MD version usually for implicit-solvent coarse-gained simulations) [117]. A simplified model of 3–4 particles per residue allowed trajectories up to 1 microsecond (μs). Soon later, a naturally occurring three-helix-bundle protein was folded from random-coil structures to RMSD of ~ 3 Å within 1 μs, by similar method [118].

Since 2000, successful atomistic folding simulations of some miniproteins in implicit solvent were reported, including: 20-residue three-stranded β-sheet Beta3s [119], 16-residue C-terminal β-hairpin from protein G [120], 23-residue BBA5 with β/αx structure [121,122], 20-residue α-helical Trp-cage [123]. Besides, 35-residue villin HP [124], 10-residue chignolin [125], and 46-residue α-helical fragment B of protein A [126] were successfully folded using explicit solvent. Plain MD or REMD were used in these works, but other enhanced sampling methods can also be used. For example, accelerated MD has been used to fold four fast-folding proteins (chignolin, Trp-cage, villin HP, and WW domain) [127].

Besides, more efficient discrete molecular dynamics (DDM) simulation can be performed based on stepwise potentials with implicit solvent [128]. Using replica-exchange DDM, six small proteins (20 ~ 60 residues) with diverse native structures have been successfully folded [129]. For the smallest three (Trp-cage, villin HP, WW domain), predictions of RMSDs between 2–3 Å can be achieved. However, we do not see wide spread use of DDM in folding studies, possibly due to using less realistic physics model.

One approach to surmount the time scale barrier is to construct Markov state models (MSMs) using many different MD trajectories. It has been successfully used for all-atom ab initio folding of small systems such as the villin HP-35, for which the most populated state has an average RMSD of 2.3 Å [130]. Combining thousands of MD simulations with explicit solvent (each trajectory up to 1 μs, totally 1.3 ms), an atomistic model of the folding of an 80-residue fragment of the λ repressor was obtained to capture dynamics on a 10 millisecond time scale [131]. Using FoldingHome distributed computing platform [132], ~3000 unfolded-initiated trajectories of implicit-solvent MD were generated for 39-residue protein NTL-9 with an experimental folding time of ~1.5 ms, two trajectories reached RMSD < 3.5 Å [133]. An alternative method is Milestoning [134,135], developed by Elber and coworkers. Milestoning samples slow processes by coarse grainning conformational space and performing large numbers of short simulations, yielding kinetics and pathways. However, these methods are too expensive to be applied in practical ab initio structure prediction, although very useful for folding mechanism studies.

With increasing computing power and force field improvements, simultaneous folding simulations of more proteins have been reported. Using special purpose computer Anton, Lindorff-Larsen et al. reported the first successful large-scale folding simulation [35]. Eleven proteins of 10–80 residues were reversibly folded to RMSD < 3.6 Å using the CHARMM22* force field in explicit water. Using REMD with our rSFF1 [136] force field, we also folded all these 11 proteins to RMSD < 3.8 Å, along with the Trp-cage Tc5b and wild-type engrafted homeodomain (EnHD) [137] (Fig. 3). By analyzing continuous trajectories tracking every replica exchange, we also found that REMD can increase the folding rate by about 6 times, through significantly (> 10 times in most cases) increases the diffusion rate on rough energy landscape. Using inexpensive GPUs and implicit solvent model (~1 μs/day can be achieved), Simmerling and coworkers reported successful folding for 16 of 17 proteins (10–92 residues) with a variety of secondary structures and topologies, although the native conformations may not be thermodynamically preferred [138].

Although ultrafast folding proteins are relatively rare, a considerable fraction of protein domains can fold within time scale of milliseconds [139,140]. Using Anton machine, Piana et al. performed eight one-millisecond (1 ms) MD simulations of ubiquitin (a very common 76-residue protein) in explicit water [141]. Starting from the folded structure, spontaneously unfolding to RMSDs of > 20Å and refolding to Cx RMSDs of 0.5 Å was observed. However, no folding events were observed in the two simulations initiated in the unfolded state, which is understandable since the estimated folding time is about 3 ms.

From above, it is clear that atomistic MD simulation of folding can be used for ab initio PSP, but it is still quite expensive and do not have significant advantage compared with fragment-assembly and MC-based methods in real PSP application. Instead, it is often used to study the folding mechanism, which is also scientifically very important.
However, when using coarse-gained force field (such as UNRES [142]), much more efficient conformational sampling can be achieved and larger protein can be folded, even in real *ab initio* PSP, rather than reproducing known structures. Especially, UNRES model with REMD simulation outperformed bioinformatics methods on predicting the correct topology of target T0663 (about 200 residues) in the CASP10 without any input from template or experiment [143,144].

### 5. Protein Structure Refinement

Currently, various bioinformatics-based PSP methods have been developed. Usually, MD-based methods are not necessary, even when a detectable template is not available. However, currently predicted protein models are often much less accurate compared with experiments, limiting their usage in some important applications including structure-based drug design [18], [145]. Thus, structure refinement from low-accuracy models to high-accurate ones is very important, although it is also a very difficult challenge [146]. A crucial motivation for developing physics-based refinement methods is that if the limits of bioinformatics-based methods were being reached, utilizing physics was essential to finally have modeling methods that rival experiment.

In 2000, refinement using explicit-solvent MD simulations that employ the locally enhanced sampling (LES) was applied to low-resolution model of a small disulfide-rich 29-residue protein CMTI-1, and improvement from 3.7 Å to 2.5 Å was reported [147]. Explicit-solvent MD simulation of Rosetta models (of 36-mer HP-36 and 65-mer S15) was also used to generate structures for subsequent ranking using the MM-PBSA free energy function [148], and structures with RMSD < 1.5 Å can be sampled. However, a systematic study (in 2001) on 12 small, single-domain proteins failed to observe successful refinement [149]. On a set of 20 proteins, explicit-solvent MD simulations of most proteins moved initial model structures further away from their native conformations, with performances worse than energy minimization using implicit-solvent potentials [150]. Several studies reveal the importance of proper scoring functions to select more native-like among structures sampled during MD [151–153], because the percentage of more native-like (improved) structures is usually <50% and may decrease as refinement progress.

Restrains can be applied to focus the conformational sampling on the vicinity of the initial model. In 2007, Chen & Brooks performed implicit-solvent REMD simulations with distance restraints between Cα atoms on five protein models, and significantly refinements were observed on three of them [154]. Later, Feig & coworkers [155,156] made significant progress, using restrained explicit-solvent MD and special structure selection and averaging protocol assisted by a statistical potential (DFIRE [157]). Their method ranked the first in the model refinement category of CASP10 [146]. Lee’s group also developed a refinement protocol based on a series of short (5 ns in total) explicit-solvent MD simulations with weak positional and distance restraints [158]. Recently, the combination of restrained MD simulations with accurate force fields is clearly useful and has been adopted by most top-ranking groups in the CASP12 refinement challenge [159]. However, the top-ranking groups are relatively conservative, yielding structures that are quite close to the initial ones.

Shaw and colleagues found that their CHARMM22* force field, which can fold a diverse set of small proteins, may not stabilize the experimental structure of a protein in long-time MD simulations, and good refinement can hardly be achieved [160]. Thus, the success of structure refinement highly relies on the force field accuracy. We recently evaluated the applicability of RSFF1 in protein structure refinement [161]. For 30 single-domain proteins from CASP8-10 refinement targets with diverse structures and a large Cα RMSD coverage of 1-9 Å, MD simulations (380 K) with weak Cα position restraints gave best structures with RMSD reduced by -0.85 Å on average. Using long-time REMD simulations with RSFF1, two homology models (TR614 and TR624) with initial RMSD > 5 Å, can be improved to RMSD < 3 Å. Results from CASP12 indicate that our approach is adventurous, and can provide significantly refined models for some targets (such as TR866, TR894 and TR944) but performs modestly overall.

### 6. Data-assisted Modeling

As described above, applications of MD in both *ab initio* structure prediction and model refinement [162] suffer from two interrelated challenges: insufficient conformational sampling and inaccurate force field. However, limited amounts of structural information can accelerate MD-based structure determination and may also improve simulation accuracy [163]. Meanwhile, experimental techniques have been developed to provide limited (low-resolution, sparse, ambiguous, or uncertain) structure information in a relatively short time and low cost, and bioinformatics techniques have also been developed to predict structure information including secondary structures and residue-residue contacts. Data-assisted modeling has become a sub-category of the CASP experiments since CASP11 (2014) [164]. It has been a long history to use data from known protein structures to guide MD folding simulation and structure prediction. In 1989, Friedrichs and Wolynes proposed the “associative memory Hamiltonians” (AMH) [165], which can learn structure features from a set of
memory proteins. With the incorporation of homologous protein(s) in the memory set and certain information based on secondary structure prediction, near-native structures of a 111-residue protein (rice cytochrome c) can be obtained starting from random structures using simulated annealing MD with a simplified model [166]. With AMH constructed from a database of non-homologous proteins, several α-helical proteins [167] and α/β proteins [168] can be folded to near-native structures (4–8 Å) using MD-based simulated annealing. Further improvements were made by the same group, especially the incorporation of water-mediated interactions [169]. After years of developments, a coarse-grained protein force field called AWSEM (associative memory, water mediated, structure and energy model) was established, which incorporates local structure biasing from fragments with known structures and similar sequences [170]. However, it seems not to be superior to popular fragment-assembly-based de novo PSP methods. Recently, a new scheme called AAWSEM (atomistic associative memory, water mediated, structure and energy model) has been developed [171]. It is an ab initio PSP method that starts from the ground up without using bioinformatics input.

MD-based methods have also been developed to incorporate distance distribution derived from the SAXS experiment as restraints, including the use of coarse-grained force field [172]. Restraints from various paramagnetic NMR experiments (including pseudo-contact shifts, residual dipolar couplings, paramagnetic relaxation enhancement, and cross-correlated relaxation) can be incorporated in computer modeling of protein 3D atomic structures, but substantial challenges remain before widespread use [173].

Chemical crosslinking mass spectrometry (XL-MS) can provide information about residue-pairs in close proximity that can be incorporated into modeling, although the data may be sparses and of low-resolution. Replica-exchange DMD simulations using the Medusa potential with distance restraints from XL-MS experiments gave lowest-energy models of 2.7 Å and 5.4 Å for FK506-binding protein and myoglobin, respectively [174]. MD simulations of a large number of proteins can also be used to find appropriate distance constraint from investigating Lys side-chain motions [175].

Dill and coworkers developed a method called Modeling Employing Limited Data (MELD) that can harness problematic experimental or theoretical data in a Bayesian framework to assist physics-based structure modeling [176,177], which can use a variety of sampling methods obeying detailed balance but implicit-solvent REMD is a good choice. Using loose physical insights (such as proteins have hydrophobic cores and secondary structures), MD simulations of protein folding can be speed up by two ~ five orders of magnitude [178]. Therefore, the MELD method can also be used for ab initio PSP, and structures of three proteins (with 97, 67, 68 residues, respectively) from CASP targets can be predicted blindly with RMSD of 2.8 Å, 1.4 Å, 1.5 Å from native structures, respectively [179].

Besides altering the potential energy function in MD, an iterative screening-after-sampling strategy can be used [180]. By selecting conformations that better fit with the low-resolution data from each cycle of MD simulations, high-quality atomic model can be achieved.

7. Summary

With rapid increasing of computer performance, as well as continuous software and force field developments, MD simulation has been increasingly used in studying biomolecular systems. In principle, it can describe the underline physics of detailed atomic interactions determining a protein structure, and potentially be more accurate than knowledge-based PSP methods. In this mini-review, we showed that all-atom MD simulations can predict structures of cyclic peptides and other peptide-based foldamers with accuracy similar to experiments. Then, we summarized some notable successes in reproducing experimental 3D structures of small proteins through ab initio folding simulations. We also described recent advancements of using MD simulations with state-of-the-art force fields in improving structure models from bioinformatics-based PSP, which is one of the most useful for real-word applications. Finally, some methodology developments and applications of using limited experimental or theoretical data to guide MD-based structure modeling were also introduced. We feel that, in the future, more sophisticated and integrative methods will be developed, including those combining different levels of structure representation (multi-scale MD simulations) and those utilizing the power of machine learning (to take advantage of large amount of data generated by MD) [181].

Declaration of Competing Interest

The authors claim no conflict of interest.

Acknowledgements

We thank the financial supports from the National Natural Science Foundation of China (21573009), and the Shenzhen Science and Technology Innovation Committee (JCYJ20170412150507046, JCYJ20170412151002616).

References

[1] Westbrook JD, Burley SK. How structural biologists and the protein data bank contributed to recent FDA new drug approvals. Structure 2019;27:211–7. https://doi.org/10.1016/j.str.2018.11.007.
[2] Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC. A threedimensional model of the myoglobin molecule obtained by x-ray analysis. Nature 1958;181:662–6. https://doi.org/10.1038/181662a0.
[3] Kendrew JC, Dickerson RE, Strandberg BE, Hart RG, Davies DR, Phillips DC, et al. Structure of Myoglobin: a three-dimensional fourier synthesis at 2 Å. Resol Nat 1960;185:422–7. https://doi.org/10.1038/185422a0.
[4] Wüthrich K. The way to NMR structures of proteins. Nat Struct Biol 2001;8:923–5. https://doi.org/10.1038/nsb1101-923.
[5] Carpenteren EP, Beis K, Cameron AD, Iwata S. Overcoming the challenges of membrane protein crystallography. Curr Opin Struct Biol 2008;18:581–6. https://doi.org/10.1016/j.sbi.2008.07.001.
[6] Shi Y. A glimpse of structural biology through X-ray crystallography. Cell 2014;159:995–1014. https://doi.org/10.1016/j.cell.2014.10.051.
[7] Miao J, Ishikawa T, Robinson IK, Muramane MM. Supplemental Materials for Beyond crystallography: diffactive imaging using coherent x-ray light sources. Science 2015;348:530–5. https://doi.org/10.1126/science.aab1394.
[8] Chen Bai X, McMullan G, SHW Scheres. How cryo-EM is revolutionizing structural biology. Trends Biochem Sci 2015;40:49–57. https://doi.org/10.1016/j.tibs.2014.10.005.
[9] Cheng Y. Membrane protein structural biology in the era of single particle cryo-EM. Curr Opin Struct Biol 2018;52:58–63. https://doi.org/10.1016/j.sbi.2018.08.008.
[10] Reuter JA, Speak DEV, Snyder MP. High-throughput sequencing technologies. Mol Cell 2015;58:586–97. https://doi.org/10.1016/j.molcel.2015.05.004.
[11] Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of nextgeneration sequencing technologies. Nat Rev Genet 2016;17:333–51. https://doi.org/10.1038/nrg.2016.49.
[12] John B, Sali A. Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res 2003;31:3982–92. https://doi.org/10.1093/nar/gkg460.
[13] Bradley P, Misura KMS, Baker D. Toward high-resolution de novo structure prediction for small proteins. Science 2005;309:1868–71. https://doi.org/10.1126/science.1113801.
[14] Rohlf CA, Strauss CEM, Misura KMS, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004;383:86–93. https://doi.org/10.1007/0-89696-6879 (04)383004-0.
[15] Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 2012;80:1715–35. https://doi.org/10.1002/prot.24085.
[16] Marks DS, Colwell LJ, Sheridan R, Hooft TA, Papagni A, Zecchina R, et al. Protein 3D structure computed from evolutionary sequence variation. PLoS One 2011;6: e28766. https://doi.org/10.1371/journal.pone.0028766.
[17] Ovchinikov S, Park H, Varghese N, Huang P-S, Pavlopoulos GA, Kim DE, et al. Protein structure determination using metagenome sequence data. Science 2017;355:294–8. https://doi.org/10.1126/science.aab4043.
[18] Zhang Y. Protein structure prediction: when is it useful? Curr Opin Struct Biol 2009;19:44–53. https://doi.org/10.1016/j.sbi.2009.02.005.
[19] Epstein CJ, Goldberger RE, Anfinson CB. The genetic control of tertiary protein structure: studies with model systems. Cold Spring Harb Symp Quant Biol 2012;28:439–49. https://doi.org/10.1101/sqb.1963.028.01.060.
Piana S, Lindorff-Larsen K, Shaw DE. Atomic-level description of ubiquitin folding.

Zhou H, Zhou Y. Distance-scaled, replica-exchange molecular dynamics. J Am Chem Soc 2014;136:9536–9. https://doi.org/10.1021/ja502735c.

Nguyen H, Maier J, Huang H, Perrone V, Simmerling C. Folding simulations for proteins with diverse topologies are accessible in days with a Physics-Based Force Field and Implicit Solvent. J Am Chem Soc 2014;136:13395–62. https://doi.org/10.1021/ja5032776.

Bogatyrevs NS, Osypov AA, Ivanov DN. KineticDB: a database of protein folding kinetics. Nucleic Acids Res 2009;37:342–6. https://doi.org/10.1093/nar/gkn596.

Lane TJ, Pande VS. Inferring the rate-length law of protein folding. Proc Natl Acad Sci 2013;8:8–12. https://doi.org/10.1073/pnas.0706606.

Piana S, Lindorf-Larsen K, Shaw DE. Atomic-level description of ubiquitin folding. Proc Natl Acad Sci 2013;110:5915–20. https://doi.org/10.1073/pnas.1218321110.

Luo A, Arulkovich P, Czaplewski C, Oldziej S, Pillardy J, Scheraga HA. A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape: application to the UNRES force field. Proc Natl Acad Sci 2002;99:5977–82. https://doi.org/10.1073/pnas.0526739.

Tazi C-H, Bai H, Taylor TJ, Lee B. Assessment of template-free modeling in CASPID and ROLL. Proteins Struct Funct Bioinforma 2014;82:57–83. https://doi.org/10.1002/prot.24470.

He Y, Mozolewska MA, Krupa P, Sieradzan AK, Wiśniewski TK, Livo A, et al. Lessons from application of the UNRES force field to predictions of structures of CASPID targets. Proc Natl Acad Sci USA 2013;110:14936–41. https://doi.org/10.1073/pnas.1313361110.

Cordon D, Kniecik S, Blazczyk M, Ekonmiou D, Kolinski A. Optimization of protein models. Wiley Interdiscip Rev Comput Mol Sci 2012;2:479–93. https://doi.org/10.1002/wcms.1090.

Nugent T, Cuzzetto D, Jones DT. Evaluation of predictions in the CASPID model refinement category. Proteins 2014;82(Suppl. 2):98–111. https://doi.org/10.1002/prot.24336.

Simmerling C, Lee MR, Ortiz AR, Kolinski A, Skolnick J, Kollman PA. Combining MONSTERS and LES/PME to Predict Protein Structure from Amino acid sequence: application to the small Protein-CMT-1. J Am Chem Soc 2000;122:8392–402. https://doi.org/10.1021/ja993119k.

Lee MR, Baker D, Kollman PA. 2.1 and 1.8 Å average Cα RMSD structure predictions on two small Proteins, HP-36 and S15. J Am Chem Soc 2001;123:1040–6. https://doi.org/10.1021/ja003150i.

Lee MR, Tasi J, Baker D, Kollman PA. Molecular dynamics in the endgame of protein structure prediction. J Mol Biol 2001;313:417–30. https://doi.org/10.1006/jmbi.2001.5032.

Chopra G, Summa CM, Levitt M. Solvent dramatically affects protein structure refinement. Proc Natl Acad Sci USA 2008;105:20239–44. https://doi.org/10.1073/pnas.0801811105.

Stumpf-Kane AW, Maksimik K, Lee MS, Feig M. Sampling of near-native protein conformations during protein structure refinement using a coarse-grained model, normal modes, and molecular dynamics simulations. Proteins Struct Funct Bioinforma 2007;70:1345–56. https://doi.org/10.1002/prot.21674.

Zhu J, Fan H, Periele X, Huang B, Mark AE. Refining homology models by combining replica-exchange molecular dynamics and statistical potentials. Proteins Struct Funct Bioinforma 2008;72:1171–88. https://doi.org/10.1002/prot.22005.

Fan H, Periele X, Mark AE. Mimicking the action of folding chaperones by Hamiltonian replica-exchange molecular dynamics simulations: application in the refinement of de novo models. Proteins 2012;80:1744–54. https://doi.org/10.1002/prot.24068.

Chen J, Brooks BL. Can molecular dynamics simulations provide high-resolution refinement of protein structure? Proteins Struct Funct Bioinforma 2007;67:522–30. https://doi.org/10.1002/prot.21345.

Mirjalili V, Feig M. Protein structure refinement through structure selection and averaging from molecular dynamics ensembles. J Chem Theory Comput 2013;9:1294–303. https://doi.org/10.1021/ct300952c.

Mirjalili V, Noyes K, Feig M. Physics-based protein structure refinement through multiple molecular dynamics trajectories and structure averaging. Proteins Struct Funct Bioinforma 2014;82:196–207. https://doi.org/10.1002/prot.24436.

Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2009;18:2714–26. https://doi.org/10.1107/s0021889809006004.

Cheng Q, Journ I, Lee J. A simple and efficient Protein structure refinement method. J Chem Theory Comput 2017;13:5146–62. https://doi.org/10.1021/acs.jctc.7b00470.

Hovan I, Olenikovskis V, Valinica H, Krystytafovych A, Saladino G, Gervasio FL. Assessment of the model refinement category in CASPID. Proteins Struct Funct Bioinforma 2018;86:352–67. https://doi.org/10.1002/prot.25409.