Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Force field development and simulations of intrinsically disordered proteins

Jing Huang[1,2] and Alexander D MacKerell Jr.[1]

Intrinsically disordered proteins (IDPs) play important roles in many physiological processes such as signal transduction and transcriptional regulation. Computer simulations that are based on empirical force fields have been increasingly used to understand the biophysics of disordered proteins. In this review, we focus on recent improvement of protein force fields, including polarizable force fields, concerning their accuracy in modeling intrinsically disordered proteins. Some recent benchmarks and applications of these force fields are also overviewed.

Addresses
[1] Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland, Baltimore, 20 Penn St., Baltimore, MD 21201, USA
[2] Laboratory of Computational Biology, National Heart, Lung and Blood Institute, National Institutes of Health, 5635 Fishers Lane, Rockville, MD 20852, USA

Corresponding author: MacKerell, Alexander D (alex@outerbanks.umaryland.edu)

Current Opinion in Structural Biology 2018, **48**:40–48

This review comes from a themed issue on **Folding and binding**

Edited by **Stefano Gianni** and **Amedeo Caflisch**

http://dx.doi.org/10.1016/j.sbi.2017.10.008

Introduction

The abundance of intrinsically disordered proteins (IDPs), which include proteins with disordered regions, in the human proteome has recently been recognized [1,2]. IDPs are characterized by the lacking of any well-defined three-dimensional tertiary structures in contrast to the common paradigm that a protein functions by folding into a single native structure. Instead, an IDP exists as an ensemble of flexible conformations that interconvert with each other, which often involves transient forming and breaking of secondary structure elements. The primary sequences of IDPs feature an enrichment of polar and charged amino acids, with decreased amounts of non-polar residues that normally drive hydrophobic core formation. The conformational flexibility of IDPs not only allows them to serve as flexible linkers between functional domains, but more importantly allows them to play essential roles in protein-protein interaction network as IDPs can adopt different conformations when binding to different partners.

The central role of IDPs in eukaryotic protein interaction networks makes them involved in many pathological conditions, especially in cancers and neurodegenerative diseases [3]. The advantage of IDPs' structural plasticity for their regulatory roles, such as signal transduction and transcriptional regulation, also makes them occur at a high frequency among tumor-related proteins such as p53 and PTEN [3]. Since IDPs can sample a large variety of conformational states, they are prone to aggregate under certain environments. The assembly and aggregation of IDPs leads to the generation of fibrils, hallmarks of many neurodegenerative diseases. Examples include α-synuclein in Parkinson's disease, the β-amyloid (Aβ) peptide and tau protein in Alzheimer's disease, and polyglutamine (polyQ) in Hungtington's disease. Although the importance of conformational dynamics has been appreciated in the computer-aided drug design (CADD), IDPs represent a very challenging case for therapeutic targeting. Instead of binding to a particular IDP conformation, a ligand needs to modulate the IDP's conformational dynamics and its interactions with binding partners.

Experimental tools to investigate IDP conformational ensembles include small-angle X-ray scattering (SAXS), nuclear magnetic resonance (NMR) and Förster resonance energy transfer (FRET) spectroscopy. However, the observables from these experiments are ensemble averaged over the interconverting conformational states of IDPs [4,5]. Even with single molecule experiments, the number of degrees of freedom for an IDP conformational ensemble still far exceeds the number of available experimental observables. To address such an underdetermined problem, theoretical models need to be introduced to extract detailed structural information from these experiments. These methods can be based on polymer physics such as the Gaussian chain model or more detailed atomistic models such as the computer simulations based on molecular mechanics force fields (FFs).

Protein force fields are empirically developed potential energy functions for polypeptides. Combined with proper sampling methods such as molecular dynamics (MD) or Monte Carlo (MC) simulations, they can be used to generate structural ensembles for any IDP without a posteriori knowledge. The atomistic details obtained

from force field-based simulations can be used to help interpret experimental results, or sometime resolve the conflicts between different experimental measurements [6••]. It is also possible to derive IDP ensembles based on mutual information of force fields and experiments. Possibilities include driving MD simulations with the guide of experimental data [7,8], or post processing force-field generated ensembles to match experimental data in a Bayesian fashion [9–13]. These atomistic models of IDP conformations serve as the starting point for structure-based drug design.

The quality of IDP ensembles, either generated completely in silico or determined jointly by combining computations and experiments, depends critically on the accuracy of underlying computational models. To this end, IDPs represent important benchmark systems for protein FFs, which were originally developed for folded proteins and are continuously under further development. In this article we will review some of the recent progress in force field development and simulations for IDPs.

Improvement of protein force fields for IDP simulations

Protein force fields are by definition the potential energy functions and corresponding parameters to describe the bonded and non-bonded interactions between the particles, typically atoms that define the amino acids, as well as the interactions between polypeptides and water. The ability for a protein force field to model these interactions is in principle transferable between folded protein and IDPs, so that any general improvement of protein FFs, though usually not directly targeting IDPs, often leads to more accurate representation of IDPs. Two types of protein FF improvements are particularly relevant for IDP simulations. The first one is to balance the propensity of the sampling of secondary structures, as the conformational dynamics of IDPs may contain frequent formation and breaking of α-helices and β-sheets. This often involves the refinement of the backbone ϕ, ψ dihedral parameters targeting short peptides that fold into α-helices such as the $(AAQAA)_3$ peptide [14] or β-hairpins (e.g. the GB1 hairpin [15] and chigolin [16]) as model systems. The second one is to improve the modeling of the balance of the protein–water and protein–protein interactions, which often results in the introduction of atom pair-specific Lenard-Jones (L-J) parameters (e.g. NBFIX in CHARMM nomenclature) in protein FFs. Useful target data include quantum mechanical data on water-model compound and model compound-model compound interactions and experimental hydration free energies [17] and, more recently, the osmotic pressures of model compounds, where the model compounds are backbone or side-chain analogs [18,19]. The balance between protein–protein and protein–water interactions is particularly important for IDPs, in which no stable hydrophobic cores are formed to bury non-polar residues. This also highlights the importance of using the correct combination of protein FF and water model in IDP simulations, as it has been shown that the equilibrium between folded and unfolded states can be modified with even a subtle change in the water model used in the simulations [20,21••]. In the remaining part of this section, we will overview recent general improvements in major protein FFs, including the Amber, CHARMM, and OPLS FFs.

Efforts from Best and Hummer to balance the secondary structure propensity for the Amber series of protein force fields led to Amber ff99SB* and ff03* [22], which corrected the bias of underestimating and overestimating the helical content in ff99SB [23] and ff03 [24], respectively. Both ff99SB* and ff03* were developed to be used together with the TIP3P water model [25]. A subsequent refinement of ff03* yielded the ff03w FF to be used with the four-site TIP4P/2005 water model [26], which has been used in a variety of IDP simulations [6••,27–29,30•]. Other Amber protein FF development included ff14SB [31], an improvement over ff99SB with new side chain dihedral parameters and empirical adjustment to the backbone ϕ energy profile. Cerutti *et al.* derived ff14ipq [32], which contains a completely new charge set using the implicitly polarized charge (IPoIQ) model [33]. The bond, angle and L-J parameters in ff14ipq were taken from ff99SB, while torsional parameters were fitted using gas phase quantum mechanics (QM) calculations at the MP2/cc-pVTZ level. A further reparametrization of bonded and non-bonded parameters based on the IPoIQ method yielded ff15ipq [34]. The ff14ipq and ff15ipq FFs were developed to be used with the TIP4P-Ew [35] and the SPC/E_b [36] water models, respectively. The AMBER-FB15 force field [37] was also developed based on ff99SB, with a focus on the optimizing the FF parameters for bonded interactions using QM RI-MP2/aug-cc-pVTZ calculations and the ForceBalance procedure [38].

The CHARMM36 (C36) protein FF [39] was published in 2012 and contains multiple improvements over its predecessor, the CHARMM22/CMAP FF (also known as CHARMM27) [40,41], including refinement of the backbone CMAP potentials and new side-chain dihedral parameters. CMAP is a two-dimensional (2D) ϕ,ψ grid-based energy correction map[42] first introduced in 2002 to improve the treatment of the protein backbone conformations. C36 is able to reproduce a variety of NMR observables for folded proteins [43], shows enhanced cooperativity of helix and hairpin formation [44], and yields high accuracy in protein structure refinement [45]. However, application of the C36 protein FF to simulate several IDPs revealed a potential deficiency that conformational states containing left handed helices were overly populated [30•]. A further refinement of the CMAP potentials was performed to address this issue, which together with the introduction of a NBFIX term for improved modeling of