Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Advances in free-energy-based simulations of protein folding and ligand binding

Alberto Perez[1], Joseph A Morrone[1], Carlos Simmerling[1,2] and Ken A Dill[1,2,3]

Free-energy-based simulations are increasingly providing the narratives about the structures, dynamics and biological mechanisms that constitute the fabric of protein science. Here, we review two recent successes. It is becoming practical: first, to fold small proteins with free-energy methods without knowing substructures and second, to compute ligand–protein binding affinities, not just their binding poses. Over the past 40 years, the timescales that can be simulated by atomistic MD are doubling every 1.3 years — which is faster than Moore's law. Thus, these advances are not simply due to the availability of faster computers. Force fields, solvation models and simulation methodology have kept pace with computing advancements, and are now quite good. At the tip of the spear recently are GPU-based computing, improved fast-solvation methods, continued advances in force fields, and conformational sampling methods that harness external information.

Addresses
[1] Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, NY 11794, United States
[2] Department of Chemistry, Stony Brook University, Stony Brook, NY 11794, United States
[3] Department of Physics and Astronomy, Stony Brook University, Stony Brook, NY 11794, United States

Corresponding author: Dill, Ken A (dill@laufercenter.org)

Current Opinion in Structural Biology 2016, **36**:25–31

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

Edited by **Ruth Nussinov** and **Jayant B Udgaonkar**

http://dx.doi.org/10.1016/j.sbi.2015.12.002

Introduction

Increasingly, our understanding of the properties and actions of proteins depends upon physics-based molecular simulations. It is of interest to model the folding of proteins into their stable structures, the binding affinities and selectivities of ligand–protein and protein–protein assemblies, as well as the solubilities and partitioning of biomolecules. The principled way to predict either static properties, or nanosecond-by-nanosecond and Angstrom-by-Angstrom detailed narratives of these actions is to utilize techniques that sample from the free-energy surface and reflect thermal populations.

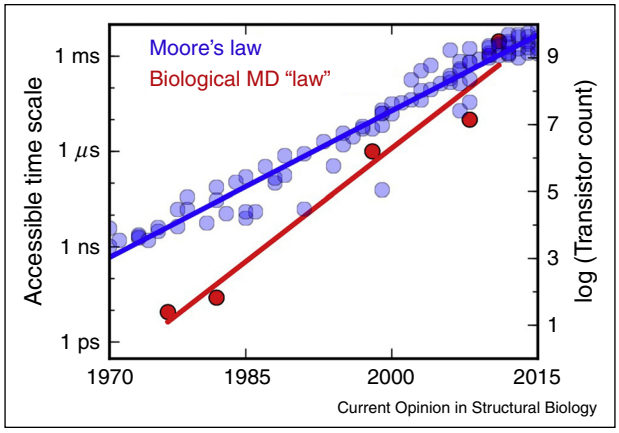
Different approaches can be taken to model biomolecules. Much can be inferred about protein structures by purely comparative modeling using the known structures in the PDB. Often, however equilibrium and kinetic information is desired. These can be inferred within a single framework using force fields and solvent models combined with sampling methods such as molecular dynamics (MD) or Monte Carlo (MC). While direct application of such methods can, in principle, properly sample populations and identify stable states, MD or MC simulations alone are usually not capable of traversing barriers and sampling rare events sufficiently to quantify the free energy differences among states. Molecular simulations can be coupled with specialized techniques for enhancing sampling or extracting information from multiple equilibrium states for this purpose, and these are often referred to as 'free-energy calculations.' Here we take the view that all methods that sample the free-energy surface are 'free-energy-based simulation,' ranging from 'brute force' MD to methods that better facilitate barrier crossings such as replica exchange MD, to techniques such as free-energy perturbation or umbrella sampling.

The obstacles to free-energy modeling have been its high computational cost and some physical inaccuracies in the energetics. However, our current opinion is that free-energy methods have become powerful both rapidly and recently. First, Figure 1 shows that advances in computers, force fields, and methodology over the past 40 years have led to faster-than-Moore's-law increases in the timescales that are now accessible to simulation. Second, here we review progress on two key problems — predicting protein native structures and ligand-binding affinities. These are but two examples and there are many others we do not have the space to cover here.

Atomistic simulations are now folding small proteins and predicting their native structures

There are recent successes in computing the native structures of small proteins by physics-based methods, for diverse folds, without direct inputs from structural databases or the need for structural alignments. DE Shaw Research (DESRES) showed that a single atomistic force field can give folding trajectories for 12 small proteins over long-time simulated trajectories in explicit water on

Figure 1



The accessible time scale for computational biology has grown faster than Moore's law of semiconductors and computing [1,2–6].

Anton, a special purpose supercomputer [6,7]. This was an important milestone in proving the power and transferability of a current force field, and in computation of the folding pathways, some of which have been confirmed experimentally [8,9•]. However, their goal was not so much to obtain accurate populations, as it was to sample multiple folding events, so their simulations were performed near the melting temperature.

A complementary study by Nguyen *et al.* was aimed at detecting native structures starting from fully extended conformations. Nguyen *et al.* used implicit-solvent and required only lab-sized computer clusters [10]. They attributed their successes to the use of GPU-based computing, and to good implicit solvent [11] and force field models [12]. Native structures were found reliably for proteins up to 50 residues, but predictions were not as consistent for longer proteins (up to 92 residues in the study). The main challenge with larger proteins was shown to be limitations of the sampling, not the force field. And, importantly, their implicit-solvent performed as well in this test as previous, more expensive, explicit modeling, implying the adequacy of fast solvation models for some protein modeling previously thought to require more computationally expensive models.

A huge challenge for finding native structures by atomistic MD is that folding times increase sharply with protein size. Experiments show that the folding time increases exponentially with the square root of the number amino acids [13]. Force field limitations and sampling inefficiency make the problem more daunting for simulation. This challenge motivates a need for free-energy methods that can search efficiently to find basins of important states, and that can sample well the populations within those basins.

Recently, an approach called MELD (Modeling Employing Limited Data) has been developed to find and sample important states efficiently by 'melding' structural or heuristic information into MD simulations [14,15]. MELD is able to use combinatorically vague and generic instructives such as: 'make a hydrophobic core' or 'make secondary structures that are consistent with those provided by webservers' or 'make a compact structure.' MELD accelerates conformational searching substantially, while at the same time preserving its critically important ability to give free energies. For example, MELD finds and samples well native structures (better than 4 Å RMSD) for 15 out of 20 small proteins, up to 92-mers, starting from fully extended states [15]. The speed advantage of MELD over brute-force MD is shown in Figure 2a.

Figure 2b shows the implication for computational structure prediction going forward: even with future Moore's-law-like advances, the severity of the exponential search problem means that pure MD will not be folding proteins bigger than 140-mer proteins for another 25 years. But, when MD is combined with external information, as is done in MELD with generic instructives, Figure 2b projects that free-energy methods will give native structures of those sizes within just 4–5 years. This protein size covers a large fraction of single domain proteins, including many biologically relevant ones like ribonuclease (134 residues), lysozyme (129), calmodulin (148) or myoglobin (154).

Free-energy methods are predicting the binding affinities of small ligands to simple proteins

Computational drug discovery is also poised to benefit from advances in free-energy methods. A traditional method for computational drug discovery has been DOCK and related algorithms [16–18]. Docking methods are fast and are often able to find correct binding site and ligand poses, but they rarely give accurate binding affinities.

Although more expensive than docking, physics-based methods have long promised to predict more accurate binding affinities, because of their better potentials and more complete conformational sampling and solvation. In a recent advance, DESRES [19,20] and others [21] have run long MD simulations, and observed the ligand seeking and finding its binding site on the protein. Such work highlights the ability of atomistic simulations to identify stable states given no prior knowledge of the binding site or pose.

Even so, it is not yet possible to sample enough unbinding events to determine rates or affinities by direct MD. Specialized free-energy methods are used to enhance or accelerate sampling, sometimes using 'alchemical'