Biotechnology Advances xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Biotechnology Advances

journal homepage: www.elsevier.com/locate/biotechadv

Research review paper

Enhanced sampling techniques in biomolecular simulations

Vojtěch Spiwok, Zoran Šućur, Petr Hošek

Department of Biochemistry and Microbiology, Institute of Chemical Technology, Prague, Technická 3, Prague 6 166 28, Czech Republic

ARTICLE INFO

Available online xxxx

Keywords:
Molecular dynamic simulation
Alchemistic simulations
Parallel tempering
Metadynamics
Free energy surface
Drug design

ABSTRACT

Biomolecular simulations are routinely used in biochemistry and molecular biology research; however, they often fail to match expectations of their impact on pharmaceutical and biotech industry. This is caused by the fact that a vast amount of computer time is required to simulate short episodes from the life of biomolecules. Several approaches have been developed to overcome this obstacle, including application of massively parallel and special purpose computers or non-conventional hardware. Methodological approaches are represented by coarse-grained models and enhanced sampling techniques. These techniques can show how the studied system behaves in long time-scales on the basis of relatively short simulations. This review presents an overview of new simulation approaches, the theory behind enhanced sampling methods and success stories of their applications with a direct impact on biotechnology or drug design.

Contents

Introduction

Biomolecular simulations, namely their fathers Martin Karplus, Michael Levitt and Arieh Warshel, were awarded Nobel prize in 2013 (Cui and Nussinov, 2014). The first of the trio, Martin Karplus, was involved in the first atomistic biomolecular simulation published in 1977 (McCammon et al., 1977). They simulated 9 ps of life of bovine pancreatic trypsin inhibitor (BPTI). This system was composed of less than 1000 atoms. Since that time we have experienced an enormous

Abbreviations: 5-HT_{2a}, 5-hydroxytryptamine receptor, type 2a; ABMD, adiabatic bias molecular dynamics; AMBER, assisted model building with energy refinement program; BPTI, bovine pancreatic trypsin inhibitor; COX, cyclooxygenase; CPU, central processing unit; CV, collective variable; EGFR, epidermal growth factor receptor; FGF, fibroblast growth factor; FGFR, fibroblast growth factor receptor; GPCR, G-protein coupled receptor; GPU, graphical processing unit; HIV, human immunodeficiency virus; mGluR2, metabotropic glutamate receptor 2; NAMD, nanoscale molecular dynamic program; RMSD, root-mean-square deviation

E-mail address: spiwokv@vscht.cz (V. Spiwok).

http://dx.doi.org/10.1016/j.biotechadv.2014.11.011
0734-9750/

Please cite this article as: Spiwok V, et al, Enhanced sampling techniques in biomolecular simulations, Biotechnol Adv (2014), http://dx.doi.org/10.1016/j.biotechadv.2014.11.011

growth of simulated time scales and system sizes. Recent simulations reach millisecond time scales (Lindorff-Larsen et al., 2011) or tens of millions of atoms (Zhao et al., 2013). This huge progress was possible mainly owing to nearly exponential growth of computer power over the decades.

The question arises whether such increase of computer power is satisfactory to make biomolecular simulation routine techniques in drug discovery or protein and enzyme design. Unfortunately, the answer is no. Today, we can simulate nanoseconds from the life of a solvated average-size protein per day on a single personal computer, microseconds on large parallel computers and milliseconds on a special hardware. In principle, we can predict the native structure of a protein by simulating its folding from the fully unfolded structure (Duan and Kollman, 1998; Lindorff-Larsen et al., 2011; Shaw et al., 2010; Snow et al., 2002). Analogously, it is possible to predict the binding mode of a ligand in a protein just by simulating a box containing the protein, ligand and solvent until the complex is formed (Dror et al., 2011). There are examples of successful simulations of protein folding or ligand binding (Dror et al., 2011; Duan and Kollman, 1998; Shaw et al., 2010; Snow et al., 2002); however, these examples are far from routine in screening large libraries of compounds or protein mutants in drug discovery or protein engineering campaigns. This situation signifies that hardware development must be complemented by design of new sophisticated simulation methods.

Hardware approaches to improve sampling

Before we present methodological methods aimed at sampling improvement, let us introduce hardware approaches. The history of biomolecular simulations has been significantly influenced by the boom of personal computers in the last decades. Computers were expensive scientific instruments in the early times, but they have evolved into today's inexpensive personal object of everyday life. Biomolecular simulations, as well as other areas of scientific computing, benefit from this development. A typical supercomputer from the 1980s contained a single or few powerful central processing units (CPUs). In the 1990s, the trend has changed from building supercomputers with a single strong CPU to combining smaller, often commodity, computers into clusters (Sterling, 2001). The world's most powerful computers today are composed of hundreds of thousands or millions of CPU cores (see www.top500.org for the biannually updated list of 500 world's fastest computers). At the same time, biologists also became prominent customers of high-performance computing centres and terminated the dominance of military industry, oil drillers and other previous users of massive computing.

Another hardware approach to increase computing power for biomolecular simulations is in application of a non-conventional hardware, such as graphical processing units (GPUs) (Pronk et al., 2013; Harvey et al., 2009). Industry of computer gaming hardware developed GPUs with enormous computing power, which can be used to speed up molecular simulations, provided that GPUs can be efficiently handled by the simulation software. Another strategy is to design special-purpose hardware, represented by Anton computer (Dror et al., 2011; Lindorff-Larsen et al., 2011; Shaw et al., 2008, 2010; Snow et al., 2002). This machine contains pieces of hardware tailored for molecular-simulation-specific calculations and is significantly faster than the general purpose computers. Numerous successful projects have also made use of computers of volunteers in distributed computing schemes, such as Folding@home project (Shirts and Pande, 2000).

Methodological approaches to improve sampling

Coarse-graining

An easy way to make simulations faster is to simplify the studied system. This is the basis of coarse grained models of biomolecular systems, which fall into the category of mesoscopic simulations. In coarse-grained simulations, a group of atoms is reduced to a single particle that represents their physico-chemical properties (Tozzini, 2005). The scheme common to many coarse-grained models is to represent four non-hydrogen atoms by one particle.

Simulations of such simplified systems are significantly faster due to two effects: first, the number of particles and especially particle–particle interactions is lower and, second, bonds vibrate with lower frequencies, which makes it is possible to increase simulation time step. Coarse-grained force fields (parameters of covalent and non-covalent interactions) were developed for proteins (de Jong et al., 2013; Shih et al., 2006), membrane components (Marrink et al., 2004; Shih et al., 2006), nucleic acids (Maciejczyk et al., 2010) and carbohydrates (Lopez et al., 2009). These models perform very well for systems where bulk properties dominate over atomic details, such as membranes and membrane–protein interactions (Potocký et al., 2014), formation of membrane nanobodies (Shih et al., 2007), formation of lipid rafts (Risselada and Marrink, 2008) and many others (Marrink and Tieleman, 2013). However, coarse-grained models lack atomic details and therefore they are not suitable for "detailed" phenomena such as binding of a ligand to a protein.

Thermodynamic-based methods

Experimental scientists in drug discovery and biotechnology work with thermodynamic parameters, such as dissociation constants of protein–ligand complexes or free energies stabilizing folded proteins. It is a great challenge to predict values of these parameters by biomolecular simulations. In order to do so, it is necessary to design a structural parameter s, further referred to as a collective variable (CV), that reaches different values in key configurations of the studied system, for example in different conformations of a protein or in different binding poses of a protein–ligand complex. Biomolecular simulation techniques such as molecular dynamic simulation and Monte Carlo method sample the studied system canonically. It is possible to simulate certain molecular system by one of these methods and then analyse the trajectory to calculate evolution of the collective variable s. Next, it is possible to calculate time spent in different configurations with different values of s. This can be simply converted to corresponding probabilities of configurations. The term "canonical sampling" means, that such probabilities are the same as the probabilities in the real macroscopic system, provided that two conditions are fulfilled: first, energies of covalent and non-covalent bonds are accurately modelled and, second, a simulation is sufficiently long. Equilibrium probabilities can be converted to the free energy surface:

$$F(s) = -kT \ln(P(s)), \quad (1)$$

where F is the free energy, P is probability, s is the collective variable (could be replaced by multi-dimensional vector $\mathbf{s}$), k is Boltzmann constant and T is thermodynamic temperature.

Determination of a model free energy surface of protein–ligand association is illustrated in Fig. 1. It is similar for protein folding simulations; by replacing the "complex" and the "dissociated state" by "folded protein" and "unfolded protein", respectively, to get a folding free energy surface. Unfortunately, on personal computers we usually cannot simulate the whole process of binding or folding because its time-scale is too long. It is even more difficult to simulate multiple folding/unfolding or binding/unbinding events, which are necessary to calculate the free energy surface. This is an opportunity for enhanced sampling techniques described below. Some enhanced sampling techniques use Eq. (1) to predict the free energy surface, whereas other methods require different approaches.

Alchemistic methods

One of the goals of the early chemists – alchemists – was to convert one element to another, usually a cheap element to gold. Elements are being converted from one element to another, at least computationally, Q3

Please cite this article as: Spiwok V, et al, Enhanced sampling techniques in biomolecular simulations, Biotechnol Adv (2014), http://dx.doi.org/10.1016/j.biotechadv.2014.11.011