Journal of Molecular Graphics and Modelling 45 (2013) 144–156

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

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

Reaching new levels of realism in modeling biological macromolecules in cellular environments

Michael Feig[d,b,*], Yuji Sugita[a,b,c]

[a] RIKEN Advanced Institute for Computational Science, 7-1-26 Minatojima-minamimachi, Chuo-ku, Kobe, Hyogo 650-0047, Japan
[b] RIKEN Quantitative Biology Center, International Medical Device Alliance (IMDA) 6F, 1-6-5 Minatojima-minamimachi, Chuo-ku, Kobe, Hyogo 650-0047, Japan
[c] Theoretical Molecular Science Laboratory, RIKEN, 2-1 Hirosawa, Wako-shi, Saitama, Wako-shi 351-0198, Japan
[d] Department of Biochemistry & Molecular Biology and Department of Chemistry, Michigan State University, 603 Wilson Road, BCH 218, East Lansing, MI 48824, United States

ARTICLE INFO

Article history:
Received 4 July 2013
Received in revised form 14 August 2013
Accepted 19 August 2013
Available online 28 August 2013

Keywords:
Crowding
Confinement
Reaction–diffusion
Coarse-graining
Implicit solvent
Brownian dynamics

ABSTRACT

An increasing number of studies are aimed at modeling cellular environments in a comprehensive and realistic fashion. A major challenge in these efforts is how to bridge spatial and temporal scales over many orders of magnitude. Furthermore, there are additional challenges in integrating different aspects ranging from questions about biomolecular stability in crowded environments to the description of reactive processes on cellular scales. In this review, recent studies with models of biomolecules in cellular environments at different levels of detail are discussed in terms of their strengths and weaknesses. In particular, atomistic models, implicit representations of cellular environments, coarse-grained and spheroidal models of biomolecules, as well as the inclusion of reactive processes via reaction–diffusion models are described. Furthermore, strategies for integrating the different models into a comprehensive description of cellular environments are discussed.

1. Introduction

Realistic models of cellular environments have long captured the imagination of biologists and physical scientists alike. A major driving force is that such models hold the promise to eventually make the genotype–phenotype connection by predicting a cell's behavior from its genomic DNA and resulting physical constituents. A particular exciting application of virtual cell models would be the rational design of new drugs to predict the effects of potential drug candidates on an entire cell instead of focusing on single targets at a time, which is the current paradigm [1,2]. A whole-cell focused drug design strategy would be able to identify potential unintended side effects early on and is likely to be much more effective for developing treatments of complex diseases such as cancer [3].

A defining step in capturing the fully complexities of biological cells were Goodsell's inspiring renderings of cellular environments based on knowledge available more than twenty years ago [4]. Since then, increasingly realistic simulations of bacterial cytoplasmic models have begun to appear [5–8] that suggest that we are rapidly moving toward comprehensive, physically and biologically accurate cellular models. A major challenge in developing such models is the need to cover a wide range of scales. Atomistic details of molecular processes occur on length scales of 0.1 nm while cellular dimensions are between 300 nm for the smallest bacterial cells and 100 μm for large eukaryotic cells. Biologically relevant time scales range from nanosecond to microsecond time scales for the internal dynamics of individual molecules to time scales of seconds to hours for entire biological processes. While a model that represents an entire cell in full atomistic detail is conceivable, it would be extremely costly to reach nanosecond time scales with today's most advanced computers and impossible to cover seconds for such a system in the foreseeable future. Instead, the development of reduced and multi-scale models is the only viable strategy today for developing physically accurate models that at the same time cover both cellular length scales and time scales of cellular-level biological processes. Such models trade reduced levels of detail for increased computational speed and the key issue is what approximations one can tolerate without comprising the overall level of realism so much that a given model loses its predictive ability. In this review a variety of recently proposed models for biomolecules in cellular environments are discussed and compared in terms of their strengths and weaknesses.

An additional issue is that most past and present research aimed at understanding biological dynamics and function tends to fall

* Corresponding author. Tel.: +1 517 432 7439.
E-mail address: feig@msu.edu (M. Feig).

1093-3263/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmgm.2013.08.017

into two camps, concentrating on either extreme of the spatial and temporal scales. On one hand, structural biology has cultivated a single-molecule, or single-complex, focused view that has led to highly predictive models. They have provided a wealth of insight into the sequence–structure–function paradigm for a large variety of biological macromolecules but with only limited consideration of the cellular environment. On the other hand, the still growing field of systems biology seeks to quantitatively describe macroscopic phenomenological observations at the cellular level with often mathematically complex but physically relatively simple models that rarely go beyond representing biomolecules as single spherical objects. There are intellectual and cultural gaps between these two communities and enhancing mutual interactions is likely to be just as important as devising computationally efficient multiscale modeling strategies. A further aim of this review is to outline how an integration of these different views could be accomplished.

In the following, we will first touch upon the different aspects to be considered when modeling biological processes at different scales in the context of cellular environments. We will then present an overview of recently applied models and discuss strategies for integration into comprehensive models of cellular environments that bridge between scales and in particular between structural and systems biology. The focus here is primarily on the modeling of cellular environments, but we will also briefly mention how to make important connections with experiments as appropriate.

2. Complexity of cellular environments

Depending on the perspective of the reader, the notion of cellular environments is likely to evoke very different ideas. To the biologist, the complexity of metabolic cycles and regulatory networks may first come to mind while structural biologists and biophysicists may think about the effects of crowding and confinement.

If we begin to think about cellular environments from a *chemical perspective*, the simplest cells, bacterial cells, consist in the interior of proteins and nucleic acids at a volume fraction of typically 20–30%, a diverse array of metabolites at different concentrations, various monoatomic ions, and water filling the remaining space [4,9,10]. Lipid membranes form the cell perimeter with a significant fraction of embedded membrane proteins [11]. In addition, cytoskeletal elements may be present to provide mechanical stability, facilitate cell motility, or to aid with developmental processes such as cell division [12]. Eukaryotic cells are complicated further by compartmentalization and a more extensive use of cytoskeletal elements [13].

From a *physical perspective*, the most defining feature of cellular environments is the effect of macromolecular crowding and confinement. Much attention has focused on how crowding and confinement affect protein folding landscapes [14–23] and protein-protein association rates [15,24–27]. The general picture from a combination of experimental, theoretical, and simulation studies is that both crowding and confinement tend to disfavor extended non-native states vs. compact, native states. The main effect at play is volume exclusion, which introduces an entropic penalty for extended states or makes them simply impossible under conditions of rigid confinement [28]. Furthermore, crowding may alter the internal flexibility or functionally relevant conformational dynamics [29–31]. This view is modulated by recent studies that indicate that native state destabilization may be possible when the crowder molecules are proteins as well so that folding and favorable protein–protein interactions can become competing processes [17,18,32,33].

From a *structural perspective*, we are approaching a time when complete structuromes for a given organism will likely become available [34], at least at the level of individual macromolecules. If one combines experimentally resolved structures and structures that can be modeled with reasonable certainty through homology [35], structures are now available for the large majority of gene products for a given organism [34,36]. This is especially true for well-ordered soluble proteins. Membrane proteins are more problematic because much fewer experimental structures are available [37,38], but as that number increases, the fraction of membrane protein structures that can be modeled will also increase [37]. Disordered regions or entire proteins that are intrinsically disordered pose another challenge that is difficult to address experimentally [39], but insight into dynamic ensembles for such systems could be obtained at least in principle using computational tools [40–42].

Nucleic acid structure is another challenge. Certain types of RNA, in particular ribosomal and transfer RNA, are well resolved but much less is known about the detailed structural organization of genomic DNA [43]. It may be ironic that although the structure of short oligomeric DNA was resolved a long time [44] and only shortly after the first protein was resolved in atomic detail [45,46], the highly complex and dynamics structure of genomic DNA with millions of base pairs is now turning out to be one of the last frontiers of structural biology [47–50].

Finally, complete cellular models have to also consider the surrounding biological membranes. Biological membrane should be modeled as mixed systems containing a variety of lipid molecules and a high concentration of membrane proteins. Membranes are highly dynamic and the lipid composition can change the elasticity of biological membranes and varies for the membrane location in a cell. The modeling of membranes is further complicated by the formation of lipid rafts that form subdomains in cellular membranes, which contain cholesterols and sphingomyelins at high concentrations. Although relatively small phospholipid bilayer systems are modeled routinely, realistic models of entire cellular membranes have not yet been established.

In the *biological context*, interactions between different macromolecules are often the key to their biological function but also the source of undesirable aggregation [51]. This results in supramolecular complex formation with varying degrees of stability. There are many examples of well-resolved complexes such as homo-oligomeric structures or multimeric functional complexes like the ribosome or RNA polymerase [52,53]. However, presumably these complexes are highly stable with limited internal dynamics since otherwise successful crystallization would have been unlikely. Other, more dynamic complexes may only be accessible via lower resolution techniques such as electron microscopy in combination with modeling [54–56]. Finally, at the level of whole cells, tomography can determine not just the structures of complexes but their locations within a cellular context [57,58]. This was demonstrated recently for *Mycoplasma pneumoniae* [59].

From a *biochemical perspective*, cells are exquisitely tuned reactors that carry out an enormous variety of chemical reactions in parallel within a very dense environment. Essentially, these reactions serve two main objectives: (1) Continuous generation of energy from varying types of metabolites absorbed from the environment; and (2) Growth and propagation through cell division. The underlying reaction networks have been the subject of many studies [60]. We are now reaching the point of being able to generate complete reaction networks where all products and reactants are accounted for in metabolic cycles and where, at the same time, all of the enzymes involved in catalyzing the various reactions are mapped to gene products for a given organism. An example was recently given for the case of *Mycoplasma genitalium*, a bacterium with the smallest known genome [61]. As functional annotations of genomes and the application of proteomics and metabolonomics studies continuously increase, it can be expected that similar complete network reconstructions will soon be reported for many other