Journal of Molecular Graphics and Modelling 58 (2015) 1–9

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

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

Complete atomistic model of a bacterial cytoplasm for integrating physics, biochemistry, and systems biology

Michael Feig[a,b,c,*], Ryuhei Harada[d,c], Takaharu Mori[c,e], Isseki Yu[e], Koichi Takahashi[f,g], Yuji Sugita[d,c,e]

[a] Department of Biochemistry & Molecular Biology, Michigan State University, East Lansing, MI 48824, United States
[b] Department of Chemistry, Michigan State University, East Lansing, MI 48824, United States
[c] Quantitative Biology Center, RIKEN, International Medical Device Alliance (IMDA) 6F, 1-6-5 Minatojima-minamimachi, Chuo-ku, Kobe, Hyogo 650-0047, Japan
[d] Advanced Institute for Computational Science, RIKEN, 7-1-26 Minatojima-minamimachi, Chuo-ku, Kobe, Hyogo 650-0047, Japan
[e] Theoretical Molecular Science Laboratory and iTHES, RIKEN, 2-1 Hirosawa, Wako-shi, Saitama 351-0198, Japan
[f] Quantitative Biology Center, RIKEN, Laboratory for Biochemical Simulation, Suita, Osaka 565-0874, Japan
[g] Institute for Advanced Biosciences, Keio University, Fujisawa 252-8520, Japan

ARTICLE INFO

Article history:
Accepted 22 February 2015
Available online 28 February 2015

Keywords:
Mycoplasma genitalium
Metabolic reaction network
Protein structure prediction
Crowding

ABSTRACT

A model for the cytoplasm of *Mycoplasma genitalium* is presented that integrates data from a variety of sources into a physically and biochemically consistent model. Based on gene annotations, core genes expected to be present in the cytoplasm were determined and a metabolic reaction network was reconstructed. The set of cytoplasmic genes and metabolites from the predicted reactions were assembled into a comprehensive atomistic model consisting of proteins with predicted structures, RNA, protein/RNA complexes, metabolites, ions, and solvent. The resulting model bridges between atomistic and cellular scales, between physical and biochemical aspects, and between structural and systems views of cellular systems and is meant as a starting point for a variety of simulation studies.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

High-throughput experiments have transformed biology and we have reached a time where it is possible to develop comprehensive models of entire biological systems [1,2]. One example is a mathematical model of the minimal bacterium *Mycoplasma genitalium* (MG) that integrates genomic, proteomic, and metabolomic data into a fully connected metabolic reaction network [3]. This parameterized model predicts metabolic fluxes under different conditions in agreement with experiments [3]. The subsequent development of a public database for such types of whole-cell models [4] suggests that many similar models will soon be developed for other organisms as well.

Reaction-network type models focus on the biochemistry in a given organism but neglect molecular details of cellular environments that are necessary to connect systems biology to the detailed mechanistic underpinnings of biological processes. Molecular-level whole-cell models promise to predict biological phenotypes from physical principles. Such models are also of great practical relevance, for example in the context of rational drug design, where it would be possible to simultaneous consider specificity and selectivity of a given drug candidate within a given complex biological environment.

Previous attempts at building molecular models of cellular systems have involved bacterial cytoplasms and synaptic vesicles [5–7]. These models primarily consisted of macromolecules, proteins and RNA molecules, at different levels of detail [8]. One study focused on the reconstruction of the metabolic environment in *Escherichia coli* [9]. However, no complete cellular model is yet available that includes all of the molecular components, proteins, RNA, protein/RNA complexes, metabolites, ions, and solvent, that are present in the whole cell or a cellular subsection. More importantly, in our opinion, no model has been reported yet that is physically and biochemically consistent, i.e. the molecular components present in the system are mapped onto complete biochemical pathways and vice versa.

Here, we are describing a comprehensive atomistic model of the cytoplasm of *MG* based on an integration of genomic, proteomic, metabolomic, and structural information and the application of bioinformatics and structure prediction tools. We chose

* Corresponding author at: Department of Biochemistry & Molecular Biology, Michigan State University, 603 Wilson Road, BCH 218, East Lansing, MI 48824, United States. Tel.: +1 517 432 7439.
E-mail address: feig@msu.edu (M. Feig).

http://dx.doi.org/10.1016/j.jmgm.2015.02.004
1093-3263/© 2015 Elsevier Inc. All rights reserved.

MG because of its minimal size, extensive genetic characterization, metabolic simplicity, and availability of proteomics and metabolomics data for *Mycoplasma pneumoniae* (*MP*), the phylogenetically closest neighbor of *MG*. The resulting model provides the most realistic view of a bacterial cytoplasm so far and represents a significant step toward a complete molecular-level whole-cell model of an entire bacterium. The model described here bridges structural biology with systems biology and can serve as a starting point for computer simulations.

2. Methods

2.1. Gene annotations

Initial annotations were taken from the comprehensive microbial resource (CMR) [10] and the Uniprot database [11]. However, a significant fraction of *MG* genes is annotated only as having 'hypothetical function' or is classified into a general functional category (such as 'hydrolase' or 'integral membrane protein') that provides only limited information about their biochemical roles. Past efforts of developing whole-cell models of *MG* have focused on reconstructing metabolic pathways with the goal of describing metabolic fluxes and overall growth rates under different conditions [2]. In such a context, non-metabolic proteins are not critical other than contributing to the overall biomass. Furthermore, it may be acceptable to compensate for missing enzymes in metabolic pathways by assuming that one of the proteins with unknown function may fill the gap without actually identifying the particular gene. Here, we attempted to achieve stringent annotations that map genes to functions and vice versa. In order to resolve missing genes and/or unclear function predictions we assumed that some proteins with related functions have promiscuous activity. Enzymatic promiscuity in *MG* has been shown experimentally for nucleoside diphosphate kinase [12] an otherwise ubiquitous gene that is absent in *MG*. The experiments suggest 'moonlighting activity' of other kinases such as pyruvate kinase that have relaxed substrate specificities and therefore can process not just the canonical ATP/ADP substrates but other nucleotides as well. Furthermore, proteomics analysis suggests that more than half of the genes in *MG* may be involved in multiple complexes suggesting multifunctional characteristics [13].

In addition to predicting function we also predicted for each *MG* gene the cellular location of the gene product and its phylogenetic context, i.e. whether genes are specific to *MG* and the closest relatives *MP* and *Mycoplasma gallisepticum*.

In a number of cases speculative annotations for previously unannotated genes were made based on remote sequence homologs and/or structural homologs, both of which are typically not considered in more conservative functional annotations. Membrane localization was determined based on annotations in the sequence database along with predictions of transmembrane helices using TMHMM [14] that revealed additional, previously unannotated, proteins that are likely membrane-bound.

2.2. Protein structure prediction

Direct experimental structural information is only available for six of the genes in *MG*: MG027, the transcription antitermination factor nusB (PDB 1Q8C); MG238, the non-essential trigger factor tig (PDB 1HXV); MG289, the extracytoplasmic thiamine binding protein (PDB 3MYU); MG354, an MG-specific, membrane-anchored protein possibly serving a structural role (PDB 1TM9); MG438, the S subunit of the type I restriction system (PDB 1YDX); MG469, the DNA replication initiation factor dnaA (PDB 2JMP). Unfortunately, none of those proteins are in the list to be included in our cytoplasmic model. Therefore, all of the structures included in our model had to be predicted via computational methods.

Template-based modeling followed a standard protocol where initial alignments between the *MG* genes and the sequence of the template structures were generated by PSI-BLAST [15]. In cases with high sequence similarity this was sufficient but in some other where alignments appeared to be unreliable we made manual adjustments to match both sequence and predicted secondary structures (from PSIPRED [16]). The final alignments were then used to generate models via MODELLER [17].

For three genes (MG278, 3′,5′-bis(diphosphate) 3′-pyrophosphohydrolase; MG366, phosphomethylpyrimidine kinase; and MG369, dihydroxyacetone kinase) we could not find suitable templates. Furthermore, their large sizes (720, 667, and 557 amino acids, respectively) made meaningful prediction via ab initio methods impossible.

In the initial model construction, large loops and missing N- or C-terminal residues were excluded. The missing parts were then added one-by-one using MODELLER's loop modeling function [18,19] in order to apply more extensive sampling to those regions. For selecting the best model we relied on MODELLER's internal DOPE [20] scoring function. In a few cases we used structural information from multiple templates. This was accomplished by building different parts of a structure based on different templates and subsequently merging those parts into a single model. While N- and C-termini are sometimes neglected in the absence of suitable structural information we included the termini, even if they involved longer regions without regular secondary structure in order to account for all of the biomass represented by the proteins and include the effect of partially disordered termini potentially extending from a given protein structure.

Loop modeling is generally reliable for up to ten residues, but there is greater uncertainty as the length of loops increases [19,21]. The same also applies for modeling N- and C-termini that are only anchored on one side by the rest of the structure. However, it is also highly likely that long loops and termini without predicted regular structure are highly dynamic so that any single predicted structure would likely not be adequate for capturing a dynamic ensemble for such regions. However, as long as a plausible structure can be predicted, it can serve as a starting structure for more extensive sampling, e.g. via molecular dynamics simulations, to better describe such dynamics.

The models obtained by MODELLER via homology modeling were subsequently subjected to a short minimization with the CHARMM all-atom force field [22] with a distance-dependent dielectric and under harmonic restraints on Cα positions. The resulting structures are usually of sufficient quality at the atomistic level to be used as a starting structure for further modeling. However, in some cases *cis*-backbone conformations (with ω near zero) appear for non-proline residues. Another potential problem is the entanglement of histidine, phenylalanine, tyrosine, or tryptophan side chains that results in intertwined rings. Neither *cis*-peptide bonds nor intertwined rings can be relieved with additional force field based minimization. Therefore all models were screened for such problems and if present the affected residues and their immediate neighbors were resampled with MODELLER until the problem was resolved.

Many of the proteins predicted here are assumed to be forming long-lived complexes and we strived to model the biologically relevant complexes as much as possible given the available structural data. We built the following heterogeneous complexes consisting of multiple different gene products: ribosome, groEL-ES, RNA polymerase, pyruvate dehydrogenase, ribonucleoside-diphosphate reductase, a complex between the phosphocarrier HPr and HPr kinase, a complex between the translation elongation factors Tu and Ts, a chaperon complex between dnaJ, grepE, and dnaK, as well

Download English Version:

https://daneshyari.com/en/article/443295

Download Persian Version:

https://daneshyari.com/article/443295

Daneshyari.com