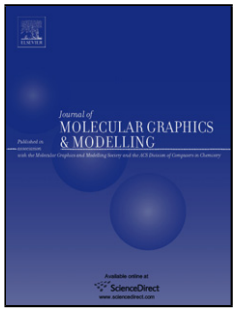

Title: Complete Atomistic Model of a Bacterial Cytoplasm for Integrating Physics, Biochemistry, and Systems Biology

Author: Michael Feig Ryuhei Harada Takaharu Mori Isseki Yu Koichi Takahashi Yuji Sugita

PII: S1093-3263(15)00038-8
DOI: http://dx.doi.org/doi:10.1016/j.jmgm.2015.02.004
Reference: JMG 6520

To appear in: *Journal of Molecular Graphics and Modelling*

Received date: 19-12-2014
Revised date: 18-2-2015
Accepted date: 22-2-2015

Please cite this article as: M. Feig, R. Harada, T. Mori, I. Yu, K. Takahashi, Y. Sugita, Complete Atomistic Model of a Bacterial Cytoplasm for Integrating Physics, Biochemistry, and Systems Biology, *Journal of Molecular Graphics and Modelling* (2015), http://dx.doi.org/10.1016/j.jmgm.2015.02.004

- Highlights
- Construction of a complete atomistic model of a bacterial cytoplasm
- Genome-scale protein structure prediction
- Correspondence between metabolic reaction network and physical composition of atomistic system

Complete Atomistic Model of a Bacterial Cytoplasm for Integrating Physics, Biochemistry, and Systems Biology

Michael Feig[ǂ,#,*] , Ryuhei Harada[+,#], Takaharu Mori[#,$], Isseki Yu[$], Koichi Takahashi[&,%], and Yuji Sugita[+,#,$]

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

*Corresponding Author: Department of Biochemistry & Molecular Biology, Michigan State University, 603 Wilson Road, BCH 218, East Lansing, MI, 48824, United States.
Tel: 1-517-432-7439 Email: feig@msu.edu

AUTHOR CONTRIBUTIONS

MF, KT, and YS designed the research; MF, RH, TM, and IY conducted the research; MF, IY, KT, and YS wrote the manuscript.

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.

KEYWORDS

Mycoplasma genitalium, metabolic reaction network, protein structure prediction, crowding

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 *E. 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 *MG* because of its minimal size, extensive genetic characterization, metabolic simplicity, and availability of proteomics and metabolomics data for *M. 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 towards 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.

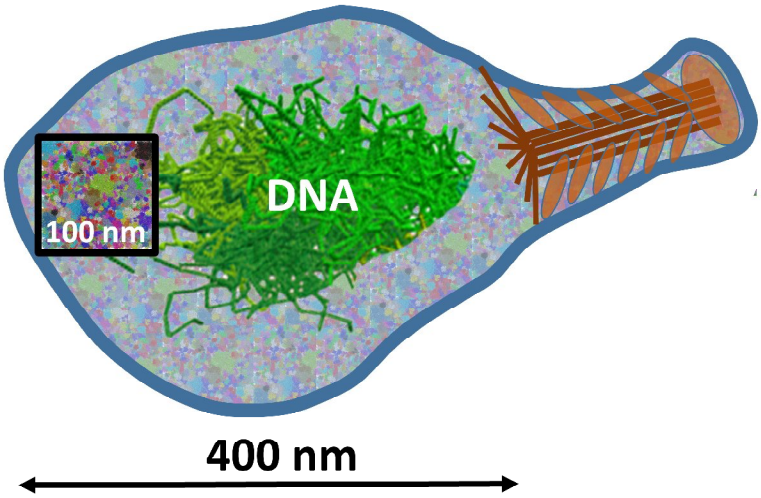

Figure 1: Schematic view of *Mycoplasma genitalium* and the cytoplasmic subsection modeled here. The attachment organelle and its internal structure is schematically illustrated in brown.

METHODS

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 un-annotated 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.

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 as complexes between aminoacyl-tRNA synthetases and their cognate tRNA where such complexes were available from template structures. In addition, we built homooligomeric complexes for a large fraction of proteins, again based on information from the template structures that were used (see Table S2). Our assumption was that the oligomeric states observed in the homologs would be preserved in *MG* as well, which may not be true in all cases. Unfortunately, comprehensive experimental data about the oligomerization states of *MG* proteins is not available while the reliable computational prediction of oligomerization is still in its infancy. In the final assembled model we included the complexes with the highest oligomerization states but also added monomers and partial complexes in cases where experimental data suggests the presence of such species. For example not just the complete ribosome was included but the 30S and 50S subunits were also present as separate particles and for many high-abundance metabolic enzymes we included monomers (names ending with a '1') in addition to dimers and tetramers (see Table S2 for details).

The final models for each macromolecule (or complex) were then minimized again using the program CHARMM[23], solvated in a TIP3 water box heated to 300K and subjected to a 20 ps molecular dynamics simulation using NAMD[24] to further relax the structures. All of the final predicted structures are depicted in Figure 2. An archive file containing the structures in PDB format is available from the corresponding author upon request.

Assembly of the Cytoplasmic System

The setup of a system under dilute conditions is relatively straightforward because macromolecules can simply be placed into a box and surrounded by solvent. In contrast, the setup of a very dense, heterogeneous system is much more challenging. One possibility would be to start with a large, more dilute system that is then subsequently shrunk by slowly reducing the box size until the desired density is reached. However, while this may be a good idea for a system consisting of a few macromolecules, the prohibitive cost for a system of the size proposed here requires a different strategy. The approach taken here involves a multi-scale procedure that assembles and equilibrates the system in several stages with increasing resolution. An overview of the assembly protocol is depicted in Figure S12. The individual steps are described in detail in the following:

1. ***Initial Assembly of Macromolecules as Spheres***: In the first step, each macromolecule was represented as a single sphere. Complexes such the ribosome are considered as a single particle resulting in 1,500 total particles that were placed randomly inside a $(100\ nm)^3$ box. A Lennard-Jones potential was used with ε=-1 kcal/mol and radii set to 1.5 times the radius of gyration of each particle to allow sufficient room for replacing the spheres with the actual molecular shape in the next step. In addition, each particle had a small charge (0.1 e) to prevent aggregation. The initial system was subjected to a short molecular dynamics simulation (100 ps) with CHARMM[23] using periodic boundary conditions and a temperature of 300K to equilibrate and randomize the particle configuration. The final system from this step is shown in Figure S13A.
2. ***Accommodation of Chain Representation of Macromolecules:*** In a second step, the spheres from the previous step were replaced with the structures of each macromolecule (or complex) at a coarse-grained level consisting of Cα atoms for proteins and P atoms for RNA. Because most macromolecules are not exactly spherical some clashes involving extended molecular parts had to be resolved. This was accomplished by randomly rotating and translating the macromolecules from their initial placement in an iterative procedure until most clashes are avoided followed by an energy-guided minimization. The energy function penalized close contacts of less than 16 Å while only allowing rigid body motions of each macromolecule to further optimize the arrangement of the macromolecules with minimal close contacts and

facilitate the transition to a fully atomistic model in the next step. The final system is shown in Figure S13B.

3. ***Replacement of Coarse-grained Macromolecules with Atomistic Representation:*** After the CG model of the cytoplasmic system was sufficiently optimized with respect to clashes between different molecules, all-atom structures for each macromolecule, using the predicted structures generated above, were superimposed onto the coarse-grained chains. The resulting system is shown in Figure S13C.

4. ***Addition of Aqueous Solvent:*** Commonly, a solvated system is generated by overlaying a pre-equilibrated water box onto a given solute. The same approach could be used here but because of the large size of the final system the time required for equilibration is a major concern. Simply placing water molecules around solutes would require equilibration of the water-solute interfaces. Since fully solvated molecular dynamics simulations were carried out for all of the macromolecules (see above), one can take advantage of the equilibrated waters when solvating the assembled cytoplasmic model. This was accomplished by rotating and translating the water molecules and counterions, where present, along with the macromolecules when fitting to the coarse-grained chains in step 3 and then adding the rotated solvent molecules unless they overlap with other solutes or with other previously added solvent molecules. In a very dense system, this is enough to fill the entire volume that is not occupied by solutes with solvent. In a more dilute system, voids will remain that have to be filled with additional water molecules in a final step. Figure S13D shows the system at the end of this step.

5. ***Addition of Metabolites***: Metabolites were added to the solvated system by randomly selecting a water molecule, placing a given metabolite with its center of mass at the water site. If a metabolite placed in this way would overlap with a macromolecular solute or with an already inserted metabolite, minor translation/rotation would be attempted to resolve the overlap and if that is not possible, a different water site would be chosen. Once a metabolite is placed comfortably with respect to the solute and other metabolites, solvent molecules that overlap with the inserted metabolite were removed from the system. Figure S13E shows the system after completion of this step.

6. ***Adjustment and addition of ions***: In step 4, Na^+ and Cl^- counterions were kept from the equilibration simulations of the single macromolecules. However, the final system after addition of metabolites does not have the correct number of ions according to Table S3. Therefore, the number of Na^+ and Cl^- ions was first adjusted to reflect the concentrations according to Table S3. Then K^+ and Mg^{2+} ions were added by replacing water molecules until the desired number of ions and charge neutrality of the entire system were reached. At this point, the initial setup of the system is complete (see Fig. S13F).

A PDB file containing the complete assembled atomistic cytoplasmic model is available from the authors upon request.

We note, that if the model constructed here is used as the starting point for simulations, it is also necessary to obtain suitable force field parameters. Protein, nucleic acid, ion, and water parameters are readily available from a number of force fields, but metabolite parameters may be more problematic. For the system constructed here almost all parameters, including the metabolites, are available from the latest CHARMM (c36 protein[22, 25] and nucleic acid force fields[26] and CGenFF[27] for metabolites) or Amber force field suites (ff14sb for proteins[28] and nucleic acids[29] and GAFF[30] for metabolites). For metabolites considered here where parameters are not directly available it is generally possible to simply combine parameters from existing compounds. For example, missing parameters for UDP-galactose can be constructed by combining existing UDP and galactose parameters.

RESULTS

A cytoplasmic model of *M. genitalium* was constructed by using experimental data together with computational tools. The target is a model of a cubic subsection of *MG* with a size of $(100\ nm)^3$. This is the largest cubic volume that would not overlap with either the genomic DNA or the cellular membrane (see Fig. 1), neither of which can be modeled reliably in atomistic detail yet. Apart from an attachment organelle, *MG* is approximately spherical with a diameter of 300-400 nm. Therefore, a volume of $(100\ nm)^3$ represents about $1/20^{th}$ of the entire cell volume. This implies that any molecule present in this subvolume would have a minimum concentration of about 1 μM.

Cytoplasmic proteins

MG has a minimal set of 525 genes in the initial genome annotation[31], 476 of which (by our count) encode for translated, functional proteins, 36 genes encode for tRNAs, three for rRNAs, and four encode for accessory ncRNAs. The transcriptome of *MP* suggests that additional ncRNAs may exist and function e.g. as antisense regulators[32]. Such ncRNAs were not considered here.

A majority of genes of *MG* are readily annotated via sequence homology to proteins in other organisms. Other genes were annotated by allowing for promiscuous activities, filling gaps in metabolic pathways, and/or by structural homology (see online methods). Because we aimed to build a model that is physically and biochemically consistent, we assigned genes to every reaction but also predicted a function for every gene if possible. The annotated list of genes in *MG* is summarized in Table 1 (see details in Table S1). Only five genes were not assigned to any functional category and only three of those are essential. Two of those are predicted to be membrane-bound and only one - MG148 – appears to encode for an essential cytoplasmic protein. So, within the uncertainties of functional predictions, our annotation of the *MG* genome is essentially complete.

To consider the minimum set of genes in a cytoplasmic model for *MG*, we excluded non-essential genes, genes encoding for membrane-bound proteins, and genes highly specific to *MG* as they are likely involved in the attachment organelle and/or interactions with host cells. The remaining genes (see Table 1) were reduced further by focusing on a cell during its quiescent phase so that neither replication nor cell division take place. Other DNA-related proteins were neglected because DNA is not part of our cytoplasmic model. Only RNA polymerase is included because it is present at high concentrations and expected to diffuse through the cytoplasm. Furthermore, all of the remaining non-membrane bound proteins that have fatty acid substrates or products were excluded because such compounds are likely localized in or near the membrane. Proteins predicted to interact with host cells or foreign DNA were not considered, and non-MG specific structural proteins were omitted because there is not enough structural information to include cytoskeletal components in our model. Finally, proteins involved in RNA processing, ribosome biogenesis, and post-translational modifications were excluded because

they were not detected in *MP*[13] and because such processes are infrequently needed once the modified macromolecules are present in a cell. The remaining genes, highlighted in Table 1, comprise the essential genes of a cytoplasmic subsection of *MG*.

Some of the corresponding gene products form stable heterooligomeric complexes. Hence, the number of freely moving macromolecular species in our cytoplasmic model is reduced to around 125. A compilation of all of the proteins, RNAs, and their complexes is given in Table S2. The resulting set of proteins is significantly more heterogeneous than most previous models of bacterial cytoplasms[6, 33, 34] but similar in complexity to a recent coarse-grained model of the *E. coli* cytoplasm[7].

The concentration of each protein species was estimated by assuming that protein copy numbers measured for *MP*[13, 35] are also valid for *MG*. Given that copy numbers fluctuate significantly within bacterial populations and under different growth conditions, this is likely a reasonable assumption. Experimental copy numbers are available for high-abundance cytoplasmic proteins in *MP* but do not cover all of the proteins assumed to be present in the cytoplasm (see Table 1). For proteins where *MP* homologs were not identified experimentally, we assume that their concentration is below the sensitivity threshold of mass spectrometry and we used a minimal concentration of one particle per $(100\ \mathrm{nm})^3$ (see Table S3).

Table 1: Number of genes according to functional categories distinguished by whether they are essential or not, specific to MG/MP, or bound to the membrane. The final set of selected cytoplasmic genes is highlighted in grey.

Classification	All	Non-Essential	Essential		
			MG-specific	Ubiquitous	
				Membrane-bound	Cytoplasmic
Replication	18	3	1	1	13
DNA Repair	8	2	0	1	5
DNA Degradation	3	1	0	0	2
DNA Recombination	5	4	0	0	1
DNA Remodeling/Stabilization	8	4	0	0	4
Cell Division	12	4	4	0	4
Transcription	18	3	3	0	12
RNA Processing	24	8	0	0	16
RNA Degradation	4	0	0	2	2
RNA Remodeling	2	0	0	0	2
Translation	67	2	3	2	60
Protein Folding	11	2	0	1	8
Ribosome Biogenesis	9	3	0	0	6
Aminoacyl tRNA Synthetase	24	0	0	3	21
Post-translational Processing	8	1	1	1	5
Signaling	3	0	0	1	2
Protein Degradation	7	2	0	1	4
Glycolysis	19	2	0	4	13
Nucleotide Metabolism	31	3	0	8	20
Lipid Metabolism	21	6	0	5	10
Sugar Metabolism	13	2	0	0	11
Amino Acid Metabolism	2	0	0	0	2
Cofactor Metabolism	27	4	1	0	22
Membrane Transport	67	16	3	48	0
Lipoprotein	12	3	3	6	0
Cell Adhesion	11	2	3	4	2
Host Cell Interaction	5	2	1	1	1
Foreign DNA Processing	4	3	0	0	1
Cytoskeleton	5	1	0	1	3
MG-Specific Function	23	10	13	0	0
Unknown	5	2	0	2	1

Modeling of cytoplasmic proteins

MG is a prime example for the divergence of sequencing and structure determination. While the genome of *MG* is known for nearly two decades, direct structural information is only available for six of its genes, none of which are in the list to be included in our cytoplasmic model. Therefore, we relied on protein structure prediction for all of the selected cytoplasmic genes. Genome-scale structure prediction has been attempted previously for *MG*[36, 37]. Since then, structure prediction methods have improved[38] and coverage of structural space in the PDB has increased significantly[39]. We were able to predict structures for all but three of the genes in our cytoplasmic set (>95%) using templates from other known structures (see Table S2). Sequence similarity to the available templates varied and therefore the expected accuracy of the homology models is also expected to vary. For homologs with high sequence similarity, models can be as close as 1-2 Å Cα RMSD from an experimental structure[40]. On the other hand, benchmark data from recent rounds of CASP suggests that models of challenging targets where templates can be identified are generally expected to have GDT_TS scores above 60% [38], meaning roughly that at least 60% of the model is within 2-4 Å RMSD of the correct structure. We expect that the models generated here fall within this range. Structures for phosphomethylpyrimidine kinase (MG366), dihydroxyacetone kinase (MG369), and guanosine-3',5'-bis(diphosphate) 3'-pyrophosphohydrolase (MG278) could not be predicted because of a lack of suitable templates. This suggests that the protein structures available today can cover most of the core biochemical functions within the limits of homology modeling.

Figure 2 shows the structures for all of the proteins and RNAs considered here obtained via homology modeling. Many of the macromolecules are assumed to be forming long-lived complexes and we modeled the biologically relevant complexes where possible given structural data from homologs. The smallest structure in this set has 70 amino acid residues (translation initiation factor IF-1), the largest is the ribosome with >7,600 amino acid residues and >4,700 nucleotides. The average number of residues per protein structure is 957 residues. Only 5% of all structures have less than 300 residues. The average molecular weight of all complexes (proteins and RNA) is 119 kDa which is larger than the value of 84 kDa in Elcock's incomplete model of the *E. coli* cytoplasm[6]. This reaffirms that the cytoplasm is dominated by large protein complexes.

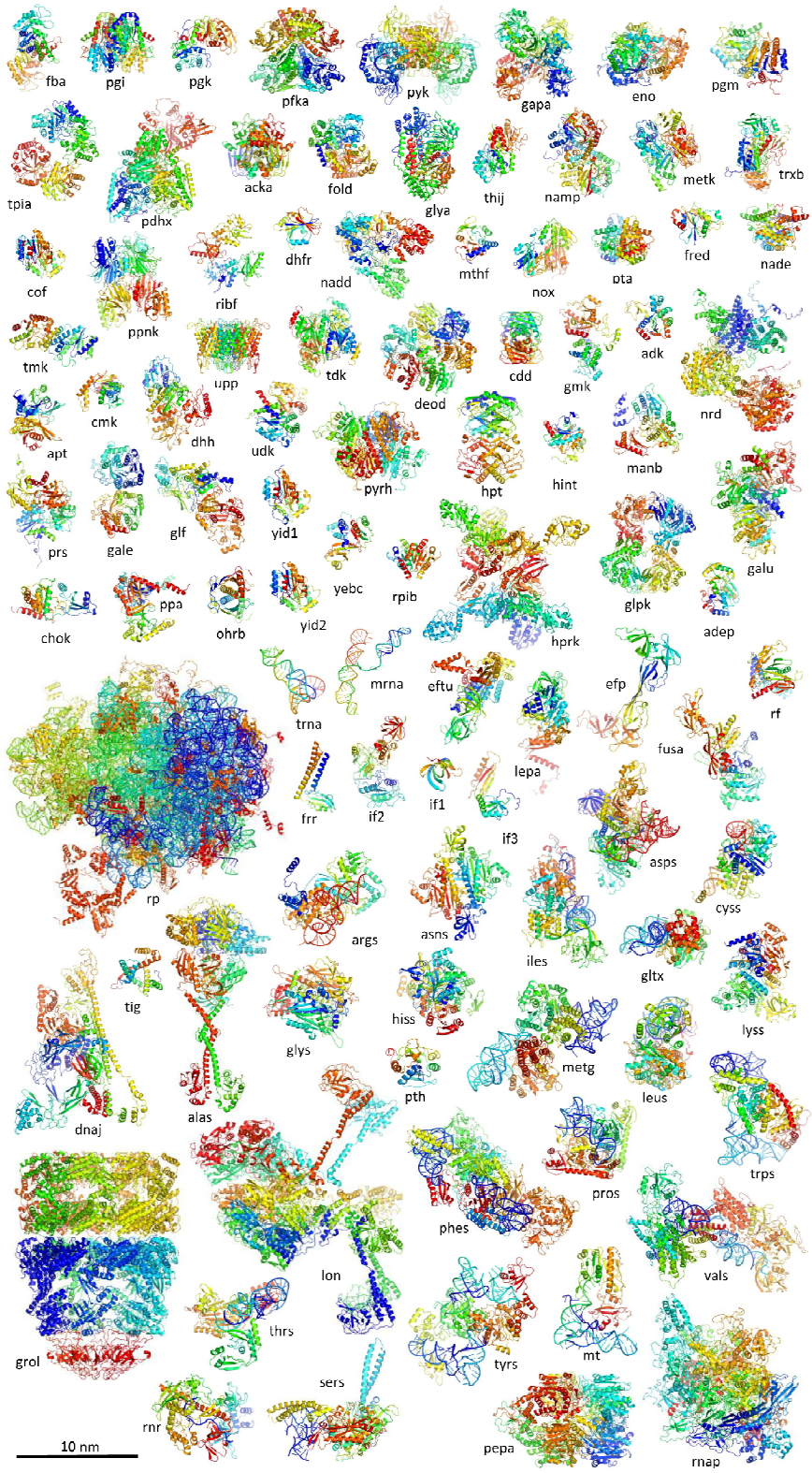

Figure 2: Structures of cytoplasmic macromolecules. Coloring is by residue index. The labels correspond to the macromolecular tags in Tables S1-S3.

Cytoplasmic reactions and metabolites

In order to build a biochemically consistent system with all of the metabolites present, we determined the list of metabolites based on the reactions carried out by the cytoplasmic proteins identified in the previous section. A biochemical reaction network was initially constructed for the entire set of genes but then trimmed to include only those reactions that involve the proteins in our cytoplasmic set. Since structures for phosphomethylpyrimidine kinase (MG366), dihydroxyacetone kinase (MG369), and guanosine-3',5'-bis(diphosphate) 3'-pyrophosphohydrolase (MG278) could not be predicted, we did not include the conversion between dihydroxyacetone and dihydroxyacetone phosphate or the metabolism of guanosine 3'-5-bis(diphosphate). Phosphomethylpyrimidine kinase would play a role in thiamine biosynthesis, but since thiamine is an essential nutrient it is likely that this gene encodes for a kinase operating on a related substrate and we are unsure about the effect of omitting this gene.

The remaining reactions are shown in Figure 3 (see also Figures S1 to S11). The pathways generally agree with previous metabolic pathway reconstructions[3, 35, 41, 42]. Different from previous studies, we did not introduce any new reactions without a matching gene to complete pathways. Instead, we attempted to assign functionally related genes that could plausibly connect pathways. Where that was not possible, we left out pathways that were considered in other reconstructions of *MG*. For example, the pentose phosphate pathway is missing in our reconstruction because one of the central enzymes, transketolase, is considered to be non-essential while another key enzyme, transaldolase, appears to be missing altogether in *MG*. In another example, two genes are predicted to be involved in non-mevalonate terpenoid precursor synthesis. However, other enzymes necessary to complete that pathway appear to be missing and therefore we assume that this pathway is not active in *MG* and it was omitted.

The resulting pathways are fragmented because we assume net influx and efflux of certain metabolites with respect to the environment. Some molecules would be imported from the extracytoplasmic environment. *MG* requires relatively rich growth media that minimally provide glucose, glcycerol, spermine, nicotinic acid, thiamine, pyridoxal, thioctic acid, riboflavin, choline, folate, coenzyme A, guanine, cytidine, adenine, fatty acids, and either amino acids or peptides[35]. However, we also assume metabolic flux to and from other parts of the cell, in

particular the membrane, where further reactions would take place not considered in our cytoplasmic model. For example, glycerol may be turned over by glycerol kinase in the cytoplasm to glycerol-3-phosphate before diffusing to enzymes near the membrane to serve as a substrate for lipid biosynthesis.

The metabolites involved in the proposed minimal cytoplasmic reactions are listed in Table S3. Concentration data for some metabolites is available from *MP*[35]. Other data was taken from *E. coli[43]*. While the use of *E. coli* data is potentially problematic because of differences in metabolism, the major pathways are similar and some uncertainty can be tolerated because metabolite concentrations vary significantly in living cells as a function of growth and environmental conditions[35]. Some of the metabolites proposed to be present based on the reactions appear to have concentrations of less than 1 μM and therefore no molecule would be present in our $(100\ nm)^3$ model. This applies to some of the co-factor related metabolites but also to some of the nucleotides.

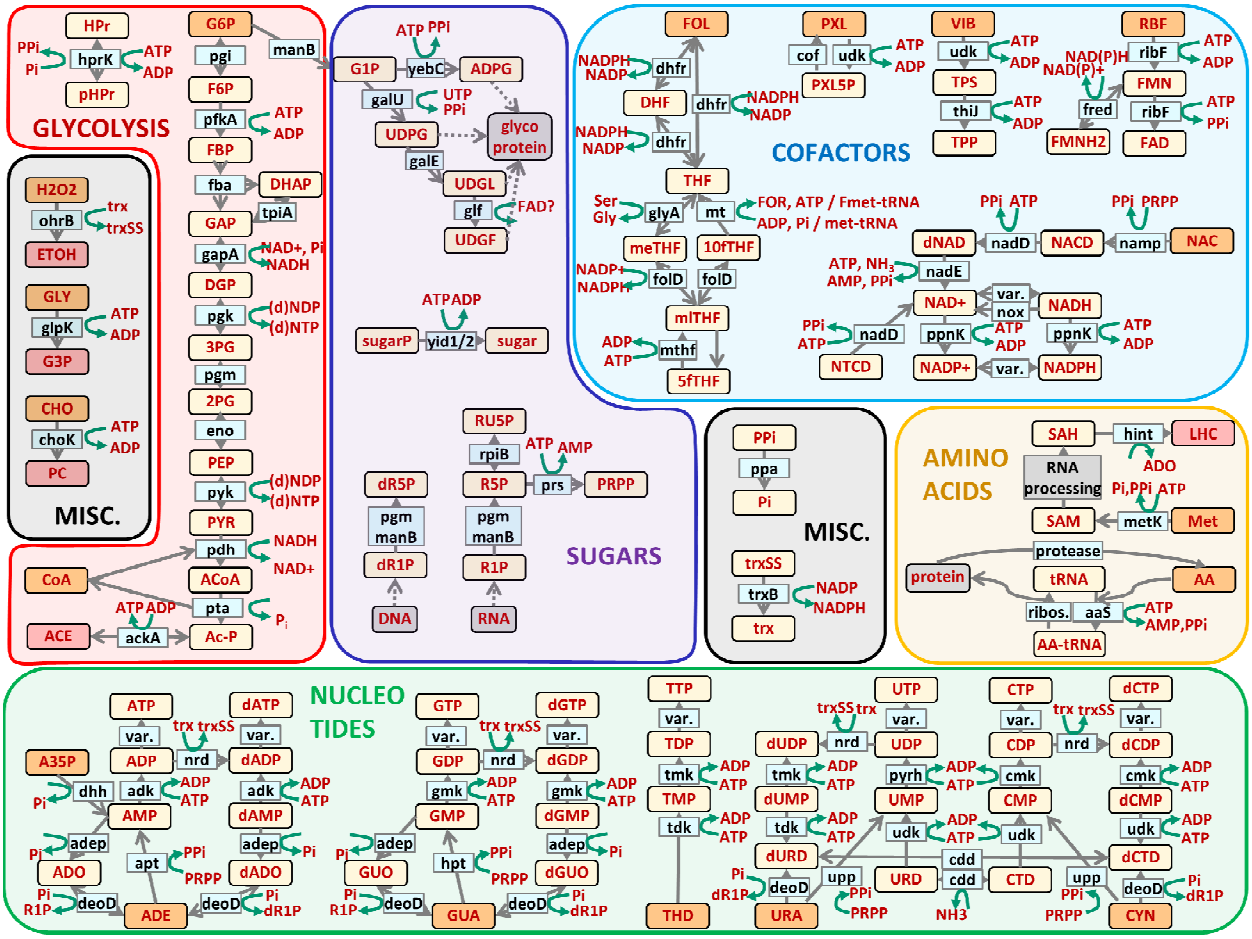

Figure 3: Predicted cytoplasmic metabolic reaction network. Compounds assumed to enter or exit the system with respect to the environment are highlighted in orange and red respectively. Colored sections represent major groups of metabolic pathways.

Components in a complete cytoplasmic model system

From the set of proteins and metabolites identified above, we constructed a cytoplasmic model. The system is essentially complete in terms of the molecular components and is consistent with the predicted reaction network. In addition to metabolic core functions, we also included all of the components necessary for translation, protein folding, and protein degradation, including aminoacyl-tRNA synthetases and GroEL/ES, and RNA polymerase. Macromolecular copy numbers were obtained by dividing copy numbers for *MP* by five under the assumption that at least half of the cell volume is occupied by either DNA or the membrane and membrane-associated proteins. Further adjustments were made to achieve a typical cellular concentration of

300 mg/ml for the proteins and nucleic acids (tRNA, rRNA, mRNA). Copy numbers for each metabolite were obtained based on their estimated concentrations. Na^+, Cl^-, Mg^{2+}, and Ca^{2+} ions were added at physiological concentrations. K^+ was used to achieve charge neutrality. The resulting K^+ concentration of about 300 mM is within range of typical values for bacteria[44]. The cytoplasmic model, including 26 M explicit water molecules, was assembled with a multi-scale protocol (see online methods). The components of the assembled system are summarized in Table 2. Details are given in Tables S3. Figure 4 shows the assembled model.

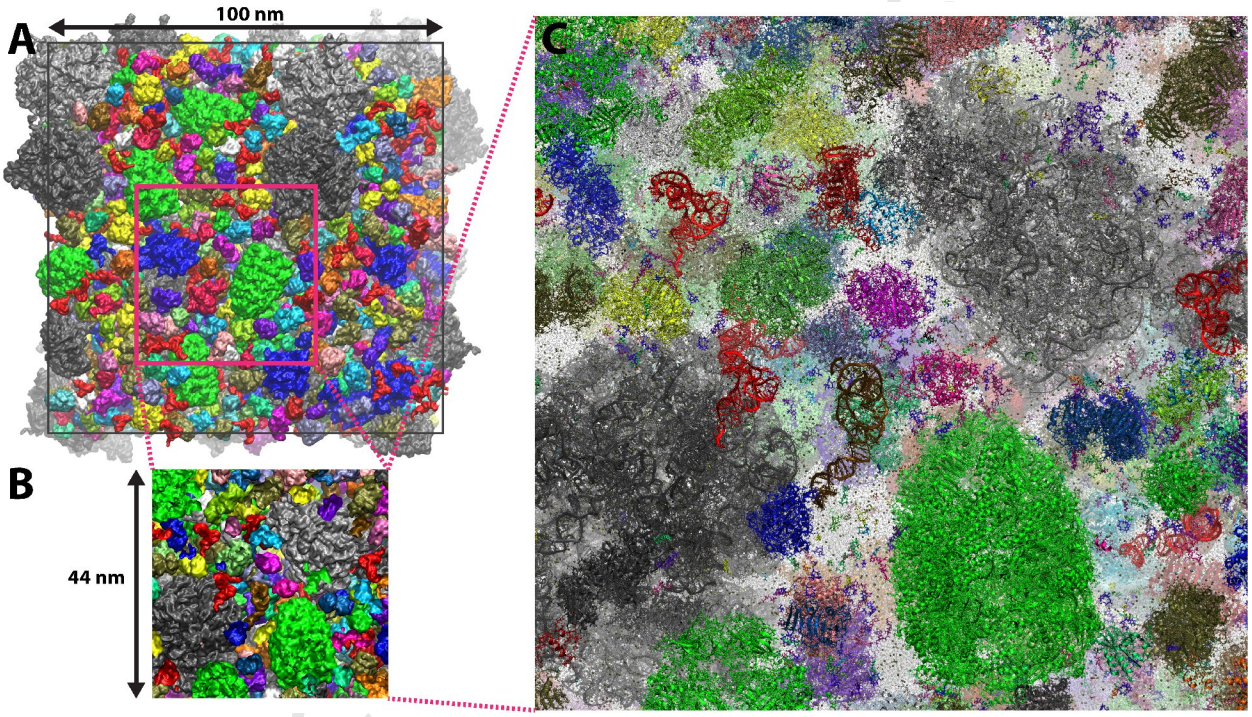

Figure 4: Complete cytoplasmic model system of *Mycoplasma genitalium*. The complete system is shown on the left, an enlarged cross-section on the right.

Table 2: Molecular components in cytoplasmic model system with selected physical parameters. In the case of the macromolecules, the copy number refers to complexes, not individual gene products.

Component	Number per $(100\ nm)^3$	Concentration		Charge [e]
		[mM]	[mg/mL]	
Metabolic proteins	992		115.35	4,931
Ribosome (proteins + rRNA)	40		107.50	-107,088
Translation Factors	76		6.15	124
Protein Folding/Degradation	33		32.02	-636
Aminoacyl tRNA synthetases	42		8.13	-728
tRNAs	275		11.15	-20,625
RNA polymerase	15		11.18	270
Amino acids	9522		2.04	-1,806
Nucleotides	16,861		13.13	-60,786
Other Metabolites	8,646		5.99	-17,652
Ethanol	903	1.5	0.07	0
H_2O_2, NH_3	18	0.03	0.001	0
Spermidine	602	1	0.15	1,806
Phosphate, Pyrophosphate	3,311	5.5	0.57	-5,719
K^+	192,619	320	12.5	192,619
Na^+	12,040	20	0.46	12,040
Cl^-	3,010	5	0.18	-3,010
Mg^{2+}	3,010	5	0.12	6,020
Ca^{2+}	120	0.2	0.01	240
Water	26,000,000		605	0

DISCUSSION

We are presenting a fully atomistic model of a bacterial cytoplasm based on *Myocplasma genitalium* that integrates experimental data with computational predictions. Focusing on a cytoplasmic subsection, it was possible to almost completely annotate the genes in this minimal bacterium and construct complete metabolic pathways where genes are mapped onto enzyme functions and *vice versa*. Molecular structures were modeled for essentially all of the cytoplasmic genes. In the resulting model, structural and systems biology converge into a cellular-scale model that is physically and biochemically consistent.

While it would be desirable to have experimental structures for every protein, confirmation of biochemical functions for every gene, and quantitative analysis of all macromolecules and metabolites in a specific organism, the integrative approach taken here that relies on data from more disparate sources in combination with computational prediction tools to fill in gaps in knowledge is a practical paradigm for developing such models that could be applied to other organisms as well. It needs to be emphasized, however, that the model presented here is without validation and remains highly speculative because of a variety of uncertainties, in particular with respect to gene function predictions, the exact metabolic reaction network, and the modeled structures macromolecules.

The model presented here is meant as a starting point for multi-scale studies of cellular systems. The molecular model can be used as an initial configuration for detailed molecular dynamics simulations, Brownian dynamics, or simulations at coarse-grained levels depending on the questions to be investigated. At the same time, the metabolic reaction network could be investigated with ODE-type kinetic models or with a spatially-explicit reaction-diffusion formalism. The integration into a dynamic description of cellular environments that spans sub-nm to sub-μm spatial scales and bridges conformational dynamics occurring during sub-μs to reactions on second time scales is now becoming possible. Using results from such simulations to predict experimental observables will be a critical next step in validating the model introduced here.

Looking further ahead, the modeling of entire cells in full molecular detail is clearly within reach. One major challenge is the modeling of bacterial nucleoids. That is becoming possible based on experimental constraints from 5C or Hi-C experiments[45, 46]. Another challenge is

the membrane envelope where structure prediction of the embedded integral membrane proteins presents the most significant obstacle[47].

ACKNOWLEDGEMENTS

Funding from RIKEN-QBIC, NIH GM092949, NIH GM084953, and NSF MCB 1330560 (to MF) and by MEXT SPIRE Supercomputational Life Science (to YS) is acknowledged. Computer resources were used at RIKEN-RICC (RIKEN Integrated Cluster of Clusters).

FUNDING

NIH GM092949, NIH GM084953, NSF MCB 1330560, RIKEN QBIC, MEXT SPIRE

REFERENCES

[1] Feig, M., Sugita, Y. Reaching New Levels of Realism in Modeling Biological Macromolecules in Cellular Environments. Journal of Molecular Graphics and Modeling. 2013, 45, 144-56.
[2] Macklin, D.N., Ruggero, N.A., Covert, M.W. The future of whole-cell modeling. Curr Opin Biotech. 2014, 28, 111-5.
[3] Karr, J.R., Sanghvi, J.C., Macklin, D.N., Gutschow, M.V., Jacobs, J.M., Bolival, B., et al. A Whole-Cell Computational Model Predicts Phenotype from Genotype. Cell. 2012, 150, 389-401.
[4] Karr, J.R., Sanghvi, J.C., Macklin, D.N., Arora, A., Covert, M.W. WholeCellKB: model organism databases for comprehensive whole-cell models. Nucleic Acids Res. 2013, 41, D787-D92.
[5] Wilhelm, B.G., Mandad, S., Truckenbrodt, S., Krohnert, K., Schafer, C., Rammner, B., et al. Composition of isolated synaptic boutons reveals the amounts of vesicle trafficking proteins. Science. 2014, 344, 1023-8.
[6] McGuffee, S.R., Elcock, A.H. Diffusion, Crowding & Protein Stability in a Dynamic Molecular Model of the Bacterial Cytoplasm. Plos Comp. Biol. 2010, 6, e1000694.
[7] Hasnain, S., McClendon, C.L., Hsu, M.T., Jacobson, M.P., Bandyopadhyay, P. A New Coarse-Grained Model for E. coli Cytoplasm: Accurate Calculation of the Diffusion Coefficient of Proteins and Observation of Anomalous Diffusion. Plos One. 2014, 9, e106466.
[8] Frembgen-Kesner, T., Elcock, A.H. Computer Simulations of the Bacterial Cytoplasm. Biophys. Rev. 2013, 5, 109-19.
[9] Cossins, B.P., Jacobson, M.P., Guallar, V. A New View of the Bacterial Cytosol Environment. Plos Comp. Biol. 2011, 7, e1002066.
[10] Peterson, J.D., Umayam, L.A., Dickinson, T., Hickey, E.K., White, O. The Comprehensive Microbial Resource. Nucleic Acids Res. 2001, 29, 123-5.
[11] Consortium, T.U. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res. 2014, 42, D191-D8.
[12] Pollack, J.D., Myers, M.A., Dandekar, T., Herrmann, R. Suspected Utility of Enzymes with Multiple Activities in the Small Genome Mycoplasma Species: The Replacement of the Missing "Household" Nucleoside Diphosphate Kinase Gene and Activity by Glycolytic Kinases. OMICS A Journal of Integrative Biology. 2002, 6, 247-58.
[13] Kuhner, S., van Noort, V., Betts, M.J., Leo-Macias, A., Batisse, C., Rode, M., et al. Proteome Organization in a Genome-Reduced Bacterium. Science. 2009, 326, 1235-40.
[14] Moller, S., Croning, M.D.R., Apweiler, R. Evaluation methods for the prediction of membrane spanning regions. Bioinformatics. 2001, 17, 646-53.
[15] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 1997, 17, 3389-402.
[16] Jones, D.T. Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices. J. Mol. Biol. 1999, 292, 195-202.
[17] Sanchez, R., Sali, A. Evaluation of Comparative Protein Structure Modeling by MODELLER-3. Proteins. 1997, Supplement 1, 50-8.
[18] Fiser, A., Do, R.K.G., Sali, A. Modeling of Loops in Protein Structures. Protein Sci. 2000, 9, 1753-73.

[19] Fiser, A., Feig, M., Brooks, C.L., III, Sali, A. Evolution and Physics in Comparative Protein Structure Modeling. Accounts Chem. Res. 2002, 35, 413-21.
[20] Shen, M.Y., Sali, A. Statistical potential for assessment and prediction of protein structures. Protein Sci. 2006, 15, 2507-24.
[21] Li, Y.H., Rata, I., Chiu, S.W., Jakobsson, E. Improving predicted protein loop structure ranking using a Pareto-optimality consensus method. BMC Struct. Biol. 2010, 10.
[22] Best, R.B., Zhu, X., Shim, J., Lopes, P.E.M., Mittal, J., Feig, M., et al. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone phi, psi and Side-Chain chi(1) and chi(2) Dihedral Angles. J. Chem. Theory Comput. 2012, 8, 3257-73.
[23] Brooks, B.R., Brooks, C.L., Mackerell, A.D., Nilsson, L., Petrella, R.J., Roux, B., et al. CHARMM: The Biomolecular Simulation Program. J. Comput. Chem. 2009, 30, 1545-614.
[24] Kale, L., Skeel, R., Bhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N., et al. NAMD2: Greater scalability for parallel molecular dynamics. J. Comput. Phys. 1999, 151, 283-312.
[25] MacKerell, A.D., Jr., Bashford, D., Bellott, M., Dunbrack, J.D., Evanseck, M.J., Field, M.J., et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. J. Phys. Chem. B. 1998, 102, 3586-616.
[26] Foloppe, N., MacKerell Jr., A.D. All-Atom Empirical Force Field for Nucleic Acids: I. Parameter Optimization Based on Small Molecule and Condensed Phase Macromolecular Target Data. J. Comput. Chem. 2000, 21, 86-104.
[27] Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., et al. CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. J. Comput. Chem. 2010, 31, 671-90.
[28] Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., Simmerling, C. Comparison of multiple amber force fields and development of improved protein backbone parameters. Proteins. 2006, 65, 712-25.
[29] Zgarbova, M., Otyepka, M., Sponer, J., Mladek, A., Banas, P., Cheatham, T.E., et al. Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. J. Chem. Theory Comput. 2011, 7, 2886-902.
[30] Wang, J.M., Wolf, R.M., Caldwell, J.W., Kollman, P.A., Case, D.A. Development and testing of a general amber force field. J. Comput. Chem. 2004, 25, 1157-74.
[31] Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., et al. The Minimal Gene Complement of Mycoplasma-Genitalium. Science. 1995, 270, 397-403.
[32] Guell, M., van Noort, V., Yus, E., Chen, W.H., Leigh-Bell, J., Michalodimitrakis, K., et al. Transcriptome Complexity in a Genome-Reduced Bacterium. Science. 2009, 326, 1268-71.
[33] Ando, T., Skolnick, J. Crowding and hydrodynamic interactions likely dominate in vivo macromolecular motion. Proc. Natl. Acad. Sci. U.S.A. 2010, 107, 18457-62.
[34] Ridgway, D., Broderick, G., Lopez-Campistrous, A., Ruaini, M., Winter, P., Hamilton, M., et al. Coarse-Grained Molecular Simulation of Diffusion and Reaction Kinetics in a Crowded Virtual Cytoplasm. Biophys. J. 2008, 94, 3748-59.
[35] Yus, E., Maier, T., Michalodimitrakis, K., van Noort, V., Yamada, T., Chen, W.H., et al. Impact of Genome Reduction on Bacterial Metabolism and Its Regulation. Science. 2009, 326, 1263-8.

[36] Kihara, D., Zhang, Y., Lu, H., Kolinski, A., Skolnick, J. Ab initio protein structure prediction to a genomic scale: Application to the Mycoplasma genitalium genome. Proc. Natl. Acad. Sci. U.S.A. 2002, 99, 5993-8.
[37] Huynen, M., Doerks, T., Eisenhaber, F., Orengo, C., Sunyaev, S., Yuan, Y., et al. Homology-Based Fold Predictions for Mycoplasma genitalium Proteins. J. Mol. Biol. 1998, 280, 323-6.
[38] Huang, Y.J.P., Mao, B.C., Aramini, J.M., Montelione, G.T. Assessment of template-based protein structure predictions in CASP10. Proteins. 2014, 82, 43-56.
[39] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., et al. The Protein Data Bank. Nucleic Acids Res. 2000, 28, 235-42.
[40] Schwede, T., Kopp, J., Guex, N., Peitsch, M.C. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res. 2004, 31, 3381-5.
[41] Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T.S., Matsuzaki, Y., Miyoshi, F., et al. E-CELL: software environment for whole-cell simulation. Bioinformatics. 1999, 15, 72-84.
[42] Suthers, P.F., Dasika, M.S., Kumar, V.S., Denisov, G., Glass, J.I., Maranas, C.D. A Genome-Scale Metabolic Reconstruction of Mycoplasma genitalium iPS189. Plos Comp. Biol. 2009, 5, e1000285.
[43] Bennett, B.D., Kimball, E.H., Gao, M., Osterhout, R., Van Dien, S.J., Rabinowitz, J.D. Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. Nat. Chem. Biol. 2009, 5, 593-9.
[44] Shabala, L., Bowman, J., Brown, J., Ross, T., McMeekin, T., Shabala, S. Ion transport and osmotic adjustment in Escherichia coli in response to ionic and non-ionic osmotica. Environmental Microbiology. 2009, 11, 137-48.
[45] Le, T.B.K., Imakaev, M.V., Mirny, L.A., Laub, M.T. High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome. Science. 2013, 342, 731-4.
[46] Umbarger, M.A., Toro, E., Wright, M.A., Porreca, G.J., Bau, D., Hong, S.-H., et al. The Three-Dimensional Architecture of a Bacterial Genome and Its Alteration by Genetic Perturbation. Miolecular Cell. 2011, 44, 252-64.
[47] Elofsson, A., von Heijne, G. Membrane protein structure: Prediction versus reality. Annu. Rev. Biochem. 2007, 76, 125-40.