Making Soup: Preparing and Validating Molecular Simulations of the Bacterial Cytoplasm

Leandro Oliveira Bortot,[†] Zahedeh Bashardanesh,[‡] and David van der Spoel[*,‡]

†*Laboratory of Biological Physics, School of Pharmaceutical Sciences of Ribeirão Preto, University of São Paulo, Ribeirão Preto, Brazil*

‡*Science for Life Laboratory, Department of Cell and Molecular Biology. Uppsala University, SE-751 05 Uppsala, Sweden*

E-mail: david.vanderspoel@icm.uu.se

Abstract

Introduction

Biomolecules move and function in an environment densely packed with high concentrations of macromolecules. The presence of macromolecules leads to steric effect due to excluded volume effect and intermolecular attrative/repulsive forces due to distributed charges on the surface of macromolecules.

Structure and dynamics of biomolecules are well characterized *in vitro*, however these studies *in vivo* are still evolving. Investigating biomolecular properties *in vivo* is possible through developments in the fields of nuclear magnetic resonance,[1,2] or fluoroscence spectroscopies.[3–5] An alternative method is to use computational models and simulation techniques. Biomolecular simulations are often carried out under dilute conditions or simple models of

macromolecular crowdings.[6–8] However, more attemps in modeling bacterial cytoplasm have been made recently.[9,10]

- How a computational model can be useful
- From a simple extrapolation in time, it has been estimated that the simulation of an *E. coli* bacterium would be possible in 25 years from now for one nanosecond with 10^{11} atoms?

Here we report on a model of *Escherichia coli* cytoplasm at atomistic level. The challenges in modeling and simulation with Molecular Dynamics are pointed and discussed and solutions are provided.

This work is the first model of *E. coli* at atomistic resolution that spans cellular dynamics on a microsecond scale.

Materials and Methods

Initial structures

The proteins and tRNA were downloaded from Protein Data Bank (PDB). We looked for the proteins structures that were either from or expressed in *E-coli*. In case of 1U22 (MetE) and 2EIP (Ppa) we used a loop-closure modelng tool based on Random Coordinate Descent (RCD) method[11] to correct the information for missing residues. The four metabolites were parametrized using GAFF and Antechamber.

Molecular Dynamics Simulation

General Simulation Setup: The proteins were simulated at 30% biomolecular mass fraction in a physiological salt concentration (0.15M NaCl). For all simulations, Amber99SB-ws force field was used[12] in combination with the TIP4P/2005 water model.[13] Electrostatic interactions were treated using the particle mesh Ewald algorithm.[14] All chemical bonds were

constrained at their equilibrium length using the LINCS algorithm[15] allowing an integration time step of 2 fs. Temperature was controlled at 310 K using the v-rescale algorithm[16] and a coupling time of 0.5 ps. The pressure was controlled at 1 bar using the Parrinello-Rahman algorithm[17] with a time constant of 10 ps.

For error analysis, each simulations were repeated three times with independent starting velocities.

All simulations were performed with Gromacs 2018. Single simulations were started from crystal conformations. The cytoplasm simulations starting conformation were taken from the equilibrated conformation of each single simulation.

Single Simulation Setup: Each component was simulated with the same parameter as the cytoplasm for $200ns$.

Analysis

Before any analysis the periodic boundary condition (pbc) artifacts have been removed. We used GROMACS tools to do the analysis. For single component simulations, first the components were made whole and jump removed and then all the atom were put inside the compact box. The same treatment were applied to the cytoplasm simulations. Additionally, each component's trajectory were extracted and fitted by rotation and translation for later rotational correlation time analysis.

A Mean Square Displacement (MSD) analysis was used to calculate the translational diffusion coefficient.[18] The diffusion coefficients were extracted by a linear fit to MSD analysis by averaging blocks with a length of $10\,ns$. In principle diffusion coefficient needs to be corrected for finite size effects[19] but due to relatively large simulation boxes this correction is negligible.

Results

Cytoplasm model

In this section we describe the rationale behind the composition of our model, which has five fractions: protein, RNA, metabolites, water and ions. We highlight that we didn't add lipids and DNAs to our model because we are considering only elements that are free to diffuse through the cytosol. We gathered data from several sources in order to build a computational model that is representative of the cytoplasm of *Escheria coli*.[9,20–22]

Protein fraction

We selected a group of eight proteins that account for 50% of the abundance of non-ribosomal proteins in the cytoplasm of Escherichia coli K-12.[?] The two most abundant proteins, TufA and MetE, account for 20% and 12% of the total protein count in E.coli K-12, respectively, while the other six proteins contribute with less than 5% each (Table 1).

Table 1: Protein fraction

Protein	Abundance in E.Coli K12
TufA	20
MetE	12
IcdA	5
AhpC	4
CspC	4
Ppa	3
GapA	2
Eno	2

RNA fraction

In order to build the RNA fraction of our cytoplasm model, we consider the following data: Since 74% of the dry weight of non-ribosomal RNAs is composed by tRNA, we chose to model the RNA presence with tRNA molecules. The protein and RNA content of the total dry weight of *E.coli* is 55% and 2.9%, respectively. That is, the total RNA weight corresponds

to 5% of the total protein weight. We incorporated this data in our model by considering the tRNA(Phe) molecule as a representative of RNAs due to the availability of a recent crystallographic structure.[23] The correct number of copies were added to the simulation box to account for the correct protein/RNA weight ratio (Table).

Metabolites fraction

The total number of metabolite molecules was calculated considering data showing that the number of metabolite molecules in the cytoplasm of *E. coli* is about 42.86 times higher than the number of proteins. We considered the most abundant molecule as representative of each metabolite class, i.e. Glutamate for amino acids, ATP for nucleotides, FBP for central carbon intermediates and Glutathione redox cofactors. The copy number for each molecule was calculated from the ratio of their experimentally observed concentration in *E. coli*.

Water fraction

The number of water molecules was calculated according to the desired biomolecular concentration, which ranges from XX to YY in biological systems such as E. coli cytoplasm. In our case, we choose the biomolecular concentration of 30%, that is, the number of molecules necessary to reach a ratio of total biomolecular mass to water mass of 30%.

Inorganic ions fraction

Finally, Mg2+ was used as counter-ions for tRNA and K+ cations were added to neutralize the charges of the simulation box. The number of water molecules was used to calculate the number of K+ and Cl- that were necessary to reach the ionic strength of 0.150 M.

Building the simulation box

All components can be put in the same simulation box by inserting each of them in random orientations in a cubic box of side L that is initially empty. However, that is not a trivial

Table 2: Cytoplasm Components

Class	Name (PDB ID)	Number
Protein	TufA (1DG1[24])	6
	MetE (1U22[25])	7
	IcdA (1P8F[26])	2
	AhpC (1YEP[27])	1
	CspC (1MJC[28])	3
	Ppa (2EIP[29])	1
	GapA (1S7C[30])	1
	Eno (1E9I[31])	1
RNA	$tRNA^{Phe}$ (4YCO[23])	5
Metabolite	GLU	1436
	ATP	144
	FBP	225
	GSH	255
Solvent	Water	306221
Inorganic Ion	K^+	4602
	Mg^{2+}	400
	Cl^-	1320

process. We need to use a box size that is big enough to allow the random insertions to succeed without structural overlapping and without creating artificial interactions between the elements, that is, we need L to be big enough so that different elements are not placed too close from each other. On the other hand, as the box gets bigger, there will be more space between the components and the system will be harder to equilibrate due to larger variations in box size.

We devised an iterative process that solves both problems simultaneously. We start with a box size L that is too small to allow all systems to fit in the box by random insertion. In our case, we started with L=30nm. Then, we allow 100 insertion trials for each element. If all trials fail for any of them, we increase the box by a dL step of 1nm and start again until all insertions succeed. In our case, all insertions succeeded after increasing L to 35nm. Additionally, instead of adding only the protein, RNA or metabolite molecules in the empty box in each trial, we actually add a droplet of water and counter ions in which the molecule of interest is embedded. Such droplets are taken from molecular dynamics simulations in

which each component was previously simulated. The benefits of using such droplets are threefold: i. it prevents artificial contacts between the components as consequence of packing in the simulation box because the water droplet acts as natural protecting layer. ii. it is a natural way to place water molecules and counter-ions in the simulation box around each component. iii. the components are already pre-equilibrated, which will help us to perform the equilibration of the whole simulation box.

In order to do this, the droplets around each component are also constructed iteractively. The number of water molecules that we must place in the simulation box is known. However, we don't know the thickness of the water layer around each component, l, that accounts for such amount of water molecules. Since l depends on a series of factors such as the shape, size and abundancy of each element, we define it iteratively. We start by taking a droplet of thickness l=3A around all elements and counting the number of water that we would add to the box. If it is smaller than the number we need, l is increased by dl=1A. If it is larger, we take a step back and reduce dl by 10 times. We can carry this process until an arbitrary precision cutoff is satisfied.

¡ IMAGE SHOWING DROPLETS, l, dl, THE EMPTY BOX AND INSERTIONS, L, dL ¿

After all droplets are successfully inserted in a simulation box by the iterative process described above, we proceed to add ions to neutralize the net charge of the system and to reach the desired ionic strength. Then, we perform energy minimization and a short simulation step of 500ps in which all molecules of the box are free to move. In this step the box shrinks to its optimum size. In our case, the simulation box shrank from the initial L=35nm to L=22.9nm. From this point, the system is ready to be submitted to the default simulation steps such as thermalization and production run (please check the materials section for details about the parameters we used).

Python scripts and all topology and structure files necessary to perform all these steps are available at github.com/dspoel... The cytoplasm model that we built here can be used as

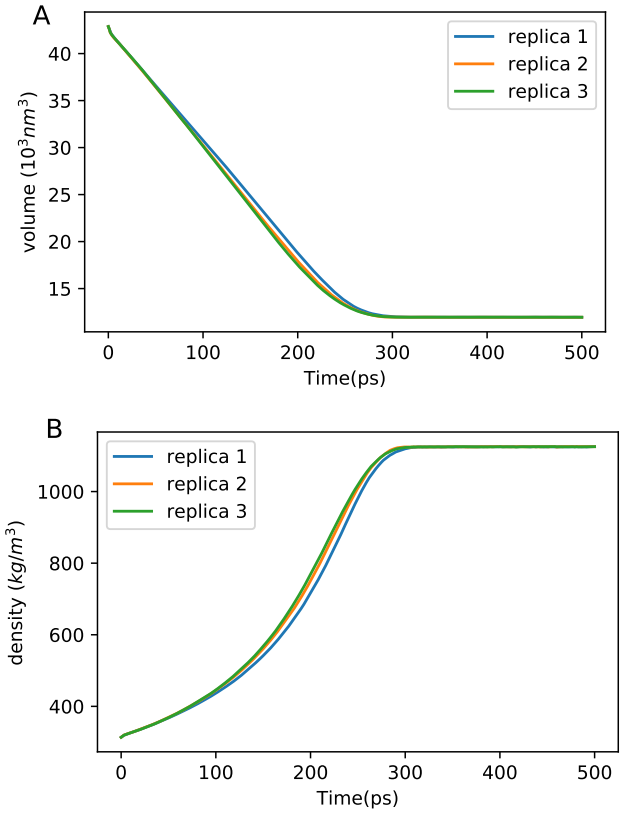

Figure 1: A) The volume of the box with length size of $L = 35\ nm$ reduces during the energy minimization in the first 300 ps B) The density of the same box increases within the first 300 ps. The repeated experiments are shown in different colors.

a tool to answer many research questions about the effects of crowding on specific systems of interest. With the files we are providing here it is possible, for example, to add a probe protein to investigate the effect of crowding on it, to add new macromolecular or small crowders, add new solvents and change the biomolecular concentration to increase or decrease the intensity of the crowding effect.

The effect of crowded systems

In order to investigate the effects of crowding in the E. coli cytoplasm model on the behavior of its elements, we constructed three boxes that have different orientations for each of its elements and submitted each to molecular dynamics simulations of 1 us. We also performed 200 ns molecular dynamics simulations for each isolated element (please check the materials section for details about the parameters).

Translational diffusion

D/D0 for all crowders

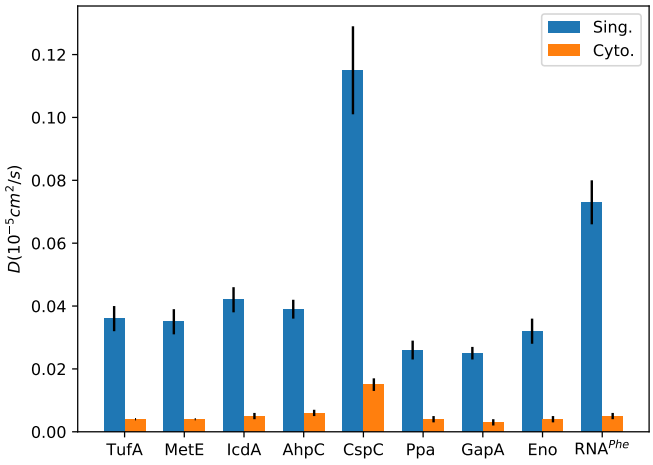

Figure 2: Diffusion coefficient of crowders from single molecule simulations (blue bars) and cytoplasm simulations (orange bars).

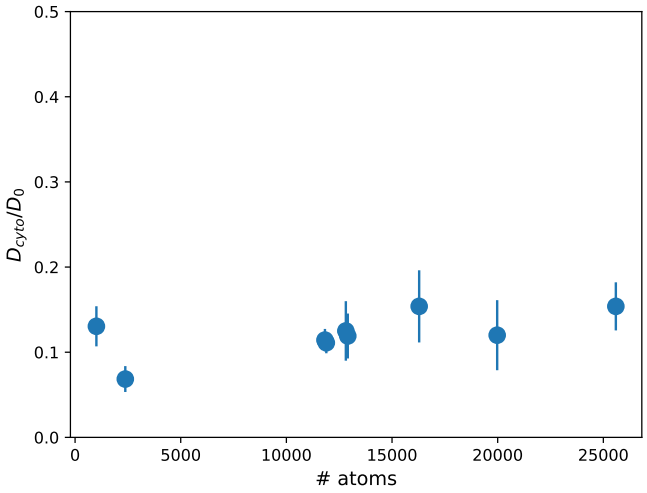

Figure 3: The ratio between diffusion coefficient obtained from cytoplasm to single molecule simulations. The crowders are sorted according to their sizes on the x-axis.

Structural integrity of the crowders

RMSD and SASA plots, show that the crowders didn't unfold

Structural dynamics

RMSF. Are there differences between the soup and single sim.?

Oligomers under crowded condition

Aggregation

We observed aggregation of the metabolites that contain phosphate, ATP and FBP, with Mg2+ that was used as initially used as counter-ion for ATP and tRNA. After running simulations of small boxes containing just these metabolites and Mg2+, we found that completely protonating their phosphate groups can prevent aggregation. Thus, we advise caution when dealing with small molecules that are highly charged, such as ATP and FBP. Trial simulations with small boxes in which these metabolites are in high concentration to verify if they'll aggregate with Mg2+ or other charged species that are present in the box. In the specific case of ATP and FBP modeled with GAFF in boxes containing Mg2+, we advise that they

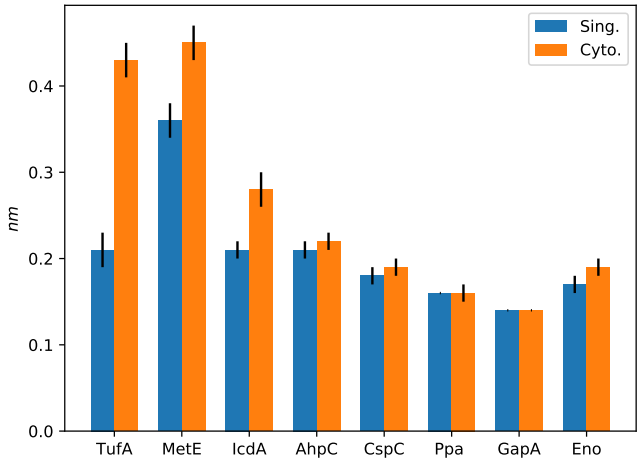

Figure 4: Root mean squared displacement (rmsd) for 8 proteins from single molecule simulation (blue bars) and cytoplasm simulation (orange bars). The average rmsd for each protein in single molecule simulation is over number of chains of proteins and three replica. The average rmsd for each protein in cytoplasm simulation are over number of chains of proteins, number of appearance and three replica. The error bars show the standard errors.

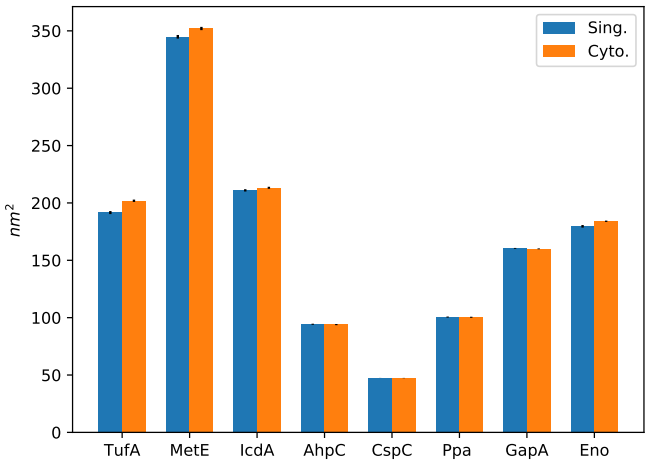

Figure 5: Solvent accessible surface area (sasa) for 8 proteins from single molecule simulation (blue bars) and cytoplasm simulation (orange bars). The average sasa for each protein in single molecule simulation is over number of chains of proteins and three replica. The average sasa for each protein in cytoplasm simulation are over number of chains of proteins, number of appearance and three replica. The error bars show the standard errors.

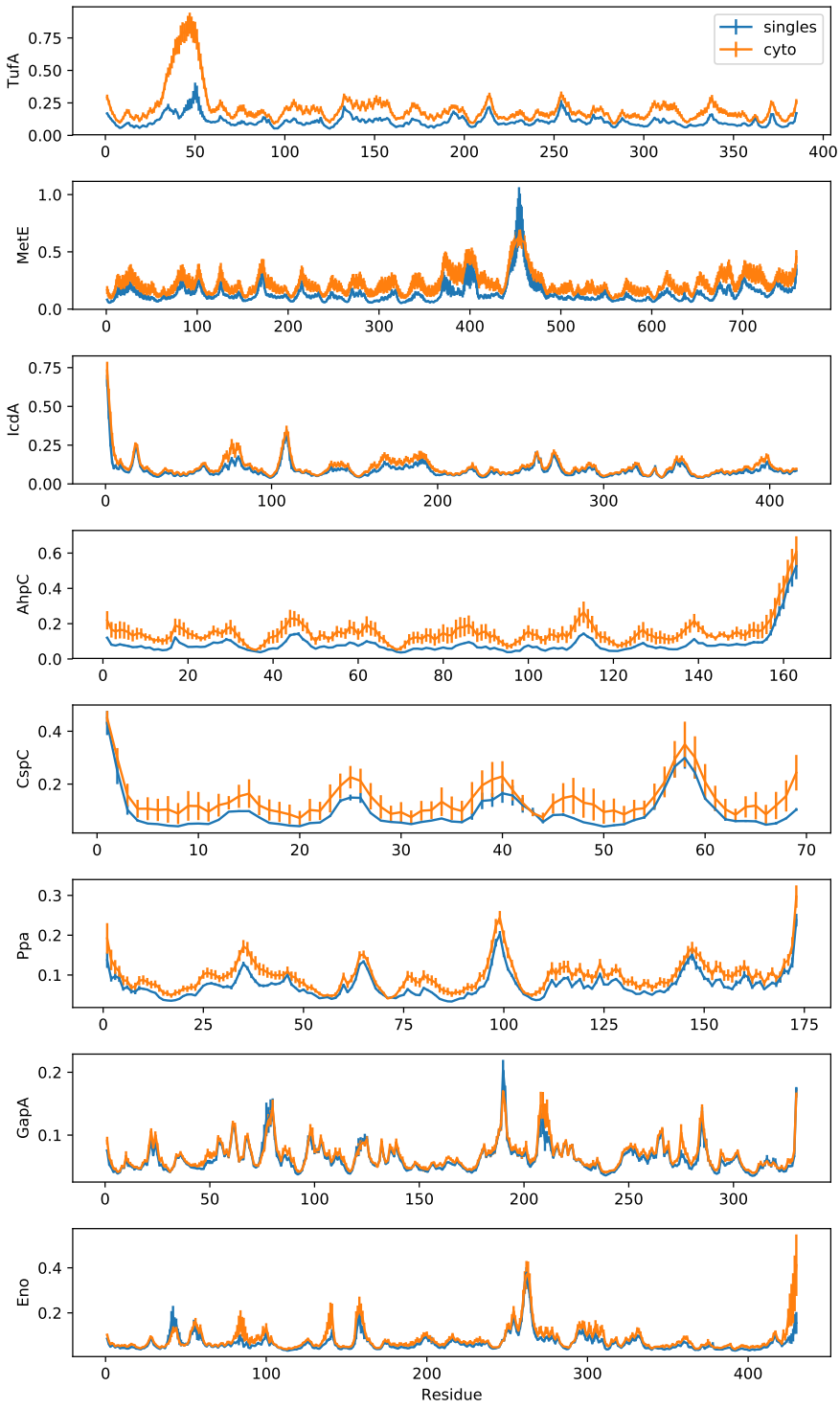

Figure 6: Root mean squared fluctuations (rmsf) for 8 crowders from single molecule simulation (blue bars) and cytoplasm simulation (orange bars). x-axis shows the residues for each chain of proteins. The average rmsf for each protein in single molecule simulation is over number of chains of proteins and three replica. The average rmsd for each protein in cytoplasm simulation are over number of chains of proteins, number of appearance and three replica. The error bars show the standard errors.

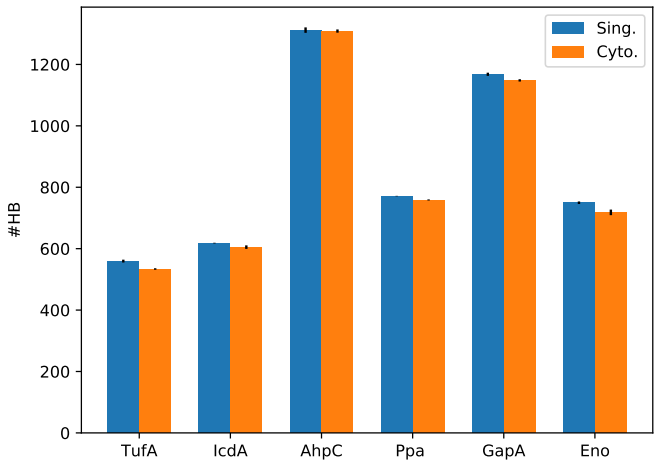

Figure 7: Intramolecular hydrogen bonds (hb) for oligomers from single molecule simulation (blue bars) and cytoplasm simulation (orange bars). The average hydrogen bonds for each oligomer in single molecule simulation is over three replica. The average hydrogen bonds for each oligomer in cytoplasm simulation is over number of appearance of the oligomer in three replica. The error bars show the standard errors.

should be completely protonated even though that is not their physiological protonation state. The topology and structure files we are providing for ATP and FBP on github are already protonated.

¡ FIGURE SHOWING THE AGGREGATION ¿ ¡ FIGURE SHOWING THAT PROTONATED PHOSPHATE DOESN'T AGGREGATE ¿

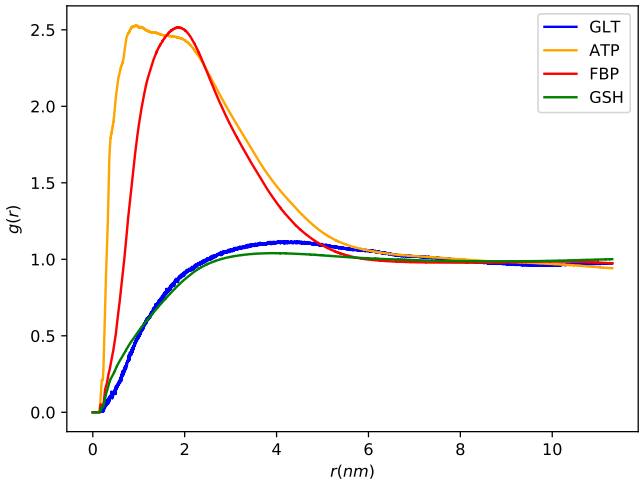

Figure 8: Radial distribution function showing the probability of finding a metabolite around RNA molecules.

Figure 9: Radial distribution function (rdf) showing the probability of finding a metabolite or an RNA around MG2+. The inset plot shows the rdf for short distance ($r < 0.6\ nm$)

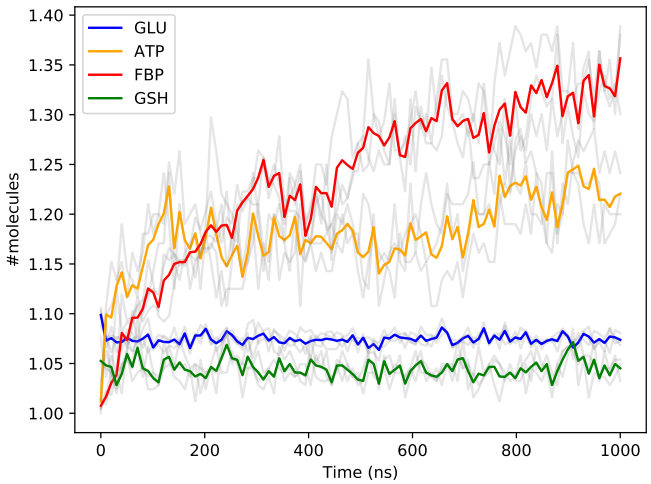

Figure 10: Number of metabolite molecules that form a cluster in cytoplasm simulations. The solid colored lines are the average of three replica (shown in grey).

Discussion

- Use of force field should be taken with care due to aggregation that was reported repeatedly. We used AMBER scaled force field for protein water interaction. In other atomistic model of Mycoplasma genitalium, the CHARRM scaled protein water interaction was used for the same reason.

- No significant structural changes (such as unfoldin or partial unfolding) can be seen from properties such as sasa (fig 5), and gyration radius (WE HAVEN'T ADD THIS FIG, SHALL WE?).

- intramoleclar protein protein interactions are not affected by the presence of crowders. This shows for oligomers (1,3,4,6,7,8 NAME THEM), the crowding doesn't cause dissociation of chains.

- intermolecular protein protein interactions can be seen in forms of rmsd (fig 4) and rmsf (fig 6). As proteins interact with their environment, their structure flutuate more

than in dilute cases. However, the effect is not large enough to cause unfolding (fig 5, and gyration)

- rmsf in all cases are higher for crowders in cytoplasm than in dilute codition
- AREN'T RMSD AND RMSF CORRELATED?
- Diffusion coefficient is significantly different, and it's independent of the size of the proteins (D cyto/D single plot in Supp. Info.). Based on the observations above (sasa, gyration, rmsd, rmsf) the drop is, to a greater amount, due to the excluded volume effect, and to a lesser amount, due electristatic interactions that has been compromised because of choice of force field.
- the greater drop in diffusion coefficient of RNA is due to confinement of RNA by metabolites (fig 9 and 10).
- NOT SURE ABOUT THIS: THE FACT THAT METABOLITES ARE POLYMERIZED DOES NOT SABOTAGE THE WHOLE SIMULATION AND WE STIL COULD ANALYSE THE SIMULATION AND GET MEANINGFUL OUTCOME, AND THIS WOULD NOT (MOST PROBABLY) BE AFFECTED BY NON-CLUSTERED METABOLITE. INDEED WE DID NOT ANALYZE CHEMICAL OR DYNAMIC PROPERTIES OF METABOLITES

References

(1) Reckel, S.; Hänsel, R.; Löhr, F.; Dötsch, V. In-cell NMR spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy* **2007**, *2*, 91–101.

(2) Pielak, G. J.; Li, C.; Miklos, A. C.; Schlesinger, A. P.; Slade, K. M.; Wang, G.-F.; Zigoneanu, I. G. Protein nuclear magnetic resonance under physiological conditions. *Biochemistry* **2008**, *48*, 226–234.

(3) Ignatova, Z.; Gierasch, L. M. Monitoring protein stability and aggregation in vivo by real-time fluorescent labeling. *Proceedings of the National Academy of Sciences* **2004**, *101*, 523–528.

(4) Xie, X. S.; Choi, P. J.; Li, G.-W.; Lee, N. K.; Lia, G. Single-molecule approach to molecular biology in living bacterial cells. *Annu. Rev. Biophys.* **2008**, *37*, 417–444.

(5) English, B. P.; Hauryliuk, V.; Sanamrad, A.; Tankov, S.; Dekker, N. H.; Elf, J. Single-molecule investigations of the stringent response machinery in living bacterial cells. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, E365–E373.

(6) Spiga, E.; Abriata, L. A.; Piazza, F.; Dal Peraro, M. Dissecting the effects of concentrated carbohydrate solutions on protein diffusion, hydration, and internal dynamics. *J. Phys. Chem. B.* **2014**, *118*, 5310–5321.

(7) Harada, R.; Sugita, Y.; Feig, M. Protein crowding affects hydration structure and dynamics. *J. Amer. Chem. Soc.* **2012**, *134*, 4842–4849.

(8) Nawrocki, G.; Wang, P.-h.; Yu, I.; Sugita, Y.; Feig, M. Slow-down in diffusion in crowded protein solutions correlates with transient cluster formation. *J. Phys. Chem. B.* **2017**, *121*, 11072–11084.

(9) McGuffee, S. R.; Elcock, A. H. Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. *PLoS Comput. Biol.* **2010**, *6*, e1000694.

(10) Yu, I.; Mori, T.; Ando, T.; Harada, R.; Jung, J.; Sugita, Y.; Feig, M. Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm. *eLife* **2016**, *5*.

() van Gunsteren, W. F. et al. Biomolecular modeling: Goals, problems, perspectives. *Angew. Chem.-Int. Edit.* **2006**, *45*, 4064–4092.

(11) Chys, P.; Chacón, P. Random coordinate descent with spinor-matrices and geometric filters for efficient loop closure. *Journal of chemical theory and computation* **2013**, *9*, 1821–1829.

(12) Best, R. B.; Zheng, W.; Mittal, J. Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association. *J. Chem. Theory Comput.* **2014**, *10*, 5113–5124.

(13) Abascal, J. L. F.; Vega, C. A general purpose model for the condensed phases of water: TIP4P/2005. *J. Chem. Phys.* **2005**, *123*, 234505.

(14) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577–8592.

(15) Hess, B.; Kutzner, C.; Van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(16) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

(17) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

(18) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford Science Publications: Oxford, 1987.

(19) Yeh, I.-C.; Hummer, G. System-size dependence of diffusion coefficients and viscosities from molecular dynamics simulations with periodic boundary conditions. *J. Phys. Chem. B.* **2004**, *108*, 15873–15879.

(20) Dong, H.; Nilsson, L.; Kurland, C. G. Co-variation of trna abundance and codon usage inescherichia coliat different growth rates. *"J. Mol. Biol."* **1996**, *260*, 649–663.

(21) 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. *"Nature Chem. Biol."* **2009**, *5*, 593–599.

(22) Link, A. J.; Robison, K.; Church, G. M. Comparing the predicted and observed properties of proteins encoded in the genome of Escherichia coli K-12. *Electrophoresis* **1997**, *18*, 1259–1313.

(23) Byrne, R. T.; Jenkins, H. T.; Peters, D. T.; Whelan, F.; Stowell, J.; Aziz, N.; Kasatsky, P.; Rodnina, M. V.; Koonin, E. V.; Konevega, A. L. Major reorientation of tRNA substrates defines specificity of dihydrouridine synthases. *Proceedings of the National Academy of Sciences* **2015**, *112*, 6033–6037.

(24) Abel, K.; Yoder, M. D.; Hilgenfeld, R.; Jurnak, F. An α to β conformational switch in EF-Tu. *Structure* **1996**, *4*, 1153–1159.

(25) Ferrer, J.-L.; Ravanel, S.; Robert, M.; Dumas, R. Crystal structures of cobalamin-independent methionine synthase complexed with zinc, homocysteine, and methyltetrahydrofolate. *Journal of Biological Chemistry* **2004**, *279*, 44235–44238.

(26) Mesecar, A. D.; Koshland Jr, D. E. Structural biology: A new model for protein stereospecificity. *Nature* **2000**, *403*, 614–616.

(27) Parsonage, D.; Youngblood, D. S.; Sarma, G. N.; Wood, Z. A.; Karplus, P. A.; Poole, L. B. Analysis of the link between enzymatic activity and oligomeric state in AhpC, a bacterial peroxiredoxin. *Biochemistry* **2005**, *44*, 10583–10592.

(28) Schindelin, H.; Jiang, W.; Inouye, M.; Heinemann, U. Crystal structure of CspA, the major cold shock protein of Escherichia coli. *Proceedings of the National Academy of Sciences* **1994**, *91*, 5119–5123.

(29) Kankare, J.; Salminen, T.; Lahti, R.; Cooperman, B.; Baykov, A.; Goldman, A. Structure of Escherichia coli inorganic pyrophosphatase at 2.2 A resolution. *Acta crystallographica. Section D, Biological crystallography* **1996**, *52*, 551–563.

(30) Shin, D.; Thor, J.; Yokota, H.; Kim, R.; Kim, S. Crystal structure of MES buffer bound form of glyceraldehyde 3-phosphate dehydrogenase from Escherichia coli. To be published.

(31) Kühnel, K.; Luisi, B. F. Crystal structure of the Escherichia coli RNA degradosome component enolase. *Journal of molecular biology* **2001**, *313*, 583–592.