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,[? ?] or fluoroscence spectroscopies.[? ? ?] 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.[?][?][?] However, more attemps in modeling bacterial cytoplasm have been made recently.[?][?]

- 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[?] 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[?] in combination with the TIP4P/2005 water model.[?] Electrostatic interactions were treated using the particle mesh Ewald algorithm.[?] All chemical bonds were constrained at their equilibrium length using the LINCS algorithm[?] allowing an integration time step of 2 fs. Temperature was controlled at 310 K using the v-rescale algorithm[?] and a coupling time of 0.5 ps. The pressure was controlled at 1 bar using the Parrinello-Rahman algorithm[?] 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.[?] 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[?] 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 DNA 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.*[?] [?] [?] [?]

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.*[?] The two most abundant proteins, TufA and MetE, account for about 20% and 12% of the total protein abundance in *E. coli, respectively, while the other six proteins contribute with less than 5% each (Table 1). Their crystallogaphic structures were used to insert them in the cytoplasm model. The number of copies of each protein in the model was calculated as the rounded up fold increase in their abundance when compared to the least abundant protein, which was set to have the copy number of 1. Then, this number was divided by the oligomeric state of each protein. This was done because the original oligomeric state for each protein was kept as reported in their crystallogaphic structure.*

RNA fraction

Transporter RNAs (tRNAs) account for 74% of the dry weight of non-ribosomal RNAs. Thus, we chose to model the RNA presence in the cytoplasm with tRNA molecules. Specifically, we considered the tRNA(Phe) molecule as a representative of tRNAs due to the availability of a recent crystallographic structure.[?] *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%*

Table 1: Absolute and cumulative fraction of the total abundance of non-ribosomal proteins in the cytosol of *E. coli* K12 to which each protein that was selected to compose our cytoplasm model corresponds to.

Protein	Fraction [%]	
	Absolute	Cumulative
TufA	19.7	19.7
MetE	11.6	31.4
IcdA	4.7	36.1
AhpC	4.1	40.2
CspC	4.0	44.2
Ppa	2.9	47.0
GapA	2.1	49.1
Eno	1.9	51.1

of the total protein weight. This protein/RNA weight ratio was used to calculate the correct number of tRNA(Phe) molecules that were added to the cytoplasm model (Table XXX).

Metabolites fraction

We considered the most abundant molecules of each metabolite class as representatives, i.e. Glutamate for amino acids, ATP for nucleotides, FBP for central carbon intermediates and Glutathione redox cofactors .[?] *The total number of 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 .*[?] *The copy number for each molecule was calculated from the ratios between their experimentally observed concentration in* E. coli .[?]

Water fraction

The number of water molecules was calculated according to the desired biomolecular concentration, which is a parameter of the cytoplasm model. This number 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% was inserted into the cytoplasm model.

Inorganic ions fraction

Finally, Mg^{2+} was used as counter-ions for tRNA and ATP. KCl was added to neutralize the charges of the simulation box and to reach the ionic strength of 0.150 mol/L by substution of randomly selected water molecules.

Table 2: Number of copies for each component of the cytoplasm model built at the biomolecular fraction of 30%

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

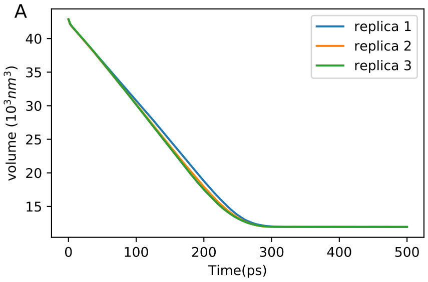
Building the simulation box

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 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 by placing them too close from each other. On the other hand, as the box gets bigger, it gets harder to equilibrate the system because the empty space between the components will induce the barostat to reduce the box volume quickly during the simulation.

We devised an iterative process that solves both problems simultaneously. We start with a box size L that is too small to allow all components to fit in the box by random insertion. In our case, we started with $L = 30$ nm. Then, we allow 100 insertion trials for each element. If all trials fail for any of them, we start the whole process again with an empty cubic box that is larger by a step size dL of 1 nm. We repeat this process until all insertions succeed. In our case, all insertions succeeded after increasing L to 35 nm. 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 equilibrated. The benefits of using such droplets are threefold: *i.* it acts as natural protecting layer that prevents artificial contacts between the components that could be created due to the random insertions. *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 cytoplasm model.

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 from the desired biomolecular fraction, which is a parameter of the model, and the total biomolecular mass of the system. 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 $l = 3$ Å around all elements and counting the number of water molecules that we would add to the box with such droplets. If it is smaller than the number we need, l is increased by $dl = 1$ Å. If it is larger, we reduce the droplet thickness by dl, reduce dl by 10 times and then increate l with the new smaller dl. We can carry this process until an arbitrary precision cutoff, such as 5%, is satisfied.

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 box size of 35 nm to 22.9 nm, which correspond to a volume change from 42875 nm^3 to 12009 nm^3. 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).



□box_big.png □box_small_.png

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.

Python scripts and all topology and structure files necessary to enable any researcher to build its own cytoplasm models are publicly available at github.com/dspoel/soup. Such models can be used as tools that answer many research questions about the effects of crowding on specific systems of interest. With the files we are providing 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, verify the effects of different temperatures 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 structural integrity and dynamics of its elements, we constructed three completely independent cytoplasm models that have different orientations for each of its elements. We submitted these systems, each composed by more than 1.5 million atoms, to molecular dynamics simulations of 1 µs. We also performed 200 ns molecular dynamics simulations for each isolated element in order to represent the non-crowded, i.e. dilute, environment (please check the materials section for details about the parameters).

Translational diffusion

Under crowded conditions, molecules are constrained to a smaller volume due to collisions with the other components of the crowded environment. This extent of this effect is evident when we compare the translational diffusion constant of each component in the cytoplasm model, $D_{cytoplasm}$, and in a dilute system, D_{dilute} (Figure **??**A). The ration of these quantities is independent of the protein size and is always close to 0.13 (Figure **??**B). However, the effect of crowding over the translation diffusion of tRNA is two times higher, indicating that its movement is specially constrained in the crowded environment.

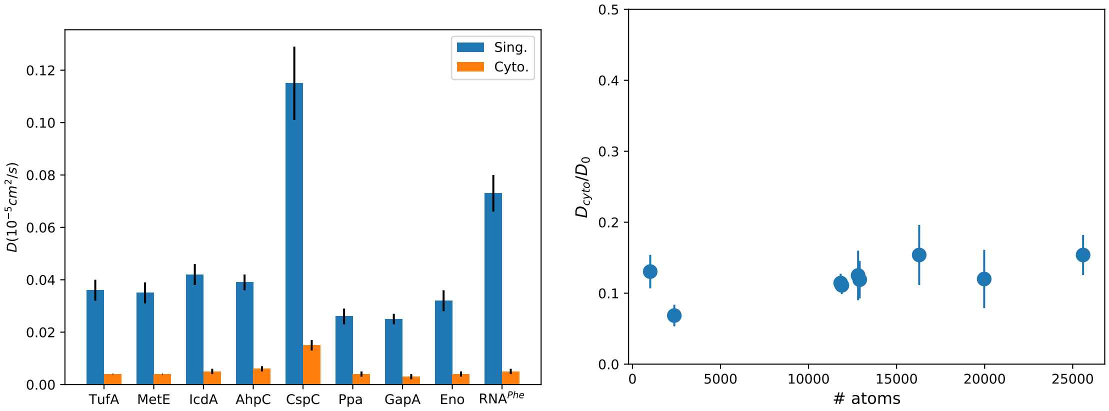


Figure 2: A)Diffusion coefficient of crowders from single molecule simulations (blue bars) and cytoplasm simulations (orange bars). B) The ratio between diffusion coefficient obtained from cytoplasm to single molecule simulations. The crowders are sorted according to their sizes on the x-axis. << Change the x-axis of D/D0 plot to kDa >>

Are we going to calculate and discuss rotational diffusion?

tRNA is aggregating

After inspecting the trajectories of the cytoplasm models, we found that the reason why the translational diffusion constant for tRNA was reduced to a greater extent than for the other components of the cytoplasm model is that it is forming aggregates with Mg^{2+}, ATP and FBP (Figure **??**). Thus, in the next sections we will ignore the RNA molecules in the analyses about the structural integrity of the crowders. Additionally, in the last section we will further investigate such aggregation and we will show a way to prevent it.

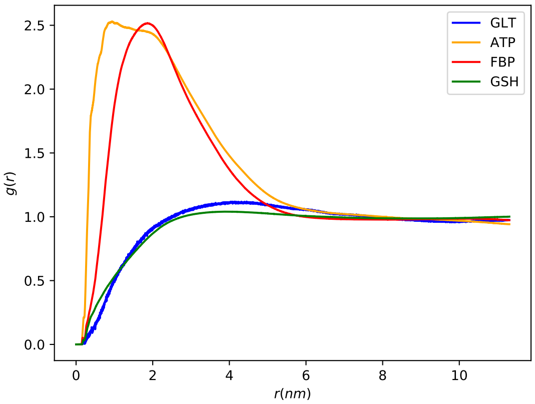


Figure 3: Radial distribution function showing the probability of finding metabolites or cations around RNA molecules in the cytoplasm model.

Structural integrity of individual chains

The structure of the individual chains of the crowders that were used in our cytoplasm model were not affected to a great extent by the crowded environment. The solvent accessible surface area (SASA) value for all crowders in the cytoplasm model is within 5% of its value under the dilute condition. The root mean square deviation (RMSD) considering only their Cα atoms show a similar behavior, except for TufA, MetE and IcdA, which have RMSD values 105, 25 and 33% higher in the crowded environment than in the dilute condition. Visual inspection of their trajectories showed that there is partial unfolding of their N- or C-terminal residues DESCRIBE CONF. CHANGES.

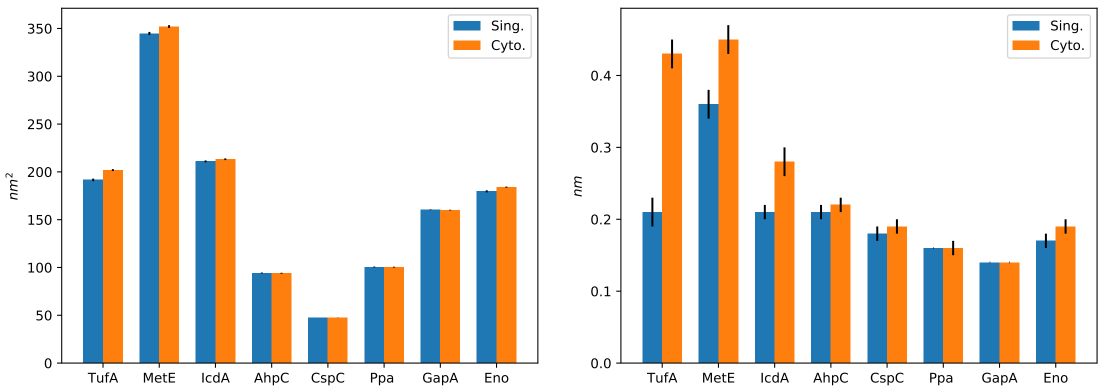


Figure 4: A) Solvent accessible surface area (sasa) for 8 proteins from single molecule simulation (blue bars) and cytoplasm simulation (orange bars). The average sasa for each protein in single molecule simulation is over number of chains of proteins and three replica. The average sasa for each protein in cytoplasm simulation are over number of chains of proteins, number of appearance and three replica. The error bars show the standard errors. B) 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.

Structural integrity of oligomers

WE STILL NEED TO MAKE THIS PLOT

Structural dynamics

Proteins are slighly more flexible in the crowded environment. Root mean square fluction (RMSF) calculations show that the flexibility profile of the proteins is not affected by crowding, except in the cases in which there are conformational changes, TufA ... as shown before (Figure ??). This is in agreement with our results showing that the structural integrity of the proteins is not significantly affected. In addition to that, our results show that the proteins are slighly more flexible in the crowded environment than in the dilute condition (Figure ??). This is likely due to the nonspecific interations and collisions that often happen under crowding conditions.[?]

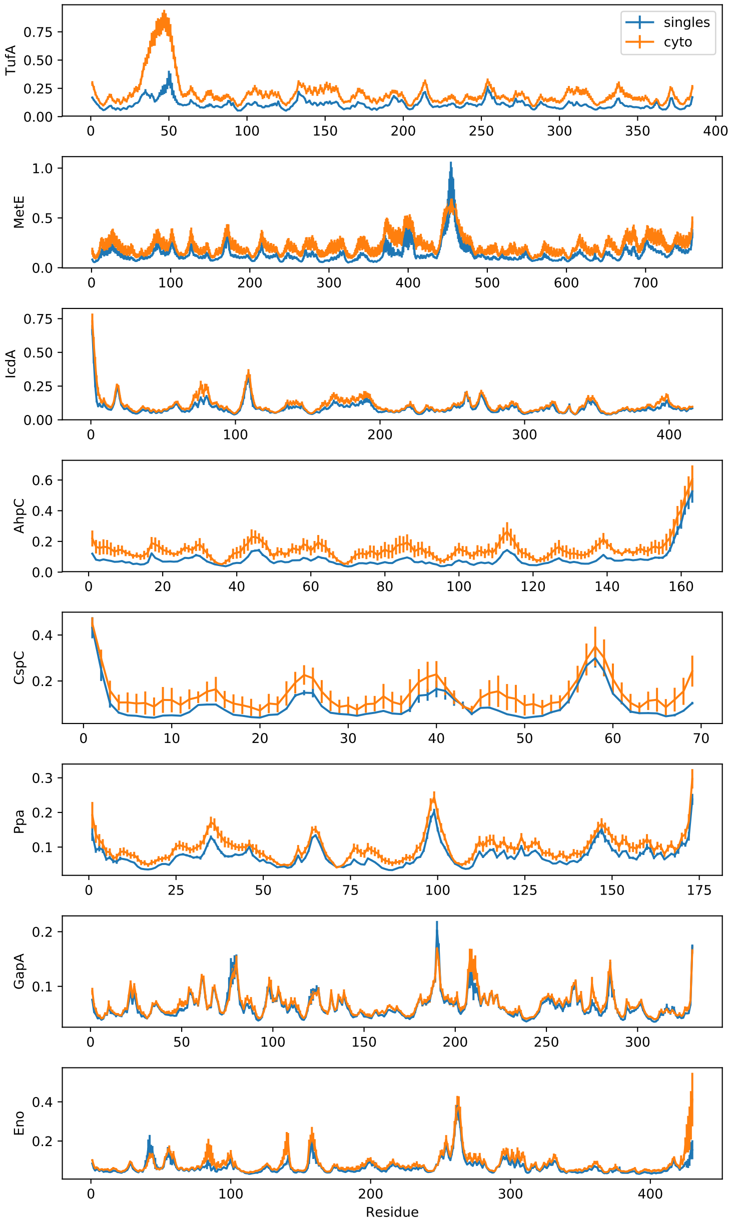


Figure 5: Root mean squared fluctuations (rmsf) for 8 crowders from single molecule simulation (blue bars) and cytoplasm simulation (orange bars). x-axis shows the residues for each chain of proteins. The average rmsf for each protein in single molecule simulation is over number of chains of proteins 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. << Make this plot dividing the plots in 2 columns and 4 rows to save space>>

Aggregation can be avoided by protonating the metabolites

The initial detection of aggregation in our cytoplasm model was due to the outlier behavior of the translational diffusion constant of tRNA when compared to the other crowders. However, tRNA is not causing such phenomenon. Instead, it is triggered by the exaggerated interaction between the phosphate-containing metabolites, ATP and FBP, with Mg^{2+}. tRNA gets involved in the aggregates because Mg^{2+} is present in the box only as counter-ion for ATP and tRNA, and so tRNA is one of the few ”sources” of Mg^{2+} in the whole system. We looked for ways of avoind aggregation with simulations of small systems composed by these metabolites in high concentration, Mg^{2+} and water (please check the methods section for details about the parameters we used).

We found that completely protonating the phosphate groups of ATP and FBP is enough prevent their aggregation with Mg^{2+} (Figure **??**). While RDFs show that ATP^{-3} and FBP^{-4} aggregate around Mg^{2+}, this is not observed for $ATPH_3$ and $FBPH_4$. We further tested this hypothesis by

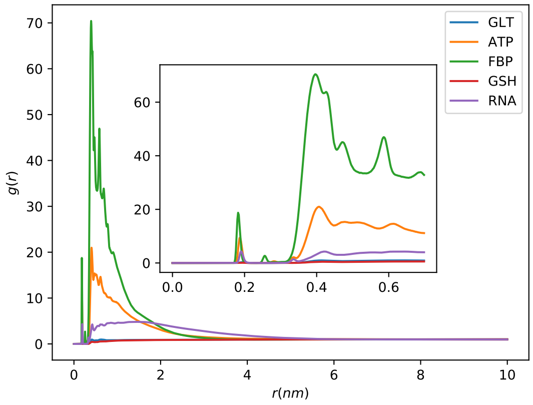


Figure 6: Radial distribution function (rdf) showing the probability of finding a metabolite or an RNA around MG2+. The inset plot shows the rdf for short distance ($r < 0.6\ nm$)

Discussion

In the specific case of ATP and FBP modeled with GAFF in boxes containing Mg^{2+}, we advise that they 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.

<< Discuss the translational diffusion results:

what is the meaning of the straight line in D/D0? Is this expected?

Is this in agreement/disagreement with which models/theories? >>

- 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

- RMSFs for cyto have higher statistical error ! (even if we have more copies in cyto.) e.g. crowder-n5

- 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

Conclusions