A SIMPLE MODEL TO STUDY A COMPLEX CYTOPLASM

{Picture of main idea}

{Logos}

Work by: Mubarrat Mahin Mursalin

Under the supervision of: Stefan Klumpp

Title
-A simple model to study the complex cytoplasm

Preface
-this work was done for the partial completion of {CITE}

Introduction

The main goal of this paper was to better understand the nature of the inside of a cell. Cells are the basic unit of life and therefor we want to study their complex behavior. Starting from an atomic level description we can build to the cellular level but we lack meaningful methods to quantitatively study systems with such detail. An accurate physical description of the inside of the cell is not only of interest to physicists, but also chemists and biologists, and progress on this topic can lead to innovations in future technologies.

To understand why a studying the inside of a single cell is of importance and interest, we can start with the publications that founded biology. In his works {CITE} {CITE} Charles Darwin describes how multiple species could have come to exist by natural selection. He hypothesized this along his travels across the world, where he was given the opportunity to witness many of these organisms. He noted many physical similarities between different species, different phenotypes. He said these different features served different purposes and they had evolved to do so, because congruent niche existed in the environment where these traits could be useful. A phenotype is beneficial to the organism, if it helps in regards to its survival and reproduction, then it will be persistent in time, and and also said phenotypes are selected for by selection. This idea was controversial, but biological evidence such as homologous structures helped prove Darwin’s hypothesis {CITE}. The question then became, what is the actual mechanism in which the information about these phenotypes are passed on to offspring.

This would only happen later on when DNA was found to be the the code for living systems (that we know of). Before that, the term cell was coined by Luenhook who developed the first microscopes {CITE}. After, cells were found to be the basic unit of all living things,and more complex life is just collections of cells. The central dogma of molecular biology has since been the paradigm for thinking in biology on the small scale {CITE}. The central dogma states that the genome of an organism is defined by its DNAs, then in cell this DNA is read and used as RNAs. RNAs then make proteins. Proteins effect what we see phenotypically, that is to say, protein content controlled by gene expression,leads to the different types of cells we see. The understanding of this is stated as trivial now, but many minds toiled for many years to achive this level of understanding. Some key discoveries along that pat include works on Nucleic acid concentrations {CITE}, the structure of DNA {CITE}, discovery of the ribosome as a machine for making proteins from RNA {CITE} are just some examples. To limit the discussion to biophysics, yet tie it to general biology, we can say that biology is concerned with the dynamics of living things, which based on their genetic and the external environment.

This is different from physics which is concerned with the motions of objects in time. The most famous early works of physics is Isaac Newton understanding an apple falls{CITE}, and his laws governing the motion of objects{CITE}. Since then physics has been concerned with finding equations of motion or dynamics, from the smallest systems to the largest. Specifically quantum mechanics{CITE} and general relativity{CITE} have been extremely successful at making predictions in the small and large scales respectively. The works of Albert Einstein that are commonly known are that of the Energy Mass equivalency, (E=MC^2). However, his lesser known work regarding the photoelectric effect and Brownian motion are what won him the Nobel prize{CITE}. There are many directions in the study of physics, and often they are not related to each other because of the difference in scales. Biophysics is the study of the motions of living objects in time. It is founded on statistical physics. In the theory section, we will use Einstein’s work on Brownian motion as a jumping off point to connect statistical physics and biological systems.

The story would be incomplete without mention of the Chemistry that connects the physical descriptions of matter to the the diversity and beauty in biology. Chemistry is concerned with the properties and behavior of matter.

Just as biology classifies organisms, Chemistry classifies chemicals by their properties. During the 19th century many experiments were done in which new materials were created from naturally occurring matter. These were verified to be pure elements gasses like oxygen and hydrogen {CITE} and metals like Ytterbium {CITE}. Properties and interactions of these new elements were tabulated, and trends could be found. These trends are best seen with the periodic table of elements{CITE}, which places all elements in a table according to similar reactivities and increasing size. Although many of the microscopic details about the elements, such as atomic number or electron configuration were not know until later, detailed descriptions of the nature of the elements, as well as their combinations were known. Chemical reactions allow for the combination and separation (or further convolutions of) different elements. The space of all these combinations is finite, yet so huge that subfield of chemistry exist just to study specific types of combinations. Naturally, biochemistry exists to quench the thirst for knowledge about how living things work based on chemicals. Furthermore, chemistry can neglect many of the microscopic details of the system, and has a language for interactions{CITE}. These facts make chemistry central in the study of how matter becomes life.

From first glance, it might seem very complicated and futile to study the microscopic structure and dynamics of the inside of a cell. As discussed, the system can be studied from three perspectives, biology chemistry and physics which all seem to provide different views. Physically, the insides of cells are a complex mess of many molecules with many shapes and sizes and compositions. Chemically, a few elements combined in a huge number of ways, react and interact to give rise to the functions we need to survive. Biologically, every part of the system is selected for, and is part of a huge network, both in a cell and an ecosystem. The time scales involved also vary from the shortest fastest quantum dynamics, to the lifetimes of entire organisms.

However, on the contrary, studying the cytoplasm (the inside of a cell), can help answer how ordinary matter becomes life. From a physical perspective, equilibrium statistical mechanics is not sufficient to describe living systems {CITE}. Biochemical reaction networks exist to describe pathways in a cell{CITE}, but strives to holistically understand cell functions have been limited. On a biological scale, a quantitative description of whole cells would set a standard for the field. Already today technologies exist which exploit our knowledge of these fields to save lives, lessen suffering, and help on our quest to understand life. Advances like molecular medicine{CITE}, gene editing{CITE}, biomedical engineering and synthetic biochemistry{CITE} are modern examples of how a multi-disciplinary approach to this topic can lead to huge success. In the paper, we will simplify the complex system of a cell into a simple physical model created from biological and chemical data. We hope to show that although many approximations have to be made to study such complex biological systems, quantitative information can still be found to the benefit of physicists, chemists and biologists.

Background & Theory

The background of this question starts with identifying cells as a structure in living things. Antonie van Leeuwenhoek, is credited with using his invention of the microscope to study all sorts of samples, some of which were living{CITE}. He established the field of microscopy by documenting, but also illustrating, what he saw through the lens. This would also set a trend in biology to classify things based on physical observations, also known as phenotypes. Leeuwenhoek also coined the term Cell after his observations of a sample resembled the cells monks reside in. This further goes to show how biology relies on phenotypic descriptions to segregate. Perhaps the most important idea to stem from microscopy is that all living things are made of cells. Although Charles Darwin's work was based on plants and animals, his ideas of evolution of species by natural selection also holds true for cellular life. Furthermore, the components of these living systems are also selected for{CITE}.

After the establishment of microscopy, scientists studied microscopic systems without the biological classifications nor chemical information we have today. Indeed, this information started from early observations of the nature of microscopic samples. One fact was that some samples were active as compared to others, which were much more inert. This helped convey the idea that microscopic life exists. Another specific observation was a distinction between life made of many cells, and single cellular organisms. Again, this was a very qualitative observation, but we now know that this distinction is actually of Eukaryotic and Prokaryotic cells.

Phenotypically Eukaryotes and Prokaryotes are very different. Not all Eukaryotes are multi cellular, but they all contain internal features that prokaryotes do not. A single cellular organism has to do all the activities associated with life, such as using energy for work consuming food and reproduction only with the machinery it has inside of it. The cells of multi-cellular organisms like you and I, have the luxury of not having to fulfill all these roles by themselves. Multi-cellular organisms can let cells specialize to have specific functions, like muscle cells which help us move or liver cells which help us metabolize. Of course these differences are based on the molecular level, that is to say, these function which the cells can do, are determined by the molecular machines they have. {CITE}

The distinction between prokaryotic and eukaryotic cells is of specific importance to us in this study. The exact composition of the inside of a cell will vary from cell to cell, even for cells of the same species. Since multi-cellular organisms are also composed of cells, a general rule of thumb is that the difference between cells of different living things will be related to how different their genomes are. As described exhaustively in biology,phonemic differences between all living things can be thought of as a (phylogenetic) tree. On the cellular level, the biggest distinction between the insides of cells come from the different domains of life on earth. The three domains bacteria, archaea, and eukaryota arose during the evolution of life, and consiquently each domain has features that are adapted for a certain lifestyle{CITE}. Neglecting arcaea, our discussion will focus on bacteria specifically because eukaryota have membrane bound organelles and are internally inhomogeneous{CITE}. It should be noted that cells themselves are membrane bound, and we are considering a cytoplasm as the mixture inside this membrane. In eukaryotic cells, the model which we describe later, can also be applied considering only systems enclosed by a membrane and not with multiple separate compartments. The inside of bacteria is homogeneous, except for a central nucleoid which is a DNA rich region. We will start thinking of the cytoplasm of a part of bacteria, inside the membrane but outside the dense nucleoid.

The magic of biology is that the information about the cell, its machines, an indeed a whole organism is written in its genome. Without going over all the discoveries in molecular biology, we will start with the work of Erwin Chargaff{CITE}. Chargaff's rules says that in a cell the concentrations of the chemicals adenine and thymine are roughly equal, as are the amounts of cytosine

and guanine. This was actually of great importance as these chemicals were known to make up DNA, a molecule found in cells, which was observed to related to cell division. Since cell divison is on of the steps in reproduction, a key activity of living things, DNA was suspected being the carrier of genetic information. This was truly described by the works of Watson and Crick {CITE}, who not only found the physical structure of the DNA molecule, but also probed its role as carrying the information which defines an organism, also known.

Deoxyribose nucleic acids (DNAs) are a double stranded polymer composed of the nucleic acids adenine thymine guanine and cytosine. The central dogma of molecular biology states that DNA is a hard code which defines an organism{CITE}. That is to say, two organisms are of the same species if they have the same genome. DNA however is not the working memory at the molecular level. That is the role of ribonucleic acids (RNAs), another polymer composed of combinations of the nucleic acids adenine thymine cytosine and uracil. While DNA is physically hard to bend and open, RNA is more flexible, which makes it easier for it to interact with other molecules{CITE}. In a process called transcription a segment of DNA is copied into RNA, which is analogous to copying information from a larger piece of text to use in a specific instance. These nucleic acid polymers are what compose the genome, or genotype of an organism.

The next step in understanding how the genotype becomes the phenotypes observed in organisms requires knowledge of how activities are done in cells. Although RNA and DNA can carry and move information, more complicated processes such as moving things and using energy require specific chemicals. Proteins are the molecules that do the large variety of work observed in living things. Proteins are also polymers but composed of amino acids, of which there are canonically 20. This allows for a large amount of combinations of amino acids to make a protein, and is what gives rise to the many functionalities of proteins. Proteins have been extensively studied because of their important functions such as applying forces{CITE} and altering reaction rates{CITE}. The process in which proteins are made starts with the reading of genetic information (specifically in the form of mRNA) and then converting this to a sequence of amino acids. This process is aptly called transcription, as the genetic information is being transcribed into a functional form{CITE}.

We can look at the information space of a living system now from an omics point of view{CITE}. The genome is based on the DNA of an organism. This DNA can code for many proteins because it be transcript for many RNAs and this layer is called the Transcriptome. At the transcriptome level, the rates of transcription are controlled by transcription factors and this allows for different levels of gene expression. Genes, which is a piece of information defined by a DNA sequence, are expressed as proteins and the collection and abundance of proteins in an organism is called its proteome. Control over the proteins, when they should function, when they should stop, is controlled by post translational modifications and this system we call the epigenome. As the huge number of proteins have a huge number of functions, the leading way of understanding this system has been the metabolic reaction network, also known as the metabolome. One can imagine that a protein may serve a function by affecting some metabolite, for example a molecular motor which does work while consuming adenosine triphosphate ATP. The collection of all these proteins, their metabolites and reaction pathways give rise to all the functions a cell can proform as well as the variety of cellular components. The omics view of the central dogma clearly shows how genetic information on the genomic level can directly lead to the physical structures and functions of living things.

Coincidentally we have just described all the constituents of our model{CITE}, simply by thinking about the central dogma. We do not consider any molecules related to the genome like DNA, as they would be located in the nucleoid. Instead fragments of DNA and RNA are in the cytoplas as Nucliotides and bases. From the transcriptome we nucleic acids like translational and messenger tRNA and mRNA, as well as proteins which control translation, called translation factors. The last constituent from the transcriptome are large molecular machines, made of both RNA (rRNA) and proteins, called

ribosomes{CITE}. Ribosomes are the location of protein synthesis from the information in mRNA. In bacteria the ribosome is made of multiple subunits (30s & 50s) which combine to make a full ribosomes (70s and 80s). This brings us to the proteome which mainly contributes proteins which can be classified by their function{CITE}. We will conisder proteins involved in tRNA synthesis and degredation, transcription, protein folding and degredation and metabolic proteins (enzymes). For a complete picture we consider the metabolome which other than the proteins requires small molecules such as adenosine triphospate. We also consider the fact that ions present in cells at high concentration, and serve many functions such as in controlling protein folding and signaling.{CITE}

From a chemistry perspective, the metabolome defines a self consistant set of reactions, and thus one can study a cellular process as a chemical reaction. For example, the microscopic details of how a protein does something can be ignored, and can be generalized into macroscopic parameters such as reaction rates.{CITE} However this gives the illusion that the metabolome can describe all the processes we see in an organism. This is not true because many of the assumptions that work on a part of the reaction network do not apply globally across the network, specifically living systems are active and out of equilibrium.{CITE} Advances such as the Nobel Prize winning work of Leland H. Hartwell, R. Timothy Hunt, and Paul M. Nurse {CITE}, showed that general features like the life cycle of cell are conserved between organisms, and selected for. This again shows that although information exists for both the microscopic physical and macroscopic biological level, but a method to directly and precisely connect the two does not yet. Ernest Rutherford is quoted as saying "All science is either physics or stamp collecting".{CITE} This statement doesn’t necessarily mean that chemistry and biology are not sufficient nor accurate. Rather it can be seen as the want to describe the amazing phenomena seen in biological systems with the precision and rigor associated with physics.

-Diffusion
- -Inspire Diffusion x
- -Ficks Laws (Macro) X
- -Einstein, X
- -Perrin X
- -Stokes Einstein, X
- -Langevin X
- -Some other (Fokker Planck)
- -Fastest possible reactions as inspiration (Smoluchouski)

It is important to remember that all these molecules, DNA RNA proteins are not living, even though they do things to make a system living. As stated before, and in contrast to biology, physics is only concerned with the motions of objects. A single molecule cannot make a system active or out of equilibrium by itself, hence the existence of active matter and collective effects. In fact, any non living object is subject to the deterministic or probabilistic laws of physics. For example Newton’s laws of motion say that things do not want to move because of inertia (mass), and accelerate due to forces
[EQ 0.1, 0.2].
This means that given the force on an object, its entire movement is known. Lagrangian mechanics extends this idea, by saying what trajectories are actually taken out of many that can be taken. The only trajectory a system takes is one that minimizses the action
[EQ 0.3 0.4]
In Hamiltonian mechanics the time evolution of a system is determined by the Hamiltonian
[EQ 0.5 0.6].
One can then imagine the trajectory (and all other possible trajectories) of a system in its initial condition. The trajectory as plotted on the coordinates of the Hamiltonian is called the phase space. Although the laws of motion are not yet consistent across length and time scales, the laws of physics are. Quantum mechanics and General relativity are very different. They refer to different processes on different objects, and thus the equations of motion for both theories are wildly different. However key concepts are the same in both

theories such as energy minimization and momentum conservation. In order to study the inside of cells, we will utilize statistical mechanics, and we will show that the key motion is from diffusion.

A microscopic description of diffusion would come from Robert Brown's work with pollen under the microscope{CITE}. Meticulously, Brown monitored the positions of individual pollen granules in water. He noticed their sporadic motion, and since then the term for this has been Brownian motion. Today, we posses a larger mathematical vocabulary and know that the trajectories Brown noticed are random walks. As one of the founding works of microscopic motion, further advancements of the details would only come with time.

Adolf Fick {CITE} studied the diffusion of a colored salt in water. Specifically English chemist Thomas Graham reported the increase of the solution and spread of the colored salt in water, with increases of temperature. Adolf Fick modeled the concept in one dimension, first describing how concentration gradients cause fluxes. Secondly, he considered the time evolution of the concentration due to the fluxes away from high concentration. Consider two boxes next to eachother along the x axis. They each have the same size but different number of particles [inlineEQ Concentration]. Just from random motion, we expect more particles to move from the box with more particles, and we call this a flux [inlineEQ Flux J]. These ideas are encapsulated in Fick's first and second laws. [EQ1.1 EQ1.2]
These laws convey the main ideas that any inhomogeneous concentration distributions will relax to a homogeneous one. This process is known as diffusion due to concentration gradients, and Fick's work gave a macroscopic description of it.

The next key works in the path to understanding diffusion came from George Gabriel Stokes. Without going into detail, he developed Stoke's Law
[EQ 2]
which describes the force of drag on a body in water. The specifics of the model he used to derive his law was that of a spherical particle in a laminar flowing Newtonian fluid with a low (zero) Reynold's number. However Stokes did not know of these terms nor details and his law was first an empirical one{CITE}. Overall this was a key finding in fluid dynamics as it stated how a moving particle is slowed down by a viscous fluid.

This brings us to the work of Albert Einstein that helped earn him his Nobel prize{CITE}. Einstein wanted a model to describe Brownian motion from microscopic details. Specifically, he believed that the random motion of the suspended particles came from collisions with surrounding particles. Einstein considered how a suspended particle would move in a given time interval in which it collides randomly with surrounding molecules. Mathematically he considered the number density of particles evolving with some timestep tau as a Taylor series expansion.
[EQ 3.1]
This required only the consideration that the particle number is conserved, and is based on the continuity equations. He then showed that this evolution of the number density is equivalent to Fick's laws of diffusion. He then described the fundamental solution to the diffusion equation as having the form
[EQ 3.2]
More importantly, Einstein was able to define the coefficient from Fick's equation in terms of mean squared displacement of the particle. Then he related this coefficent again to measurable physical quantities like the temperature.

Almost simultaneously, Marian Smoluchowski developed another explanation of the observations of Brownian motion{CITE}. Combining the Einsteins previous work on the Kinetic theory of molecules, he concluded that diffusion coefficient from Fick's law is determined by the temperature and the mobility of the particle
[EQ 3.3]
Combining the works of Stokes's Einstein and Smoluchowski, we arrive at two key relations

[EQ 3.4, EQ 3.5]
We now have intuition about the diffusion coefficient, which is dependent on the size of the molecule the viscosity of the surrounding fluid and the temperature of the system. Further more, we know from the fundamental solution the the diffusion equation, that on average the particle does not move
[eq 3.6]
but the mean squared displacement of the particle
[eq 3.7]
increases linearly with time. The implications of these works are actually still understated. They not only described the motions of microscopic objects, but also proved further the existence of atoms{CITE}.

One more key milestone in the study of diffusion and statistical mechanics was the development of the Langevin equation of motion{CITE}. Paul Langevin wanted a macroscopic description of the motion of a particle with some random driving force. This was analogous to a macroscopic description of diffusion as the motion of a diffusing particle is caused by random collisions with microscopic particles from its surroundings. To quantify this random force Langevin considered that the noise should obey two properties
[EQ 4.1, 4.2]
That is to say that on average the force is random and thus zero, and has no particle particle nor time correlations due to the delta functions. Considering only a non-interacting particle surrounded by a microscopic fluid he wrote the equation of motion in the form
[EQ 4.3.0]
With this he was able to show the distribution of velocities was fixed
[4.3.1]
However with only a random noise term the second moment of the distribution increases with time
[4.3.2]
However if a drag term is introduced in the exact form
[EQ 4.4]
then the mean position of the particle is zero and mean squared displacement is related to the strength of the noise, which is the Stokes-Einstein relation. The drag force can also be derived as a random force but is also velocity dependent as more collisions will occur in the direction of travel. The implications of this is quite profound. It says that due to temperature, nature will give you a random force which pushes you, but one cant take advantage of it because the same random process will also hinder you. The Stokes-Einstein relation is only one example of this, and the generalized idea is called a fluctuation dissipation theorem{CITE}. With the Langevin equation of motion, we now have a method to drive particles which we want to diffuse. This innovation was huge in the field of statistical mechanics, as we developed a stochastic differential equation of motion.

-Solutions and Mixtures
- -Some basics (why water solutions)
- -Mixing
- -Solutions

Caught up in our discussion of the motion of the molecules we neglected what the fluid these objects exist in. In equation

-Complex Mixtures
- -Reaction Diffusion
- -LLPS
- -Glasses

-Theory specific to this work
- -The cytoplasm as spheres
- -Our interaction model
- -Measures used in analysis

Methods

Making Soup; Building a cytoplasm

- We started from the work of Feig et al as a complete annotated database of the contents of a piece of cytoplasm. This paper had a database which contained many constituents with concentrations. It was supplemented with data from PaxDB and UniProt DB which are proteomics and genomics databases respectively. This was done to account for proteins in the Feig et al data which were in PDB format. Our final database has 452 constituents with some species having a minimum concentration of 1um. From this database we can generalize our average cytoplasm to be in a box of (100nm)^3 with 94.81% filled with water molecules.

- From the model cytoplasm we also created a test system. Our aim was to create an analogous system to the full cytoplasm, which means keeping physical qualities such as mass and radii abundance, while simplifying the system. The main quantities of the cytoplasm constituents were charge and radius so the test system was modeled using 9 molecule types corresponding to 3 radii small medium and large and 3 types of charges positive negative and neutral. The abundances of the constituents were made to represent volume fractions in the real cytoplasm.

The Simulation Engine

-Data from our database was given as an input to a simulation engine built on the HOOMD-Blue library on Python. HOOMD-Blue was selected for its ease of use, good documentation, and for its ability to run molecular dynamics simulations of GPUs.

-The simulation box was constructed according to occupied volume fraction. The total volume of the box was set so that the molecules from database would occupy a specified volume fraction. To create the system, a scaled up box was filled up with molecules in the database. These molecules were placed randomly with the PACMOL software implemented in python via the Mbuild library.

-Multiple steps were taken to ensure that while each instance of the simulation was different with the same constituents, each instance was also thoroughly equilibrated before any quantities of interest were recorded. Specifically, the enlarged box filled with molecules was shrunk to the correct size. After compression, the system was allowed to equilibrate with potentials that were both cut off at a small radius, and modified to raise the potential energy minimum. This was rationalized as erroneous results would only result if there was significant overlap of the spheres representing the molecules, and this would not happen from the cutoff nor the boosted potential. The pressure and potential energy was monitored during the equilibration to verify that the system had stabilized.

-Molecular dynamics simulations of this equilibrated system was run as the main computer experiment in this study. The specific data that was given to the simulation was a simulation box filled with molecules, an n-n pairwise interaction matrix with n as the number of species which scales the LJ epsilon parameter, physical parameters of temperature viscosity and ionic strength, and simulation constraints such as the timestep the total time and how often to write data to the trajectory. Very specifically, at each timestep the simulation box was initialized. For each particle the force was calculated as generated from pairwise LJ and Yukawa interactions with every other particle in the cutoff radius. During said timestep, the force and a random noise term drove the particles, while a viscous dampening slowed them as according to brownian dynamics. The usual simulation was done with in a box of size (100nm)^3, with a timestep of 0.1 femtoseconds, at 30C and 150mM ionic strength and a viscosity of 1 (water). HOOMD-Blue wrote the trajectory at a given rate as a binary gsd file. To minimize file sizes, the simulation was broken into 2 time regimes,

slow and fast. 2 trajectories were created from 1 simulation, with 2 write rates, with the slow writer ending after a certain number of steps. The data was stitched together from both trajectories during analysis.

-To monitor the accuracy of the simulation, the kinetic, potential and total energies along with the kinetic temperature and pressure were monitored. This was done by extracting these quantities along with others at each trajectory step. Stability was determined when the total energy, potential energy and pressure stabilized during equilibration.

-To monitor the mean squared displacement of the particles from the trajectory, the Freud library was used. The library is able to calculate the MSD for each particle in the trajectory over a sliding time window, which decreases computational cost. For a certain species the MSD was averaged over all particles of said species. Other quantities derived from the MSD v time such as log MSD v log time was also calculated from the averaged MSDs.

Results

Representations of the Model
-Plots of Cytoplasm Composition
-Plots of the interactome

Short Time Scale Dynamics

Discussion

Conclusions

-The model is okay, need better tools and theory to study chaotic statistical systems
- -Pros and Cons of the model
- -The Limiting case of explicit water and ions
- -HPMC as an alternative simulation

-Living cells need active control over the cytoplasm
- -Cytoplasm is chaotic
- -Cell must have control to avoid equilibrium
  - -EQ is often death or disease, examples
- -Relation between controlled dynamics and controlled glassy state