UNIVERSITÉ LIBRE DE BRUXELLES

TRAN-F501

INTERNSHIP - 201819

Project: A stochastic simulation system for protein aggregation

Supervisor:
Tom LENAERTS
Author:
Prateeba RUGGOO

December 4, 2019

Contents

1 Introduction

Diseases like Alzheimer and Parkinson disease are the result of proteins aggregating into large fractal structures that hinder the cell function or even destroy them. Understanding how aggregates are formed and change over time is important to understand when they become harmful and how maybe treatments affect aggregate formation.

The goal of this Internship is to implement a simulation system to study aggregation between proteins and is performed in collaboration with the Switch lab in the KU Leuven, who has an extensive expertise in studying aggregation and related diseases.

1.1 Mathematical modeling

The principle task is to develop a method for simulating the time evolution of the N quantities $\{X_i\}$, knowing only their initial values $\{X_i^{(0)}\}$, the form of the M reactions $\{R_\mu\}$ and the values of the reaction parameters $\{c_\mu\}$. There are two fundamental approaches to the mathematical modelling of chemical reactions.

1. Deterministic models which are based on differential equations.
2. Stochastic simulations where the fundamental principle is that molecular reactions are essentially random processes, i.e it is impossible to say with complete certainty the time at which the next reaction will occur. This approach uses basic Newtonian physics and thermodynamics to arrive at a form defined as the propensity function that gives the probability a_μ of reaction μ occurring in the time interval $(t, t + \delta t)$.

2 Gillespie's stochastic framework of chemical kinetics

Definition 1. *Problem definition: We are given a volume V containing molecules of N chemically active species $S_i (i = 1, \ldots, N)$. Let $X_i \equiv$ current number of molecules of chemical species $S_i \in V, (i = 1, 2, \ldots, N)$ and let $R_\mu (\mu = 1, \ldots, M)$ be the chemical reactions in which the species S_i can participate. Each reaction R_μ is characterized by a numerical reaction parameter c_μ. The goal is to simulate the trajectories of the N chemically active species S_i and predict which reaction will occur at each time step according to the correct probability distribution.*

Definition 2. *The reaction parameter c_μ is defined so that $c_\mu \delta t$ gives the probability that a randomly chosen molecule of chemical species S reacts during the time interval $[t, t + \delta t]$ where t is time and δt is an infinitesimally small time step.*

Definition 3. *The probability that exactly one reaction μ occurs during the infinitesimal time interval $[t, t + \delta t]$ is equal to $S(t)k\delta t$ where $S(t)$ is the number of chemical species S at time t.*

Definition 4. *State of the system is defined by the number of molecules of each species and changes discretely whenever one of the reactions is executed.*

2.1 Gillespie's Direct Method

Given the problem defined above, the Gillespie's Direct Method answers two questions :

1. Which reaction occurs next ?
2. When does it occur ?

2.1.1 Gillespie's Direct Method formulas

1. Probability density $P(\mu, \tau)$ that the next reaction is μ and that it occurs at time τ is given by :
   $P(\mu, \tau)d\tau = a_\mu exp(-\tau \sum_j a_j)d\tau$.
2. Probability that the next reaction is reaction μ is given by :
   $Pr(Reaction = \mu) = a_\mu / \sum_j a_j$.
3. Probability distribution for times
   $P(\tau)d\tau = (\sum_j a_j)exp(-\tau \sum_j a_j)d\tau$.

2.1.2 Gillespie's Direct Method : Pseudo code

Algorithm 1 Gillespie's Direct Method

Input: N chemically active species $S_i, \{X_i^{(0)}\}$ initial values of each species S_i, the set R of chemical reactions and the reaction parameter c_μ for each reaction.
Output: Sample trajectory of a chemical process in the stochastic framework.
while !(simulation time exceeded) **do**
1. Initialization: Set initial number of molecules in the system, set $t \leftarrow 0$.
2. Calculate the propensity function, $a_i, \forall i$.
3. Choose μ according to the distribution in 2.
4. Choose τ according to an exponential with parameter $\sum_j a_j$ as in 3.
5. Update the number of molecules to reflect execution of reaction μ. Set $t \leftarrow t + \tau$.
6. Go to step 2.

end while

2.1.3 Gillespie's Direct Method : Example

The source code of the Direct method simulation model can be found at Direct-method-github. The reaction set is the example used in Gillespie's paper [Gil76].

$$X \xrightarrow{c_1} Y \tag{1}$$
$$Y \xrightarrow{c_2} X \tag{2}$$
$$2X \xrightarrow{c_3} Z \tag{3}$$
$$Z \xrightarrow{c_4} 2X \tag{4}$$
$$W + X \xrightarrow{c_5} 2X \tag{5}$$
$$2X \xrightarrow{c_6} W + X \tag{6}$$

where the reaction parameters $c_i \in (i = 1, 2, \ldots, 6)$ is equal to [1,1,2,1,2,1] respectively and the initial number of each species is = 10.

Example 5. *Generating sample trajectories of a chemical process in the stochastic framework : Direct Method*

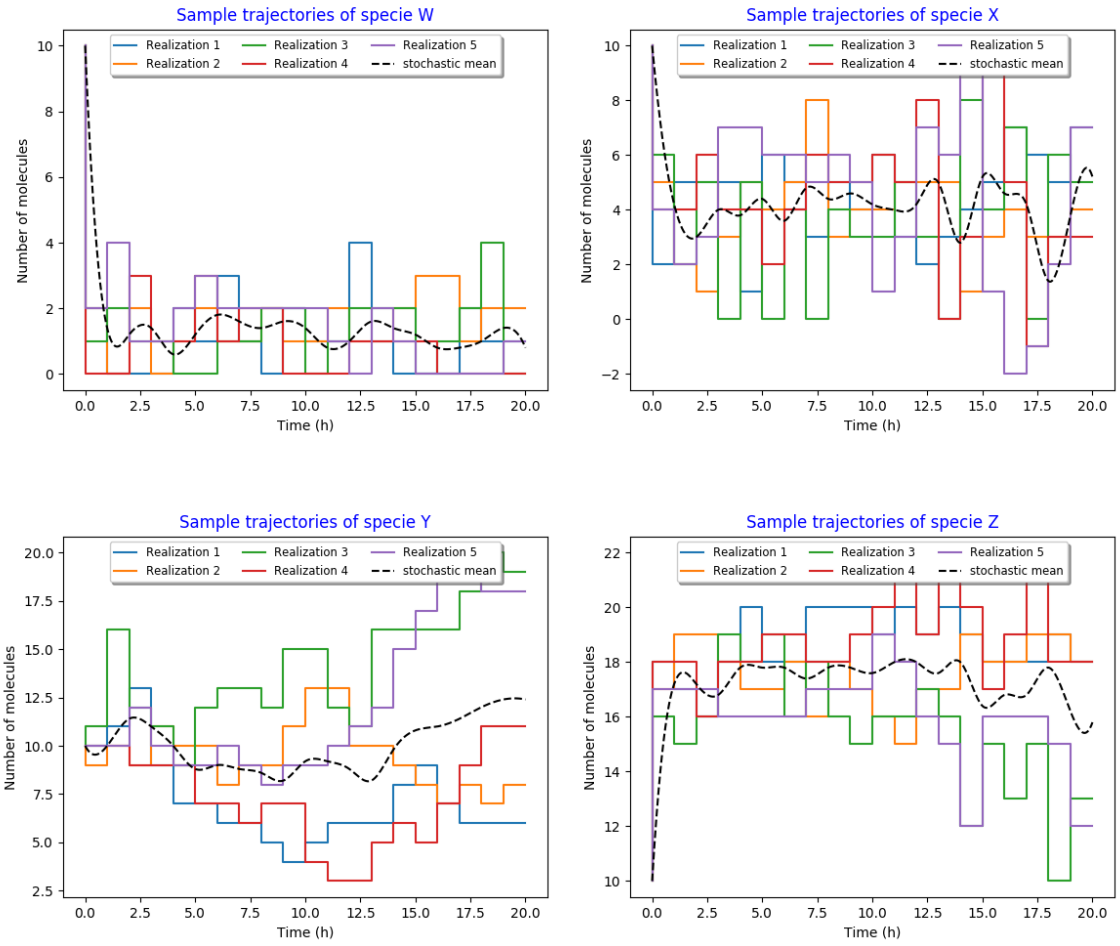


Figure 1: The x-axis denotes the duration time of the simulation and the y-axis denotes the number of molecules of each species S_i at each time step.

2.2 Gillespie's Next Reaction Method

The goal of the next reaction method is to do away with the above processes and update only the data that has been modified. The main idea is to generate a putative time τ_i for each reaction i that will occur and choose the reaction μ whose putative time τ_μ is least. In order to minimize the computation time, the Next Reaction Method is implemented using specific data structures. To make the sampling of reactions more efficient, an indexed priority queue is used to store the reaction times. On the other hand, to make the recomputation of propensities more efficient, a dependency graph is used. This dependency graph tells which reaction propensities to update after a particular reaction has fired.

2.2.1 Gillespie's Next Reaction Method : Pseudo code

2

Algorithm 2 Gillespie's Next Reaction Method

Input: N chemically active species $S_i, \{X_i^{(0)}\}$ initial values of each species S_i, the set R of chemical reactions and the reaction parameter c_μ for each reaction.

Output: Sample trajectory of a chemical process in the stochastic framework.

while !(simulation time exceeded) **do**

1. Initialization :
   - 1.1 set initial number of molecules, set $t \leftarrow 0$, generate a dependency graph G.
   - 1.2. calculate the propensity function, $a_i, \forall i$.
   - 1.3. for each i, generate a putative time, τ_i, according to an exponential distribution with parameter a_i.
   - 1.4. store the τ_i values in an indexed priority queue P.
2. Let μ be the reaction whose putative time, τ_μ, is least.
3. Let τ be τ_μ.
4. Update the number of molecules to reflect execution of reaction μ. Set $t \leftarrow \tau$.
5. For each edge(μ, α) in the dependency graph G,
   - 5.1 update a_α.
   - 5.2 if $\alpha \neq \mu$, set $\tau_{alpha} \leftarrow (a_{\alpha,old}/a_{\alpha,new})(\tau_\alpha - t) + t$.
   - 5.3 if $\alpha = \mu$, generate a random number, ρ, according to an exponential distribution with parameter a_μ and set $\tau_\alpha \leftarrow \tau + t$.
   - 5.4 Update the old τ_α value in P.
6. Go to step 2.

end while

2.2.2 Gillespie's Next Reaction Method : Example

The source code for the Next Reaction Method can be found at Next-Reaction-Method-github. The reaction set is the example used in the following paper [GB00].

$$A + B \xrightarrow{k_1} C \tag{7}$$

$$B + C \xrightarrow{k_2} D \tag{8}$$

$$D + E \xrightarrow{k_3} E + F \tag{9}$$

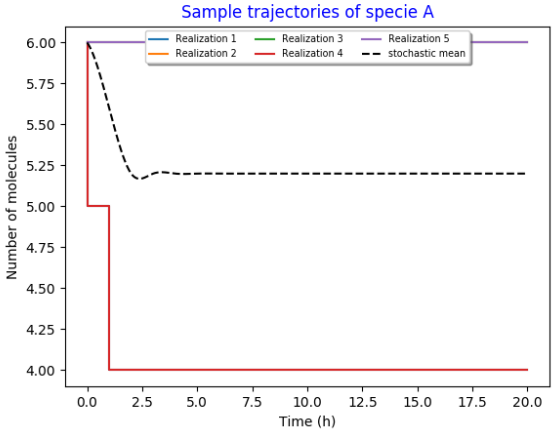
$$F \xrightarrow{k_4} D + G \tag{10}$$

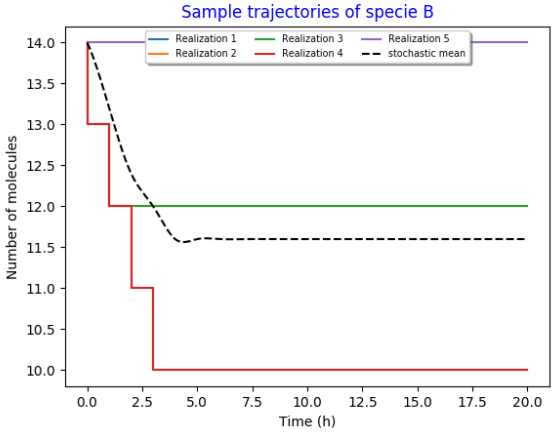
$$E + G \xrightarrow{k_5} A \tag{11}$$

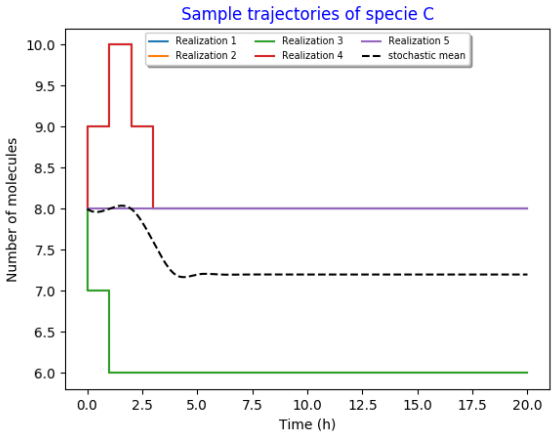
$$\tag{12}$$

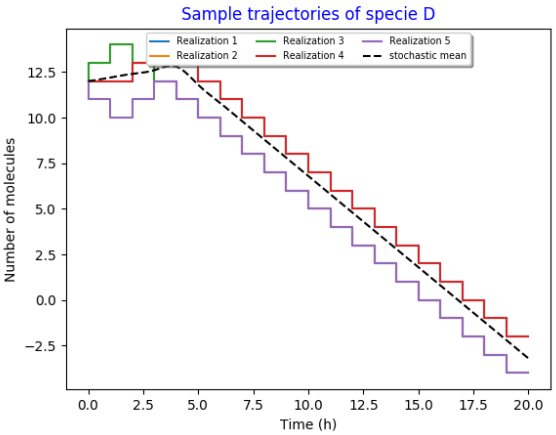
where the reaction parameters $k_\mu \in (i = 1, 2, \ldots, 5)$ is equal to [1,2,2,2,1] respectively and the initial number of each species is = 10.

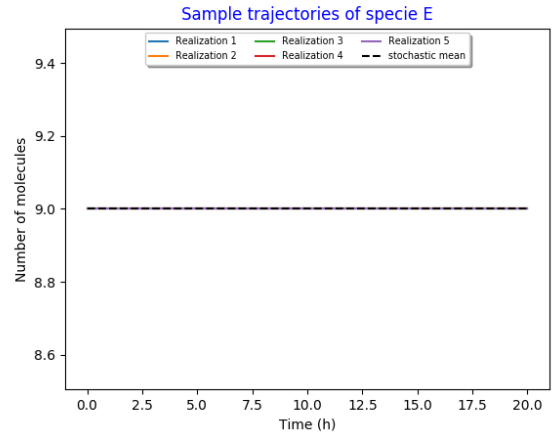
Example 6. *Generating sample trajectories of a chemical process in the stochastic framework : Next Reaction Method*

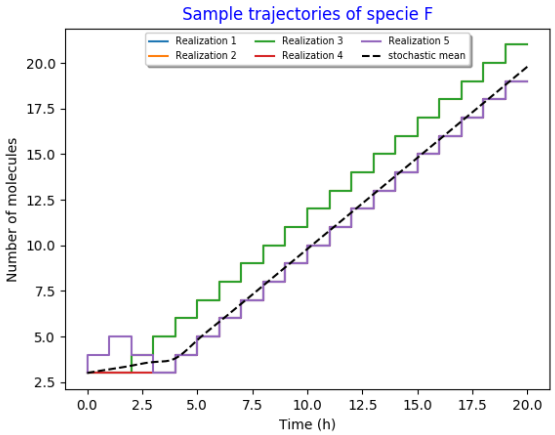

Sample trajectories of specie A
Realization 1
Realization 2
Realization 3
Realization 4
Realization 5
stochastic mean
Number of molecules
Time (h)

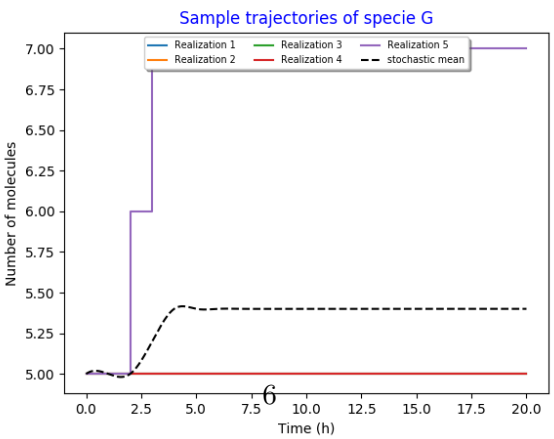

Sample trajectories of specie B
Realization 1
Realization 2
Realization 3
Realization 4
Realization 5
stochastic mean
Number of molecules
Time (h)


Sample trajectories of specie C
Realization 1
Realization 2
Realization 3
Realization 4
Realization 5
stochastic mean
Number of molecules
Time (h)


Sample trajectories of specie D
Realization 1
Realization 2
Realization 3
Realization 4
Realization 5
stochastic mean
Number of molecules
Time (h)


Sample trajectories of specie E
Realization 1
Realization 2
Realization 3
Realization 4
Realization 5
stochastic mean
Number of molecules
Time (h)


Sample trajectories of specie F
Realization 1
Realization 2
Realization 3
Realization 4
Realization 5
stochastic mean
Number of molecules
Time (h)


Sample trajectories of specie G
Realization 1
Realization 2
Realization 3
Realization 4
Realization 5
stochastic mean
Number of molecules
Time (h)

3 Molecular mechanisms of protein aggregation

Until now, all what was done is using random reaction parameters for each reaction in the stochastic simulation. However, in order to study the protein aggregation, understanding the kinetics of the aggregation is fundamental. In the following paper [MKA$^+$16], a protocol is described on how to make full use of kinetic descriptions derived from a model of the underlying microscopic reactions that make up the aggregation network. The fitted parameters are therefore meaningful and correspond to physical properties of the system, such as nucleus sizes, the kinetic rate constants of individual reactions and saturation concentrations.

The key idea is to analyze a large data set of multiple kinetic traces at different reagent concentrations simultaneously,with a single rate law and yield strong mechanistic constraints that will be helpful in choosing the appropriate models.

3.1 Obtaining qualitative constraints : Half times

Definition 7. *Half-time is defined as the time at which the signal has reached half of its final plateau value.*

Given a dataset, several fundamental models can be considered for the fitting process to yield the rate constants. In order to choose the best model, adding some constraints based on the concentration dependence of the aggregation reaction is a solution. By considering how the half times scale with monomer concentration and how this scaling depends on the monomer concentration, one can obtain constraints on possible mechanisms.

3.1.1 Extracting half times

Algorithm 3 Extracting half times algorithm

Input: Normalised data of aggregate concentration in units of fluorescence.
Input: Monomer concentration for each curve.
Output: Scaling exponent.
for each curve **do**
1. Select the middle part of the curve.
2. Fit a middle line through the selected middle part.
3. Compute the point at which the fitted line crosses the value 0.5

end for

The slope of Log(half time) versus Log(monomer concentration) also known as the scaling exponent gives insight about the dominant process of fibre multiplication. This can in turn can be used to decide on possible models for the fitting.

Very generally :

1. A negative curvature in the double logarithmic plot (i.e., when the slope becomes steeper at higher monomer concentrations and therefore the process is more monomer dependent) is indicative of competition between several processes in parallel.

2. A positive curvature, in contrast (i.e., a flattening of the curve at higher monomer concentrations and thus a decrease in monomer dependence), suggests the presence of a saturation effect in a serial process or, in rare cases, at monomer concentrations close to solubility, it can be due to a change in nucleus size.

The behavior of half-times with varying monomer concentration is, therefore, a good first guide to narrowing down the number of possible models. It limits the number of acceptable reaction networks by determining the reaction order of the dominant process and probing for competition or saturation effects. The model for fitting needs to be chosen to reflect these findings.
A flowchart to help decide on probable models using scaling exponents is shown section 3.2.

3.2 From Half-times to models

[MKA+16]

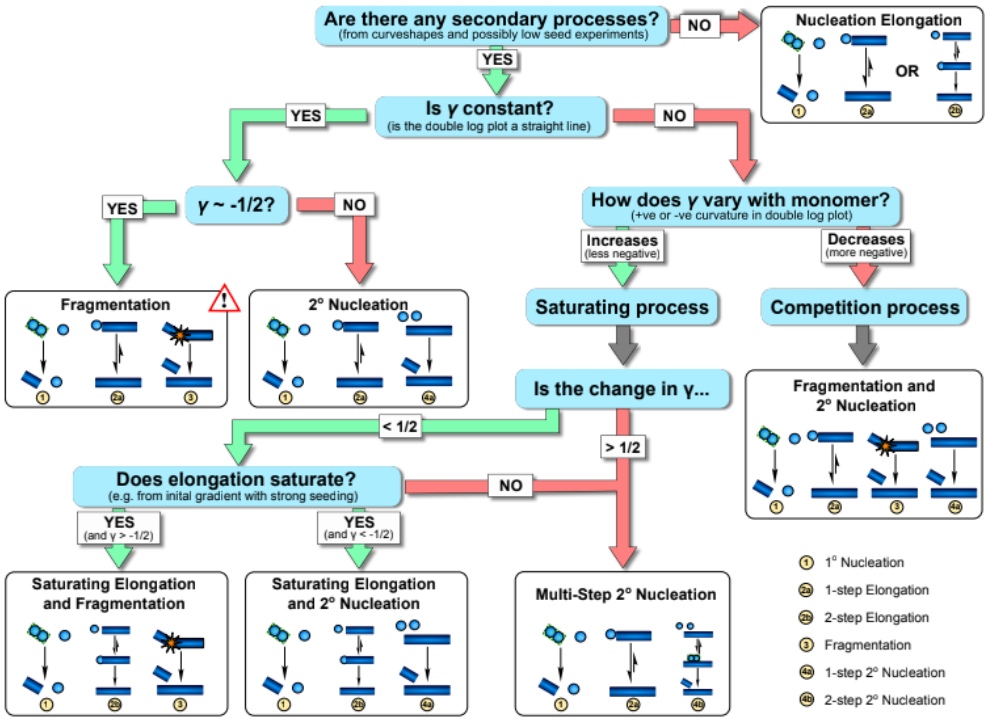


Figure 3: The curvature of the double logarithmic plots and the value of their slopes gives insights into which aggregation mechanisms are dominant. The flowchart illustrates the decision process to arrive at a likely model, based on the half time plots.

.

3.3 Global fitting

The Global fitting approach allows fitting large data sets, under a variety of conditions, usually a number of monomer concentrations, simultaneously to a single rate law. This enforces a relationship between experimental curves in which the free fitting parameters, such as rate constants and reaction orders, must be equal for all curves. It ensures that

the microscopic model has the correct dependency not only on time, but also on the other parameters that are varied, such as the monomer concentration. Only in this manner can we obtain sufficient constraints to distinguish between the various complex models describing different aggregation reaction networks.

The fitting process minimizes the mean squared residual error (MRE), given by

$$\frac{1}{N}\sum_{i=0}^{N}(y_i - f(t_i))^2$$

where N is the number of data points, y_i is the measured value at time point t_i and $f(t_i)$ is the model function evaluated at that time point.

Mathematically, this represents a search for the global minimum on an n-dimensional energy landscape, where n is the number of free fitting parameters.

4 Results

4.1 Amylofit's estimated rate constants in Gillespie's stochastic framework

Given the following dataset :

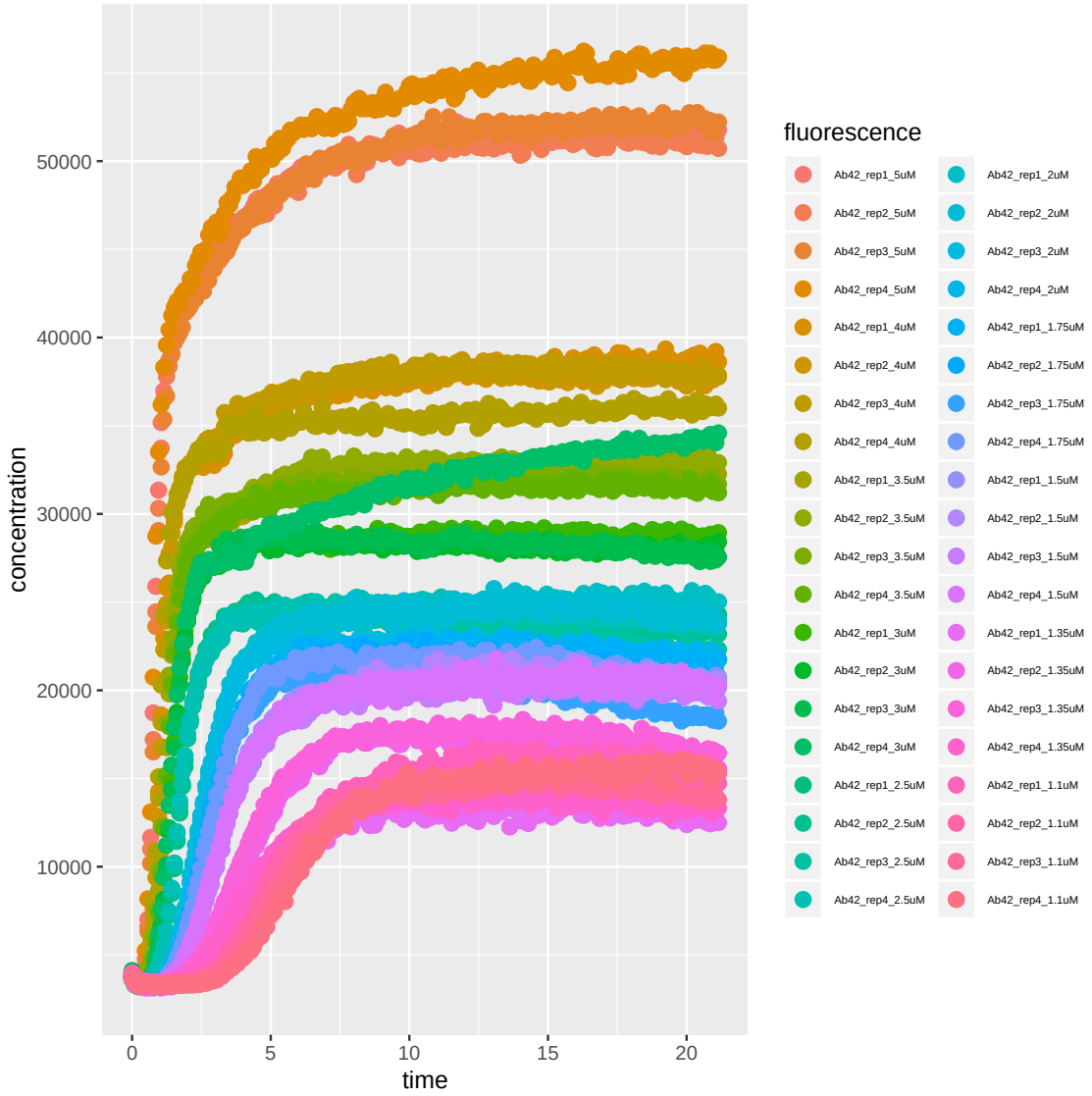


Figure 4: The x-axis denotes the time step and the y-axis denotes the fluorescence intensity of each experiment.

1. The data is first normalised using the following formula :

$$y_{norm,i} = (1 - M_{0,frac}) \frac{y_i - y_{baseline}}{y_{plateau} - y_{baseline}} + M_{0,frac}$$

where :

(a) y_i is the original value of the ith data point.

(b) $y_{norm,i}$ is its normalized value.

(c) $y_{baseline}$ is the average value of the data at the baseline.

(d) $y_{plateau}$ is the average value of the data at the plateau.

(e) $M_{0,frac}$ is the relative initial concentration of aggregates.

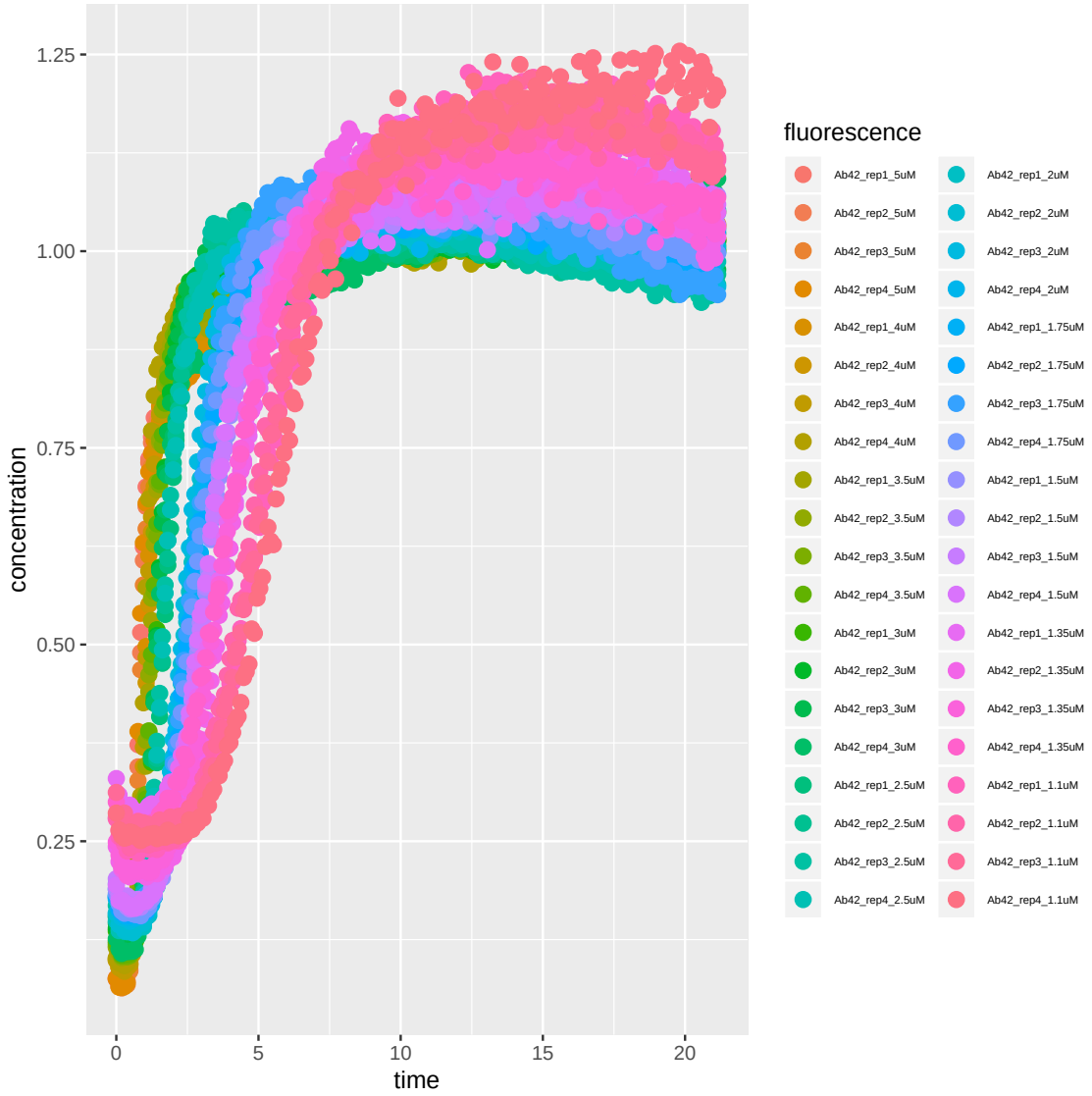


Figure 5: The x-axis denotes the time step and the y-axis denotes the normalised fluorescence intensity of each experiment.

2. Secondly, the half time of each curve is computed.

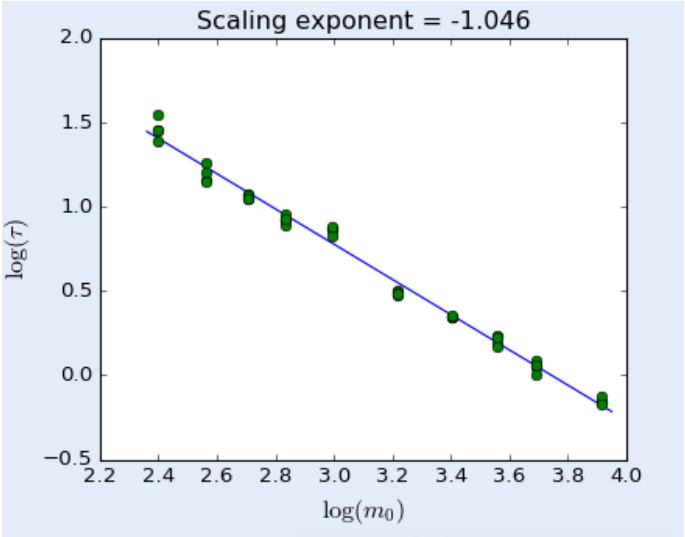


Figure 6: The x-axis denotes the initial monomer concentration of the middle part of each curve and y-axis denotes the half time of each curve.

3. Thirdly, the model is chosen according to the scaling exponent and the flowchart in section 3.2. For the current dataset and according to the scaling exponent, the model that will fit the best is the Secondary nucleation.

4. The fourth step involves global fitting the data as described in section 3.3.

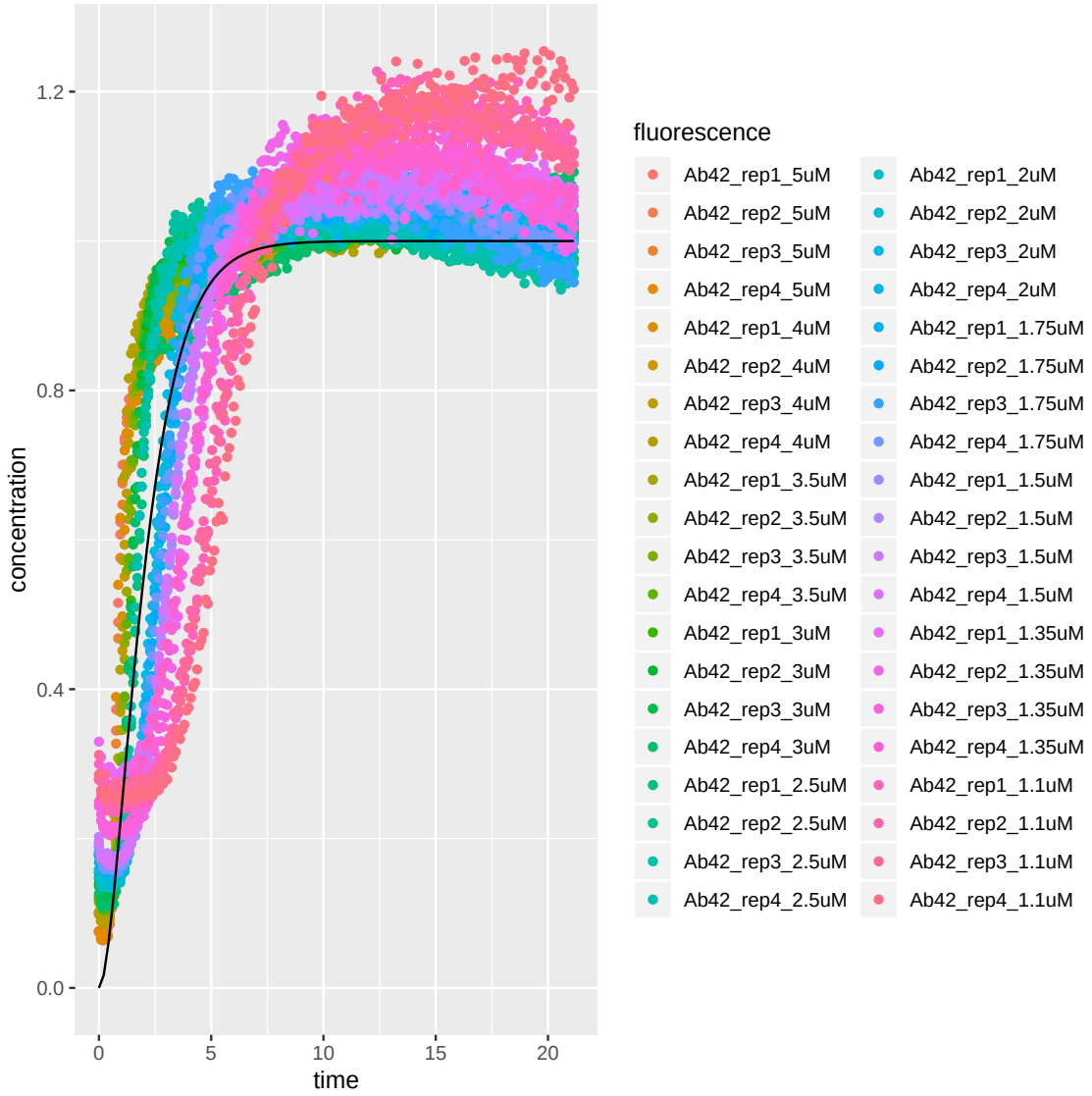


Figure 7: The x-axis denotes the time step and the y-axis denotes the relative aggregate concentration of each experiment. The continued black line represents the fit found.

The fitted paramaters are :

(a) $k_+k_n = 4.28 + 11$ in $conc^{n_c}time^{-2}$

(b) $k_+k_2 = 636$ in $conc^{n_2^-1}time^{-2}$

(c) $n_c = 2$

(d) $n_2 = 2$

5. The last step consists of using the above fitted parameters in Gillespie's stochastic framework in order to observe the behaviour of each component of the model.

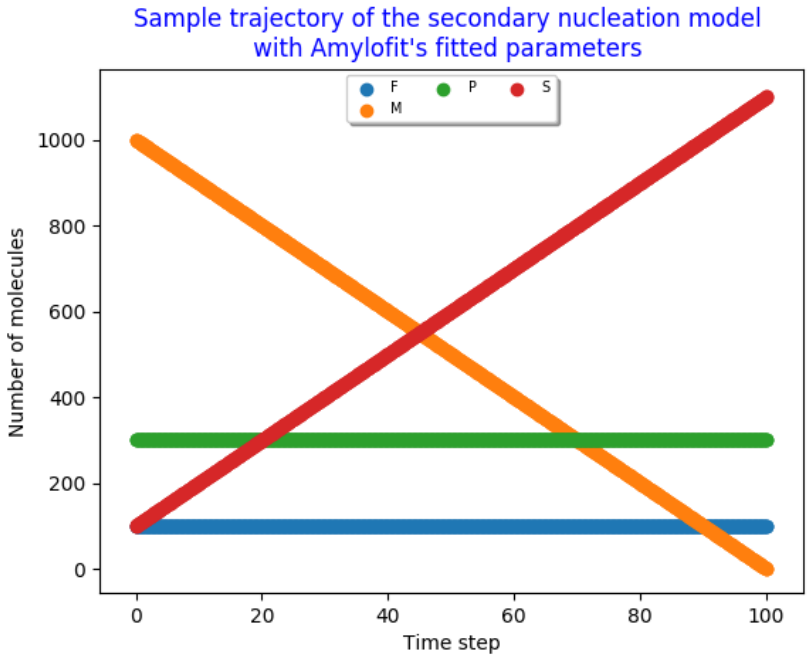


Figure 8: The x-axis denotes the time step y-axis denotes the number of molecules of each component where F is the created fibril ends. $F_0 = 100$. P is the number of primary nuclei. $P_0 = 300$. S represents the secondary nuclei. $S_0 = 100$. M is the number of free monomers in the system. $M_0 = 1000$

4.2 Implemented simulations's estimated rate constants in Gillespie's stochastic framework

1. Firstly the curves are normalised and the half times are computed.

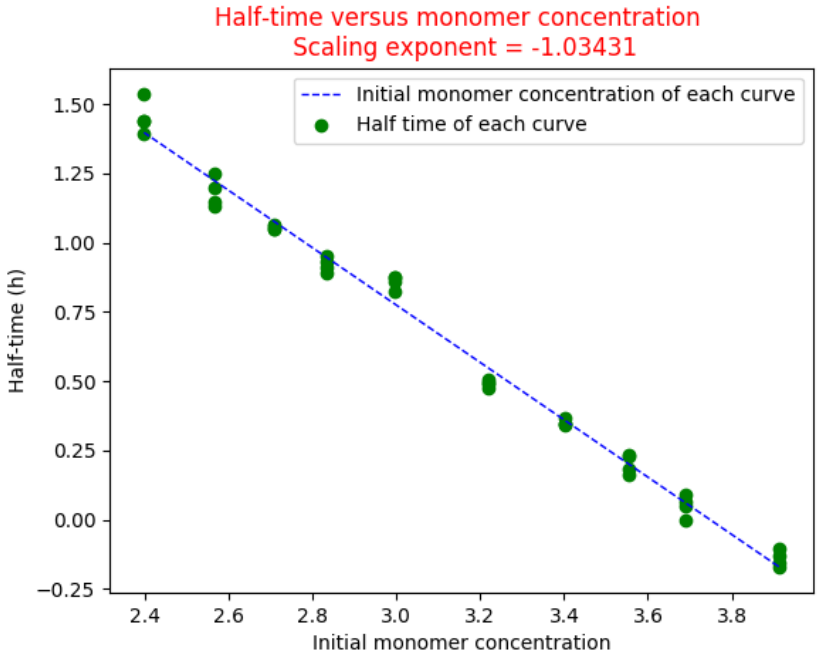


Figure 9: The x-axis denotes the initial monomer concentration of the middle part of each curve and y-axis denotes the half time of each curve.

2. Secondly, the data is fitted according to the Secondary nucleation model following the flowchart in section 3.2.

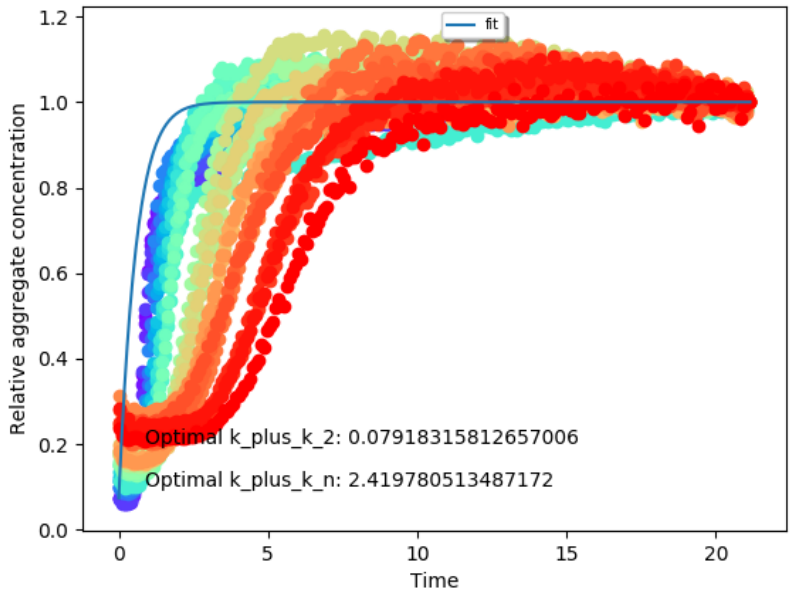


Figure 10: The x-axis denotes the time step and the y-axis denotes the relative aggregate concentration of each experiement. The continued black line represents the fit found.

4.3 Implementation

In the following github repository, simulator the different implementations can be found :

1. The Gillespie's direct method.
2. The Gillespie's Next reaction method.
   To compile :
   (a) Browse to G_next_reaction folder.
   (b) Compile with the make command.
   (c) The result of the simulation is saved in the results folder. Some of the models mentioned in the [MKA+16] is implemented the include/models.h folder.
3. The source code to plot the double logarithmic plot of a given dataset and can be compiled as follows :
   (a) Browse to the Fitting folder.
   (b) Compile with the make command.
   (c) Execute with : ./main path-to-data-file path-to-monomer-concentration-file -n(if the data is already normalised).
   (d) The double logarithmic plot is then saved in the results folder.

4. The Global fitting of the kinetic rates according to a given dataset can be fitted by doing :
   (a) Browse to the Protein-kinetics folder.
   (b) Execute with : python main.py.
   (c) The result is saved in the results folder.

5 Conclusion

The goal of this work was to study aggregation dynamics and develop a stochastic simulation system for protein aggregation and to provide scripts for visualization and analysis of the results produced by the simulation.

At the term of this internship, a minimal system that mimics what is known from experiments has been implemented, thus, finding ways to fit the known fluorescence's data and find the actual kinetic rates. Secondly a stochastic simulation system was implemented to study protein aggregation based on the Gillespie's next reaction method. Lastly the fitted kinetic rates are plugged into the stochastic simulation to observe the behavior of the simulator. This work was done in hopes of providing a stepping stone for further research and project applications in that context.

References

[GB00] Michael A. Gibson and Jehoshua Bruck. Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels. *The Journal of Physical Chemistry A*, 104(9):1876–1889, March 2000.

[Gil76] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, December 1976.

[MKA+16] Georg Meisl, Julius B Kirkegaard, Paolo Arosio, Thomas C T Michaels, Michele Vendruscolo, Christopher M Dobson, Sara Linse, and Tuomas P J Knowles. Molecular mechanisms of protein aggregation from global fitting of kinetic models. *Nature Protocols*, 11(2):252–272, February 2016.