Computational inference of mechanisms of biological self-assembly through static light scattering

Yiyang Guo

Advisor: Russell Schwartz

Abstract

Understanding biomolecular self-assembly processes is important to understanding cellular biochemistry but challenging due to their extremely small space and time scales. Current experimental techniques are not yet able to directly reveal the dynamics of such processes, but rather, provide discrete data points of high-level observables from which we must indirectly infer system dynamics. One such experimental method is static light scattering, which gives a time series of average molecular weights of the reaction complex. A direct question that follows is that how much we can confidently infer about the true underlying assembly mechanisms given static light scattering curves. This project discusses some of the techniques involved in this challenge. Specifically, we looked at questions being asked about the mechanisms of common self-assembly systems – nucleation, with specific application to filamentous assemblies. We developed formal specifications of local rule based models that can be efficiently simulated/sampled. We developed heuristics to determine or falsify model parameters through simulation optimization to determine when it is possible to accurately discover assembly mechanisms from such data.

Motivation and Background

Self assembly system is a process where collections of units spontaneously assemble into some structure under specific conditions. It is essential in many biological processes including virus capsid, actin filaments, etc. and has been widely studied as a biophysical system for their biomedical significance. [2] However, exact biomolecular mechanisms underlying these self-assembly systems are hard to understand. Current experimental techniques in vitro are not yet able to directly observe assemblies in detail. Computational modeling and simulation have been applied to fill the gap between true biological mechanisms and in vitro experimental observables. This project discusses the challenges involved in this process.

Related Works

For computational modeling and simulation, many works have been dedicated to address large, concurrent, and stochastic aspect of the biomolecular systems. In general, it comes down to a tradeoff between simulation efficiency and level of model abstractions. The more simple and abstract the models are, the more efficiently they can be simulated. Meanwhile, the level of abstractions directly limits the expressiveness of the model in terms of biological mechanisms. This hierarchy of expressiveness is shown to be crucial when we infer self-assembly mechanisms in this project. Thus, here we briefly summarize the major modeling and simulation methods in this domain.

(1) Differential Equations

This is the one of the most coarse-grained descriptions of molecular dynamics. The dynamics are characterized by a differential equation system, where the solution (when solvable) gives an averaged particle distributions under the reaction dynamics.

(2) Local rule-based models and Gillespie algorithm

Local rule-based models formulate behaviors of individual particles, instead of population of particles, as it is in differential equation method. But it assumes the molecules in the system are "well-mixed", meaning that they are uniformly distributed spatially. This implies reaction propensity solely depends on the reactants, but not the distance or contact energy between reactants. The whole dynamics are characterized by a set of reaction rules and individual molecules are tracked during simulation. At every time step, one reaction is sampled, with distribution specified by the reaction rules (can be viewed as a continuous markov chain where individual transition rate is given by propensity = reaction rate * number of particle combinations that the reaction rule applies). Due to their fine granularity and capability handle combinatorial complexity of possible particles in the system, many modeling languages have been developed for local rule-based models, including Kappa[10], and BioNetGen[11].

(3) Green's Function Reaction Dynamics/Spatial Next-Reaction Method[12]

These two methods capture spatial dynamics of the molecules. Unlike local rule-based models and Gillespie algorithm, where individual particle's position is assumed to be uniform, they keep track of enough spatial information of particles such that individual contact of particles and thereby their reaction can be correctly sampled.

Our approach

In this project, we utilize existing modeling and simulation framework to study explore the capabilities and limitations of computational inference in this domain. We limit our study to filament assemblies and whether they are limited by nucleation, a phase where monomers of the assemblies form a small structure with distinct thermodynamic status. We also focus on one specific type of experimental observables that are cheap and commonly practiced by experimental biophysicist: static light scattering. Our work can be divided into two steps:

(1) We model hypothesized biological mechanisms (whether the assembly is nucleation limited or not, the size of nucleation) in a language that can be efficiently simulated for their static light scattering observables.

(2) Using this framework of turning a hypothesis to the simulated experimental observables, we study whether our hypotheses can be confirmed or falsified through parameter fitting via simulation optimization. If models of two hypotheses, under some parameters, exhibit aligning simulated experimental observables, there is no hope for us to verify or falsify the corresponding hypotheses. On the other

hand, if two models exhibit divergent simulation results, it gives us some confidence to conclude on the hypothesis that aligns with experimental data.

Method and Experimental details

(1) Modeling and Simulation:

We utilize BioNetGen[11], a local rule-based modeling language and an accompanied simulator based out of Gillespie algorithm. The experimental observables, static light scattering curve we want to derive from the simulation is given by the following. Given parameter set p, the curve over time t is: $R(t,p) = k\ c\ S(t,p)$, where k is scaling factor, c is concentration of monomers, and $S(t,p)$ is the average assembly size at time t [3].

The main difficulties of this step was that BioNetGen is only able to specify finitely many different particles and there is no direct way to encode parameterized particles (e.g. filament of size n). We solved this by embedding static light scattering observables into the reaction rules; instead of deriving the scattering curve from trajectories of particles distribution, we directly . The key observation that leads to this is that the change incurred in the static light scattering data by one reaction, can be locally determined by their reaction rule. Therefore we can simply create dummy particles that keep track of the drop/increase in scattering data per reaction inside the rules.

Other technicalities include: every reaction rule in BioNetGen can only have at most 2 reactants. Those were resolved by transforming into well-formed models that can be shown to have the exact same semantics under the simulation algorithm. The full model of nucleation limited filament assembly can be found on the homepage of this report.

(2) Inferring mechanisms through parameter fitting via simulation optimization:

Given two parameterized models, we want to determine if they could possibly appear the same from the static light scattering curves. However, stochastic nature of self-assemblies requires the conclusion to account for the fact that many different trajectories can be derived from single model.

Here we adopt the most simple and obvious approach: we average the static light scattering curves from multiple trials of simulation and see if we can fit one model's averaged curve to another 's under some parameters. We define the fitness of two curves as rooted mean of pointwise squared distance[3]. We attempt to find the parameters that gives the best fit. The objective function under this setting of parameter optimization problem cannot be accurately evaluated and takes time to be estimated through simulation(need to run trials of simulation and average the curves). We use generic optimizer toolbox provided by MATLAB to scan for the parameters that yield the better fit. Although this approach has arguably many drawback (explained in future work section), it has been used in many simulation optimization problems and shown to effectively capture the dynamics [4].

Results

1. Contrast group

To establish the number of simulation trials in order to capture the dynamics precisely enough, we conduct contrast group experiment: we set non nucleation limited model as ground truth model and simulate for their averaged curve. We instantiate the parameter fitting framework to see if the exact same model is able to fit the same ground truth model with parameters close to ground truth. We found 20 trials of simulation for ground truth and 10 trials for simulation optimization gives an averaged curve that is learnable of itself.

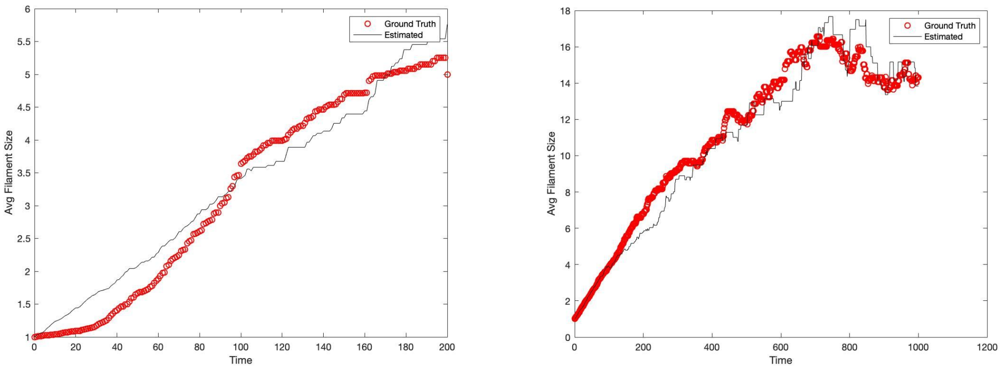

Figure 1: Parameter fitting to the same model (Number of simulation trials 10/20)

2. Nucleation limited versus non nucleation limited one

The experiment shows under low concentration (total number of monomers = 50), simulation optimization between nucleation limited model and non nucleation limited model never yield a good fit. This implies that we can relatively confidently conclude about nucleation limitation from static light scattering experiments.

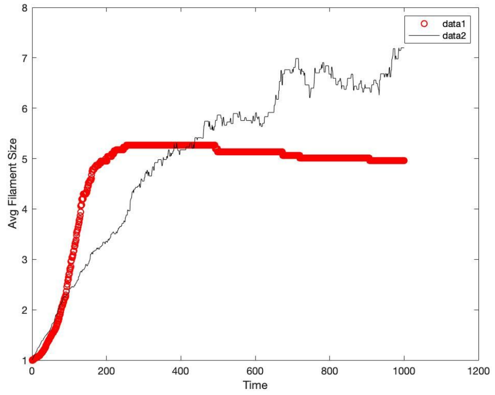

Figure 2(a): Nucleation vs non-nucleation (Low concentration, total number of monomers = 50)

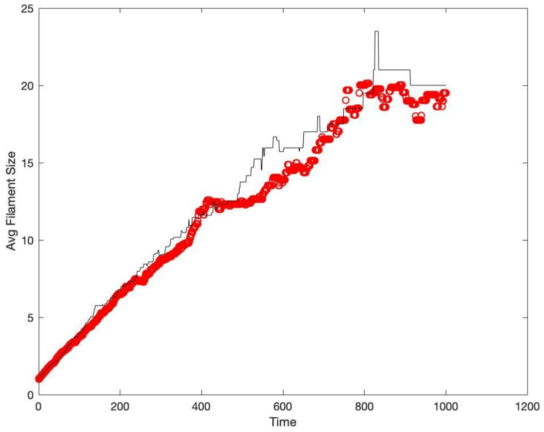

Figure 2(b): Nucleation vs non-nucleation (High concentration, total number of monomers = 200)

However under high concentration (total number of monomers = 200), the experiment shows a good fit between the two models. By observing the simulation traces, we conjecture this is due to combinatorially large of monomer pairs dominates the reaction propensity, over the nucleation guided reactions during the simulation dynamics. However, it is unclear whether it is truly the case for high concentration in vitro experiments; or our local rule-based models are not expressive enough to capture the difference.

3. Nucleation size = 2 and nucleation size = 4

The experiment shows the size of nucleation cannot be accurately estimated through static light scattering. Figure 3 shows two trials of successful parameter fitting between two models of different nucleation sizes.

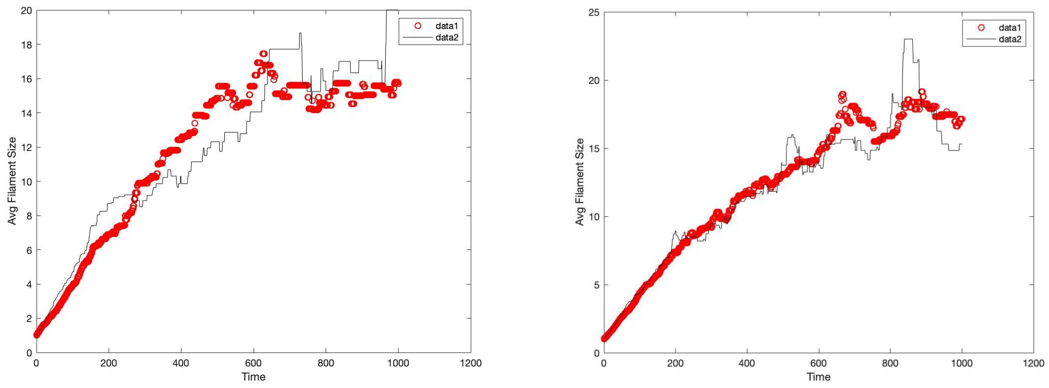

Figure 3: Nucleation size = 2 vs. size = 4

Conclusions

1. Under low concentration of monomers, nucleation limited filament assemblies are observably divergent from those that are not. However, under high concentration, nucleation limited filament assemblies cannot be distinguished from those that are not, in our modeling and simulation setting. Whether the divergence simply doesn't exist under high concentration, or is lost during the modeling abstraction, cannot be concluded from this project.
2. The size of nucleation cannot be estimated precisely from static light scattering curve.

Future Work

Future directions of this project include the following aspects:

1. Further justify the conclusions drawn in this project
    a. Confidence bounds derived from number of simulations

       In this project we averaged trial of simulation curve to characterize the dynamics of the model. Yet it is unclear theoretically how many curves we should sample to get a relatively confident data of the dynamics. Since the distribution we sample from at every time step is fully specified, some sort of measurement of confidence bounds should be derivable from the number of simulations.

    b. Fitting time series data [8]

For time series data like static light scattering, it is not well-justified to take the average of trials of simulations to get a characterization of the dynamics. It is fairly plausible that some dynamics can stochastically flip to several different "phases" with very different observables. In this case, averaging out the simulation trials would not make sense since it would essentially come from different dimensions.

2. Explore difference paradigmes of computational inference of biological mechanisms

The main issue with current approach is that the computational inference relies heavily on domain knowledge for biology. For example, when we conduct parameter fitting,the result is very sensitive to the time frame we choose to fit to. Currently, this solely relies on biological observation of "the stable status of the dynamics". There is a potential to separate domain knowledge embedding and computational inference to conduct computational inference in a more systematic way. Possible directions include probabilistic program synthesis, Markov chain identification [7] [9].

Reflections

- Even the ideas that seem straightforward and easy to execute incur many drawbacks and take time to address those issues. I had several experiences throughout the semester where I thought I can implement some method right away, but I ran into many problems, including both empirical issues (e.g. the modeling language has many limitations/bugs), and purely theoretical flaws that I didn't think it through before executing it. I learned the importance of executing the ideas as soon as possible.

- Keeping a milestone schedule was a lot more helpful than I imagined. When I drafted the milestones at the end of last semester, I wasn't sure how that would help me carry out the project, since many things were undefined, or out of my knowledge at that moment to figure out. But throughout the semester, having a milestone schedule, and referring back/updating it was really effective. For one part, keeping the whole picture of the project was really effective for me to orient the project at the initial phases. For another part, having a schedule forced me to balance between going off random ideas and settling down and working through feasible directions.

Special thanks to my advisor, Prof. Russell Schwartz for providing advice throughout the semester as well as the freedom to experiment e new ideas. I would like to thank 15-300/400 for providing guidance throughout the year. It wouldn't have been a rewarding experience if it wasn't for their valuable advising and support.

References

[1] Smith GR, Xie L, Lee B, Schwartz R. Applying molecular crowding models to simulations of virus capsid assembly in vitro. Biophys J. 2014;106(1):310–320.

[2] Sweeney B, Zhang T, Schwartz R. Exploring the parameter space of complex self-assembly through virus capsid models. Biophys J. 2008;94(3):772–783.

[3] Lu Xie, Gregory R. Smith, Xian Feng, Russell Schwartz, Surveying Capsid Assembly Pathways through Simulation-Based Data Fitting, Biophysical Journal, Volume 103, Issue 7, 2012

[4] Zhang T, Schwartz R. Simulation study of the contribution of oligomer/oligomer binding to capsid assembly kinetics. Biophys J. 2005;90(1):57–64.

[5] Stanley CB, Perevozchikova T, Berthelier V. Structural formation of huntingtin exon 1 aggregates probed by small-angle neutron scattering. Biophys J. 2011;100(10):2504–2512.

[6] Thomas M, Schwartz R. A method for efficient Bayesian optimization of self-assembly systems from scattering data. BMC Syst Biol. 2018;12(1):65. Published 2018 Jun 8.

[7] Castelletti, Federico; Consonni, Guido; Della Vedova, Marco L.; Peluso, Stefano. Learning Markov Equivalence Classes of Directed Acyclic Graphs: An Objective Bayes Approach. Bayesian Anal. 13 (2018), no. 4, 1235--1260.

[8] A. Saad, Feras & E. Freer, Cameron & L. Ackerman, Nathanael & Mansinghka, Vikash. A Family of Exact Goodness-of-Fit Tests for High-Dimensional Discrete Distributions. 2019

[9] Feras A. Saad, Marco F. Cusumano-Towner, Ulrich Schaechtle, Martin C. Rinard, and Vikash K. Mansinghka. 2019. Bayesian synthesis of probabilistic programs for automatic data modeling. Proc. ACM Program. Lang. 3, POPL, 2019

[10] https://kappalanguage.org

[11] Harris, L. A. et al. BioNetGen 2.2: advances in rule-based modeling. Bioinformatics 32, 3366–3368 (2016)

[12] http://gfrd.org/overview