Improved Model Fitting for Complex Self-Assembly Reaction Networks

Lu Xie[1,3], Gregory R. Smith[2], Marcus Thomas[2], and Russell Schwartz[2, 3]

Abstract — **Self-assembly is a crucial component of nearly every cellular function, yet quantitative modeling of self-assembly remains primitive due to the experimental challenges of monitoring assembly reactions and the computational challenges of accurately and efficiently simulating them. Coarse-grained rule-based models have made accurate and efficient sampling of reaction trajectories possible but learning parameters for real systems remains challenging given the limits of experimental assays of assembly progress. We describe advances in model inference, specifically exploring the potential of improved fitting algorithms and improved data sources to more accurately and efficiently learn correct assembly models from bulk experimental measures of assembly progress. Exploration of simulated virus capsid assembly data suggests that better algorithms and better data sources can each independently lead to more accurate and precise model fits, although the advantage of better algorithms diminishes with richer data. Application of the methods to real viral data provides novel insights into pathway selection in unprecedented detail as well as a platform for exploring the effects of changes to better mimic the cellular assembly environment versus the *in vitro* conditions under which kinetic data is gathered.**

Keywords **— Self-Assembly, Virus Capsid, Stochastic Simulation, Rule-based Models Data Fitting, Optimization.**

I. Motivation

VIRUS capsid assembly has long been a model system for general macromolecular assembly due to its high complexity and relative experimental tractability. Nonetheless, detailed quantitative understanding of subunit-level assembly pathways has remained elusive. There are no experimental methods to observe fine-scale assembly dynamics directly, only to monitor bulk assembly *in vitro*. Simulation methods provide a window into the unobservable fine details of assembly, but are hindered by the huge potential pathway space of even small complexes. We have previously developed methods to simulate realistic scales and parameters ranges of capsid assembly through the use of coarse-grained, rule-based models [1] and their combination with fast stochastic sampling algorithms [2]. We later addressed the lack of direct data through simulation-based model fitting to static light scattering (SLS) measurements of bulk assembly *in vitro* [3,4]. This approach made it possible for the first time to model subunit-level pathway space for real capsids and to explore how pathway usage might vary in more realistic models of the intracellular environment [5]. Nonetheless, limited data sources and uncertain fit quality call into question the reliability of the model inferences.

Acknowledgements: This work was supported by NIH grant 1R01AI076318.

[1]Joint Carnegie Mellon – University of Pittsburgh Ph.D Program in Computational Biology and [2]Department of Biological Sciences and [3]Computational Biology Department, Carnegie Mellon University.

II. Modeling and Methods

We develop local rule models for viral capsids and simulate them via stochastic sampling as in our prior work. We use these to explore three experimental assays in current use: SLS, time-resolved non-covalent mass spectrometry (NCMS), and dynamic light scattering (DLS). We use simulated capsid models with artificially chosen rate parameters to simulate idealized data from each source. We then fit model parameters to simulated data using either our prior gradient-based algorithms or variants on derivative-free optimization (DFO) to minimize root mean square deviation between the model and data. We evaluate quality of fit by accuracy of inferred parameters and reaction trajectories. We also apply the methods to real SLS data to assess fit quality and explore pathway space in real viruses.

III. Discussion

Our results indicate that learning accurate models of complex assembly reaction networks is feasible via simulation-based data fitting. Richer data for monitoring bulk assembly can yield substantial improvements in fit quality over past work, although the best algorithms can learn generally accurate models even from older SLS data. Our results suggest that further work on experimental methods and algorithms is needed. Nonetheless, they show that these approaches already have enormous and largely unappreciated potential for exploring a critical but still poorly handled aspect of cellular reaction networks.

References

[1] Schwartz R, Shor PW, Prevelige B, Berger B (1998) Local rules simulation of the kinetics of virus capsid self-assembly. *Biophys J.* **75**, 2626-2636.

[2] Zhang T, Rohlfs R, Schwartz R (2005) Implementation of a discrete event simulator for biological self-assembly systems. *Proc. 37th Winter Simulation Conf.* pp. 2223-2231

[3] Kumar MS, Schwartz R (2010) A parameter estimation technique for stochastic self-assembly systems and its application to human papillomavirus self-assembly. *Phys. Biol*. **7**, 45005-45016.

[4] Xie, L, Smith, GR, Feng, X, Schwartz, R (2012) Surveying capsid assembly pathways through simulation-based data fitting. *Biophys. J.* **103**, 1545-1554.

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