Carnegie Mellon University
Mellon College of Science

Thesis

Submitted in partial fulfillment of the requirements for the degree of Ph.D. in Biological Sciences

Unraveling the Role of Cellular Factors in Viral Capsid Formation
Title

Presented by Gregory Robert Smith

Accepted by the Department of Biological Sciences

Major Professor Russell Schwartz Date 3/13/15

Department Head John Wooley Date 3/16/15

Approved by the MCS College Council

Dean Frederick J. Kuhlman Date 3/16/15

Unraveling the Role of Cellular Factors in Viral Capsid Formation

by Gregory Robert Smith

Submitted to the Department of Biological Sciences

in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

March 3rd, 2015

Thesis Advisor: Russell Schwartz

Thesis Committee: Dr. Dannie Durand

Dr. Frederick Lanni

Dr. Alex Evilevitch

Unraveling the Role of Cellular Factors in Viral Capsid Formation

By

Gregory Robert Smith

Department of Biological Sciences

Carnegie Mellon University

Pittsburgh, Pennsylvania

Abstract

Understanding the mechanisms of virus capsid assembly has been an important research objective over the past few decades. Determining critical points along the pathways by which virus capsids form could prove extremely beneficial in producing more stable DNA vectors or pinpointing targets for antiviral therapy. The inability of current experimental technology to address this objective has resulted in a need for alternative approaches. Theoretical and computational studies offer an unprecedented opportunity for detailed examination of capsid assembly. The Schwartz Lab has previously developed a discrete event stochastic simulator to model virus assembly based upon local rules detailing the geometry and interaction kinetics of individual capsid subunits. Applying numerical optimization methods to learn kinetic rate parameters that fit simulation output to *in vitro* static light scattering data has been a successful avenue to understand the details of virus assembly systems; however, information describing *in vitro* assembly processes does not necessarily translate to real virus assembly pathways *in vivo*. There are a number of important distinctions between experimental and realistic assembly environments that must be addressed to produce an accurate model.

This thesis will describe work expanding upon previous parameter estimation algorithms for more complex data over three model icosahedral virus systems: human papillomavirus (HPV), hepatitis B virus (HBV) and cowpea chlorotic mottle virus (CCMV). Then it will consider two important modifications to assembly environment to more accurately reflect *in vivo* conditions: macromolecular crowding and the presence of nucleic acid about which viruses may assemble. The results of this work led to a number of surprising revelations about the variability in potential assembly rates and mechanisms discovered and insight into how assembly mechanisms are affected by changes in concentration, fluctuations in kinetic rates and adjustments to the assembly environment.

Acknowledgements

I would first like to thank my adviser, Dr. Russell Schwartz, for all the guidance and support he has provided me throughout my time at Carnegie Mellon. When I was still in the Master's Program in Computational Biology, it was taking his course on biological modeling and simulation that inspired me to want to continue into a PhD program and study this fascinating subject. Dr. Schwartz's door has always been open for any questions and concerns and ideas I have had and his insight into my research as well as preparation for my future career goals have been extremely helpful.

I would also like to thank the other members of my thesis committee Dr. Frederick Lanni, Dr. Dannie Durand and Dr. Alex Evilevitch for their suggestions and input into my thesis work as well as their careful reading of my thesis. Their often challenging questions have greatly improved the quality of my research and have served as great preparation for conference and career presentations.

I would further like to thank the other professors at Carnegie Mellon and the previous institutions I attended for the knowledge they have provided, the training to be a better student and researcher and for opening my eyes to so many interesting subjects. My research has required input and understanding of a diverse array of scientific disciplines and none of this would have been possible without the skills and adaptability they instilled in me.

I would like to thank all of the past and present members of the Schwartz Lab. I have thoroughly enjoyed being a part of the research environment in the lab and have been extremely lucky to be part of a group that not only promotes great collaboration, but also makes me look forward to going to work every day. I want to give special thanks to Lu Xie who has also been

working on the virus capsid assembly project for the past five years and his work has been instrumental to mine. I will also give special thanks to Byoungkoo Lee, Rupinder Khandpur, Xian Feng, Han Lai and Marcus Thomas who have all worked on the capsid assembly project during my time in the lab. I would also like to thank lab members from before my time - Tiequan Zhang, Rori Rholfs and Blake Sweeney - who all laid the groundwork for the capsid assembly simulations detailed in this thesis. Once a member of the Schwartz Lab, always a member of the Schwartz Lab.

I would like to thank David Wu and Adam Zlotnick for the experimental data used in this thesis work as well as the Lane Center for Computational Biology for the use of its computing cluster. I would also like to thank NSF and NIH for their funding support of this project.

Lastly, I would like to thank my friends and family for all of their love and support and especially my parents whose personal sacrifices to provide a better future for me I will never truly be able to repay. They have always nurtured my desire for knowledge and excitement for understanding the complexities of the world in which we live. That excitement is in large part the reason I have always loved science and am so thankful that I have been given such amazing opportunities to do work about which I am so passionate.

Table of Contents

Chapter 1: Introduction

From the beginnings of modern biological research, the study and utilization of viruses have played a huge role in increasing the understanding of disease, genetics and how cells function. Viruses have become an effective tool as DNA vectors (1) both for their use to study the specific function of genes as well as its potential for medical therapy (2). Viral infection leads to millions of deaths each year and countless more stricken with disease. Furthermore, it is estimated that twenty percent of all cancer is the result of viral infection (3). The 2014 Ebola epidemic is a perfect example of the numerous impacts of a viral outbreak, not only in the loss of lives directly resulting from viral infection but also long-term societal impacts, deterioration of public health from starvation or failure to treat other illnesses, economic impact from decreased productivity, and the potential for political instability (4). While vaccination has been a successful route at combating a number of viral outbreaks, promoting a targeted immune response against specific viral antigens is not always possible. The rapid mutation rate of many viruses, such as Human Immunodeficiency Virus (HIV) (5,6), or the variety of strains for a specific virus (7), as is the case for influenza, require alternative approaches to be developed to address viral infection.

There is great need, then, to develop a more detailed understanding of the different stages of the viral life cycle, and from a clinical perspective, to determine what are the critical points in that life cycle to potentially target for inhibition. Researchers have targeted a number of potential points during viral replication (8) including viral attachment to the host cell (9-10), viral entry (11-13), viral polymerase function (14-16), viral capsid assembly (17-18), viral capsid maturation (19) and viral exocytosis (20) among many others. My research has focused on the

assembly of the virus capsid, and specifically understanding what are the potential assembly pathways by which capsids form inside living cells.

1.1 Virus Structure and Assembly

Viruses are composed of an external capsid shell which consists of proteins encoded by the genetic material enclosed within. These proteins produce a geometry determined by virus-specific bonding angles and interactions between proteins. The resulting geometries can vary greatly, from helical shapes with only one form of capsid subunit to icosahedral shapes with a variety of confirmations of capsid subunits to more complicated viral structures such as bacteriophages which contain components with both icosahedral and helical symmetry. Icosahedral viruses are further classified based upon their triangulation (T) number. A T=1 capsid consists of 60 capsid protein subunits forming a simple dodecahedron. A T=3 capsid consists of 60*3 = 180 subunits forming a truncated icosahedron, a T=4 capsid consists of 60*4 = 240 subunits forming a truncated rhombic triacontahedron and so forth. Smaller viruses will generally have a single layer of protein capsid, while larger viruses may contain two separate capsid layers as is the case with bluetongue virus (BTV). Viral genetic material can be in the form of double stranded or single stranded DNA, double stranded RNA or single stranded RNA. Single stranded RNA (ssRNA) viruses can be further classified into (+)-sense and (-)-sense RNA viruses where (+)-sense RNA viruses are in the same sense as mRNA and (-)-sense RNA viruses are in the opposite sense. Viral proteins essential for replication may also be included within the capsid depending upon the complexity of the virus. Many viruses will also contain a lipid envelope surrounding the capsid to mirror the cellular membrane to assist with viral attachment. A variety of surface proteins may also be present on the capsid or envelope surface to facilitate interaction with receptors on the surface of target cells. Likely the most well-known example of

this is the surface receptors for influenza: hemagglutinin and neuraminidase, the specific forms of which dictate the strain description for the virus.

The assembly of the capsid from individual capsid proteins is a critical step in the viral life cycle; however, ascertaining detailed information about the steps between single monomers and completed capsid is very difficult. Current experimental techniques are still very limited in their ability to see time-dependent details of assembly progression. *In vitro* studies have so far been the extent of the capability to examine dynamics of viral assembly. Even these studies, however, cannot directly measure the formation of viral capsids *in vitro*. Instead, what is available are techniques that provide an approximate progression of capsid assembly over the time of the experiment, such as light scattering, fluorescence correlation spectroscopy or certain mass spectrometry methods. Because of this gap in current experimental capabilities, computational simulation methods provide an alternative approach to addressing this problem and have, to date, offered an unprecedented view into the details of capsid assembly pathways.

Virus capsid assembly is viewed as a classic model for self-assembly systems as individual capsid protein subunits can spontaneously assemble into completed structures in an appropriate solution. Many of the insights learned from researching virus assembly can be applied to other self-assembly systems, whether they are cytoskeletal filaments, lipid membranes or even malformed proteins assembling into insoluble amyloid fibers in prion-related disorders. Simplistically, this problem can be resolved purely through simulating or experimentally measuring individual capsid proteins spontaneously forming completed capsid structures; however, *in vivo* assembly paints a far more intricate picture. To truly understand how viruses assemble in their natural environment, many other factors have to be considered. Some factors are related to the solution in which the capsids are assembling, such as the level of

macromolecular crowding present at the site of assembly, or the presence of cellular proteins that can either play a promoting, as in the case of chaperone proteins, or inhibitory role. Another major factor that affects virus assembly is the presence of nucleic acid. While some viruses, most notably double stranded DNA viruses, have their genetic material inserted into the capsid via a motor protein following completion of the capsid, many other viruses will form their capsids around their genetic material. A further major advantage of computational simulation methods, then, is the ability to consider a variety of these major changes in assembly environmental conditions, both individually and in concert, which is exceedingly difficult experimentally.

1.2 Discrete Event Simulation

1.2.1 Assembly Simulation Techniques

Attempts to understand the biophysics of molecular assembly systems have long included simulation techniques (21,22), especially for problems that need to compensate for a lack of experimental evidence. This is particularly the case with regards to virus capsid assembly where simulation methods have made it possible to infer various emergent properties of hypothetical assembly models (23–26), to explore the effects of perturbations in parameter spaces (25,27–29), and to examine possible assembly pathways and mechanisms accessible to theoretical models (27,30–33). A wide variety of simulation techniques have been utilized ranging from ordinary differential equation (ODE) models (21-26), coarse-grained Brownian dynamics (BD) models (27-33), and discrete stochastic simulation algorithm (SSA) models (34-39). ODE simulations treat the concentration of each potential intermediate in capsid assembly as a variable and proceed to model each changing concentration over time. BD models take a far more detailed approach to individual molecular motion and interaction measuring changes in individual particle location over time. In order to scale to the complexity of the virus capsid, generally some

amount of course graining is necessary to limit computational cost. SSA simulations take an alternative discrete event approach as opposed to the other continuous simulation methods discussed. Here, each capsid protein is treated as a monomer and individual bond forming or breaking events between monomers or resulting intermediates are treated as discrete events within the simulator.

While each method has its own distinct advantages and disadvantages, the SSA and ODE approaches share a number of technical advantages over BD models: simulating large numbers of trajectories needed for computing pathway statistics, efficiency over a wide range of parameter values, and parameterization in terms of small sets of reaction rate constants. From a computational efficiency perspective, ODE models are far more capable than SSA models of handling large quantities of individual monomers in capsid assembly simulations, as the complexity of SSA models are dependent upon the size of the monomer pool; however, ODE model complexity is dependent upon the number of intermediates whose concentrations must be maintained and updated. Furthermore, the differential equations themselves become increasingly complex as more potential intermediates are introduced. This is exceedingly problematic for modeling capsid assembly where there can be astronomical numbers of potential subunit combinations present as intermediates between monomers and completed capsid. Lastly, as with BD models but unlike ODE models, SSA methods can examine assembly details at the single-particle interaction level, which is essential for pathway analysis. Because of these advantages, the Schwartz Lab has chosen SSA as the best approach to tackling this important problem.

1.2.2 Model Viruses Studied

The viruses studied in the present work are hepatitis B virus (HBV), cowpea chlorotic mottle virus (CCMV) and human papillomavirus (HPV), all three of which are model

icosahedral viruses with pre-existing static light scattering data. HBV is a circular DNA virus with a not fully formed double strand. The virus is enveloped by a lipid membrane covering an icosahedral capsid. The capsid consists of 120 dimeric subunits, which exist in one of two binding conformations of equal numbers (40). The first conformation constructs pentamers while the second conformation forms trimers in between the pentamers with the end result forming a T=4 capsid. Interestingly, while HBV is a DNA virus, the HBV capsid assembles around its pre-genomic RNA and a polymerase protein and only following encapsidation is the RNA reverse transcribed into a final DNA form. CCMV is a (+)-sense single-stranded RNA virus, which is non-enveloped and whose capsid consists of 180 identical subunits which are present in three binding conformations in equal numbers. The first conformation forms 12 pentamers while the second and third conformations interact with each other to form 20 hexamers (41). Connecting these pentamers and hexamers forms the complete T=3 Capsid. As with HBV, the CCMV capsid assembles around RNA and interactions between nucleotides and capsid proteins play an important role in the assembly process. HPV is a double-stranded DNA virus that is not enveloped by a lipid shell and its DNA is inserted into the viral capsid following assembly completion via a motor protein. Its outer capsid consists of 72 equivalent star-shaped capsomers named L1 (42). L1 is itself a pentamer of smaller subunits and can exist in two separate binding conformations. Sixty of the L1 capsomers are present in the first conformation and the remaining twelve are present in a second conformation. In the first conformation, L1 behaves as a pseudo-hexamer and interacts with 6 other capsomers, one of which is in the second conformation. In the second conformation, L1 interacts with only 5 other capsomers, all of which are in the first conformation. This interesting combination of conformations leads to a resulting capsid which can be described as a T=7 dextro icosahedral lattice.

1.2.3 Lab Simulation Technique

The simulations described in this work are conducted by Discrete Event Simulator of Self-Assembly (DESSA) (33), a rules-based discrete event stochastic simulator designed to model the process of capsid assembly from individual subunit building blocks through a sequence of single association and dissociation events into completed capsids. Simulated assembly is governed by simple biochemical rule sets specifying the geometries of the subunits, three-dimensional positioning of binding sites and the specificities and on- and off-rates of binding events between binding sites. With respect to the viruses studied (CCMV, HBV and HPV) all subunits that compose the final capsid structure for each virus are the same; however, different conformations of these subunits are necessary to form the completed capsid structure. The combination of conformations, their shapes, and their binding site specificities imply an overall geometry to the completed capsid. In DESSA simulations, for example, HPV consists of 60 pentamers and 12 pseudo-hexamers. All 72 small oligomers have the same chemical composition, but different conformations lead to different binding sites and thus different binding interactions.

A schematic representation of the local rules for each capsid model in the simulator is shown in Figure 1.1. In the viral systems studied, the individual subunits can represent either individual coat proteins or small stable oligomers of coat proteins. In the case of HPV, the subunit corresponds to a pentamer of HPV capsid proteins, which experimental evidence has shown to be the basic unit of assembly (43,44). For CCMV and HBV, dimers of coat proteins are selected as the individual subunits, as the experimental data also involved *in vitro* assembly from

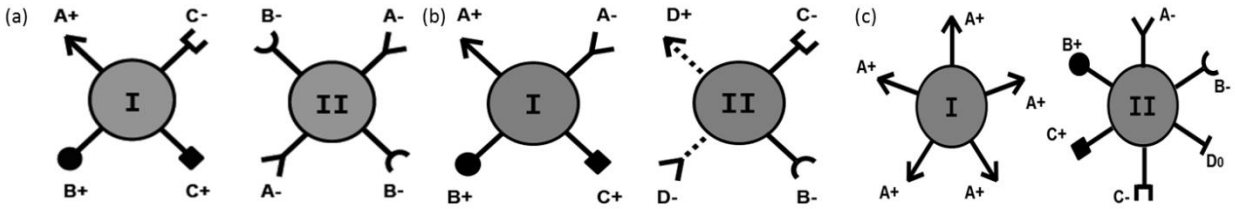

Figure 1.1. Cartoon schematics of local rules models for CCMV (a), HBV (b) and HPV (c). Each rule specifies a set of binding sites occupying a desired position in space relative to the center of the subunit, chosen to yield the correct capsid geometry for each given virus. Each binding site has specificity for binding a specific other binding site. Sites labeled *N*+ bind those labeled *N*- and sites labeled N_0 bind to other cites labeled N_0 for any *N*. For example, a site labeled *C*+ binds one labeled *C*- and a site labeled D_0 binds a D_0 site on another copy of the same subunit. In addition, each site has a defined on- and off-rate, specified in the local rules for a simulation.

coat dimers (45-48). For a given simulation run, DESSA samples among all possible bond formation (association) and breaking (dissociation) events at each step in the simulation using a variant of the stochastic simulation algorithm (SSA) (49) developed to accelerate runtime and reduce memory overhead for systems such as viral capsids with large numbers of distinct intermediates (50). For each potential reaction event, a corresponding event time is calculated by sampling an exponential distribution with mean value the given kinetic rate for that interaction. The minimum-time event among all possible events is selected based on these sampled times, yielding a kinetically correct sample of possible trajectories from among the full ensemble implied by the rule set. Once an event is selected, the simulator evaluates the reaction to ensure it is not sterically hindered and, if not, implements it and updates the event queue with times for any possible reactions enabled by the new event. A simulation ends when either all potential events have been exhausted or a predetermined time limit has been reached.

1.2.4 *Learning Kinetic Rate Parameters*

One of the major hurdles to overcome in developing biologically-relevant simulations is to find appropriate input values for the local rules model. While the geometries of the virus capsid proteins studied are well known, the kinetic rates dictating bond forming and breaking

events cannot be directly measured by any contemporary technology. A variety of approaches have been implemented to address this lack of direct measurement, such as using simulation models to find ranges of rates resulting in successful assembly (34,35), using analytical fits of small numbers of parameters to fit light-scattering data (36), determining approximations for rate parameters from structural models (38,39), applying further pathway constraints such as those imposed by virus-specific capsid protein-nucleic acid interactions (37), applying global estimates of the free energy of assembly to infer averaged rate approximations (36), or scanning potential parameter spaces to constrain the range of possible virus assembly behaviors (25,26,28,29,33,51,52). That said, precise values for rate constants are still not known for any real viral system and simulation work suggests that assembly pathways and mechanisms are highly sensitive to changes in rate parameters (33). This suggests that approximations are not a viable solution and more exact determination of kinetic rates must be achieved in order to yield faithful reproductions of real assembly behavior.

In prior work, the Schwartz Lab proposed an alternative strategy of using optimization algorithms to learn kinetic rates that fit DESSA simulation output to *in vitro* data (53). Parameter estimation when applied to a computational or mathematical model is represented as an optimization problem over the parameter space with the goal to minimize the deviation between the model and some measurable behavior of the real system. This process then affords an otherwise potentially impossible opportunity to learn parameters that cannot be computed analytically or determined experimentally. This prior work sought to learn kinetic rate parameters that minimized deviation between indirect measures of assembly progress, in this case *in vitro* static light scattering data, and projections of those measures implied by simulation results.

To address the particular difficulties of capsid assembly systems, the method derived by the Schwartz lab uses a novel data fitting algorithm that interpolates between response surface (54) and quasigradient (55) approximations to allow rapid convergence in smooth regions of the objective function while still robustly handling bumpier regions. While this algorithm was used specifically with light scattering data, it is not dependent upon that single data type and can be applied to a variety of time-dependent assembly data sources. The algorithm was tested on a single light scattering curve measuring HPV assembly progression from Casini et al. (43) and demonstrated that a set of physically plausible on and off rates could be learned and provide a good fit to experimental data. It was then still to be determined, however, if the method could be generalized to other virus systems or if HPV was a special case with a relatively simplistic objective surface. Furthermore, the assembly mechanism learned for HPV was primarily a simplistic monomer-based addition model where individual simulation subunits would add on to growing capsid structures one at a time until the capsid structure was completed. It was still to be determined if this mechanism was the norm with regards to assembly pathways or if more complex pathways were possible and, if so, could learned parameters recreate these more complex pathways.

1.3 Transition from *in vitro* Data to *in vivo* Understanding

While designing viral models that accurately reflect capsid assembly measured by *in vitro* light scattering experiments is an important step, this is not the end goal. Even a perfectly faithful model of assembly *in vitro* may yield only limited insight into the natural assembly of the virus. This is because *in vitro* assembly environments are very different from the environment of a living cell in which a virus would normally assemble. The cellular environment poses a multitude of obstacles and advantages which are not present in a test tube. The cytoplasm is

extremely crowded with proteins, lipids, nucleic acid and other macromolecules that can greatly hinder motion, which reduces both the rate of forming new bonds, because of a decreased rate of diffusion of molecules through the crowded environment, and breaking existing bonds, because of excluded volume effects ensuring increased molecular compound stability. Recent studies have shown how crucial including crowding information can be, specifically with respect to HIV (56).

Another important issue is the presence of chaperone proteins in the cytoplasm which have been found to aid in a multitude of parts of the viral life cycle, from nucleic acid replication to transport and assembly (57,58). A viral protein NCp7 which acts as a chaperone has been shown to be a major factor in HIV-1 genome replication and the binding of NCp7 to the viral RNA helps the capsid proteins recognize around which RNA to assemble (59). There will also be a discrepancy with regards to the amount of viral proteins present in the cytoplasm compared to a specific amount available in an *in vitro* experiment. This can arise both in the ability of a host cell to produce the necessary proteins for viral assembly as well as where the proteins are localized within the cell. This could be disadvantageous to the virus if it is localized poorly or potentially advantageous if proteins are localized near other important cellular machinery necessary for viral assembly. In the case of flock house virus (FHV), dense patches of viral capsomers form near sites of RNA replication inside cells (60). This advantageous increase in capsid protein concentration would show higher assembly rates than a standard diffuse *in vitro* system.

A final major missing piece from current experiments is the presence of nucleic acid in capsid assembly. Not all viruses require their genetic material at the time of capsid assembly; however, it is an essential attribute of many, especially in viruses which assemble around RNA

as is the case for both HBV and CCMV. RNA is a negatively charged molecule and its presence can make a large difference in the thermodynamics of the assembly environment, beyond the potential direct interactions that may also aid in the assembly process (61-63). In the case of CCMV, the first 25 amino acids of the N-terminus of its capsid proteins contain an arginine-rich motif which can bind to RNA (64). The deletion of this motif from the capsid proteins caused malformed capsids to assemble *in vivo*.

In principle, computational methods provide a way to bridge this gap between *in vitro* and *in vivo* environments by allowing us to learn interaction parameters of assembly proteins from the *in vitro* system then observe how their behavior changes when transferred to a more faithful computational model of the environment *in vivo*. Accurately accounting for the myriad differences between the two environments is not a simple task, however, especially when many of the individual differences are still imperfectly understood. The extreme complexity of the true system and the uncertainty regarding which factors actually affect assembly pathways suggest the prudence of a bottom-up approach: studying individual factors of interest and determining whether they are likely to, singly or in combination, substantially alter assembly mechanisms. This thesis will, in part, seek to address two of these major differences in assembly environment: the presence of RNA during capsid assembly and increased macromolecular crowding, both of which result in fascinating changes to the rate and mechanisms of assembly.

1.3.1 RNA-Capsid Protein Interaction

The role of RNA during capsid assembly is a hotly contested source for debate at the intersection of virology, computational biology and physical chemistry. Experimental evidence, already very limited when it comes to capsid assembly, is even more limited when examining assembly around RNA. This lack of evidence raises a number of questions that are by no means

trivial to answer: How does the presence of RNA change the electrostatics of the assembly environment? Does direct interaction between RNA and capsid proteins act as a recruiting or nucleating influence in assembly? What is the structure of the RNA inside the capsid and does the RNA fill the entire space within the capsid or is it confined further? The answers to these questions most certainly vary between viruses, especially as the complexity of the system increases. For my work in tackling this problem, I use the relatively simplistic system of CCMV whose assembly *in vivo* is solely around its RNA without the extra complications of other viral proteins. The HBV capsid, for example, assembles around both its pre-genomic RNA and a polymerase protein which will later be necessary to reverse-transcribe the RNA into a final DNA form and, while the polymerase protein is rather small with a low net charge compared to the RNA or capsid surface, the interaction between the RNA and the polymerase protein within the capsid protein is not known nor how their interaction may alter the conformation of the RNA and thus modify the electrostatics of the capsid assembly process.

Even in the comparatively straightforward case, as with CCMV, of a capsid assembling solely around its RNA, there is great ongoing debate about the relative importance of sequence specificity in RNA packaging. Evidence in certain ssRNA viral systems such as bacteriophage MS2 and satellite tobacco necrosis virus (STNV) support the importance of specific packaging signals on the RNA that act to recruit capsid proteins and potentially initiate capsid assembly (63,65). In each case, researchers have shown that inserting capsid proteins into an *in vitro* environment of STNV RNA causes a conformation change in the RNA strand allowing for the RNA to fit within the volume of the resulting capsid (65), something they claim would not be possible under only electrostatic interactions between capsid protein and RNA. Specific packaging signals are known for a number of viruses, including STNV and MS2, the deletion of

which results in competitive inhibition of proper RNA packaging when mixed with other alternative RNA strands of similar shape and size (63,65,66).

Despite this presence of specific packaging signals in a number of ssRNA viruses, other experimental evidence shows a surprising flexibility in what can be encapsidated provided certain charge and size requirements are met (67,68). In experiments conducted with CCMV viral capsids, researchers were able to encapsidate not only true CCMV RNA strands, but also RNA strands for similar viruses, even showing a stronger packaging affinity for brome mosaic virus (BMV) RNA1 than homologous CCMV RNA. Another interesting study highlighted the ability of BMV capsids to package negatively charged gold nanoparticles further suggesting that in certain viruses, packaging is not entirely specific (69). In virus studies which have highlighted this non-specific interaction between capsid proteins and charged polymers, an alternate hypothesis on what is most necessary for polymer encapsidation is proposed based upon the size and charge of the polymer. A proper molar ratio of capsid proteins:RNA around 6:1 was found to be most advantageous to ensure rapid assembly and complete encapsidation (68). In fact, in experiments to measure CCMV capsid assembly around longer RNA strands, larger capsids would form even causing a change in geometry to a larger triangulation number (68,70). At a certain RNA length, however, the geometry of the capsid could not be modified further and instead multiple capsids would form along the length of the RNA (68). While this is an extreme example in *in vitro* conditions designed to promote capsid assembly, this once again highlights the diversity in what can be packaged in certain viral systems.

While both extreme beliefs in the debate over electrostatics versus specific unique-to-virus interactions have merit and experimental evidence to support their findings, I believe the truth is an important middle ground. Showing the ability of capsid structures to assemble around

a variety of charged particles *in vitro* does not truly imply that in an *in vivo* system under real assembly conditions, the same would be true. Indeed, cells are very complex crowded structures full of potential competitors, especially native RNA, for assembly packaging and yet capsids still form properly during infection. On the other extreme, packaging signals are proven to exist in numerous viruses and there is certainly an important role to be played by binding interactions between capsid proteins and RNA; however, recruitment seems to be the most likely role for packaging signals with other direct effects having less experimental support at this time.

Since experimental approaches cannot quantify the effect of RNA on capsid assembly, various theoretical and computational approaches have been applied. A number of attempts have been made to devise analytical models that accurately reflect the interaction of RNA with capsid proteins (28,34,61,71) and calculating the free energy of the assembly reaction system as well as individual capsid protein binding events. This is often tackled by breaking calculations of free energy into individual contributions theorized to be most relevant or with the greatest impact to the system with extra emphasis on the coulomb interactions between charged particles and entropic effects due to the confinement of the RNA within the capsid structure. Computational scientists have attempted to include RNA strands in coarse-grained assembly simulations as well (72-75) and have shown some success in corroborating experimental results pertaining to specific elements of capsid assembly, for example the optimal ratio of capsid proteins to RNA to ensure RNA packaging (72). To account for the increased complexity of the RNA strand and its complex interaction with the assembling capsid, the overall simulation must be very coarse-grained and loses the detail necessary to investigate the pathways by which the capsid itself is assembling.

An interesting middle ground is offered by applying an analytical model of the effect of RNA on the free energy of the capsid assembly system as fast corrections to best-fit kinetic rates entered as parameters in a discrete event simulator. This approach has the advantage of utilizing a form of simulation tailored to learning assembly pathway details while still accounting for the electrostatic effects of the RNA on the capsid assembly process. Furthermore, by modeling the system to reflect the high concentrations of capsid protein in the near vicinity of the RNA immediately prior to assembly, this approach also accounts for the ability of RNA packaging signals to localize capsid proteins to the site of assembly.

1.3.2 Macromolecular Crowding

Macromolecular crowding is another factor that sharply distinguishes intracellular from typical *in vitro* systems and that is well known to influence kinetics and thermodynamics of numerous macromolecular assembly processes (76-79). Predicting the effects of nonspecific crowding on any particular system is extremely difficult because crowding can both inhibit growth, by slowing diffusion and impeding binding, and promote growth, by providing an entropic benefit to coalescing into more compact assembled forms due to excluded volume effects. Experimental evidence with regards to HIV capsid assembly highlights the complexity of this issue very well (56,80). Increasing levels of macromolecular crowding can limit drug effectiveness in inhibiting assembly (80). This is in part because of increased stability of growing assemblies but also increases in competitive inhibition limit the ability of the drug to interact with capsid proteins. Studies have also shown that increasing crowding levels can actually speed up the rate of assembly (56) despite the negative effects increased crowding can impose. Computational modeling studies have also shown the impact of crowding on capsid assembly (81-83) corroborating the complex nature of this impact, although the effects of crowding on

complex pathway selection is not characterized in sufficient detail to explore how assembly pathways of any particular virus may be affected by intracellular levels of crowding.

Different computational approaches have been taken to address the effects of crowding with some groups favoring highly detailed crowding models of ensembles of cellular crowders to avoid omitting possibly relevant factors (82,83), whereas others favor incorporating the minimal detail necessary to yield observed crowding behavior and argue that simple models can accurately mimic effects of far more complicated ensembles expected in nature (84). Although the latter work does not explain the precise parameters a simple uniform-crowder model should have to accurately mimic a cell-like ensemble of distinct crowders, it does suggest that scanning a range of possible total crowding levels in a simple uniform model can stand in well for a much higher dimensional search of possible combinations of crowder sizes and concentrations that might be present in any actual system.

Because of the size and complexity of viral model systems, it is advantageous to start with a simple model and incorporate only complexity shown to be most relevant to quantifying the effect of macromolecular crowding on assembly reactions. To that end, a simulator was previously constructed in the Schwartz Lab to simulate the effects of macromolecular crowding on simple assembly reactions. This crowding simulator, Three Dimensional Stochastic Off-Lattice Model (3DSOLM) (85), previously designed by a former graduate student in the lab, Byoungkoo Lee, utilized Green's function reaction dynamics (GFRD) to test effects of varying levels of macromolecular crowding on a generic homodimerization test system. Based upon the results of this simulator, a regression model was developed to accurately predict changes in equilibrium constant based upon varying input parameters, including macromolecular crowding (86). These modifications to equilibrium constant could then be applied to the rate parameter

inputs for DESSA (33) allowing for detailed capsid assembly pathway analysis while accounting for a wide range of potential crowding levels without compromising runtime.

1.4 Contributions of this Thesis

This thesis seeks to make major steps into understanding the complex nature of viral capsid assembly *in vivo*. Utilizing the in-house assembly simulator DESSA (33), I helped to expand upon the lab's parameter estimation technique (53) to function on any virus system. This parameter estimation algorithm was used to learn kinetic rate parameters for three icosahedral viruses, CCMV, HBV and HPV that best fit simulation output to *in vitro* light scattering data. I then designed a variety of data analysis methods to gain greater insight into the assembly pathways utilized by each virus during the simulations. These data analysis methods showed a significant difference between HPV whose assembly was non-nucleation-limited and primarily based upon individual monomer addition, and CCMV and HBV, both of which exhibited nucleation-limited assembly and a far more diverse set of assembly pathways.

My focus then shifted to modifying the assembly models to more accurately reflect the *in vivo* conditions seen during real virus assembly. This work focused on two such major modifications: the presence of RNA around which certain capsids such as CCMV assemble and the effects of macromolecular crowding on assembly. In order to continue utilizing the advantages of DESSA for detailed pathway analysis, I addressed each of these modifications as a series of corrections to the kinetic rate parameters in the local rules models for each virus. In the case of the effect of RNA on capsid assembly, I developed an analytical model derived in part from Flory theory (87,88) that breaks the RNA effect into four separate modifications to kinetic rates based upon RNA-RNA interaction, RNA compression within the capsid, RNA-capsid protein interaction and the increased concentration of capsid proteins around the RNA following

recruitment to the assembly location. I then ran simulations analyzing how each calculated effect or combination of effects would modify assembly rate and pathways. I found that while individual effects could be beneficial or detrimental to overall assembly rate, it was in fact the combination of all positive and negative effects that would lead to the most efficient capsid assembly, suggesting a system highly evolved to its natural assembly conditions.

In order to address the effects of macromolecular crowding on capsid assembly, I utilized a regression model based upon a previously developed crowding simulator 3DSOLM in the Schwartz Lab which calculates changes in equilibrium constant depending upon different levels of crowding agents in solution. I then applied these changes as modifications to the best fit kinetic rates for each virus I study: CCMV, HBV and HPV. I found a surprising distinction in the effects of increasing levels of crowding between nucleation-limited and non-nucleation-limited assembly pathways. In the case of non-nucleation-limited HPV, increased levels of crowding only served to decrease rate and yield of capsid assembly simulations; however, the effects were complicated for both HBV and CCMV. Despite initial decreases in assembly rate and yield at low crowding levels, as crowding continued to increase, assembly rate for CCMV and HBV began to rebound and increase as well. I theorized that despite crowding serving to decrease the rate of diffusion and thus slow down individual binding reactions, the extra stability of the nucleation step during assembly serves as a greater overall advantage to the assembly process. Lastly, I sought to combine RNA and crowding effects as a first attempt to examine multiple major *in vivo* environmental changes together in one simulation. While I once again found a complicated effect of increasing levels of macromolecular crowding on CCMV, now modeled to reflect the presence of RNA at the time of assembly, the differences in the crowding effects between hollow capsids and capsids with RNA served to raise as many questions as answers.

Chapter 2: Applying Parameter Estimation to Fitting Simulation Output to *In Vitro* Data[1]

2.1 Introduction[1]

The first major contribution of this thesis comprises the improvement of the in-house parameter estimation algorithm and subsequent development of data analysis techniques to gain insight into the assembly pathways of multiple icosahedral virus systems. In collaboration with another graduate student in the lab, Lu Xie, I modified the previously designed parameter estimation code (53) to more effectively handle the computational challenges specific to learning capsid models as well as provide flexibility to the algorithm to be applicable to any icosahedral capsid system. Algorithmic improvements focused on extending the algorithm to simultaneously fit multiple data curves to reduce the potential for overfitting, and replacing computationally costly large-scale simulations with a much higher quantity of parallelized smaller simulations greatly improving run-time. The improved parameter estimation algorithm was applied to fitting static light scattering curves of hepatitis B virus (HBV) and cowpea chlorotic mottle virus (CCMV), as well as refitting the human papillomavirus (HPV) system. Following the data fitting, I conducted detailed data analysis into the pathways of each virus through design and generation of mass fraction plots, binding frequency tables and movies of capsid assembly pathways. The results of this analysis suggest great diversity in assembly mechanisms between the viruses and insight into not only the important binding pathways for each virus, but in some cases, critical intermediates that play a pivotal role during assembly. This chapter demonstrates

[1] This chapter is based upon work published in Xie et al. (2012). Surveying capsid assembly pathways through simulation-based data fitting. Biophys. J. 103:1545–1554.

the ability of discrete event simulation to fit *in vitro* data to infer unknown kinetic parameters and provide biologically relevant insight into capsid assembly mechanisms.

2.2 Methods

2.2.1 Measuring Quality of Fit

The major priority in the data fitting algorithm is to learn a set of rate parameters such that computational results in the form of a simulated light scattering data set provide a good fit to a set of true experimental light scattering curves. One of the outputs of the DESSA simulator is a time series detailing the numbers of oligomers of each size present as a function of time in the simulation (33). Following the work of Casini et al. (43), one can approximate the light scattering curve produced by any given parameter set p over time τ as follows:

$$R(\tau,p) = k \times c \times \left(\sum_{i=1}^{n} N_i(\tau,p) \times i^2 \right) \Big/ \left(\sum_{i=1}^{n} N_i(\tau,p) \times i \right) = k \times c \times S(\tau,p) \qquad (2.1)$$

Here $R(\tau,p)$ is the value of a simulated light scattering curve at time point τ with parameter set p, n is the number of subunits in a full viral capsid, c is the concentration of subunits, $N_i(\tau,p)$ is the number of assemblies consisting of i subunits at time τ for parameter set p, and k is a scaling factor. For sake of simplification, one can introduce the notation $S(\tau,p)$ for the average assembly size. For this work, the parameter set p consists of a set of on- and off-rates for all possible binding interactions described by the local rule set. $R(\tau,p)$ is estimated for each parameter set p by averages over a set of trajectories in order to minimize stochastic noise, as described below. In contrast to the lab's prior work (53), it is assumed that it is necessary to fit p and possibly k to a set of true light scattering curves $E_1, \ldots, E_m$ representing measurements of assembly at m distinct concentrations. Quality of fit of the corresponding simulated curves $R_1(\tau,p), \ldots, R_m(\tau,p)$

is measured for a given parameter set p in terms of a root mean square deviation (RMSD)

$$\Phi(p) = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{t_j}\sum_{\tau=1}^{t_j}\left(E_j(\tau) - k\cdot c_j \times S_j(\tau, p)\right)^2\right)} \qquad (2.2)$$

where the deviation is computed over a series of discrete time points 0, …, t_j, where t_j is the total time measured in curve j.

As mentioned above, fitting may also require the scaling factor k, which is handled in a unique matter for the three viruses depending on how the data is reported. In the case of HPV, the scaling factor is given as $k = 7.04 \times 10^{-8}$ (43) and that value for k is used consistently. In contrast, no scaling factor value is given in the HBV and CCMV experimental studies and thus k has to be treated as an additional unknown to be learned. In order to learn k, it is assumed that no assembly has occurred at the initial time point of the light scattering experiments following the addition of assembly subunits. This allows the use of the approximation $\Phi(0, p) = k \times c$. Then, one can seek a best-fit value k via:

$$\hat{k} = \arg\min_k \Phi(k \mid p) = \arg\min_k \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{t_j}\sum_{\tau=1}^{t_j}\left(E_j^*(\tau) + k\times c_j - k\times c_j \times S_j(\tau \mid p)\right)^2\right)} \qquad (2.3)$$

Here E^* is the modified experimental curve that is shifted along the Y-axis with $E^*(0) = 0$, so $E^* + k \times c$ starts at $k \times c$. By finding the zero of the derivative with respect to k, the maximum likelihood estimation can be derived by:

$$\hat{k} = \left|\sum_{j=1}^{m}\left(\frac{1}{t_j}\sum_{\tau=1}^{t_j}E_j^*(\tau)\times c_j \times \left(S_j(\tau \mid p) - 1\right)\right)\right| \Big/ \left|\sum_{j=1}^{m}\left(\frac{1}{t_j}\sum_{\tau=1}^{t_j}c_j^2\left(S_j(\tau \mid p) - 1\right)^2\right)\right| \qquad (2.4)$$

Next, the fitting strategy must be varied for the three viruses dependent upon how the data

is reported. In the case of HPV, there is a single known *k* therefore fitting of *k* is omitted. For HBV, it is assumed that there is a single unknown *k* that is shared across all curves thus this value is learned prior to multi-curve fitting. Finally for CCMV, the strategy must be modified in a couple ways. First, a 2.5 second artificial "lag phase" is added for each curve to account for uncertainty in the timing of the beginning of assembly. This is especially important with CCMV because of its fast assembly initiation compared to other viruses. Second, due to CCMV curve normalization, each curve must be fit with a separate *k* and then average the RMSD across curves:

$$\Phi(p) = \frac{1}{m}\sum_{j=1}^{m}\sqrt{\frac{1}{t_j}\sum_{\tau=1}^{t_j}\left(E_j(\tau) - k_j \cdot c_j \times S_j(\tau, p)\right)^2} \tag{2.5}$$

Here, *m* is again the number of curves to be fit simultaneously and k_j is the scaling factor for the j^{th} curve derived by applying Eq. 2.3 and 2.4 to a single curve at a time.

2.2.2 Data Fitting Algorithm

Data fitting was performed by an improved version of the optimization algorithm described in Kumar and Schwartz (53). This method is based on a search algorithm constructed to locally optimize parameter sets by the quality-of-fit measure Φ, which is in turn used as the basis for a heuristic global search. Next, the global search begins with an initial scan of a parameter space reduced by the assumption that all on-rates in the system are equal and similarly all off-rates are equal. This is followed by a series of local searches, each search using the local optimum of the previous search as an initial estimate in an expanded parameter space allowing two previously equal off-rates to vary independently. This process is then repeated until each off-rate is independently fit. While individual off-rates are fit, a single on-rate is used for all binding sites to

constrain the parameter space and maintain computational tractability. In the previous work with HPV (53), the reverse was done, fitting independent on-rates while using a single fit off-rate. The alternative was chosen in this work in order to model binding rate as diffusion-limited and thus essentially equivalent between binding sites on atomically equivalent though potentially structurally-different subunits. That said, as is observed below, the new method does infer comparable HPV free energies and pathways to the prior study.

The order in which independent rates are introduced is determined manually for each virus. For HPV, after learning a single on- and off-rate for all binding sites, the off-rates of the four sites are broken into two groups (A | BCD), then three groups (A | BC | D), and finally four (A | B | C | D). For HBV the steps by which groups are subdivided are AD | BC, A | D | BC, and A | D | B | C. The scheme for CCMV is AB | C, and A | B | C.

Local optimization is conducted by estimating the quality of fit Φ at a series of nearby grid points in parameter space. The points are then used to determine a gradient vector which is used to predict a local optimum by steepest descent. The method also fits a quadratic response surface with which a local optimum of that surface can also be predicted. The method then interpolates between these two approximations via a weighting factor that is adjusted to favor the response surface prediction when the method improved the fit on the previous step or favor the steepest descent prediction when the method failed to produce an improved fit. This approach applies similar reasoning as the Levenberg-Marquardt optimization method (89,90) which interpolates between two data fitting techniques adjusting between algorithms more suitable to smooth or rough regions in the parameter space. At each step, estimates are computed for the current grid size by using the previous best guess, variations of plus or minus one grid step in each individual parameter, and variations of plus or minus the grid size in each pair of parameters.

Despite the algorithmic advantages of utilizing stochastic methods for large assembly system simulations (50), the noise introduced can be problematic for data-fitting. There are two ways to counteract this noise: either run larger individual simulations or run larger numbers of simulations. Previous work had relied on the first approach; however, the latter allows for better use of parallel computing resources and overall a reduction in computational runtime. Therefore, in this study, a relatively small number of subunits are used per simulation (720, 600 and 450 for HPV, HBV and CCMV, respectively) with numerous trajectories averaged for each estimation of Φ. This required 120 simulation trajectories per sampled point and 750 to evaluate a new local minimum. All the replica simulations needed at each step are automatically submitted and distributed to a compute cluster with 20 Xeon E5520 quad-core processors.

2.2.3 Data

Data for the three viral systems was gathered from prior studies by (43) (HPV), (36) (HBV), and (47) (CCMV). In each case, data consists of 90 degree light scattering measurements of time progress of *in vitro* capsid assembly systems. For HPV, light scattering data was gathered from purified L1 coat protein in citrate buffer with 0.5 M NaCl at pH 5.20 for 250 minutes per experiment (43). For HBV, data was gathered from stock solutions of Cp149 coat dimers in 0.1 M sodium bicarbonate and 5 mM DTT at pH 9.5 for 600 seconds per experiment (36). For CCMV, experiments were performed on solutions of coat dimers in 200 mM sodium citrate and 1 M NaCl, with data reported for 300 seconds per experiment (47). For HPV, data fitting was applied to three curves corresponding to capsomer concentrations of 0.53 μM, 0.72 μM, and 0.80 μM. For HBV, fitting was applied to coat dimer concentrations of 5.4 μM, 8.2 μM and 10.8 μM. For CCMV, fitting was applied to coat dimer concentrations of 14.1, 15.6 and 18.75 μM. In the case of HPV, data was provided directly by David Wu in electronic format, while for HBV and

CCMV, data was derived from the appropriate figures in the references (Fig. 4 *C* in (36) and Fig. 1 *B* in (47), respectively).

2.2.4 Data Analysis Techniques

A variety of analysis and visualization methods were applied to study simulation trajectories individually and in aggregate. First, I derived summaries of assembly pathway in the form of tables of frequencies of possible binding interactions averaged across all trajectories and time-series of mass fractions of different sizes of species versus time for single trajectories. To construct binding frequency tables, I counted all association events that occur across all repetitions of each simulation input and produced a matrix with one row or column for each subunit contained in a completed capsid. I placed in position (*i*,*j*) the count of all binding events involving an assembly of *j* subunits producing an assembly of *i* subunits. I then scale each row by the total number of association events involving the production of an assembly of *i* subunits, so that each position in the matrix contains a frequency between 0 and 1.

To construct mass fraction plots, I recorded after each simulation event the quantity of each assembly size (from monomer to complete capsid) currently present in the simulator. I then scaled each of these assembly size counts by the number of subunits present in the assembly (e.g., a pentamer is scaled by five as the "mass" of a pentamer in each simulation would be five times that of a monomer). I finally normalized each value such that it represents the frequency, between 0 and 1, instead of the total count of subunits in each assembly size. I plotted this value versus simulation time for each potential assembly size. I further generated movie files of specific assembly pathways from individual capsid protein to completed structure. This was accomplished by modifying the simulator to log all specific details about all bond forming and

breaking events including the ids of the assemblies involved in the events. A graph is then constructed of all interactions between individual assembly IDs treating each assembly as a node with forward and reverse connections to other assembly nodes. From this, a pathway is constructed in reverse starting with the final completed capsid structure and backtracking until a single subunit is reached. To ensure simplicity and the best likelihood of retrieving useful pathway information, the largest possible intermediate is selected among all choices in each backtracking step. The stochasticity of DESSA is dependent upon an initial random seed. If that random seed is logged and reused, the exact same simulation will occur. Thus, following construction of the pathway of assembly ids involved in formation of the completed capsid, the same simulation is run again with the purpose of logging, for each assembly ID present in the pathway, the three-dimensional spatial location of each subunit in the assembly. When plotted, this view of an assembly intermediate along the pathway to a completed capsid serves as a frame of the resulting movie. When compiled together, the completed movie shows frame by frame all steps in a single pathway.

2.3 Results

2.3.1 Fitting Simulation Output to Experimental Data

For each virus system, parameter estimation begins with an initial scan of a wide range of potential rate parameters within a simplified parameter space, in which all on-rates are assumed to be equal and all off-rates are assumed to be equal. Following this was the pseudo-global optimization for each capsid system starting from the point in parameter space of the lowest RMSD from each initial two-parameter scan followed by successive local searches performed over larger spaces of reverse rates as described in Section 2.2.2.

Table 2.1: Best-fit rate parameters and corresponding free energies for the three virus models. Rows correspond to the labeled binding sites of the rule sets of Fig. 1.

Virus	Site type	On-rate ($M^{-1}s^{-1}$)	Off-rate (s^{-1})	Free energy (kCal/mol)
HPV	A	1.4×10^3	0.12	-5.6
	B	1.4×10^3	0.11	-5.6
	C	1.4×10^3	0.12	-5.6
	D	1.4×10^3	0.13	-5.5
HBV	A	1.4×10^6	1.2×10^5	-1.5
	B	1.4×10^6	1.4×10^5	-1.4
	C	1.4×10^6	1.4×10^5	-1.4
	D	1.4×10^6	1.2×10^5	-1.5
CCMV	A	1.2×10^6	3.0×10^4	-2.2
	B	1.2×10^6	3.0×10^4	-2.2
	C	1.2×10^6	3.9×10^4	-2.0

Table 2.1 shows the final fit parameters for HPV, HBV and CCMV. Despite changes in the model relative to prior work in the lab (53) by assuming a shared on-rate rather than a shared off-rate, the search ended with comparable parameter estimations for HPV. As in that prior work, similar free energies between binding sites were learned, with each now inferred to have a free energy of -5.5 to -5.6 kcal/mol. This is somewhat stronger than prior free energy estimates of capsid protein binding in other viral systems, which have varied between roughly -2.8 kcal/mol for HBV (34) to -4.4 kcal/mol for phage P22 (35). Nevertheless, free energy estimates of roughly

-5.5 kcal/mol are still physically plausible. Figure 2.1(a) shows the quality of fit of the best-fit model to the true light scattering data from Casini et al. (43). Overall, the parameter estimation algorithm produced a good fit to both short and long timescale features of the curves, although there is some underestimation of the 0.72 μM curve and some overestimation of the 0.80 μM curve. There is also some difference in the initial slope following the lag phase with the slope of the true data tending to be sharper, though this might be an artifact of the sensitivity of the light scattering experiments with regards to small oligomers.

The final parameter fits for HBV detail a substantially weaker interaction than those found for HPV, in the range of -1.4 to -1.5 kcal/mol. On average, there are two binding interactions per dimer in a complete capsid so the average free energy per dimer in a complete capsid would be about -2.9 kcal/mol which is well in line with previous estimates by Ceres and Zlotnick (34). One important note is the rate of assembly is much higher with respect to HBV assembly than in HPV assembly giving credence to the argument of Zlotnick et al. that weak capsomer interactions are necessary for efficient assembly. The relatively weak interactions suggest an assembly process with far more trial-and-error where assembly pathways involve repeatedly attaching and detaching additional subunits or potentially small oligomers until stable substructures are formed and can act as ratchets forcing the assembly process forward. Figure 2.1(b) shows the best-fit curves to the *in vitro* HBV light scattering data. Again, the model produces a generally good fit to all phases of assembly despite some underestimation of the 5.4 and 10.8 μM curves and overestimation of the 8.2 μM curve. The final fits for CCMV displayed a free energy range between the strong extremes of HPV and HBV, from -2.0 to -2.2 kcal/mol. The free energies and kinetic rates are closer to those of HBV which suggests the possibility of more similarity in assembly mechanism. Certainly, CCMV too will likely be more dependent

upon a trial-and-error approach to assembly. Figure 2.1(c) shows the best-fit curves to the *in vitro* CCMV light scattering data. The model once again produces a good fit to both early and late stages of the data with some minor overestimation in the 14.1 and 15.6 μM cases with very minor underestimation in the 18.75 μM case.

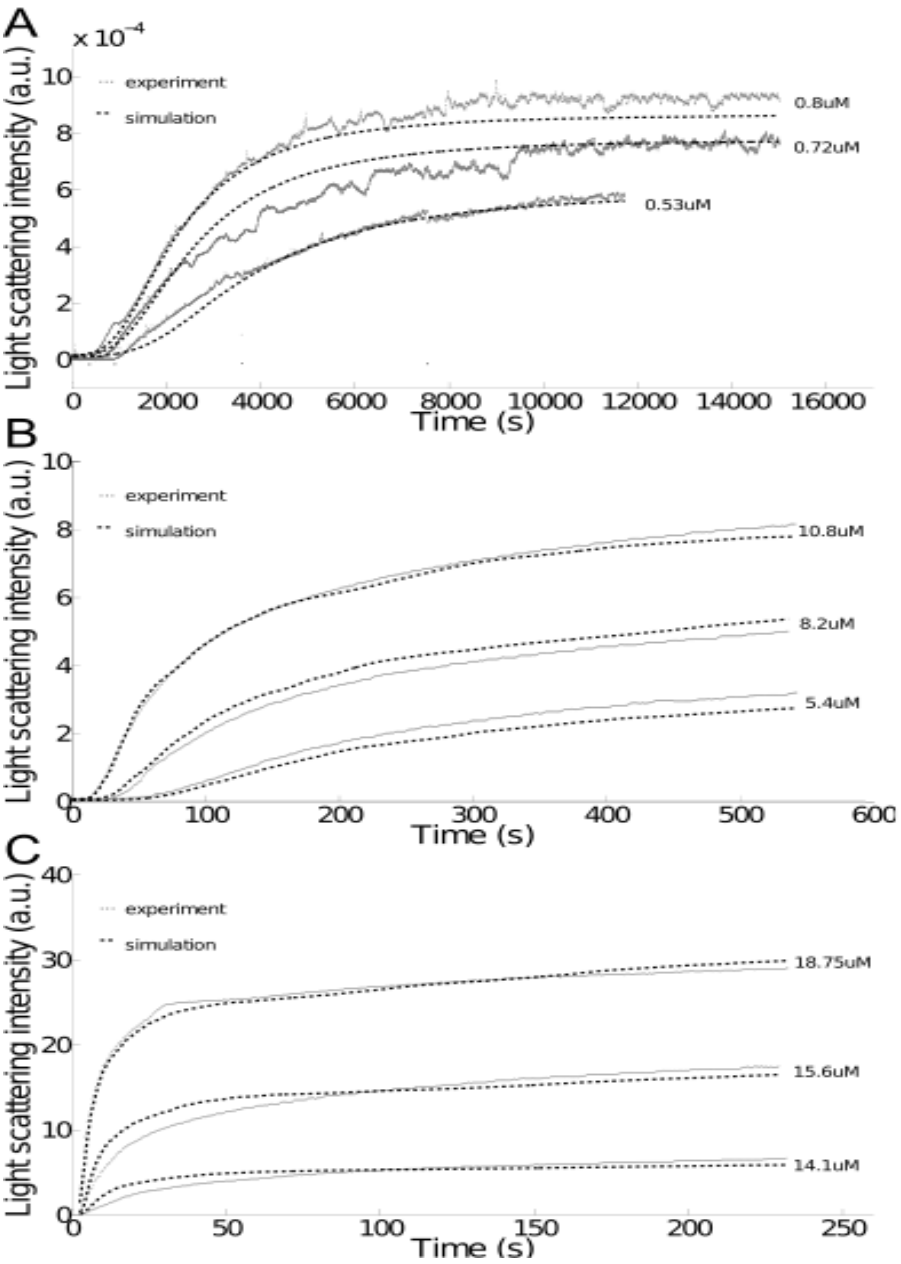

FIGURE 2.1: Best fits of experimental and model light scattering curves for the three capsid systems. Each figure shows three experimental curves measured at different concentrations (solid line) and the corresponding best-fit model curves (dashed line). (*A*) HPV. (*B*) HBV. (*C*) CCMV.

2.3.2 Analyzing Assembly Pathways

One of the major goals of this study is to learn details of the assembly pathways utilized by a diverse array of icosahedral viruses to determine if simulation experiments could display with consistency more complicated viral pathways than simple monomer-based addition. To this end, I constructed mass fraction plots examining distributions of intermediates as functions of time. Figure 2.2(a) shows one such mass fraction plot from a single sample trajectory for HPV at 0.80 μM. The figure shows a very complex array of intermediates assembling with seemingly no preference for building up any specific pools of oligomers. Furthermore, there is no pronounced depletion of larger oligomers, as one would expect for nucleation-limited assembly. Instead, the mass fraction plot shows a gradual increase in the size of the intermediates present in the simulations over time as the original pool of monomers decreases at a relatively linear pace. This pattern is consistent with a non-nucleation-limited assembly model dominated by successive addition of individual monomers. Further evidence in support of this model can be found in the binding frequency tables in Figures 2.3,(a)-(c), which offer another perspective on assembly pathways utilized in the simulator by plotting the frequency with which any given reactant oligomer sizes are used to assemble any given product oligomer size. These figures confirm that at each concentration measured, bond association events are largely driven by single capsomer addition, with rare exceptions occurring primarily in assembly reactions only involving small oligomers. These rare exceptions involving non-monomer-based assembly reactions do occur in slightly increased frequency under the higher capsid protein concentrations studied, although these occurrences are still quite rare.

I conducted further analysis of HPV assembly by developing a movie of a single pathway from monomer to completed capsid. This movie treats a 3D representation of each assembly

intermediate along the pathway as a single frame in the full movie, although for the purposes of this document, I can only show specific important frames that represent the overall nature of the pathway visualized. Figure 2.4 shows four such frames for a single representative pathway of HPV assembly at a capsomer concentration of 0.72 μM. The movie as a whole is very short, with only 119 frames representing 119 events between the first assembly reaction of two monomers and the formation of the final capsid. This suggests an assembly process not very reliant on trial and error and indeed, as is seen in the mass fraction plot, intermediates that are formed tend to stay assembled for long periods in the simulation without breaking down or building to form new larger intermediates. This specific assembly pathway shown takes roughly 5000 seconds of simulation time which is a much longer time scale than the other two viruses studied. Further, there is little overall consistency in the geometry of how the building capsid grows. Individual subunits are added to the existing structure building extra triangles one at a time, as seen in Figures 2.4,(a) and (b), at early stages of assembly, although their overall placement is fairly random. This randomness can lead to numerous situations in which intermediates produce holes that are very difficult to fill with a dwindling pool of monomers and an increased number of similarly stuck intermediate structures; a few such holes can be seen in Figure 2.4(c). In the case of this assembly pathway, enough monomers were present still in the simulation to afford continued gradual assembly and eventually a completed HPV capsid is formed in Figure 2.4(d). In nucleation-limited growth models, the elongation phase of capsid assembly tends to be exceedingly rapid and the percentage of overall assembly time devoted to elongation following assembly of stable intermediate structures is very small. That is the exact opposite of what is seen here, where 79.6 percent of the assembly time is between the formation of an 18mer and the final completed capsid.

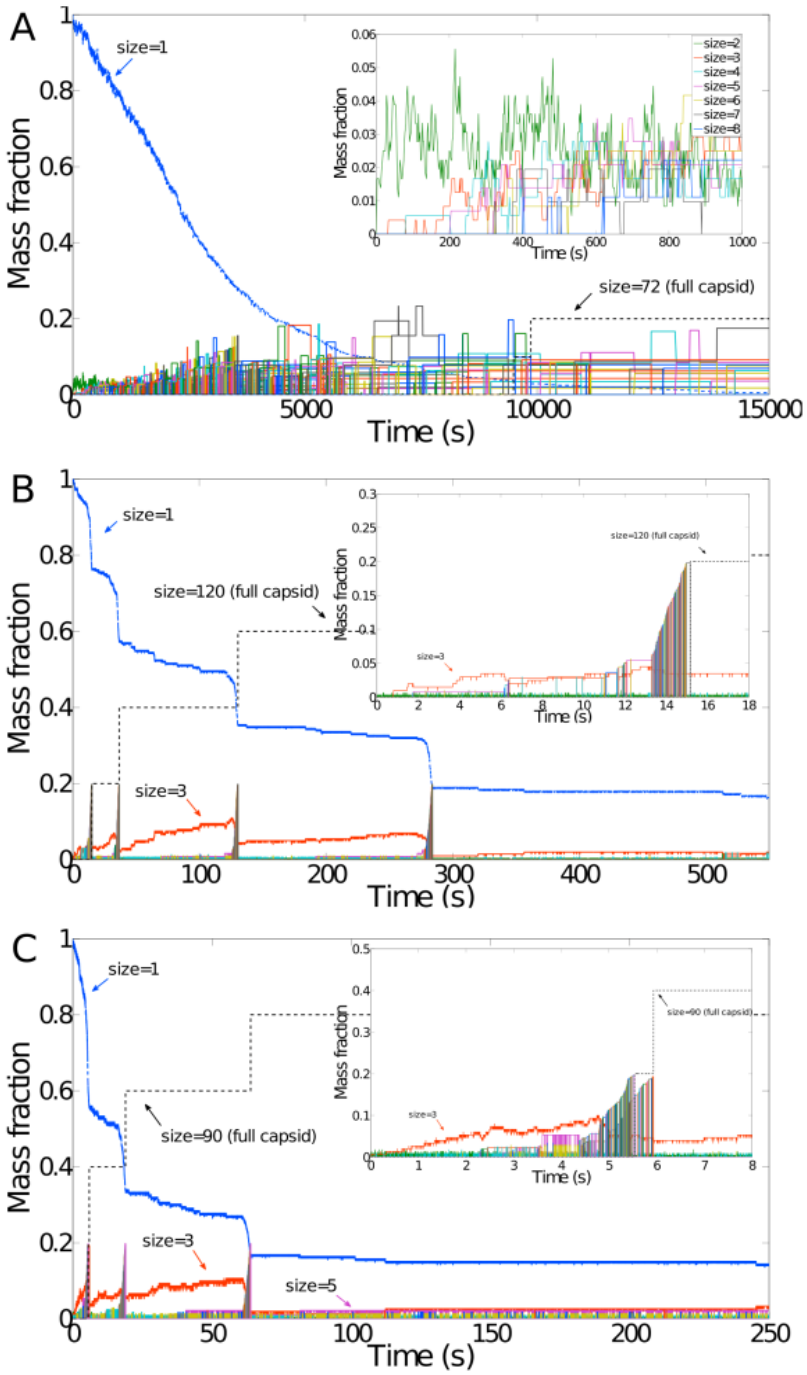

FIGURE 2.2: Mass fractions of intermediates versus time for sample trajectories of the three capsid systems. Each curve corresponds to a single size of intermediate species. Insets in each plot magnify early stages of the reaction. While all possible intermediate sizes are plotted, for simplicity a key is provided only for sizes 2-8. (a) Mass fractions versus time for 720 HPV capsomer subunits at 0.80 μM. (b) Mass fractions versus time for 600 HBV dimer subunits at 10.8 μM. (c) Mass fractions versus time for 450 CCMV dimer subunits at 14.1 μM.

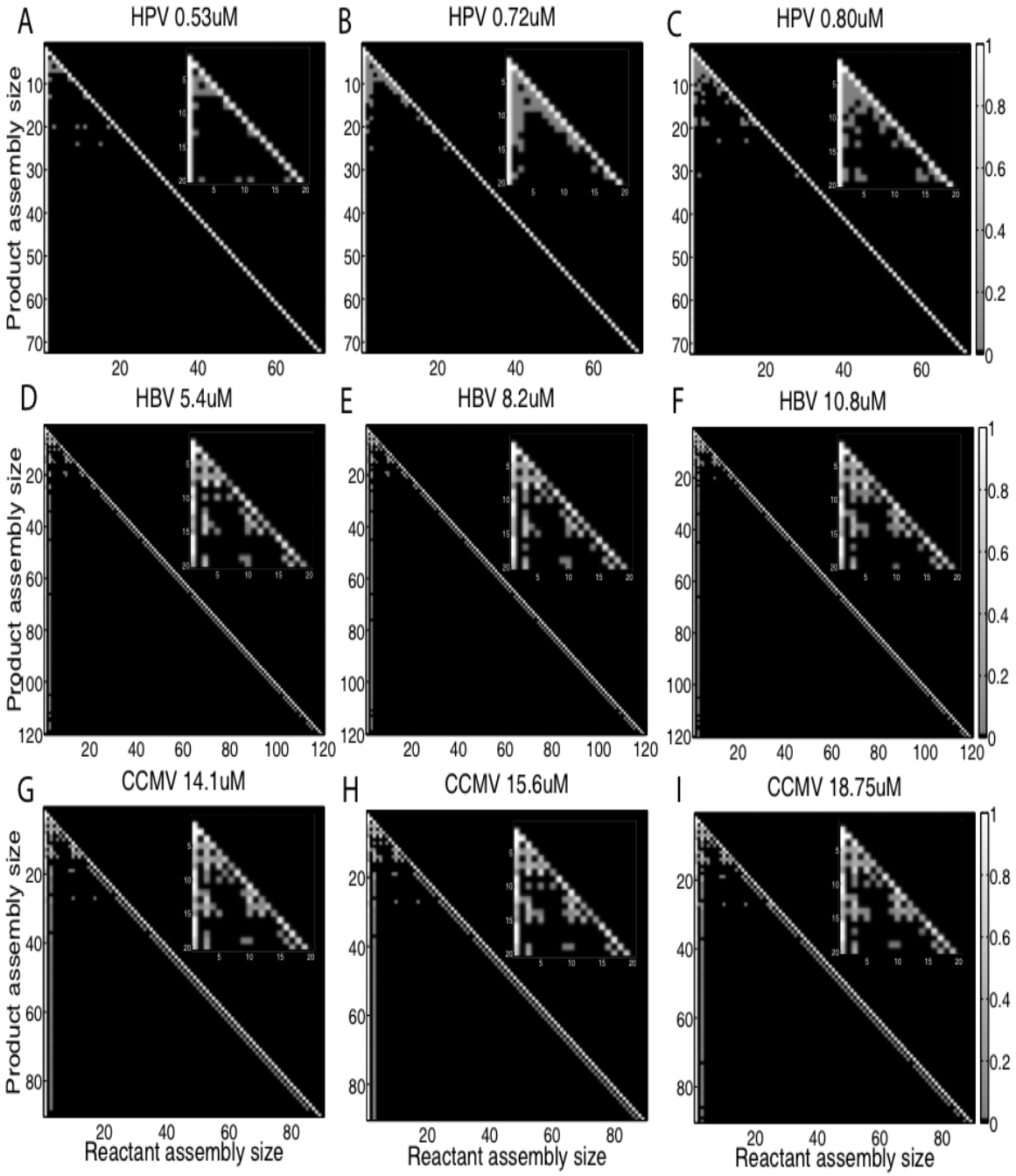

FIGURE 2.3: Visualization of reaction usage for viral assembly reactions. Each plot is organized into a set of rows and columns such that the shading of the box in row *i* and column *j* denotes the fraction of oligomers of size *i* produced by a binding reaction involving an oligomer of size *j*. Lighter colors denote higher frequencies and darker colors lower frequencies, with black representing a reaction unobserved in the course of the simulations. Each subfigure shows the full plot of all possible assembly sizes, with an inset in the upper-right magnifying a portion of the upper-left region of the full plot corresponding to reactions producing smaller oligomers. (a) HPV at 0.53 μM. (b) HPV at 0.72 μM. (c) HPV at 0.80 μM. (d) HBV at 5.4 μM. (e) HBV at 8.2 μM. (f) HBV at 10.8 μM. (g) CCMV at 14.1 μM. (h) CCMV at 15.6 μM. (i) CCMV at 18.75 μM.

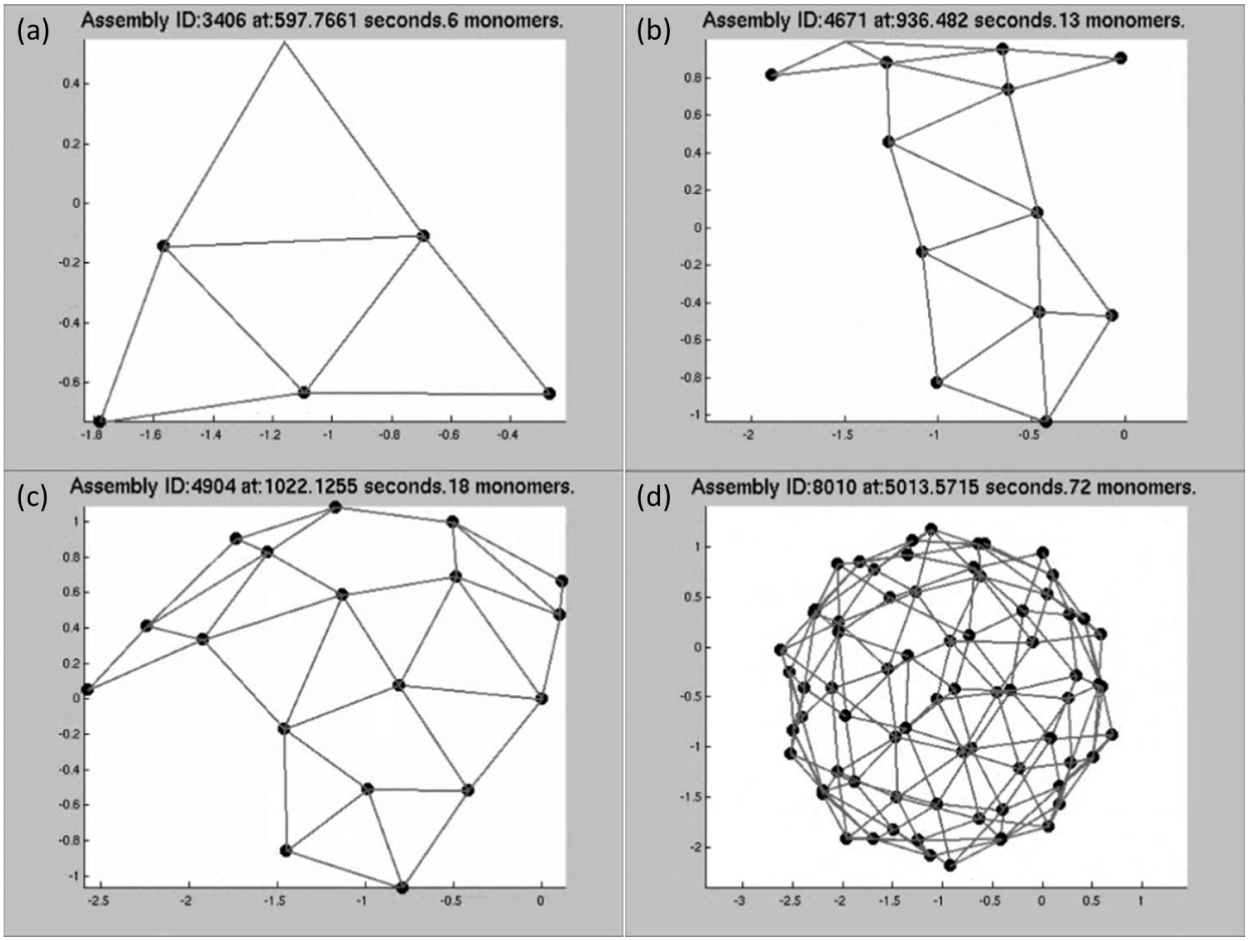

Figure 2.4. Individual frames of a movie following an assembly trajectory of a HPV capsid at 0.72 μM. Assembly information for each frame is listed in the header for each sub-image.

Detailed analysis of the pathways utilized during HBV assembly simulation provides a very different picture compared to HPV. Figure 2.2(b) shows a mass fraction plot detailing a sample HBV assembly simulation trajectory for the best-fit parameters at a capsid protein concentration of 10.8 μM. Unlike the gradual loss of monomers and gradual gain of increasingly larger intermediates seen in the HPV plot, the most prominent feature of the HBV plot is a set of "spikes" in which the system rapidly cycles through a series of increasingly larger intermediate assemblies resulting in the formation of a new capsid. These spikes are the signature of nucleation-limited growth where each spike is sparked by the formation of some small intermediate defined as the nucleus. After the nucleus is formed, the capsid forms around this

intermediate very rapidly in a short elongation phase. Another major difference between the HPV and HBV plots is the presence of a pool of free trimers (hexamers of coat protein) which forms quickly and persists through the simulation. Monomers and trimers of subunits (dimers and hexamers of coat proteins) seem to reach equilibrium with one another early in the simulation and readjust to a new equilibrium following each nucleation peak and resultant production of another completed capsid. The drops in mass fraction for the pools of monomers and trimers during each nucleation spike suggest that monomers and trimers are both utilized as building blocks during assembly, contrary to the almost exclusivity of monomer-based addition in HPV assembly. This is further confirmed by the binding frequency tables in Figures 2.3,(d)-(f). At each concentration of HBV capsomers examined, both monomers and trimers are used seemingly at random for individual steps of the assembly. Overall monomers are favored but trimers are used as assembly reactants 10-20% of the time at most elongation steps. Because of the increased trial-and-error nature of HBV assembly, especially with regards to monomer-based assembly where the attachment and detachment of free individual subunits to assemblies is a very large percent of all assembly reactions, this 10-20% figure for the likelihood of trimer-based addition is far underestimating the role of trimers in effective assembly growth. Regardless, it is clear that this system cannot be described by a single pathway but instead by an ensemble of distinct assembly pathways. The production of small oligomers examined in the insets to Figures 2.3,(d)-(f), show even more complexity with more frequent usage of trimers and even pentamers during assembly. In fact the most common method of assembling an octamer in HBV simulations is via the binding of a trimer and pentamer. Changes in capsomer concentration did not seem to have any effect on binding frequencies.

Despite all of this insight, it is still difficult to determine a precise nucleation step from

mass fraction plots and binding frequency tables alone. To further assist in this process, I examined HBV assembly pathways in great detail by generating movies describing every bond association and dissociation event between individual subunit and completed capsid. Figure 2.5 shows important frames from one such movie of a single trajectory of HBV capsid assembly at a capsomer concentration of 8.2 μM.

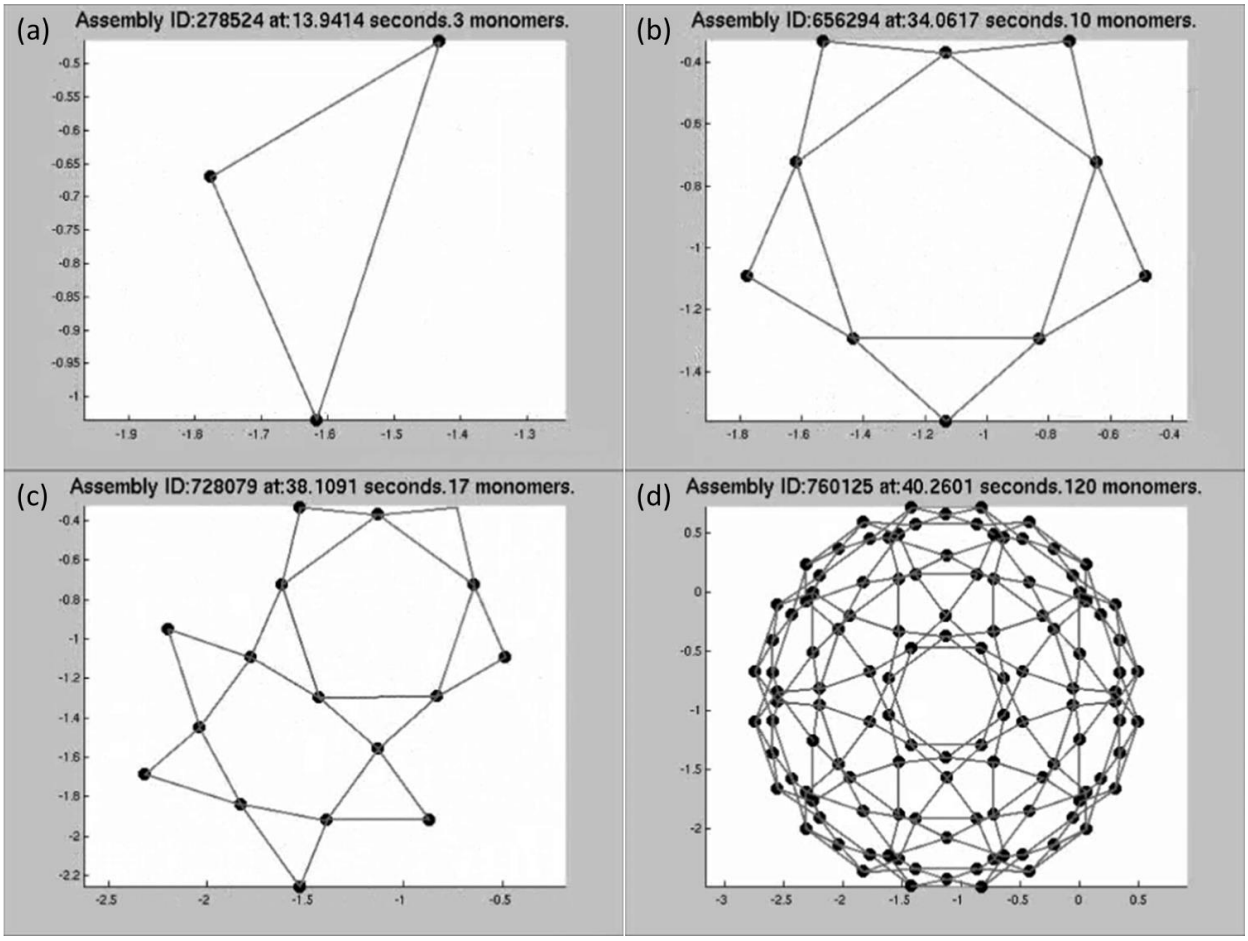

Figure 2.5. Individual frames of a movie following an assembly trajectory of a HBV capsid at 8.2 μM. Assembly information for each frame is listed in the header for each sub-image.

There are numerous differences in the pathway seen here for HBV compared to the HPV pathway shown in Figure 2.4. First, the assembly time from individual subunit to completed capsid is only just over 40 seconds compared to roughly 5000 seconds for the HPV capsid. Despite the more rapid assembly time, the HPV movie only consisted of 119 frames representing

119 specific bond association and dissociation events prior to a completed capsid forming. In contrast, there are over 6200 frames in the HBV movie; that is, 6200 bond association and dissociation reactions are necessary to assemble a capsid consisting of only 120 subunits. This dramatic difference highlights the trial-and-error nature of HBV capsid assembly where weaker interactions between subunits allow for consistent rearrangement of capsid subunits as necessary to produce resulting completed structures. This rearrangement helps limit issues with kinetic trapping that are so pervasive during HPV assembly. Another interesting note with regards to HBV assembly comes from the building blocks present. In Figure 2.5(a), formation of a small trimer is seen as an important early step in assembly and a larger 10mer structure is constructed from a ring of trimers as seen in Figure 2.5(b). However, this stable pentagon of triangles does not by itself instigate rapid assembly of the remainder of the capsid. It is instead when a further hexagon structure is added, as seen in Figure 2.5(c), that the remainder of capsid assembly occurs rapidly. In fact, the elongation phase between the 17mer, which includes the hexagon structure, and the completed capsid only accounts for 5.3 percent of overall assembly time, another tell-tale sign of nucleation-driven assembly. Recall that the time to assemble the 72 subunit completed HPV capsid from a similarly-sized 18mer required almost 80 percent of the assembly time.

Examining the assembly process for CCMV in greater detail shows a model that is qualitatively far more similar to that of HBV than HPV. Figure 2.2(c) shows mass fractions of CCMV intermediates at a capsomer concentration of 14.1 μM. The most striking feature of this mass fraction plot is again the spikes indicative of nucleation-limited growth. There is also a standing pool of trimers present during the CCMV assembly as with HBV. What is different from HBV, however, is that there is also a small standing pool of pentamers later in the reaction.

Both HBV and CCMV display a similar process of gradual accumulation of trimers in the lag phase between nucleations until an appropriate equilibrium is met between monomers and trimers, at which point the next nucleation peak occurs and another capsid is subsequently assembled while the pool of trimers is rapidly depleted. Binding frequency tables further corroborate the similarities between CCMV and HBV assembly pathways. Figures 2.3, (g)-(i), show binding frequency tables for CCMV assembly at each capsomer concentration studied and, as with HBV, monomer and trimer addition provide an ensemble of potential assembly pathways that are utilized at each point during the assembly process. Assembly of smaller oligomers again show greater assembly variety with trimer and pentamer addition at a higher than normal frequency. I do note that there also appears to be a single frequently used step involving large (10mer and 17mer) oligomers that is seen less frequently with HBV. Pathway usage does appear to be insensitive to changes in capsomer concentration as there are no major changes in binding frequency seen in Figures 2.3, (g)-(i).

As with HPV and HBV, I developed movies of individual pathways for CCMV capsid assembly. Figure 2.6 displays frames from one such representative movie of CCMV capsid assembly at a capsomer concentration of 15.6 μM. This specific assembly pathway again shares far more in common with HBV than with HPV. CCMV assembly time is the fastest yet at only 5.671 seconds until the capsid was completed. Despite this further decrease in assembly time, 2400 association and dissociation events were required to form the 90 subunit completed structure again showing the trial-and-error nature of CCMV assembly. The assembly of a trimer and a dimer at an early point to form two triangles, as seen in Figure 2.6(a), serves as an initial building block for a 12mer including a stable pentagon structure shown in Figure 2.6(b). This pentagon is again not enough to drive the nucleation spike, however. Instead, it is the formation

of an intermediate containing a hexagon, as in Figure 2.6(c), that begins the rapid assembly of the remainder of the capsid. As with HBV, the elongation phase is only a small fraction of the overall assembly time, here just 24.8 percent of the simulation time. I mentioned previously the role of a 10mer and a 17mer as larger intermediates involved in assembly reactions. The likely 10mer utilized in these reactions is the stable pentagon wrapped in triangles seen in Figure 2.6(b). Similarly, the 17mer is likely the nucleus seen in Figure 2.6(c) consisting of a hexagon bound to a stable pentagon with both wrapped in triangles. This complex structure shows up time and time again in simulations for both HBV and CCMV and lends credence to the belief that the formation of this hexagon is the essential nucleation event that drives CCMV and HBV capsid assembly.

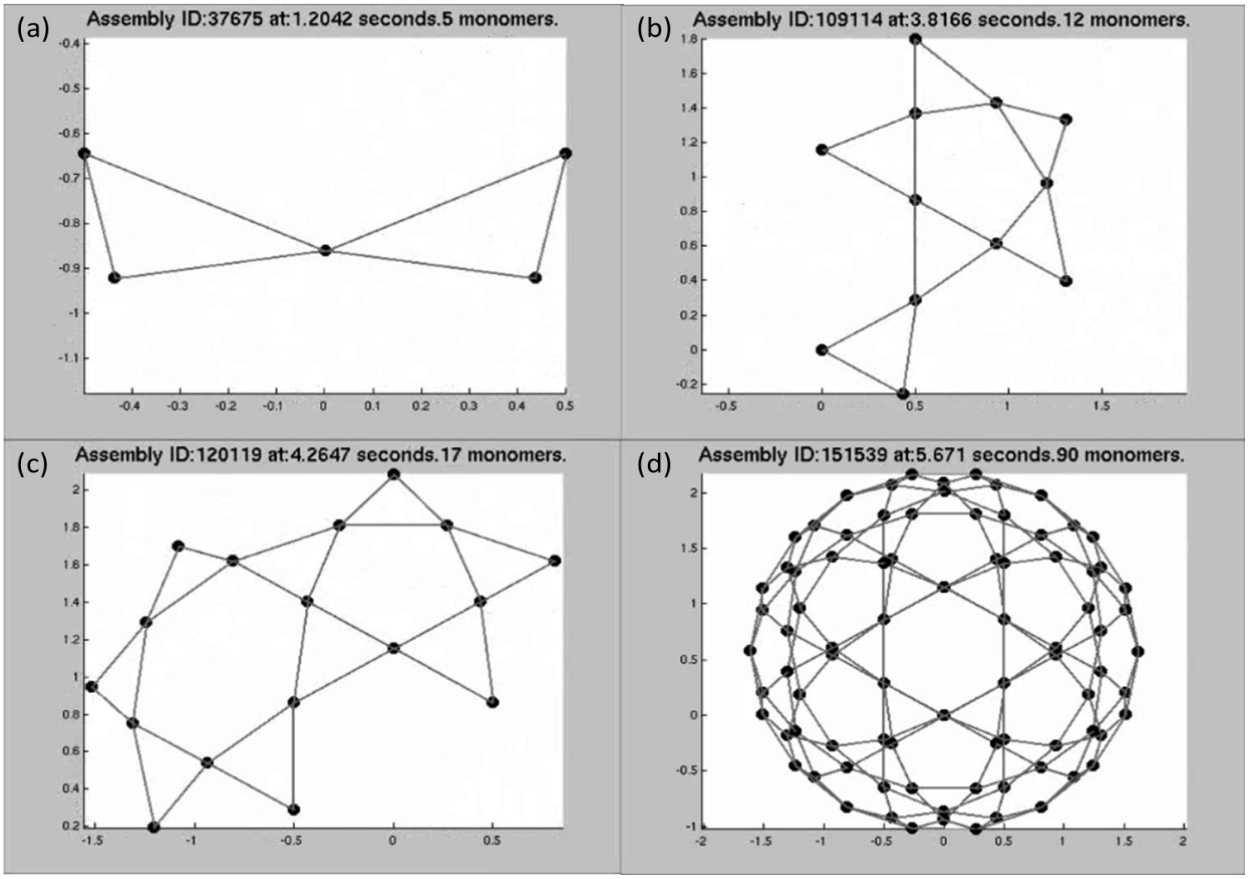

Figure 2.6. Individual frames of a movie following an assembly trajectory of a CCMV capsid at 15.6 μM. Assembly information for each frame is listed in the header for each sub-image.

2.4 Discussion

The work described in this chapter provides an important step forward both in the capabilities of parameter estimation to learn biologically plausible kinetic rate parameters of virus assembly as well as an unprecedented ability to analyze the pathways by which virus capsids assemble. The parameter estimation technique utilized here produced good fits between experimental and simulated light scattering curves for each virus studied representing a wide range in icosahedral structure and assembly rate. This technique even provided good fits when fitting a common parameter set to multiple experimental curves at different capsomer concentrations, a necessary further complexity to reduce problems with data overfitting and redundancy of solutions. Furthermore, these results highlight the ability of simulation models to display diversity in assembly between structurally similar viruses. The inferred models show a surprising variability in assembly behaviors, including nucleation-limited or non-nucleation-limited assembly, monomer-based or hierarchical oligomer-based assembly, and assembly consistent with either a single well-defined pathway or assembly that is driven by a large ensemble of potential pathways used singly or in combination. HBV and CCMV share many similarities in assembly pathways despite differences in icosahedral geometry of the capsid and sizable differences in learned best-fit rate parameters. HPV on the other hand shows very different behavior in both rates and pathways. This study, then, suggests that virus capsid assembly as a whole cannot be easily inferred from studying any single virus system. Instead, diversity in virus structure and assembly rate leads to great diversity in assembly strategies demonstrating the necessity of detailed specific examination of any virus system of interest.

The assembly pathway results described here raise several important questions about each virus studied. First, the lack of nucleation-limited assembly for HPV has been a surprising result

of both the work in this chapter as well as earlier work in the lab on this system (53). Theoretical models have long championed the role of nucleation-limited growth in preventing kinetic traps from blocking assembly completion (25,27-29,51) and simulations of HPV show numerous examples of kinetically trapped intermediates. One of the starkest contrasts between HPV assembly and the assembly of HBV and CCMV is the far fewer association and dissociation reactions that occur along a pathway to assemble a completed capsid. So little trial-and-error in HPV assembly pathways means that once an intermediate is formed, it is very unlikely for it to break apart. Once the pool of monomers is depleted by the formation of enough kinetically trapped intermediates, it becomes impossible for further completed structures to assemble. Indeed, the yield of completed HPV capsids in assembly simulations is very low compared to the other viruses studied; a fact that is in keeping with the *in vitro* HPV data used to fit rate parameters. This lack of nucleation-limited growth for HPV may reflect an inability of the model to learn the correct methods by which HPV assembles *in vivo*, perhaps because the *in vitro* experiments and thus the simulations lack some important feature that is essential to true HPV assembly. One such explanation for this distinction is that the *in vitro* HPV assembly experiments did not include a chaperone protein that plays a crucial role in the assembly process *in vivo* (91). Perhaps, if the functionality of this protein is included *in vitro* or *in silico*, overall assembly rate for HPV would be much faster and nucleation-limited growth would be observed, although this is pure speculation.

Following the prior work on HPV, there was still the possibility that the parameter estimation technique was unable to find a set of rate parameters that would produce nucleation-limited assembly. This has definitively been refuted with the results from HBV and CCMV, which both show nucleation-limited capsid assembly and a wide array of potential assembly

pathways. This second fact is especially intriguing highlighting the seemingly random sampling from a variety of potential assembly reactions that occur at each step in the simulation involving incorporation of either individual monomers or small oligomers or even on occasion much larger intermediates. This seemingly stochastic sampling among possible pathways is in contrast to standard models for biochemical systems which usually include a defined pathway that must be followed. Results suggesting that the assumption of a defined pathway for self-assembly is incorrect can have important implications for the use of simplified theoretical models to describe and reason about such systems. Furthermore, diversity in assembly pathways suggests an absence of any specific individual reactions necessary for capsid assembly, which can have practical implications for targeting capsid assembly via antiviral drugs (92-94). Nevertheless, while there may not be any specific reactions necessary for assembly of the viruses studied, in the cases of HBV and CCMV, examining what is necessary for nucleation to occur might still produce promising drug-targeting leads. The prevailing conclusion by examining numerous individual pathways that result in completed HBV and CCMV capsids is that the formation of a hexagon of subunits is essential for the rapid elongation and completion of assembly. Now this does not suggest that the formation of a single hexamer is the nucleation step. In fact, this hexagon structure cannot be characterized by any one size of intermediate. Instead, the formation of a variety of potential intermediate sizes can be the nucleation step provided a hexagon is included in this intermediate structure. Future structural studies to analyze the chemical nature of this hexagon will potentially lead to very interesting results about why this structure is so pivotal to HBV and CCMV capsid assembly. A final observation is the robustness of pathway usage across a range of capsomer concentrations. This suggests that the best-fit rate parameters for all three viruses studied are at positions in the parameter space that are relatively insensitive to

perturbation. This is in contrast to hypotheses from previous theoretical studies (33) that suggested that pathway usage could be quite sensitive to even small changes in rate parameters. This robustness could potentially be a general feature of viral assembly systems where evolution selected over time for the production of structures that are impervious to certain modifications of assembly environment, something that could be advantageous in living cells which can produce a number of modifications to assembly environment. This is mere speculation and far more examples of pathway robustness in other virus systems would be necessary for any well-founded conclusions on this matter.

There are still open questions about the quality of the fits from the parameter estimation technique employed here as well as the accuracy of the inferred assembly pathways resulting from these fits. While overall, the results show great capability fitting simulated results to multiple *in vitro* light scattering curves, there still are some imperfections; for example, in the initial slope of the HPV curves following the lag phase. This will continue to be an issue as there is no direct experimental way to measure the true kinetic rates or to directly observe assembly pathways. Previous *in vitro* studies have noted the presence of specific intermediates during the assembly process, but without the context of the position of that specific intermediate in a larger assembly pathway, it is exceedingly difficult to use such data to draw any meaningful conclusions about capsid assembly. Therefore, it cannot be proven whether any deviations in the data fitting are only minor inaccuracies or if they produce fundamentally incorrect models of assembly. Also, by the nature of the heuristic algorithm employed to tackle the expansive parameter space, it is quite possible that there are locations in the parameter space that could produce an even better fit to the *in vitro* data suggesting the method was trapped in a local optimum instead of finding the global best fit. Even assuming the inferred models are correct,

they are nonetheless correct models of *in vitro* assembly systems and not of the viruses *in vivo*. Transitioning from models trained to analyze pathways of *in vitro* assembly systems to models that more accurately reflect *in vivo* conditions will be the main challenge tackled in the remainder of this thesis. Specifically, I will examine two such major aspects of the *in vivo* environment that are distinct from the *in vitro* experiments and known to affect assembly progression: non-specific macromolecular crowding (95), and viral nucleic acid (72).

Chapter 3: Modeling the Effect of RNA on Capsid Assembly Pathways[2]

3.1 Introduction

The next major contribution of this thesis consists of considering the first of two major changes to the virus capsid assembly environment to better reflect *in vivo* conditions: nucleic acid. In this chapter, I utilize a combination of simulation and analytical modeling to attempt to project the effects of RNA on CCMV assembly. I seek to use a model of purified coat protein assembly learned from *in vitro* light scattering data, as discussed in Chapter 2, and then build in fast corrections to that model to account for likely influences lacking from the experimental data. This approach provides a way to take advantage of the high efficiency of the stochastic simulation methods, and the data-fitting technology they enable, to gather statistics on detailed pathway usage of a data-fit model of a real virus.

CCMV has four RNA strands that can potentially be encapsidated: RNA1 and RNA2, which are each encapsidated individually, and RNAs 3 and 4, which are encapsidated together. I specifically use RNA1 for the model but make no explicit assumptions about specific packaging signals. To mod^{2}el electrostatic effects of RNA, I apply a Flory theory (87) to calculate free energy changes induced by encapsidating a charged RNA polymer. I separately examine four individual effects of RNA on assembly: the energy of RNA-RNA interactions during packaging, the entropic cost of confining the RNA polymer within the CCMV capsid, the energy of charged interaction between the capsid proteins and the RNA polymer, and the increased local concentration enabled by packing coat monomers on a single RNA strand. I apply analytical

[2] This chapter is based upon work in the paper Smith et al. (2014). Modeling the effect of RNA on capsid assembly pathways via stochastic simulation. In Review.

models of these effects to adjust kinetic rate parameters experimentally fit to bulk CCMV assembly data and examine how these different RNA effects, both individually and in concert, modify CCMV capsid assembly. The results reveal complex interplay balance of influences that collectively greatly enhance assembly while altering both kinetics and pathway usage relative to the *in vitro* model.

3.2 Methods

3.2.1 Capsid simulation method

The lab has previously developed a rules-based discrete event stochastic simulator called Discrete Event Simulator of Self-Assembly (DESSA) (33) to model the process of capsid assembly from individual subunit building blocks through individual association and dissociation events into completed capsids. Simulated assembly is governed by simple biochemical rule sets specifying the geometries of the subunits, three-dimensional positioning of binding sites and the specificities and on- and off-rates of binding events between binding sites. DESSA samples among all possible bond formation (association) and breaking (dissociation) events at each step in the simulation using a variant of the stochastic simulation algorithm (SSA) (49,50). I model CCMV assembly via dimers of coat proteins using model parameters computationally fit to *in vitro* assembly data from purified CCMV coat dimers (47), as derived in Chapter 2 and detailed in Xie et al (96). I specifically examine CCMV encapsidating RNA 1, which consists of 3171 nucleotides.

3.2.2 Modeling Nucleic Acid Effects

To better understand the specific ways in which RNA might influence CCMV capsid assembly, I subdivided estimation of RNA effects on model rate constants into four contributions: 1) RNA-

RNA interactions, 2) Entropy of RNA chain compression, 3) RNA-protein interactions, and 4) local concentration of coat on the RNA. While analytical models of each effect involve various approximations and assumptions, they provide sufficient guidance to give a theoretical projection of how these forces might act individually and in concert to alter overall kinetics and pathways of assembly. RNA-protein interactions and changes in concentration both cause an increase in equilibrium constant for CCMV, whereas RNA compression and RNA-RNA interaction cause a decrease in equilibrium constant. I apply the RNA-RNA and RNA-protein interaction effects to both the on and off rates evenly, while the effect of RNA compression and changes in concentration are applied solely to the on rates. These corrections are then applied to the previous best fit rate parameter values to represent the individual nucleic acid effects. I consider each contribution in turn:

3.2.2.1 RNA-RNA Interactions. I treat the RNA within the capsid as a polymer of individual Kuhn segments of RNA, each of Kuhn length *b*. Each Kuhn segment represents the length of nucleotides necessary to be freely jointed in solution. For semi-flexible or worm-like chains (WLC), the Kuhn length is approximately equal to double the persistence length of the polymer (97). The persistence length of RNA in a 1M monovalent salt solution has been estimated to be 1.3 nm (98), and thus b = 2.6 nm. The RNA is not solely confined to the entire interior of the capsid, however. The RNA forms a thin layer inside the capsid wall, in the case of CCMV RNA 1, of thickness $D \approx 1.18\ nm$, a value extrapolated from similar studies on other virus systems (99). I then calculate the Flory free energy for the polymer chain via the formula (88)

$$F_{RNA} \approx k_B T\left(\frac{l^2}{Nb^2} + \frac{1}{2}v\frac{N^2}{V}\right) \tag{3.1}$$

accounting for entropic and excluded volume interaction effects respectively. The excluded volume interaction term is reduced to reflect the presence of counterions on the RNA strand. *N* is the number of Kuhn segments, *V* is the pervaded volume of the RNA calculated as the volume within the capsid filled by the RNA:

$$V = \frac{4}{3}\pi(R^3 - (R-D)^3) \tag{3.2}$$

Here *R* = 10.5 nm is the inner radius of the capsid. *l* is defined to be the end-to-end distance of the RNA polymer which we calculate to be the maximum distance traversed within the volume of the RNA inside the capsid, i.e. $l = \pi R \approx 32.99$ nm. I define *v* as the excluded volume of a given Kuhn segment of the RNA and calculate the excluded volume following (100) via the formula:

$$v \approx \frac{\pi b^3}{48} + \frac{64\pi\lambda_B^3\lambda_D^2}{b^2} \approx 3.537nm^3 \tag{3.3}$$

where $\lambda_B \approx .7nm$ is the Bjerrum length and $\lambda_D \approx .3nm$ is the Debye screening length in 1M monovalent salt (100). Thus, for CCMV capsids assembling around RNA 1, which is a single strand of 3171 nucleotides, $N \approx 414.5$ and $F_{RNA} \approx k_B T(.388 + 208.40) = k_B T(208.788)$. This energy yields a multiplicative equilibrium constant $K_{EQ}(34)$ by $e^{\frac{-208.788}{180}} = .3135$. One would expect this energy to correspond to effects on both on- and off-rates. I approximate these kinetic effects by attributing the equilibrium change equally to on and off rates, multiplying the rate of binding events by $\sqrt{.3135} = .5599$ and dividing the rate of dissociation events by .5599 as well.

3.2.2.2 Entropy of RNA Chain Compression. I calculate the free energy change from RNA compression as a separate entropic cost from that described above (88). Recall that the RNA is not solely confined to the entire interior of the capsid (99), and so I must constrict the volume confinement further to reflect the nature of the thin layer of RNA within the capsid surface. Thus, the free energy of confinement is calculated by

$$F_{conf} \approx k_B T \frac{Nb^2}{R^2 - (R-D)^2} \approx k_B T(119.81) \tag{3.4}$$

Here, R is the interior radius of the capsid and $Nb^2/(R^2-(R-D)^2)$ represents the number of compression blobs (88) for the RNA polymer confined inside the capsid. Following the same method as above (34), I convert this into a K_{EQ} scaling factor of $e^{\frac{-119.81}{180}} = .5140$ which is attributed solely to on rates.

3.2.2.3 Free Energy of RNA-Protein Interactions. To understand the free energy associated with RNA-protein interaction, I have to take into consideration the attractive force between the negatively-charged RNA polymer and the positively charged capsid proteins. Here instead of treating Kuhn segments as the basic monomers, I will treat a single nucleotide as a monomer as that reflects a single charge interaction with the capsid proteins. I will determine the energetics of this interaction via the Flory theory of an adsorbed chain to a charged surface assuming that monomers (nucleotides) are uniformly distributed within the thickness of the RNA inside the capsid (88). Thus, I calculate the free energy associated with RNA-Protein charge interaction following (88):

$$F_{ch} \approx -k_B T \delta M \frac{r}{D} \tag{3.5}$$

Here, δ represents the adsorption energy per nucleotide-capsid protein contact, which we set at 0.5 to reflect the presence of counterions on the RNA polymer, *M* = 3171 is the number of nucleotides in the RNA polymer, and *r = .34 nm* is the length of a single RNA nucleotide. Thus, $F_{ch} \approx -k_B T \frac{1}{2} 3171(\frac{.34}{1.18}) = k_B T(-456.84)$. I convert this into a K_{EQ} scaling factor of $e^{\frac{456.84}{180}} = 12.6543$ which is applied evenly between the on and off rates, multiplying the rate for bond forming events by 3.5573 and dividing the rate for bond breaking events by 3.5573 as well.

3.2.2.4 Local Coat Protein Concentration by RNA. The *in vitro* capsid assembly experiments previously conducted to learn simulation parameters have capsid protein concentrations between 5 and 20 μM (96). I would expect the effective concentration to be vastly higher for coat proteins aggregated onto an initially disordered RNA strand. I estimate this local concentration by treating the RNA as a polymer in a weak confinement regime (WCR) (101). I use the formula $R_3(radius) = aN^{\nu_3}$ where *a* is the length of an individual nucleotide, 3.4Å, *N* is the number of nucleotides in the polymer and ν_3 is the Flory exponent, which is set to 0.6 in a good solvent.

In the case of CCMV, $R_3 = 428.7426$ Å and then the corresponding volume is $V = 3.013 \times 10^8$ Å^3. Previously, I used in simulations an experimental concentration of 15.6 μM of CCMV capsid proteins (96). This reflects 9.39432×10^{21} molecules/m^3. For a volume of 3.013×10^8 Å^3, one would have 3.1013 dimers at 15.6 μM. In order to have 90 dimers present in that volume, one would need an increase in concentration to $\frac{90}{3.1013} \times 15.6\mu M = 452.71\mu M$. To reflect this increase, I multiply the on-rates by $\frac{90}{3.1013} = 29.02$ in the simulation input.

Table 3.1. Corrections to equilibrium constants, on-rates, and off-rates for all possible combinations of four RNA effects. For brevity, I use a four digit binary code to reflect the combinations of effects. 0000 represents a hollow capsid with no RNA effects modeled, while 1111 represents all four effects included. The first digit represents the free energy of RNA-RNA interactions, the second digit represents compression of the RNA strand inside the capsid, the third digit represents RNA-capsid protein interactions, and the fourth digit represents local concentration of capsid proteins on the RNA polymer.

CCMV1	$K_{eq}(mol^{-1}m^2)$	k_+ **(mol^{-1}m^{2}s^{-1})**	k_- **(s^{-1})**
0000 – Hollow	1	1	1
1000 – RNA-RNA	.3135	.5599	1.786
0100 – Compression	.5140	.5140	1
0010 – RNA-Protein	12.65	3.557	.2812
0001 – Concentration	29.02	29.02	1
1100	.1611	.2878	1.786
1010	3.966	1.992	.5022
1001	9.098	16.25	1.786
0110	6.502	1.828	.2812
0101	14.92	14.92	1
0011	367.1	103.2	.2812
1110	2.038	1.024	.5022
1101	4.676	8.352	1.786
1011	115.1	57.80	.5022
0111	188.7	53.06	.2812
1111	59.15	29.71	.5022

3.2.3 Simulation Experiments and Analysis

I produced corrected parameter files, dividing effects into four categories, free energy of the RNA (RNA-RNA interaction), RNA compression, RNA-protein interactions and concentration changes. I further examined all possible combinations of these individual factors. I summarize the possible combinations of effects and the resulting corrections to the reaction rate parameters of the simulations in Table 3.1, using a four-digit binary code to represent the combination of effects in any given parameter domain. The first digit represents RNA-RNA interaction, the second digit represents RNA compression, the third digit represents RNA-protein interaction and the fourth digit represents increased protein concentration. A value of 1 means that effect is turned on and a value of 0 means that effect has been turned off.

I ran two hundred simulation trajectories for CCMV for each of the 16 combinations of effects. In previous work, I included enough capsid proteins to produce multiple completed capsids (96); however, in this case, since the specific process being examined is of one capsid assembling about its RNA, I limited the number of initial subunits to an amount just sufficient to assemble a single capsid, 90 subunits in the case of CCMV. Each simulation ends when either a completed capsid consisting of all initial subunits is formed, or a time limit of 100 seconds has been reached. This predetermined time limit was empirically determined to allow simulations to reach a state of pseudoequilibrium. As in Chapter 2, I applied a variety of data analysis techniques previously described in Section 2.2.4 to study simulation trajectories individually and in aggregate, including simulated light scattering curves, binding frequency tables, mass fraction plots and movies showing detailed assembly pathways.

3.3 Results

Table 3.1 shows the corrections to the equilibrium constant as well as on and off rates under the individual changes based upon the free energy of the RNA strand (RNA-RNA self-interaction), RNA compression, RNA-protein interaction and increased concentration, as well as different potential combinations of these effects. One would expect local concentration increase and RNA-protein interaction to yield net positive (assembly-promoting) contributions and RNA-RNA and RNA compression effects to yield net negative (assembly-inhibiting) contributions. I examined all sixteen possible combinations of presence or absence of the four effects. First, in section 3.3.1, I will focus on four representative combinations: A) hollow capsid (no RNA effects, code 0000), B) the two negative effects (RNA-RNA interaction and RNA compression, code 1100), C) the two positive effects (RNA-protein interaction and local concentration, code 0011), and D) the combination of all four RNA effects (code 1111). The remaining twelve combinations each behaves similarly to one of the four representative combinations discussed here and will be described in section 3.3.2.

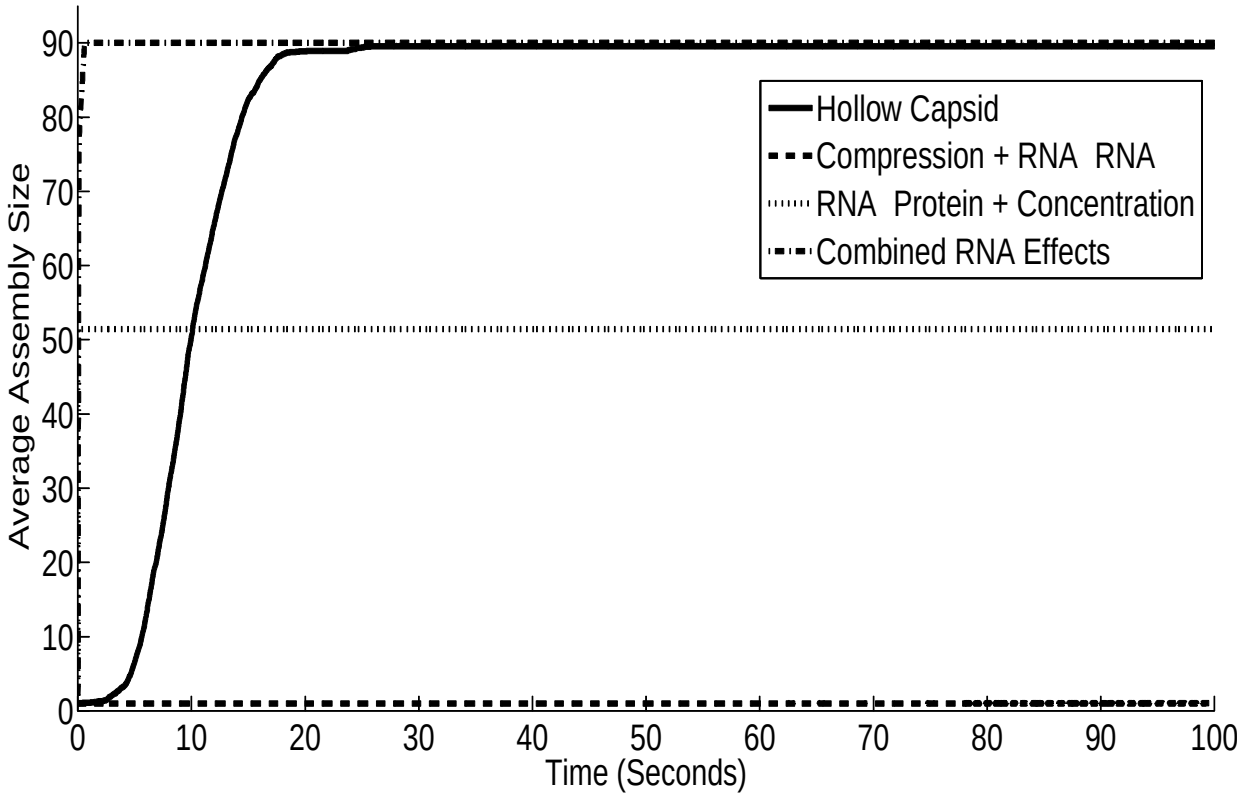

Figure 3.1. Simulated light scattering curves for CCMV capsid assembly under different representative combinations of RNA effects: hollow capsid, the two negative RNA effects (RNA Compression + RNA-RNA), the two positive RNA effects (RNA-Protein + Local Concentration) and the combination of all four RNA effects.

3.3.1 Four Representative RNA Effects

3.3.1.1 RNA effects on bulk assembly kinetics. Figure 3.1 shows simulated light scattering curves for CCMV under the four representative combinations of RNA effects. The hollow capsid curve behaves similarly to previous studies under no RNA effects (96), consistent with an initial lag phase followed by a sudden nucleation event then rapid growth to completion. The combined RNA effects curve also goes to completion but on a far faster time scale with no observable lag phase. In contrast, the other two cases in Figure 3.1 show unsuccessful assembly, but via two different failure modes. Combining the two negative effects, RNA compression and RNA-RNA interaction, abolishes any significant assembly, resulting in a simulated light scattering curve that remains barely above the origin. Combining the two positive effects, RNA-protein interaction

and increased concentration, produces rapid growth but plateaus at approximately half complete. This is a profile consistent with a regime observed when nucleation-limited growth breaks down, in which large intermediates assemble quickly and deplete the pool of free monomers, leading to a kinetically trapped system of partially assembled shells.

3.3.1.2 RNA effects on individual assembly trajectories. I next examined individual trajectories for each parameter combination to better understand how low-level interactions give rise to altered assembly pathways and thus the high-level kinetic profiles observed in Figure 3.1. I constructed mass fraction plots from individual simulation runs for each combination of RNA effects. Each mass fraction plot shows the fraction of capsid proteins in assemblies of each potential size at any given point during the simulation trajectory. Figure 3.2 shows mass fraction plots for the four representative cases: hollow capsid (Figure 3.2(a)), combined RNA effects (Figure 3.2(b)), RNA compression + RNA-RNA interaction (Figure 3.2(c)) and RNA-protein interaction + increased concentration (Figure 3.2(d)). Figure 3.2(a) initially shows a prolonged latency period where only small assemblies are produced, notably a pool of trimers (shown in red) building up in concert with a slow decline in the overall pool of monomers (shown in blue). This is followed by a sudden cascade of events then rapid completion of the capsid, consistent with a nucleation-limited growth mechanism observed for this model in prior studies (97). Figure 3.2(b) shows that the combination of all four events yields a qualitatively similar growth profile but on a timescale approximately 200-fold faster.

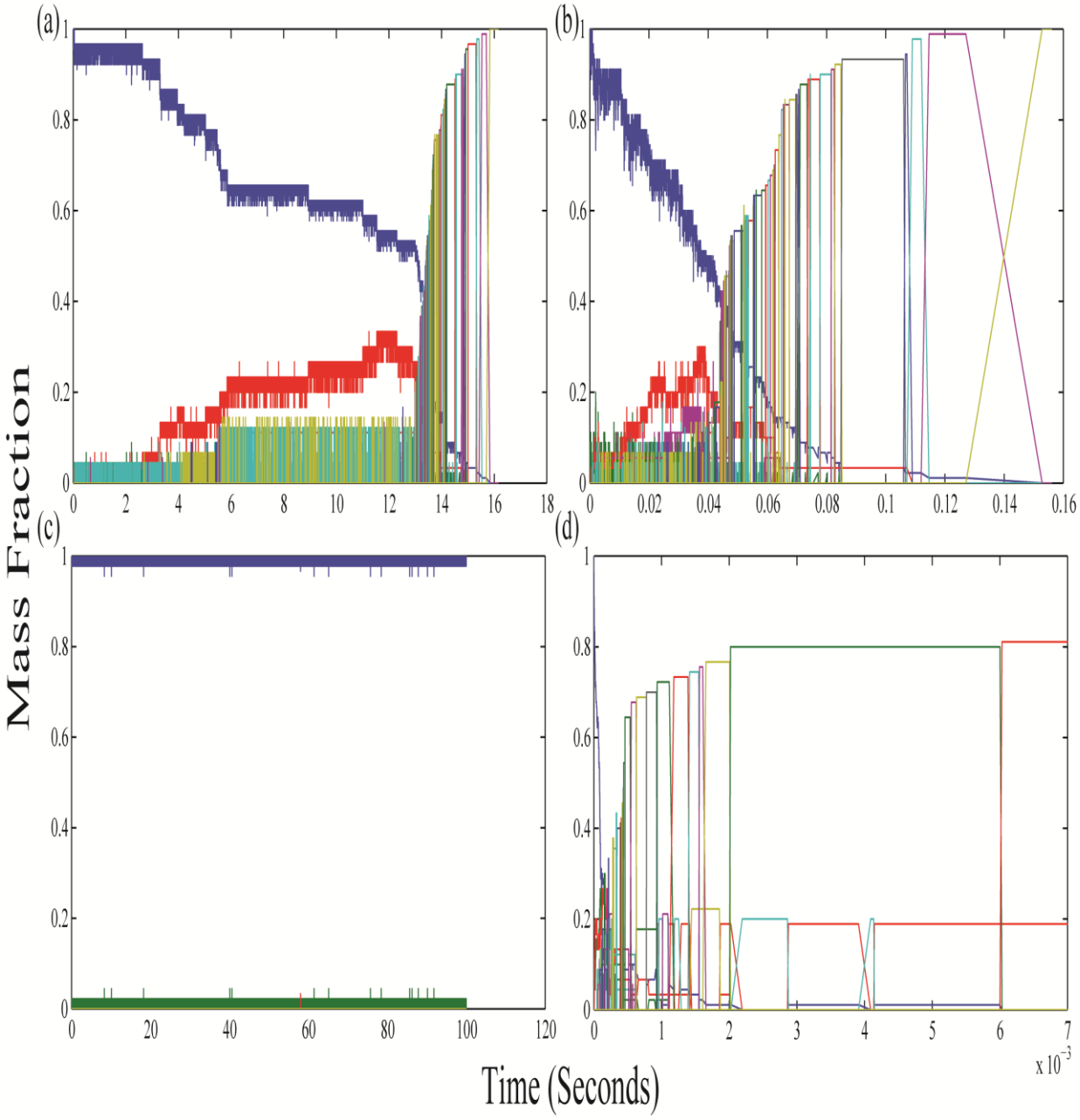

Figure 3.2. Mass fraction plots for (a) hollow CCMV capsid assembly, (b) CCMV capsid assembly with all combined RNA effects, (c) CCMV capsid assembly under both negative effects (1100), and (d) CCMV capsid assembly under both positive effects (0011). Each plot measures the mass fraction of each potential assembly size from individual monomers to completed capsids at each time point in a single simulation run. Note that the time axis is on a different scale for each plot due to the very different timescales of the assembly reaction under the different effects models. Additional combinations of parameters are examined in Figures 3.11-3.13 and seem to group approximately into one of the four paradigms observed here: A) slow, successful assembly; B) fast, successful assembly; C) lack of assembly; or D) fast, kinetically trapped assembly.

One notable feature of this profile is that the lag period and elongation period are approximately equal in length, which would normally be expected to abolish nucleation-limited growth and move a system into a kinetically trapped regime. I further note that an additional pool of pentamers builds up in the lag phase and drains away in the elongation phase, alongside the trimer pool that was observed in the empty capsid case. Figure 3.2(c) shows a failure to assemble when only the two negative RNA effects are applied, with no oligomer larger than a dimer ever appearing. Figure 3.2(d) shows that the combination of positive effects yields a profile in which elongation-like growth appears almost immediately without a discernible lag phase or nucleation step, leading to the formation of multiple large oligomers. While some manage to combine, the system ends up in a kinetically trapped end state with two large oligomers, a 73mer and a 17mer, that are unable to combine. Additional scenarios appear in Figures 3.11-3.13. As with the simulated light scattering profiles, each combination of effects appears to group qualitatively with one of the four representative scenarios in Figure 3.2.

To more clearly illustrate these virtual single-particle assembly pathways, I have also constructed movie files for the four representative scenarios. For the purposes of this document, I will only be able to show individual screen shots of important points during the movies. Each movie represents a single simulation trajectory following one pathway from single subunit to completed capsid. The movies described in Figures 3.3-3.6 are taken from the same simulations that produced the mass fraction plots in Figure 3.2 (Figures 3.3 and 3.2(a) are from the simulation and the same is true for Figures 3.4 and 3.2(b), Figures 3.5 and 3.2(c) and Figures 3.6 and 3.2(d)) for ease of comparison.

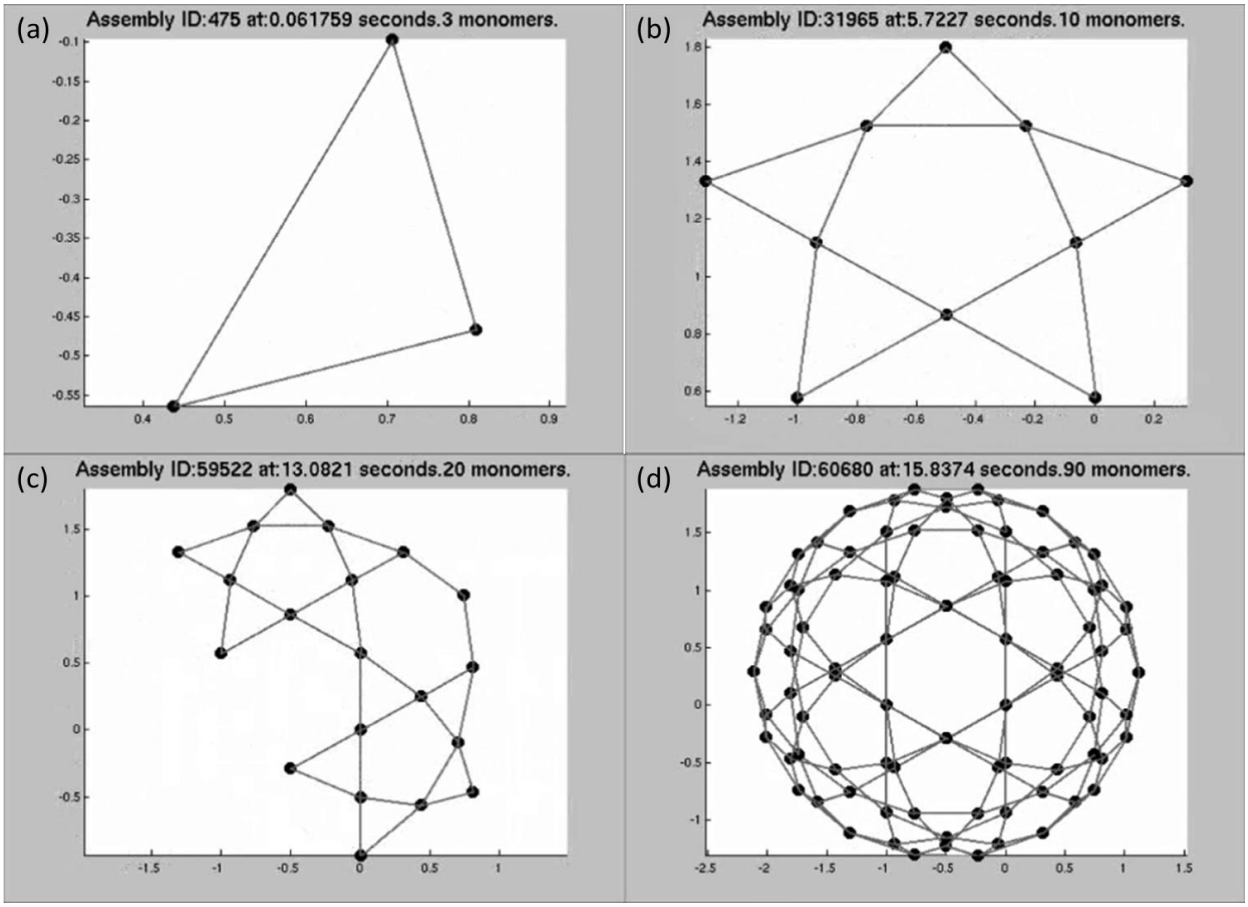

Figure 3.3. Individual frames of a movie following the same assembly trajectory of a hollow CCMV capsid shown in Figure 3.2(a). Assembly information for each frame is listed in the header for each sub-image.

Figure 3.3 displays screenshots of a movie following the assembly of a hollow CCMV capsid as in Figure 3.2(a). The movie shows a process of extensive trial-and-error, consistent with the weak bonding mechanisms predicted by numerous prior simulation studies as a way of ensuring nucleation-limited growth (25,36,102). This is highlighted by the distinct assembly IDs listed in Figures 3.3,(a) and (b). A new assembly ID is created after every bond forming and breaking event. Despite not much progress in the overall assembly of the CCMV capsid in the first two panels, the simulation has gone through over 31,000 assembly IDs each one representing a bond forming or breaking event in the simulation. While not all of these events are directly involved in this specific pathway, this is a telling example of the trial and error in these

assembly simulations. A distinct nucleation step is clearly observed, but in an unexpected way. While one would expect that a stable oligomer would be the key to touching off nucleation, stable pentamer oligomers do form in the simulation, as seen in Figure 3.3(b), but do not make effective templates for further polymerization. Elongation is touched off only when two of these pentagon structures bond together and subsequently close to form a 20-mer containing a hexagon structure shown in Figure 3.3(c), which provides an effective template for continued low-order elongation reactions. Following the assembly of this stable nucleation step, less than 1100 further simulation events occur, out of over 60,000 total, before the final capsid structure is completed as shown in Figure 3.3(d). Manual examination of other trajectories showed that this specific oligomerization is only one of a diverse array of pathways visualized, however, with nucleation seemingly not dependent on a specific size of intermediate but instead on a specific hexagon substructure, which could first appear as part of a variety of intermediate sizes.

Figure 3.4 follows assembly of a model capsid under combined RNA effects. Relative to Figure 3.3, the first striking difference is how much faster assembly is both with respect to simulation time and number of events along the path to assembly completion: less than 5000 total simulation events occur prior to the completed capsid forming. The mechanism is both faster and more directed, largely because the far slower off-rate reduces the trial-and-error process seen in the previous movie. Nucleation occurs just following 0.04 seconds as the first hexagon is closed, yielding a 20mer that quickly binds with another 20mer, the result of which can be seen in Figure 3.4(c). While binding of large oligomers is still rare, they occur much more frequently in this scenario, as quantified in the next section. As mentioned previously with respect to the mass fraction plots, the elongation period takes up a far larger percentage of the overall simulation time under the combined RNA effects case. Despite overall seeing far faster

assembly of CCMV capsids, nucleation occurs just prior to Figure 3.4(c) at .045 seconds, while the capsid as a whole is not completed until .153 seconds into the simulation. This implies that roughly 70 percent of the simulation time occurs following nucleation compared to just over 17 percent of the simulation time in the hollow CCMV capsid case.

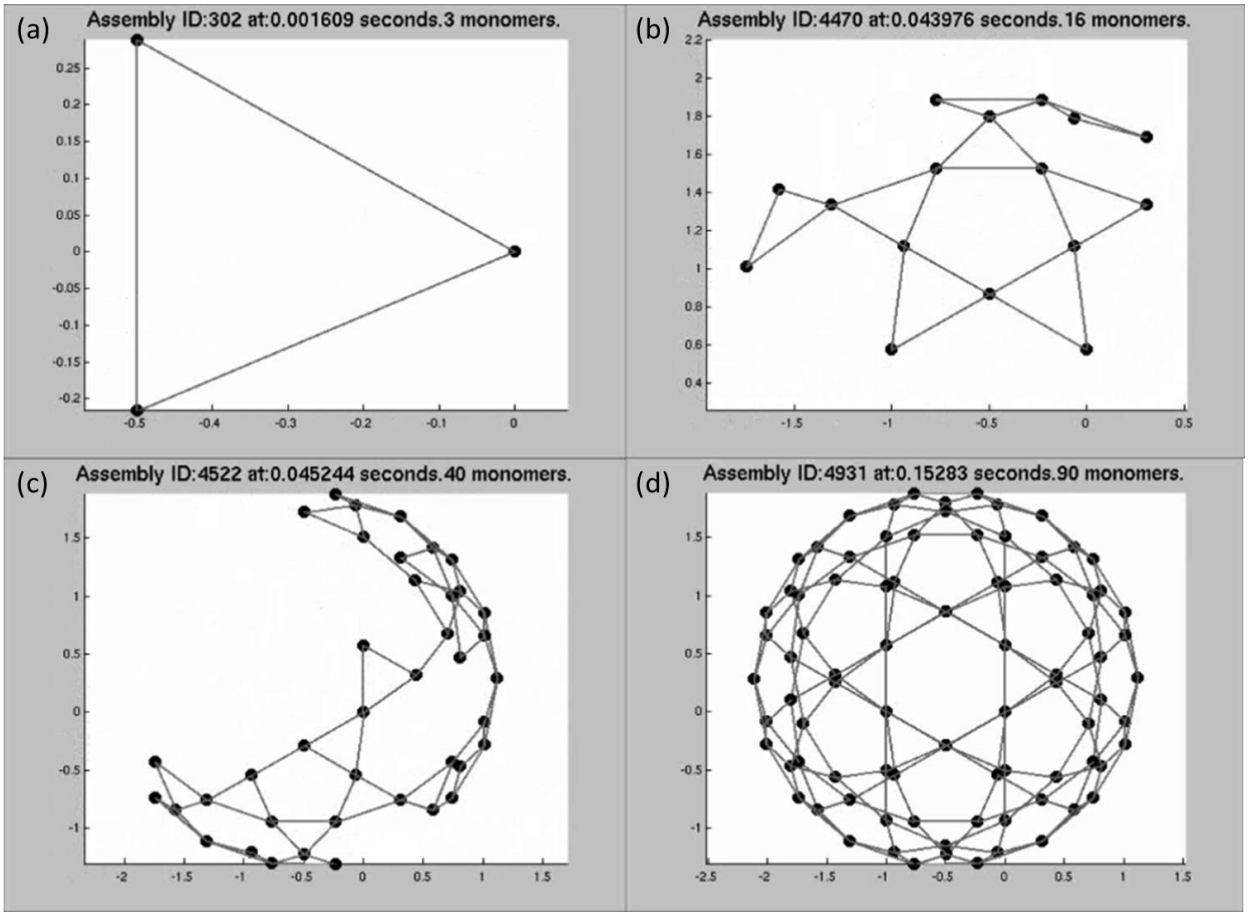

Figure 3.4. Individual frames of a movie following the same assembly trajectory of a CCMV capsid under combined RNA effects shown in Figure 3.2(b). Assembly information for each frame is listed in the header for each sub-image.

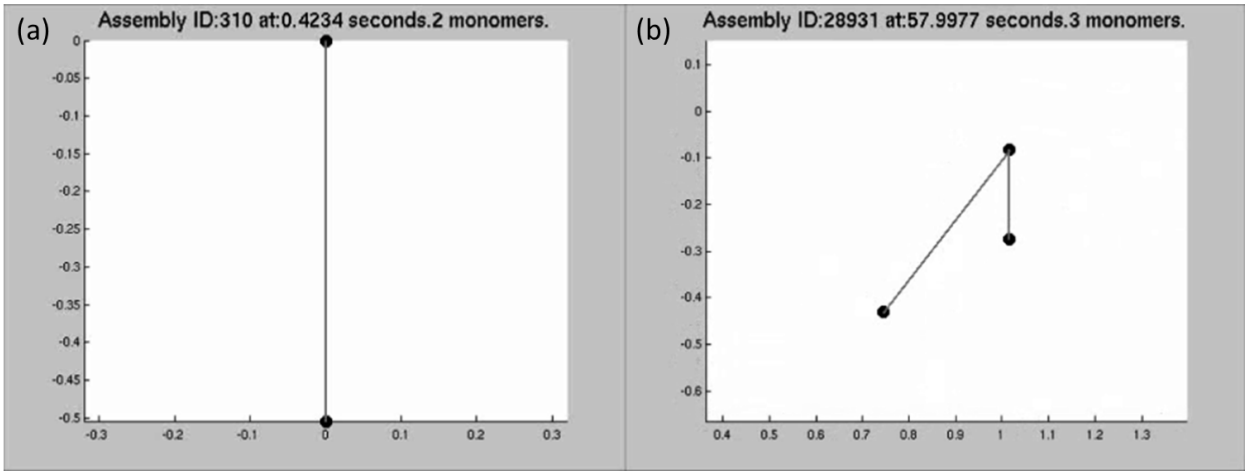

Figure 3.5. Individual frames of a movie following the same assembly trajectory of a CCMV capsid under both negative effects shown in Figure 3.2(c). Assembly information for each frame is listed in the header for each sub-image.

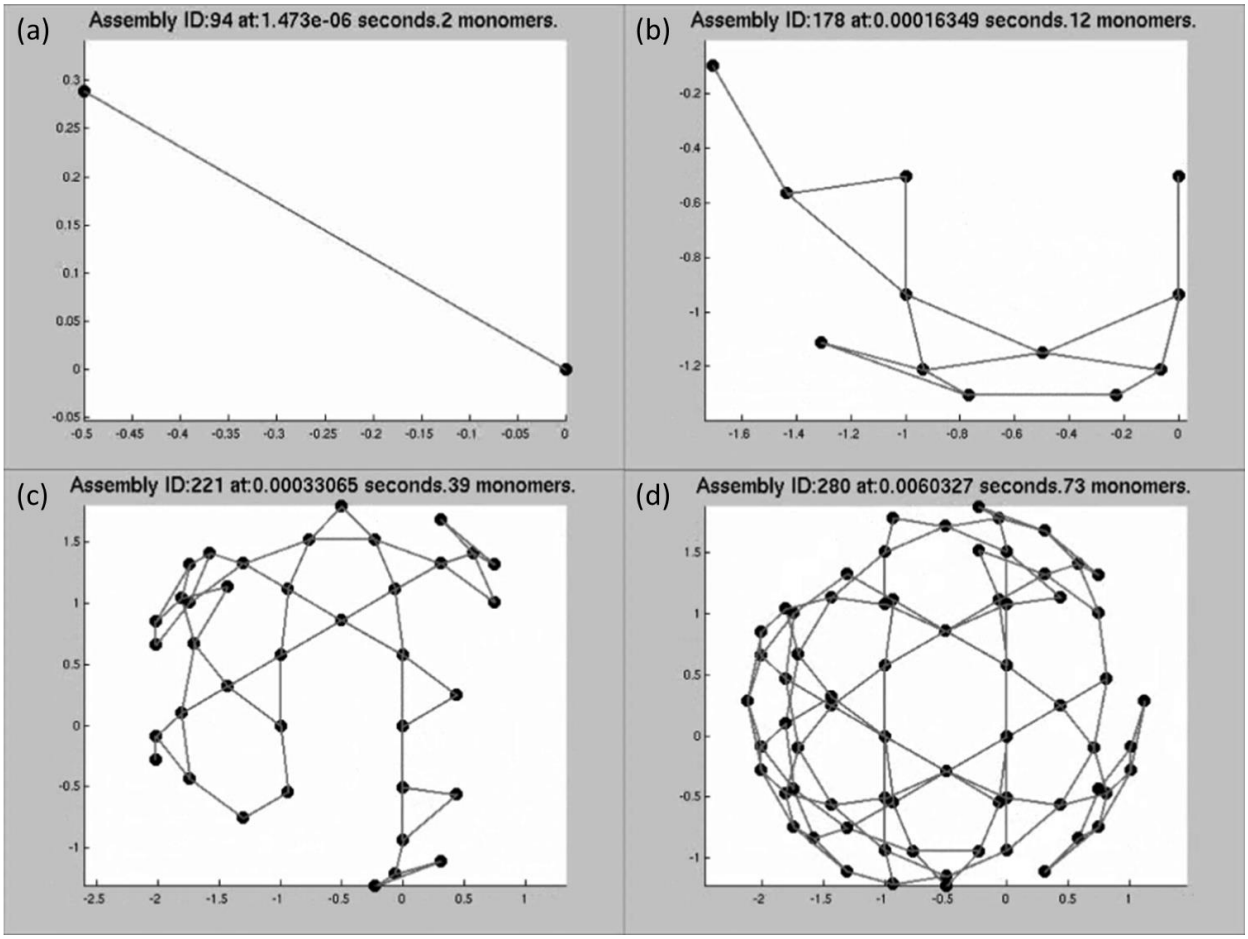

Figure 3.6. Individual frames of a movie following the same assembly trajectory of a CCMV capsid under both positive effects shown in Figure 3.2(d). Assembly information for each frame is listed in the header for each sub-image.

Figure 3.5 shows a trajectory of unsuccessful assembly under the two negative RNA effects, failing to produce any oligomer larger than a trimer on the timescale examined. The high ratio of off- to on-rates means that small oligomers are far more likely to fall apart than grow, making formation of a stable nucleus too rare to observe on this timescale. Figure 3.6 shows combined positive effects, yielding even more rapid initial stages of assembly than the combined RNA effects case. The rapid nucleation, however, leads to two large intermediates that are sterically incompatible but unable to break apart, locking the system into a kinetically trapped state. One can note that the trial-and-error process of this assembly is almost completely removed as only 280 simulation events occur prior to reaching this kinetically trapped 73mer. Examining the intermediates constructed during the simulation, there is far less structure to what is made. Instead, what is shown is the rapid addition of any subunits or intermediates that will bind together until nothing else fits in an appropriate puzzle piece, as can be seen in Figure 3.6(d) where multiple holes in the capsid structure remain to be filled without the necessary intermediates to fill them.

3.3.1.3 Averaged pathway usage across trajectories. To further understand the effect of RNA on CCMV assembly pathways, I constructed binding frequency tables quantifying average pathway usage for each potential combination of RNA effects. For each effect combination, the corresponding binding frequency table shows the frequency with which each assembly size is used as a reactant in producing any given larger assembly size (e.g., the frequency with which a dimer is a reactant for a reaction producing a pentamer). Figure 3.7 presents binding frequency tables for the four effect combinations cases detailed above: hollow capsid (Figure 3.7(a)), combined RNA effects (Figure 3.7(b)), negative effects only (Figure 3.7(c)) and positive effects only (Figure 3.7(d)). The remaining cases appear in Figures 3.14-3.16.

Hollow capsid assembly (Figure 3.7(a)) shows two major kinds of elongation reaction: a dominant monomer-addition reaction and a rarer trimer-addition alternative. Assembly under combined RNA effects (Figure 3.7(b)) shows both an increase in use of trimer reactions and the emergence of a third pathway of pentamer-addition reactions. Early stages of assembly yield more complex combinations of oligomer reactions for both conditions, but again with a wider variety in the presence of all RNA effects. Figure 3.7(c) shows near complete abolition of assembly, with all reactions using only monomer and dimer reactants. Over the two hundred simulations run, the largest intermediate ever formed is a tetramer. The combination of positive effects (Figure 3.7(d)) once again yields a far more complex picture involving a diverse array of nearly every potential binding reaction for producing small-to-medium oligomers but complete loss of complete or near-complete assemblies. This profile is typical of kinetically trapped domains, where numerous oligomers form and interact but eventually get locked into a domain of sterically incompatible, irreversible partially-assembled forms.

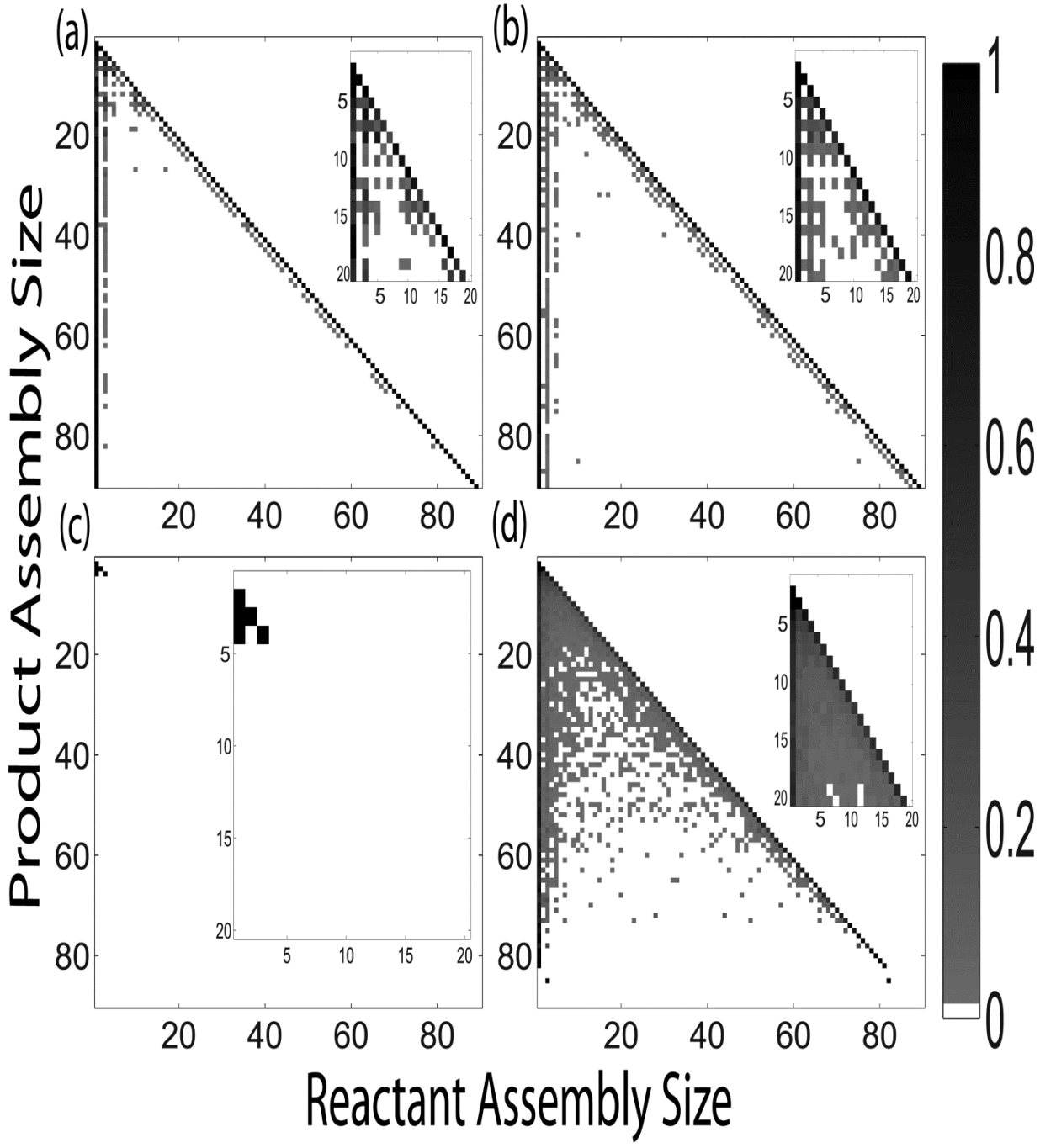

Figure 3.7. Binding frequency tables for (a) hollow CCMV capsid assembly (0000), (b) CCMV capsid assembly with all combined RNA effects (1111), (c) CCMV capsid assembly under both negative RNA effects (1100), and (d) CCMV capsid assembly under both positive RNA effects (0011). In each plot, each row corresponds to a product size and each column to reactant sizes that produce that product. Intensity of the pixel in the corresponding position shows the frequency with which the given reactant size is used to produce the given product size. Insets within each plot expand the upper-left corner of the main plot, corresponding to products of size 20 or smaller, to better visualize pathways involved in production of small oligomers.

3.3.2 Analysis of Remaining Twelve RNA Effect Combinations

3.3.2.1 RNA Effects on Bulk Assembly Kinetics. Figures 3.8-3.10 show simulated light scattering curves for CCMV under different combinations of RNA effects. Each part of the figure includes the cases of no effects as well as all effects to better understand the transition between hollow capsids and the presence of RNA. Because of the large difference in time scales between reactions, each plot is shown in two versions showing a slow timescale (part (a)) and a fast timescale (part (b)). Figure 3.8 examines each effect individually, i.e., simulated curves representing a hollow capsid, a capsid under all RNA effects, and curves for each individual effect: RNA-RNA interaction, RNA compression, RNA-protein interaction and increased local protein concentration. Figure 3.8(a) shows the full 100 seconds of simulation and Figure 3.8(b) shows just the first second. As expected given the scaling factors of the on- and off-rates (Table 3.1), the effect of RNA compression moderately reduces the speed of capsid assembly, although assembly is still achieved. The RNA-RNA interaction effect alone prevents any large intermediates from being formed, with nothing above an 8mer assembled in any simulation run. On the other hand, the effects of RNA-protein interaction and increased protein concentration both dramatically increase the rate of capsid assembly. Both of these simulated curves show similar kinetics to the combined RNA effects curve, which also shows a far faster assembly rate than the hollow capsid curve.

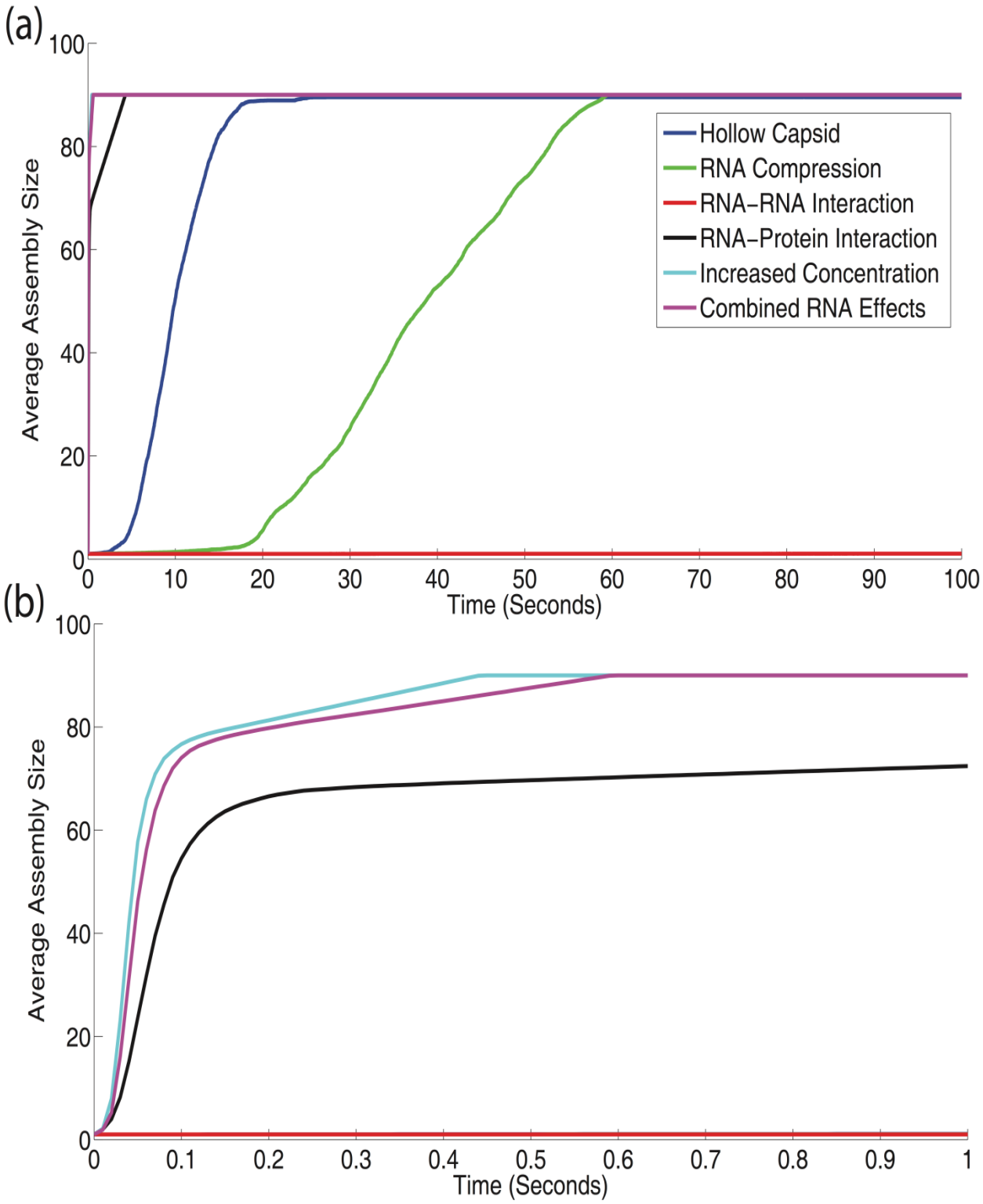

Figure 3.8. Simulated light scattering curves for CCMV capsid assembly under all individual RNA effects as well as the hollow capsid and combined RNA effects case. Figure S.8(a) shows the entire simulation time course while Figure 3.8(b) shows the first second.

Figure 3.9 examines each combination of two effects, as well as the baseline curves for hollow capsid and combined RNA effects, with Figure 3.9(a) showing 100 seconds and Figure 3.9(b) just the first five seconds. The majority of combinations behave as one would expect just based upon the scaling factors provided in Table 3.1. Addition of the negative factors of RNA compression and RNA-RNA interaction slow down assembly rate whereas the addition of the positive factors RNA-protein interaction and increased concentration increase assembly rate. The combination of the effects of RNA compression and RNA-RNA interaction still fails to produce larger intermediates, with tetramers now being the largest product assembled in the simulation runs. It is interesting to note that although the combination of RNA-RNA interaction and increased local protein concentration decreases equilibrium constants for assembly reactions, the overall assembly process is still faster than that of a hollow capsid. This is likely due to the different ways in which these effects are applied. This combination involves speeding up both on- and off-rates and while increasing the off-rate does increase the instability of assembled products, increasing the on-rate does more than enough to offset that. One surprising combination is that of RNA-protein interaction and increased local concentration. The two positive effects on capsid assembly, when combined, abolish assembly of capsids in the simulations. While large intermediates can sometimes be constructed, increasing the on-rate to the extent of this combination while also decreasing the off-rate inevitably leads to kinetically trapped intermediates without the ability to reform into completed capsids. This is a good example of the complexity and sometimes counter-intuitive nature inherent in self-assembly systems. Combining one positive and one negative effect results in systems that reliably go to completion, with each scenario yielding kinetics between those of the empty capsid and all-effects models. The one partial exception is the combination of RNA chain compression with

local concentration increase, which shows an initially greater lag than the all-effects model but ultimately goes to completion slightly faster than the model of all four effects.

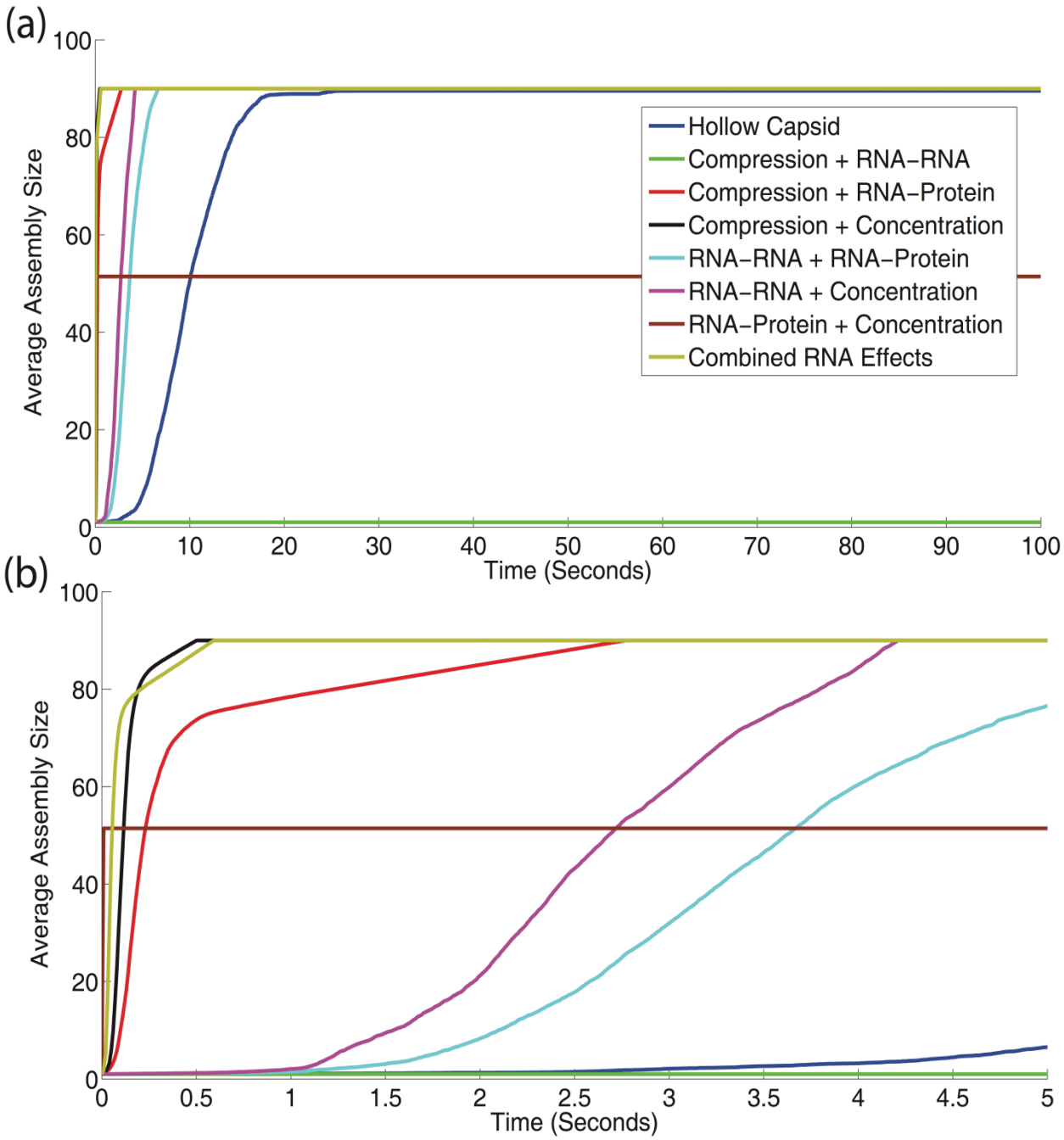

Figure 3.9. Simulated light scattering curves for CCMV capsid assembly under all combinations of two RNA effects as well as the hollow capsid and combined RNA effects case. Figure 3.5(a) shows the entire simulation time course while Figure 3.5(b) shows the first five seconds.

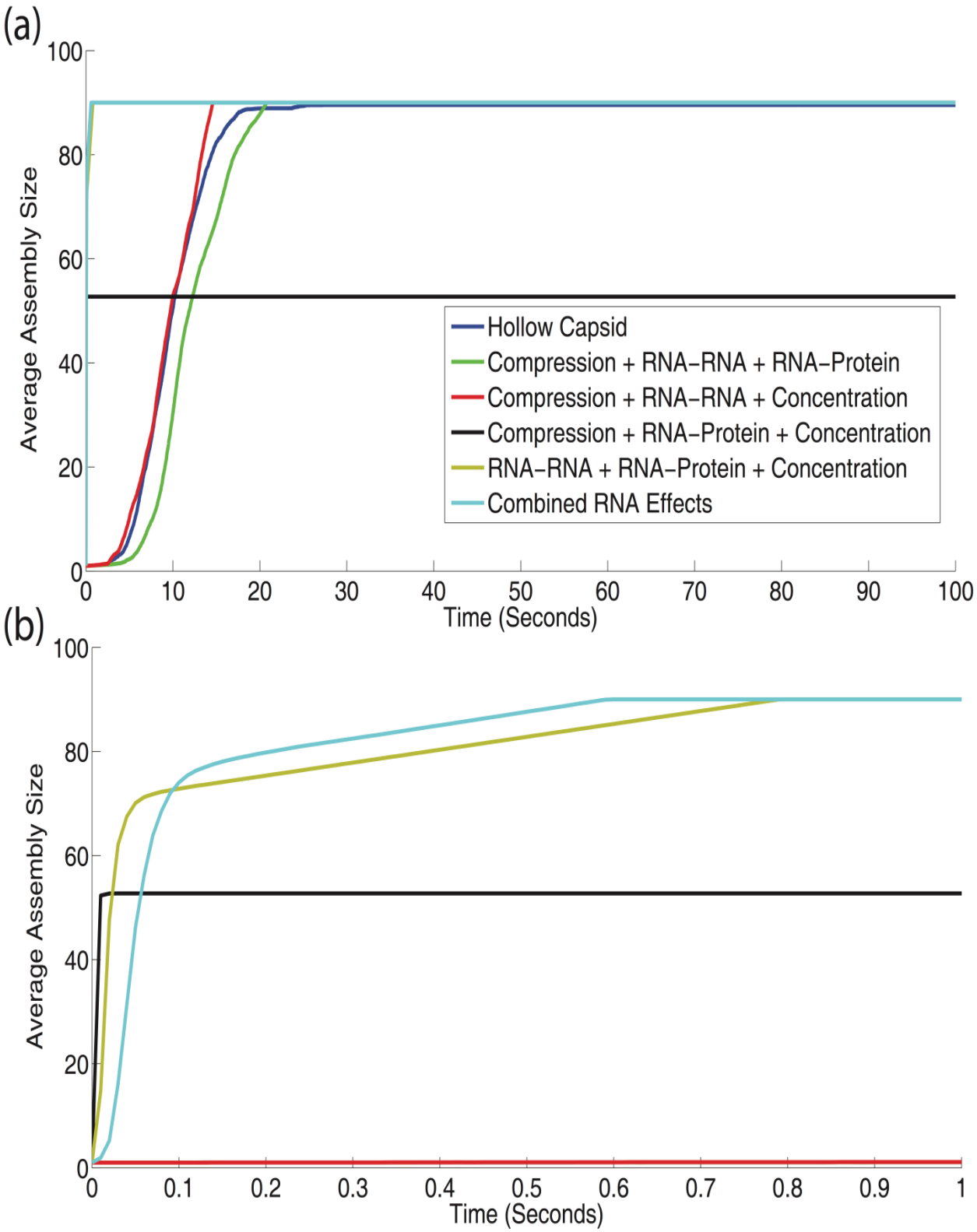

Figure 3.10. Simulated light scattering curves for CCMV capsid assembly under all combinations of three RNA effects as well as the hollow capsid and combined RNA effects case. Figure 3.6(a) shows the entire simulation time course while Figure 3.6(b) shows the first second.

Figure 3.10 then examines each combination of three effects as well as the baseline curves. Figure 3.10(a) again shows 100 seconds of simulation time, while Figure 3.10(b) shows just the first second. In each case where two negative effects are combined with a single positive effect, RNA-RNA + compression + RNA-protein and RNA-RNA + compression + concentration, the simulated curve is very similar to that of the hollow capsid. Once again, combining RNA-protein interaction and increased concentration effects without a strong negative effect included still does not result in completed capsids, shown here in the case of compression + RNA-protein + concentration; however, when combined with the RNA-RNA interaction effect, capsids are once again assembled. It should be noted that this combination of three effects, while having higher equilibrium constants for the assembly reactions and a faster initial rate of assembly, requires a longer average timeframe to completely assemble capsids than the combined effects case.

3.3.2.2 RNA Effects on Individual Assembly Trajectories. I next examined individual simulations in detail to explore fine details of pathway usage in single trajectories for all potential combinations of RNA effects. Figures 3.11-3.13 provide mass fraction plots showing evolution of counts of distinct oligomer sizes versus time for the twelve scenarios omitted from the previous section.

Figure 3.11 provides mass fraction plots showing trajectories for each individual RNA effect in isolation. Figure 3.11(a) shows the RNA-RNA interaction effect, which prevents all larger oligomers from forming. Only one trimer forms during the simulation. Some tetramers transiently form but are unstable and break down before they can seed further growth. In each of the other three cases, capsids still form, though there are some important distinctions. In Figure 3.11(b), the negative effect of RNA compression slows down overall assembly rate as well as

greatly extending the length of the early stages of capsid assembly following the nucleation event. Figure 3.11, (c) and (d), show a similar assembly process for the positive effects of RNA-Protein interaction and increased protein concentration respectively. In both of these cases, the nucleation rate and overall assembly rate are greatly increased both in absolute terms and relative to the elongation rate.

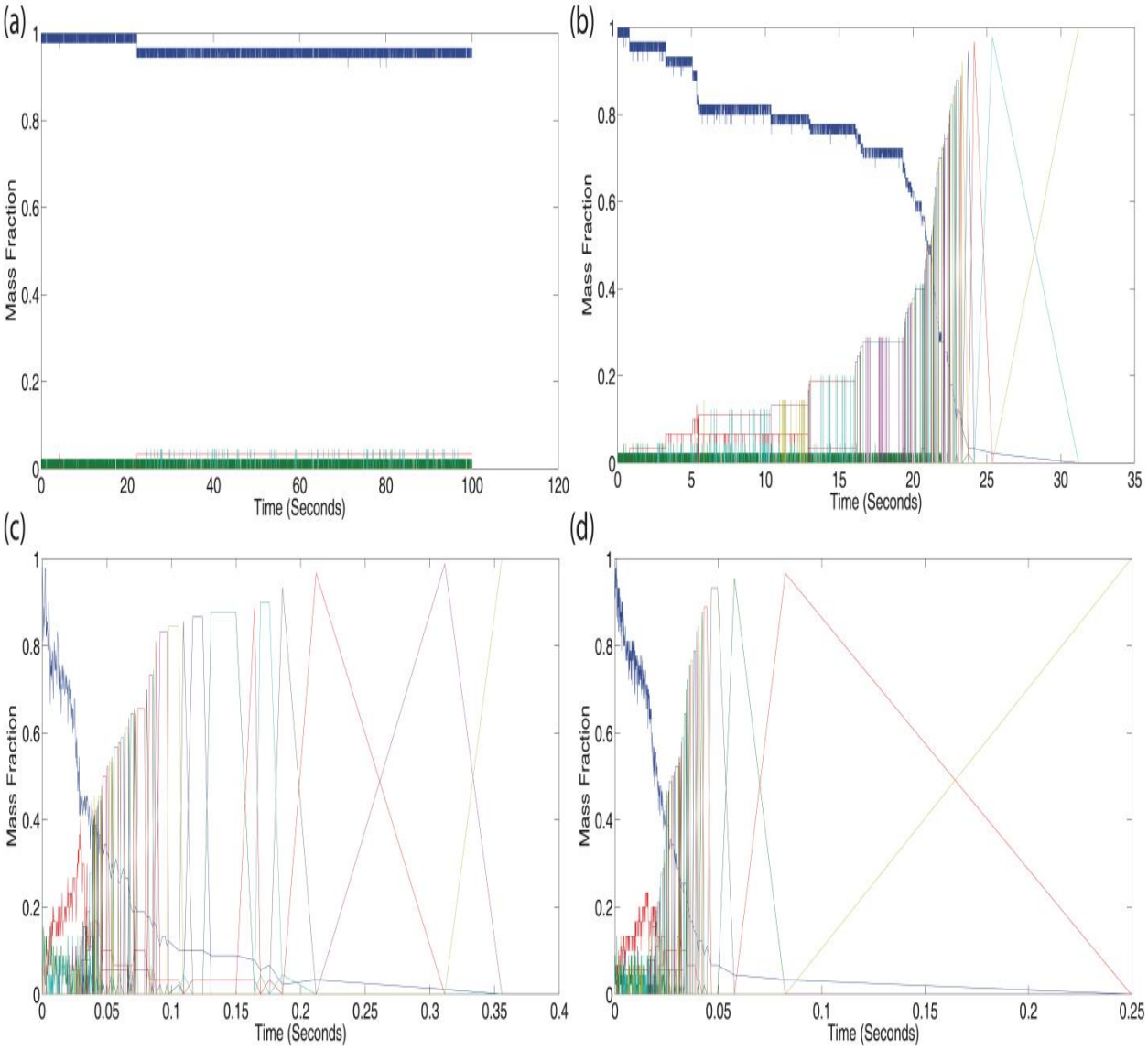

Figure 3.11. Mass fraction plots for CCMV capsid assembly with individual RNA effects: (a) RNA-RNA interaction, (b) RNA compression, (c) RNA-protein interaction, (d) increased protein concentration. The negative RNA-RNA interaction effect prevents any large intermediates from being formed, while the other three still allow for capsids to be completed.

Figure 3.12 examines the influence of pairs of RNA effects on assembly, with each subfigure covering one scenario pairing a positive and a negative effect. For each of these figures I use the same binary four digit code to explain the effect combinations as described in Table 3.1. The first digit represents RNA-RNA interaction, the second digit represents RNA compression, the third digit represents RNA-protein interaction and the fourth digit represents increased protein concentration. A value of 1 means that effect is turned on and a value of 0 means that effect has been turned off. Figure 3.12(a) shows effect combination 1010, which is RNA-RNA interaction combined with RNA-protein interaction, which behaves similarly to hollow capsid assembly, but with a moderately increased rate of assembly and a slightly more gradual relative assembly rate following the nucleation event. Figure 3.12(b) shows effect combination 1001, RNA-RNA interaction combined with increased concentration. Again assembly rate is moderately increased over the hollow capsid case, but with significantly altered behavior. Here, again, there is a distinct lag in the earlier stages of assembly following the nucleation event, indicative of a need for a second high-order nucleation-like event to allow elongation to proceed. Oddly, progress seems to depend on interaction of a 36mer with a trimer, a reaction step that we observed frequently, but not always, in simulation runs under these conditions where an apparent second lag occurs. This same point in assembly seems to be important in the similar case of Figure 3.12(b) where a lag is seen until assembly of a 39mer. Figure 3.12, (c) and (d), show similar mass fraction plots to one another. In each case, RNA compression is paired with a single positive effect, RNA-protein interaction (0110) and increased concentration (0101) respectively. Both behave as slightly slowed down versions of the single positive effect plots in Figure 3.12, (c) and (d), respectively. Because of the weaker nature of the RNA compression effect, this minor change makes sense.

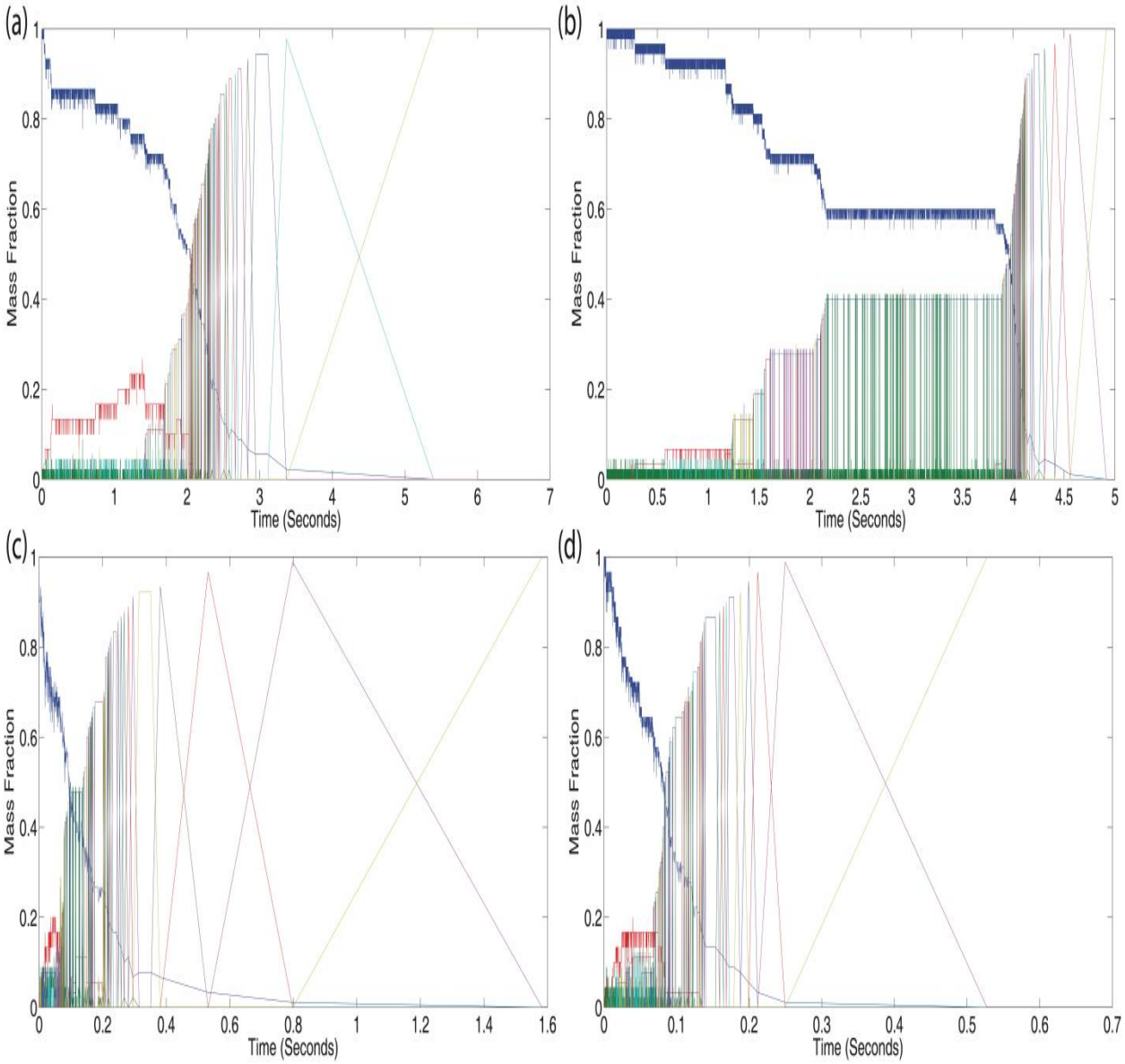

Figure 3.12. Mass fraction plots for CCMV capsid assembly with combinations of two RNA effects. Combinations are described by a four digit binary code where a 1 means an effect has been turned on and a 0 means an effect has been turned off: (a) is 1010, (b) is 1001, (c) is 0110, (d) is 0101.

Figure 3.13 examines the remaining cases, combinations of three different RNA effects, using the same binary code as described previously. Figure 3.13, (a) and (b) each show both negative RNA effects combined with one of the two positive RNA effects, RNA-protein interaction (1110) or increased concentration (1101). In both cases, the mass fraction plot is similar to that of the hollow capsid, although at a slightly faster rate. On the other hand, the

effect combination 1011 of RNA-RNA interaction and the two positive effects of RNA-protein interaction and increased concentration in Figure 3.13(c), behaves much more like the combined effects case in Figure 3.2(b), although the assembly rate is still faster and the nucleation peak is still much sharper, as with the hollow capsids case.

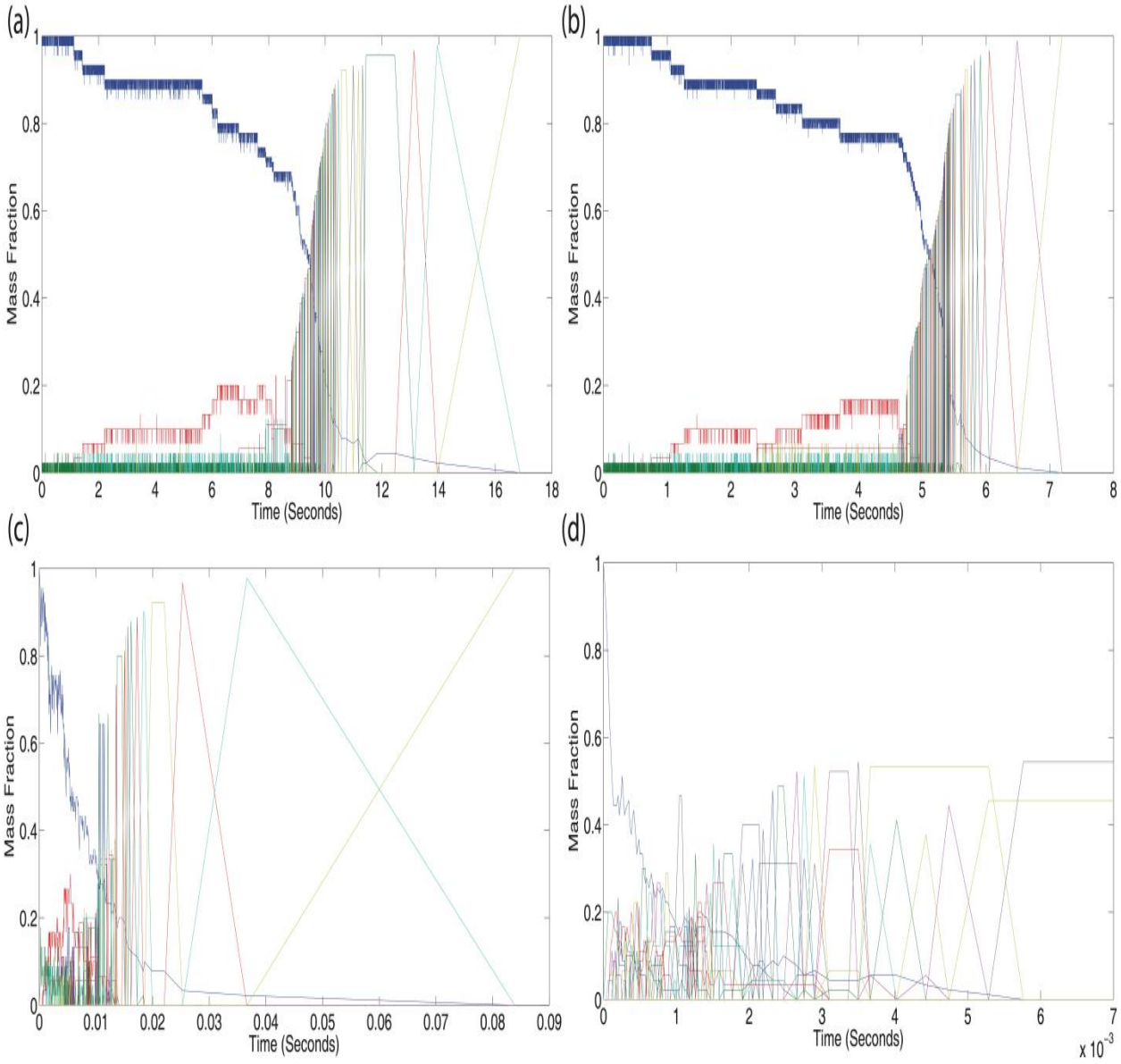

Figure 3.13. Mass fraction plots for CCMV capsid assembly with combinations of three RNA effects. Combinations are described by a four digit binary code where a 1 means an effect has been turned on and a 0 means an effect has been turned off: (a) is 1110, (b) is 1101, (c) is 1011, (d) is 0111.

Lastly, Figure 3.13(d) shows the effects combination 0111 involving RNA compression, RNA-protein interaction and increased concentration. The figure shows kinetic trapping; similar to what is observed in the case of only positive effects, with elevated on-rates resulting in multiple nucleation events happening near simultaneously. Without a fast off-rate to counterbalance this, two stable intermediates are formed, a 49mer and a 41mer.

3.3.2.3 Averaged Pathway Usage across Trajectories for all Combinations of RNA Effects. I next examine aggregate pathway usage for the twelve remaining scenarios as quantified in binding frequency tables. Each table shows, for a given scenario, the frequency with which each possible reactant size is used to make each possible product size, averaged over 200 trajectories.

Figure 3.14 provides binding frequency tables for CCMV capsid assembly under single RNA effects. Figure 3.14(a) shows once again that the RNA-RNA interaction effect alone abolishes capsid production. Over the two hundred simulation runs for this case, a 8mer was the largest assembly produced. In contrast, the RNA compression case in Figure 3.14(b) exhibits a very similar set of pathways as the hollow capsid case, which is to be expected since RNA compression has only a minor negative effect on the on rates. Figure 3.14, (c) and (d), show the two positive effects, RNA-protein interaction and increased concentration, both of which behave similarly and more in line with the combined RNA effects of Figure 3.7(b). In both Figures 3.14, (c) and (d), the pentamer-based assembly pathway is prominent, although the RNA-protein interaction case shows an increased variety in pathways utilized compared to both the increased local concentration and the combined effects cases.

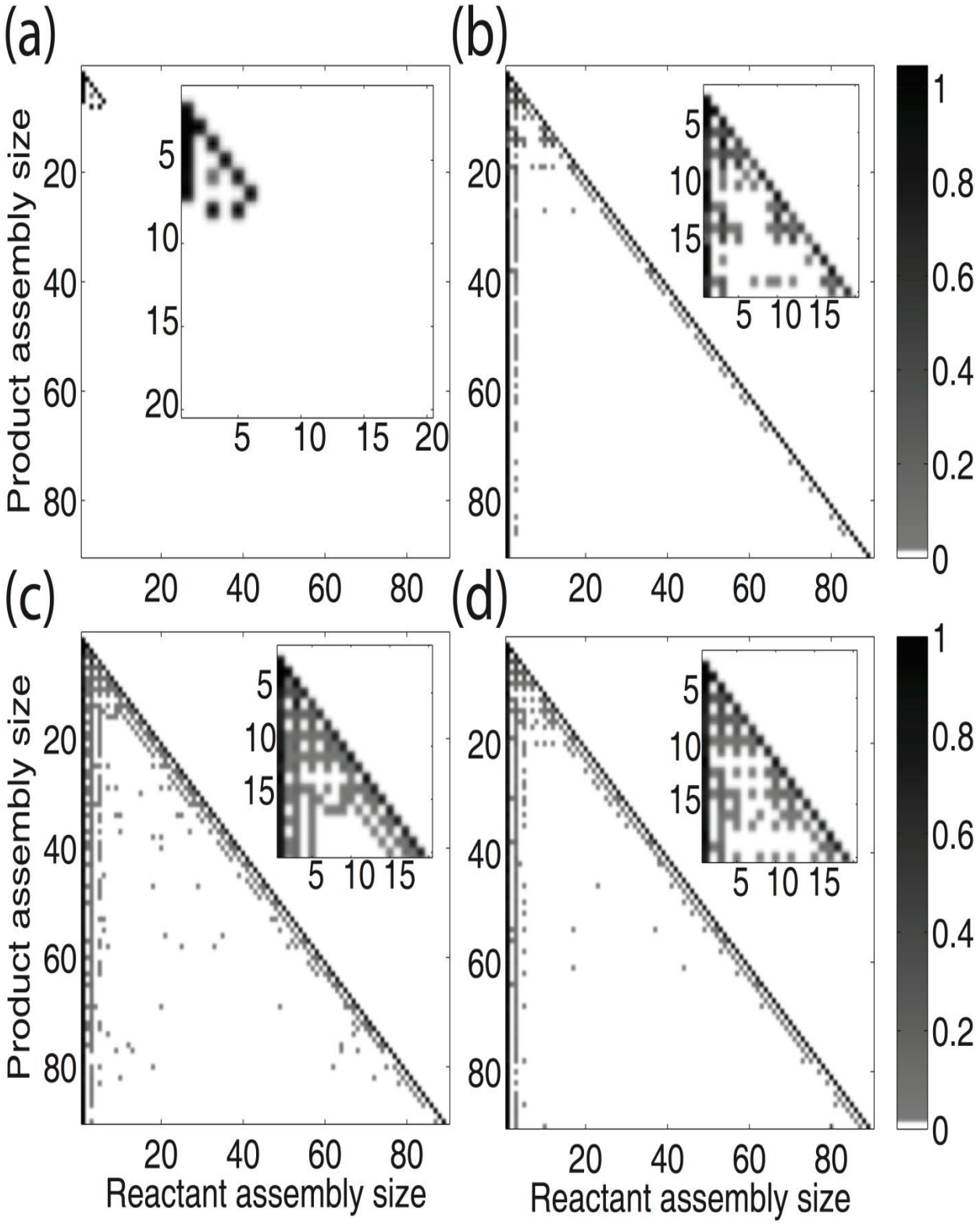

Figure 3.14. Binding frequency tables for CCMV capsid assembly with individual RNA effects: (a) RNA-RNA interaction, (b) RNA compression, (c) RNA-protein interaction, (d) increased concentration.

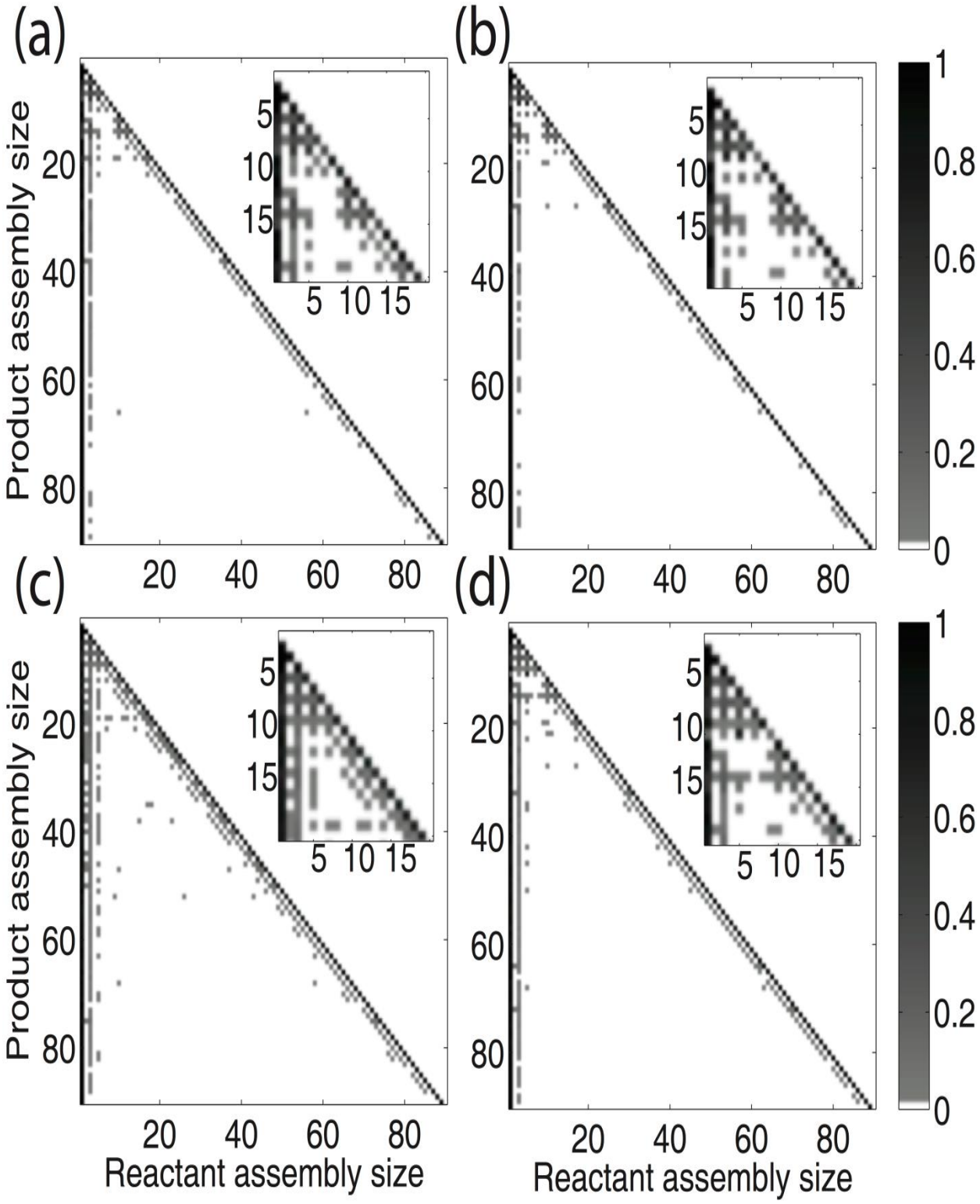

Figure 3.15. Binding frequency tables for CCMV capsid assembly with combinations of two RNA effects. Combinations are described by a four digit binary code where a 1 means an effect has been turned on and a 0 means an effect has been turned off: (a) is 1010, (b) is 1001, (c) is 0110, (d) is 0101.

Figure 3.15 shows binding frequency tables for CCMV capsid assembly under two RNA effects. I again use the same coding scheme as we did for the mass fraction plots and Table 3.1. Coupling RNA-RNA interaction with either single positive effect produces binding frequency tables very similar to that of a hollow capsid. Figures 3.15, (a) and (b), show the combination of RNA-RNA interaction with RNA protein interaction (1010) and increased concentration (1001), respectively. Both only have two main elongation pathways: monomer and trimer addition. It is interesting to note that there is little difference in the binding frequency plots despite a significant difference in their mass fraction plots. There is no change in binding frequencies that would seem to provide insight into the second lag phase seen in the 1001 effect combination. Figure 3.15(c), RNA compression and RNA-protein interaction (0110), and Figure 3.15(d), RNA compression and increased local concentration (0101), show similar pathway utilization to figure 3.14, (c) and (d), which again makes sense as RNA compression has only a minor negative effect to the on-rate.

Figure 3.16 displays binding frequency tables of CCMV capsid assembly under combinations of three RNA effects, again with the same binary code previously described. Figure 3.16, (a) and (b), show combinations involving both negative effects and a single positive effect: RNA-RNA interaction, RNA compression, and RNA-protein interaction (1110) or RNA-RNA interaction, RNA compression, and increased local concentration (1101), respectively. In each case, the binding frequency table shows pathways similar to those of the original hollow capsid case with growth primarily based upon monomer and trimer addition. In contrast, Figure 3.16(c) shows the combination of RNA-RNA interaction, RNA-protein interaction and increased local concentration (1011), which yields pathways very similar to the combined effects case, with potentially slightly more diversity in the lower assembly sizes.

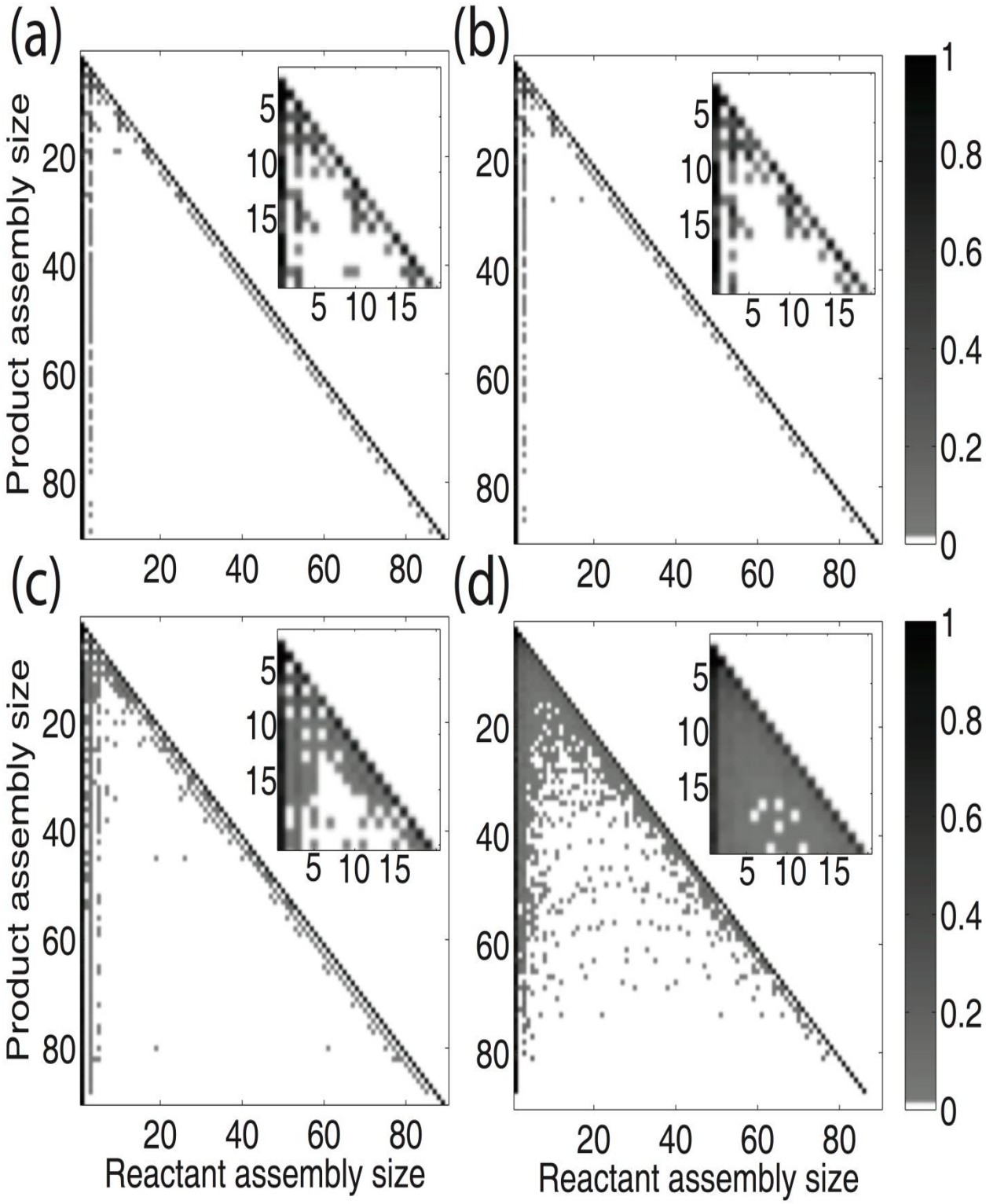

Figure 3.16. Binding frequency tables for CCMV capsid assembly with combinations of three RNA effects. Combinations are described by a four digit binary code where a 1 means an effect has been turned on and a 0 means an effect has been turned off: (a) is 1110, (b) is 1101, (c) is 1011, (d) is 0111.

Lastly, Figure 3.16(d) shows the combination of RNA compression, RNA-protein interaction and increased local concentration (0111), which is still unable to produce completed capsids but again shows a wide array of binding reactions as was seen in the other kinetically trapped case, the combination of just the two positive effects.

The results of the binding frequency tables can be grouped into four major potential outcomes. There are two cases in which no large intermediates are produced and thus no real pathway information gained (1000, 1100). Two further cases result in kinetic trapping with a wide array of binding reactions among small-to-medium oligomers but failure to produce complete capsids (0011, 0111). Of the cases where completed capsids are produced, binding reactions are either primarily based on monomer and trimer addition (0000, 0100, 1010, 1001, 1110, 1101) or pathways exist in which a third option of pentamer addition becomes frequent (0010, 0001, 0110, 0101, 1011, 1111). The appearance of this third elongation pathway is typically associated with fast, successful assembly while the two-pathway systems tend to assemble more slowly. While this is a simplification, as there is still diversity within each of these classifications, it provides an approximate classification of the range of pathways that can be seen under the different combinations of RNA effects. It is also worth noting, that even the simpler pathway sets actually describe ensembles of many possible pathways corresponding to potential addition of multiple possible oligomers and likely multiple possible binding sites at each step of elongation.

3.4 Discussion

In this work, I have applied a combination of simulation, analytical modeling, and data-fitting to explore how nucleic acid may act to alter kinetics and pathway usage in viral capsid

self-assembly with CCMV assembly on RNA1 as a model system. Using theoretical models of various expected contributions of nucleic acid to assembly, one can make modifications to simulations of the RNA-free experimental system to reflect RNA effects without losing the simulation speed or detail essential to these pathway studies. This approach provides a way to bypass inherent limitations of both purely experimental and purely theoretical studies, using simulation-based model-fitting to learn quantitative models of interactions of purified coat proteins and then computationally project how the activity of those coat proteins would be altered under regimes in which direct experimental observation is infeasible. I break down the effect of the presence of RNA into four categories: RNA-RNA interactions, RNA chain compression, RNA- protein interactions, and local concentration increases of coat protein on nucleic acid. The resulting simulations provide an unprecedented level of detail in understanding how nucleic acid might influence fine-scale assembly pathways of a real virus capsid system, revealing potentially significant pathway changes between *in vitro* and *in vivo* systems and demonstrating how a balance of growth-promoting and growth-inhibiting forces can act synergistically to greatly accelerate growth without producing the kinetic trapping one would expect from prior theoretical models.

Nucleation-limited growth was long ago established as a key feature of capsid assembly (103) and simulation (25-27,29) and theory (28) have repeatedly suggested that it is crucial to robust capsid production. Nucleation theory provides a conceptual basis for understanding some of my key observations. Simulations of empty capsids, with parameters learned by direct fitting to *in vitro* data, clearly show nucleation-limited growth, as one would expect from the relatively weak binding interactions of the learned model and resulting extensive trial-and-error before the capsid can build a stable intermediate. This observation is likewise consistent with prior theory

(103) and repeated simulation observations (25-27,29). It is, then, unsurprising that incorporating those features of nucleic acid that tend to inhibit assembly will abolish growth. The theory predicts that the capsid should exist in a domain in which a rare nucleation event is needed to drive assembly forward, and biasing against growth will make these rare events essentially non-existent on the normal timescale of assembly. It is similarly unsurprising, in the light of nucleation theory, that incorporating only RNA effects that promote assembly likewise abolishes capsid formation. As theoretical studies have long suggested, promoting stronger binding will tend to drive such a system into a kinetically trapped domain, in which excessively rapid nucleation relative to elongation leads to multiple nucleation events before individual capsids can complete, exhausting free subunits and trapping the system in a partially assembled condition, exactly as I observed.

What is surprising, then, is that combining all of the RNA effects leads to growth that is orders of magnitude faster than assembly of the empty capsid, yet just as successful. One would naively expect that a large net shift toward more stable binding would drive the system into the kinetically trapped domain and in fact abolish assembly of complete capsids. The combined RNA effect model indeed yields much stronger free energies of binding, much faster kinetics, and far less trial-and-error than the empty capsid model, yet it produces capsids with just as much success. This observation is a seeming paradox to theoretical models, yet empirically, this is known to be correct. My results show, for the first time, how the seeming paradox is resolved at the level of fine-scale interactions. Specifically, the RNA effect model produces a striking balance of effects, with positive RNA effects greatly accelerating both pre-nucleation and post-nucleation growth rates while negative effects disproportionately slow prenucleation steps and prevent the multiple-nucleation phenomenon that normally leads to kinetically trapped domains.

These trends are reinforced by a novel accumulation of pentamers pre-nucleation that then drives a new pentamer-addition pathway during elongation. Collectively, these effects lead to reliable capsid completion with approximately 200-fold speedup relative to the empty capsid model.

Like all theoretical work, this study does depend on numerous assumptions and approximations in the simulation models, the process of fitting model parameters to experimental data, and in the analytical theories used to estimate corrections to those parameters to account for the presence of nucleic acid. The fine details of the resulting simulations need to be considered speculative and considered a source of hypotheses rather than a definitive statement about how CCMV assembles on RNA. Nonetheless, it provides projections at a level of detail unavailable by any prior method, revealing emergent effects of the models that would not have been predictable by any other method and revealing a possible answer for an important puzzle in reconciling theory and empirical observation about virus assembly *in vitro* versus *in vivo*. While these models do not provide certainty about true CCMV assembly pathways, neither can any other method, and these models do provide better inference than any alternative yet available. I hope these models can inspire future work, beyond the scope of my own theoretical study, to evaluate testable features of these models on real capsid assembly systems.

In addition to providing insight into an important but puzzling feature of the assembly of RNA viruses, the work helps illustrate some of the continuing value of simulation methods and the synthesis of simulation, experiment, and analytical theory. The simulation approach provides a unique window into fine-scale reaction processes, such as capsid assembly, by providing a platform in which we can observe and precisely quantify aspects of assembly one would have no technology to observe experimentally. Model-fitting to experimental data makes it possible to bring this capability out of the realm of abstract theory of generic capsids to prediction about

experimentally unobservable features of specific real systems. Analytical theory lets us take that contribution a step further, to project how the system might behave in environments one cannot monitor experimentally at all. While I have applied those capabilities here to a specific question of how nucleic acid might influence assembly of spherical capsids, this combination of techniques can be expected to apply to numerous other complex systems for which theory, simulation, and experimental methods individually are limited.

Chapter 4: Applying Macromolecular Crowding Models to Capsid Assembly Simulations[3]

4.1 Introduction

My next major thesis contribution focuses on the impact of macromolecular crowding on capsid assembly pathways. To this end, I combined two separate modeling approaches—one for simulating capsid assembly and one for simulating the effects of macromolecular crowding on simple assembly reactions—to explore ranges of possible crowding effects on model capsid assembly systems. To achieve the high efficiency needed to model large numbers of trajectories for systems with often very slow rate-limiting nucleation reactions, I relied on stochastic simulation models of capsid assembly (30,32,33), trained to fit light scattering data on real *in vitro* capsid assembly systems (53,96). To model crowding without compromising runtime, I extended an approach using test reactions run in a comparatively slow space-aware diffusion model (85) to train regression models one c[3]an use to estimate crowding effects on kinetics of a wide range of parameter values (86). To deal with uncertainty in the total crowding level likely to be encountered by any real viral system, I applied this model across a broad range of total crowding levels, from 0% to 45% total excluded volume. The result is a dual-scale simulation that offers the efficiency of the prior rule-based models, needed for detailed pathway analysis, combined with physical representations of a simple particle model of macromolecular crowding. I used these tools to project possible effects of increased crowding on three virus systems analyzed in the prior work: CCMV, HPV, and HBV. The remainder of this chapter describes the computational approach in greater detail, reports apparent effects of increasing levels of

[3] This chapter is based upon work published in Smith et al. 2014. Applying molecular crowding models to simulations of virus capsid assembly in vitro. Biophys. J. 106:310-320.

computationally simulated crowding on each system, and uses these results to draw some conclusions about how crowding may influence these specific viruses or viral assembly generically.

4.2 Methods

4.2.1 Capsid simulation method

The Schwartz Lab has previously developed a rules-based discrete event stochastic simulator called Discrete Event Simulator of Self-Assembly (DESSA) (33) to model the process of capsid assembly from individual subunit building blocks through individual association and dissociation events into completed capsids. Simulated assembly is governed by simple biochemical rule sets specifying the geometries of the subunits, three-dimensional positioning of binding sites, and the specificities and on- and off-rates of binding events between binding sites. DESSA samples among all possible bond formation (association) and breaking (dissociation) events at each step in the simulation using a variant of the stochastic simulation algorithm (49,50). More details of the DESSA simulator and its application are provided in Section 1.2.3. I assumed kinetic rates in the uncrowded case to be those derived from a previous study of these viruses using a numerical optimization scheme to fit parameters to minimize the root mean-square deviation between *in vitro* experimental static light scattering data and light scattering curves generated based upon the DESSA simulation output (96). This parameter estimation method is described in more detail in Sections 2.2.1 and 2.2.2.

4.2.2 Modeling the crowding effect

4.2.2.1 The 3DSOLM Crowding Simulator. To simulate potential effects of crowding on capsid assembly, I used a strategy previously developed to quickly estimate corrections to equilibrium constants to account for crowding in complex assembly models. This method was intended to

address the problem that simplified stochastic models used in large-scale assembly simulations cannot model crowding effects, while the more realistic explicit particle models that can represent crowding are too slow to handle the large numbers of particles and simulation trajectories required for these studies. The method first uses off-lattice particle models implemented with Greens Function Reaction Dynamics (GFRD) (104) to test effects of varying levels of a homogeneous non-specific crowding agent on a generic homodimerization test system (85). GFRD relies on the observation that if a particle is found at some position $x = 0$ at time $t = 0$ then its position at a later time t', provided it has not collided with any other particles in the intervening time, can be treated as a Gaussian random variable representing the probability density of possible positions to which it may have diffused in time t':

$$P(x,t') = \sqrt{\frac{1}{4\pi Dt'}} e^{-x^2/4Dt'} \qquad (4.1)$$

, where D is the diffusion coefficient from the Stokes-Einstein diffusion equation $D = \frac{kT}{6\pi\eta r}$. By noting that it is extremely unlikely for a Gaussian random variable to be more than a small number of standard deviations from its mean, one can treat each particle over a short time scale as if it were defined by a sphere of possible positions centered on the mean of the Gaussian and with radius proportional to its standard deviation. This is defined as the diffusion limit sphere of the particle, and the radius of the sphere is R_{diff}. One can then perform a fast simulation of particle diffusion by treating it as a discrete event process that tracks growth of diffusion limit spheres according to the three dimensional random walk model $R_{diff(3D)}(\Delta t) = 3\sqrt{6D\Delta t}$ for given time interval (Δt) and diffusion coefficient of a particle (D), and only evaluates particle position when the diffusion limit spheres of two particles come into contact. At that time, new positions

are sampled for the two particles within their respective diffusion limit spheres and their radii are reset to zero.

At the start of a simulation, one computes the time at which each pair of particles' diffusion limit spheres will collide. This is first done by calculating the distance between two particles, say A and B:

$$d_{AB} = R_{diff,A}(t'-t_A) + R_{diff,B}(t'-t_B) = 3\sqrt{6D(t'-t_A)} + 3\sqrt{6D(t'-t_B)}\,, \qquad (4.2)$$

where t_A and t_B are the times at which the most recent positions of particles A and B were updated. Calculating radius of the diffusion limit sphere varies between monomers, dimers and inert particles in the simulation space:

$$R_{diff(3d)monomer} = 3\sqrt{6\Delta t\frac{kT}{6\pi\eta r_{monomer}}} = 3\sqrt{\Delta t\frac{kT}{\pi\eta r_{monomer}}} = C\sqrt{\Delta t} \qquad (4.3)$$

$$R_{diff(3d)dimer} = 3\sqrt{6\Delta t\frac{kT}{6\pi\eta r_{dimer}}} = 3\sqrt{\Delta t\frac{kT}{\pi\eta\alpha^{1/3} r_{monomer}}} = C\frac{\sqrt{\Delta t}}{\alpha^{1/6}} \qquad (4.4)$$

$$R_{diff(3D)inertparticle} = 3\sqrt{6\Delta t\frac{kT}{6\pi\eta r_{inertparticle}}} = 3\sqrt{\Delta t\frac{kT}{\pi\eta\beta^{1/3} r_{monomer}}} = C\frac{\sqrt{\Delta t}}{\beta^{1/6}} \qquad (4.5)$$

Here I define $C = 3\sqrt{\frac{kT}{\pi\eta r_{monomer}}}$, α is the volume ratio of dimer to monomer and β is the another volume ratio of inert particle to reactant monomer. Plugging these values into the equation for d_{AB}, one can determine:

$$\frac{d_{AB}}{C} = \frac{\sqrt{t'-t_A}}{n_A^{1/6}} + \frac{\sqrt{t'-t_B}}{n_B^{1/6}}, \qquad (4.6)$$

where $n_A = n_B = 1$ if A and B are monomers, $n_A = n_B = \alpha$ if A and B are dimers and $n_A = n_B = \beta$ if A and B are crowding agents particles. Therefore, one can calculate the collision time of two diffusion limit spheres as the solution to the above equation:

$$t' = \left(\frac{-d_1 + d_2\sqrt{1 + d_3 d_2^2 - d_1^2 d_3}}{(d_2 + d_1)(d_2 - d_1)} \right)^2 (d_{AB}/C)^2 + t_A \text{ if } d_1 \neq d_2 \text{ or } n_A \neq n_B \qquad (4.7)$$

$$t' = \frac{(d_{AB}/C)^2 n^{1/3}}{4} + \frac{t_A + t_B}{2} + \frac{(t_A - t_B)^2}{4(d_{AB}/C)^2 n^{1/3}} \quad \text{if } n_A = n_B = n, \qquad (4.8)$$

where $d_1 = n_A^{-1/6}, d_2 = n_B^{-1/6}, d_3 = \frac{t_B - t_A}{(d_{AB}/C)^2}$.

Each step of the simulation consists of finding two colliding diffusion limit spheres, determining the position of the particles based upon their position distributions. Upon such collisions, one tests if the spheres are within a predefined distance tolerance and, if so, one treats their interaction as a collision and allows for possible dimerization reactions with probability specified by a binding probability parameter *B*. If a binding event occurs, one replaces the particles with a dimer centered at the average of their two positions.

One simultaneously allows for the possibility of any dimer breaking into two monomers. A potential dissociation reaction is associated with each dimer, treated as an exponential decay process with mean time defined by a simulation parameter *M*. After each collision or reaction event, one begins growing new diffusion limit spheres around the affected particles at size zero and calculates the times to their next collision events, adding these into an event queue for the full simulation. The event queue is then queried to return the next event, collision of two diffusion limit spheres or dissociation of a dimer, to occur in the simulation.

4.2.2.2 Regression Model of Crowding Effects. The output of the crowding simulations can be converted into a simulated K_{eq} value for the reversible reaction of two monomers binding to form a dimer:

$$K_{eq} = \frac{D_{eq} * V}{(M_0 - 2D_{eq})^2} [molecules^{-1} m^3], \quad (4.9)$$

where D_{eq} is the concentration of dimers at the quasi-equilibrium state, M_0 is the initial number of monomers and V is the total volume of the simulation space. After running crowding simulations for a range of tunable parameters, the method then learns a regression model of this equilibrium constant as a function of the simulation parameters that can accurately predict assembly behavior at untested parameter values (86). The computationally trivial regression model then becomes a surrogate for the computationally costly particle models in assembly simulation trajectories. The five tunable parameters for this model are the volume ratio of dimer to monomer (α); the total concentration of molecules (crowding agents plus capsomers) present in solution (C); the binding probability upon collision between two reactants (B); the mean time (inverse rate) of the dissociation reaction (M); and the diffusion coefficient (D). In prior work, the following regression model is fit for this dimerization test system:

$$K_{eq} = 10^{-19}[\frac{B}{.7}][\frac{M}{2.9*10^{-5}s}][\frac{D}{4.63*10^{-11}m^2 s^{-1}}][-.0094 + .0874C + .0148\alpha - .173C\alpha - .0334C^2 - .0059\alpha^2 + .1314C\alpha^2 + .018C^2\alpha + .0842C^3 - .0017\alpha^3 - .0468C\alpha^3 + .0272C^2\alpha^2 - .1962C^3\alpha + .2298C^4 + .0017\alpha^4 + .0069C\alpha^4 - .0187C^2\alpha^3 + .127C^3\alpha^2 - .35C^4\alpha + .4148C^5 - .0003\alpha^5] \quad (4.10)$$

The degree of the model with respect to α and C was fit empirically to minimize root mean square deviation between training simulations and the predictions of the regression model as assessed by leave-one-out cross-validation.

In the present work, I extend this model to estimate the kinetic effects of increasing macromolecular crowding on the dimer system using input parameter sets specific to each virus. The regression model is then used to adjust kinetic rate constants learned for the *in vitro* virus assembly systems learned in previous work in the lab (96) in order to estimate rate constants for each virus at varying levels of non-specific macromolecular crowding. To derive α values for each virus, I used assembly prediction models generated from *PDBePISA* (105), which calculates the stable assemblies possible based upon Protein Data Bank (PDB) files of capsid monomers. I applied *PDBePISA* to estimate excluded volumes of assembly subunits – pentamers for HPV and dimers for HBV and CCMV – as well as dimers of these basic subunits for each simulation using the Voss Volume Voxelator (106). Using this method, I was able to calculate the following dimer:monomer ratios for our corresponding viruses: 2.004:1 for CCMV, 2.005:1 for HBV and 2.000:1 for HPV. I also used volumes of single subunits (16188 $Å^3$ for CCMV, 28,848 $Å^3$ for HBV and 212,075 $Å^3$ for HPV) and the molar concentrations of capsid subunits present in the *in vitro* experiments to estimate the percentage of solution volume consisting of capsid proteins in the true *in vitro* systems. This calculation gives us percentages of .03% for CCMV and HBV and .01% for HPV. Thus, the amount of experimental volume consisting of capsomers is nearly negligible in the true *in vitro* systems, allowing us to treat them as effectively uncrowded. To account for that baseline crowding level in our regression models, I treated the total concentration *C* of molecules in the crowding simulations to be the coat subunit volume calculated above plus a simulated crowding agent volume. Due to the extreme difficulty of accurately predicting crowding effects at any specific crowding level, I examined a broad range of simulated crowding, from 0 to 45% in increments of 5%. Binding probability *B* and mean dissociation time *M* were inferred from the bond association and dissociation rates previously

learned for each virus by the parameter estimation algorithm (96). *B* is scaled from a base value of .7 dependent on if the bond forming time is faster or slower than the average. *M* is the inverse of the best-fit dissociation rates learned from the previous study (96) for each binding site event. The diffusion coefficient $D = 4.63*10^{-11} m^2 s^{-1}$ is calculated from the Stokes-Einstein diffusion equation as described in previous work (86). I then used the derived crowding regression models to compute corrected rates for forward and reverse binding at each binding site in each capsid system, using the parameters learned from the *in vitro* data (96) as representative of assembly at negligible (*C*=0%) total crowding.

To accomplish this, I calculated equilibrium constants for crowding agents ranging from 0% to 45% of simulation space for each virus using the regression-derived formula above. To convert those K_{eq} estimates into kinetic rates, explicit particle crowding simulations for each crowding level were run, which were assumed to be equivalent for each virus, to model a diffusion-limited binding mechanism. From the estimated K_{eq} and k_+ values, I calculated the corresponding k_- values by the relation $k_- = k_+ / K_{eq}$.

Table 4.1 shows relative crowding corrections for equilibrium and rate constants for each virus at crowding levels from 0% to 45%. By crowding level, I refer through the remainder of the chapter to the percentage of nonspecific crowding (0% and 45%) rather than the total crowding to avoid confusion due to the slight differences in volume contributed by the viral assembly subunits themselves. True total crowding levels will be between one and three-hundredths of a percent higher than these stated crowding levels, depending upon the virus. The table shows a general trend toward increased Keq with an increasing crowding level, concurrent with simultaneous decreases in both k+ and k- with increasing crowding. This effect is similar to

that observed in prior uses of these crowding models (85,86). The corrections in Table 4.1 provide scaling factors needed to adjust the previous *in vitro* best-fit kinetic rates (96) to more accurately model increasing crowding levels for each virus. The resulting crowding-corrected k_+ and k_- values for each binding site , provided in Table 4.2, then served as the input rates for stochastic simulations of capsid assembly at each crowding level via the DESSA assembly simulator.

Table 4.1. Crowding-corrections for equibilibrium constants, on rates, and off rates for HPV, CCMV and HBV generated by the GFRD simulator and regression model. The corrections calculated here are applied to the best-fit *in vitro* parameters of our capsid assembly simulator to reflect increasingly crowded assembly conditions.

	Crowding Level	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%
HPV	$K_{eq}(10^{-24}mol^{-1}m^2)$	1	1.26	1.39	1.54	1.87	2.54	3.88	6.46	11.9	26.8
	k_+ (10^{-16} $mol^{-1}m^2s^{-1}$)	1	.683	.474	.340	.252	.193	.147	.106	.069	.041
	k_- (10^{-8} s^{-1})	1	.543	.341	.221	.135	.076	.038	.016	.006	.002
CCMV	$K_{eq}(10^{-24}mol^{-1}m^2)$	1	1.32	1.45	1.65	1.96	2.68	4.04	6.60	12.0	27.1
	k_+ (10^{-16} $mol^{-1}m^2s^{-1}$)	1	.683	.474	.340	.252	.193	.147	.106	.069	.041
	k_- (10^{-8} s^{-1})	1	.517	.326	.206	.128	.072	.036	.016	.006	.002
HBV	$K_{eq}(10^{-24}mol^{-1}m^2)$	1	1.34	1.47	1.69	1.98	2.71	4.06	6.63	12.0	27.1
	k_+ (10^{-16} $mol^{-1}m^2s^{-1}$)	1	.683	.474	.340	.252	.193	.147	.106	.069	.041
	k_- (10^{-8} s^{-1})	1	.508	.322	.201	.127	.071	.036	.016	.006	.002

Table 4.2. Best-fit association and dissociation rates for each binding site of each capsid model estimated by correcting rates inferred in vitro to reflect simulated crowding levels from 0% and 45%. These rate parameters act as inputs for the capsid assembly simulations.

	Crowding Level	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%
HPV	*A*+ ($10^3M^{-1}s^{-1}$)	1.4	.9562	.6636	.4760	.3528	.2702	.2058	.1484	.0966	.0574
	A- (s^{-1})	.12	.0652	.0409	.0032	.0162	.0091	.0046	.0019	.0007	.0002
	B+ ($10^3M^{-1}s^{-1}$)	1.4	.9562	.6636	.4760	.3528	.2702	.2058	.1484	.0966	.0574
	B- (s^{-1})	.11	.0597	.0375	.0243	.0149	.0084	.0042	.0018	.0007	.0002
	C+ ($10^3M^{-1}s^{-1}$)	1.4	.9562	.6636	.4760	.3528	.2702	.2058	.1484	.0966	.0574
	C- (s^{-1})	.12	.0652	.0409	.0032	.0162	.0091	.0046	.0019	.0007	.0002
	D+ ($10^3M^{-1}s^{-1}$)	1.4	.9562	.6636	.4760	.3528	.2702	.2058	.1484	.0966	.0574
	D- (s^{-1})	.13	.0706	.0443	.0287	.0176	.0099	.0049	.0021	.0008	.0003
CCMV	*A*+ ($10^6M^{-1}s^{-1}$)	1.2	.8196	.5688	.4080	.3024	.2316	.1764	.1272	.0828	.0492
	A- (10^4s^{-1})	3.0	1.551	.9780	.6180	.3840	.2160	.1080	.0480	.0180	.0060
	B+ ($10^6M^{-1}s^{-1}$)	1.2	.8196	.5688	.4080	.3024	.2316	.1764	.1272	.0828	.0492
	B- (10^4s^{-1})	3.0	1.551	.9780	.6180	.3840	.2160	.1080	.0480	.0180	.0060
	C+ ($10^6M^{-1}s^{-1}$)	1.2	.8196	.5688	.4080	.3024	.2316	.1764	.1272	.0828	.0492
	C- (10^4s^{-1})	3.9	2.016	1.271	.8034	.4992	.2808	.1404	.0624	.0234	.0078
HBV	*A*+ ($10^6M^{-1}s^{-1}$)	1.4	.9562	.6636	.4760	.3528	.2702	.2058	.1484	.0966	.0574
	A- (10^5s^{-1})	1.2	.6096	.3864	.2412	.1524	.0852	.0432	.0192	.0072	.0024
	B+ ($10^6M^{-1}s^{-1}$)	1.4	.9562	.6636	.4760	.3528	.2702	.2058	.1484	.0966	.0574
	B- (10^5s^{-1})	1.4	.7112	.4508	.2814	.1778	.0994	.0504	.0224	.0084	.0028
	C+ ($10^6M^{-1}s^{-1}$)	1.4	.9562	.6636	.4760	.3528	.2702	.2058	.1484	.0966	.0574
	C- (10^5s^{-1})	1.4	.7112	.4508	.2814	.1778	.0994	.0504	.0224	.0084	.0028
	D+ ($10^6M^{-1}s^{-1}$)	1.4	.9562	.6636	.4760	.3528	.2702	.2058	.1484	.0966	.0574
	D- (10^5s^{-1})	1.2	.6096	.3864	.2412	.1524	.0852	.0432	.0192	.0072	.0024

For each virus, I produced crowding-corrected parameter files for crowding agent levels from 0% to 45% in increments of 5%, with rate parameters determined relative to the best-fit *in vitro* parameters for all three viruses, presumed to represent 0% nonspecific molecular crowding. I then ran 100 simulation trajectories for each virus at each crowding level, to allow for adequate sampling given the stochasticity of the simulator. For each simulation, I followed the same protocols as previous work in the lab (96). For HBV and CCMV, each simulation was begun using enough initial free subunits to generate five complete capsids per simulation: 450 subunits for CCMV and 600 subunits for HBV. For HPV, to ensure in the 0% crowding agent case a greater likelihood of producing at least one completed capsid, each simulation was begun using enough subunits to generate 10 complete capsids per simulation: 720 subunits total. Each simulation ends when all capsids have been formed, there are no events left in the simulation queue, or a predetermined simulation time limit empirically determined to allow simulations to go to pseudoequilibrium is reached. For these simulations, the time limit for CCMV and HBV was set to 1000 s, whereas the time limit for HPV was set to 150,000 s because of the far greater time required to assemble an HPV capsid *in vitro* and *in silico*. As in previous chapters, I applied a variety of data analysis techniques previously described in section 2.2.4 to study simulation trajectories individually and in aggregate, including simulated light scattering curves, binding frequency tables, mass fraction plots and movies showing detailed assembly pathways.

4.3 Results

4.3.1 Effects of crowding on bulk kinetics of simulated assembly

I ran replicates of assembly simulations for each of the three viruses and 10 crowding levels. Figure 4.1 shows simulated light scattering curves for CCMV, HBV, and HPV at each crowding concentration, averaged over 100 trajectories per crowding level.

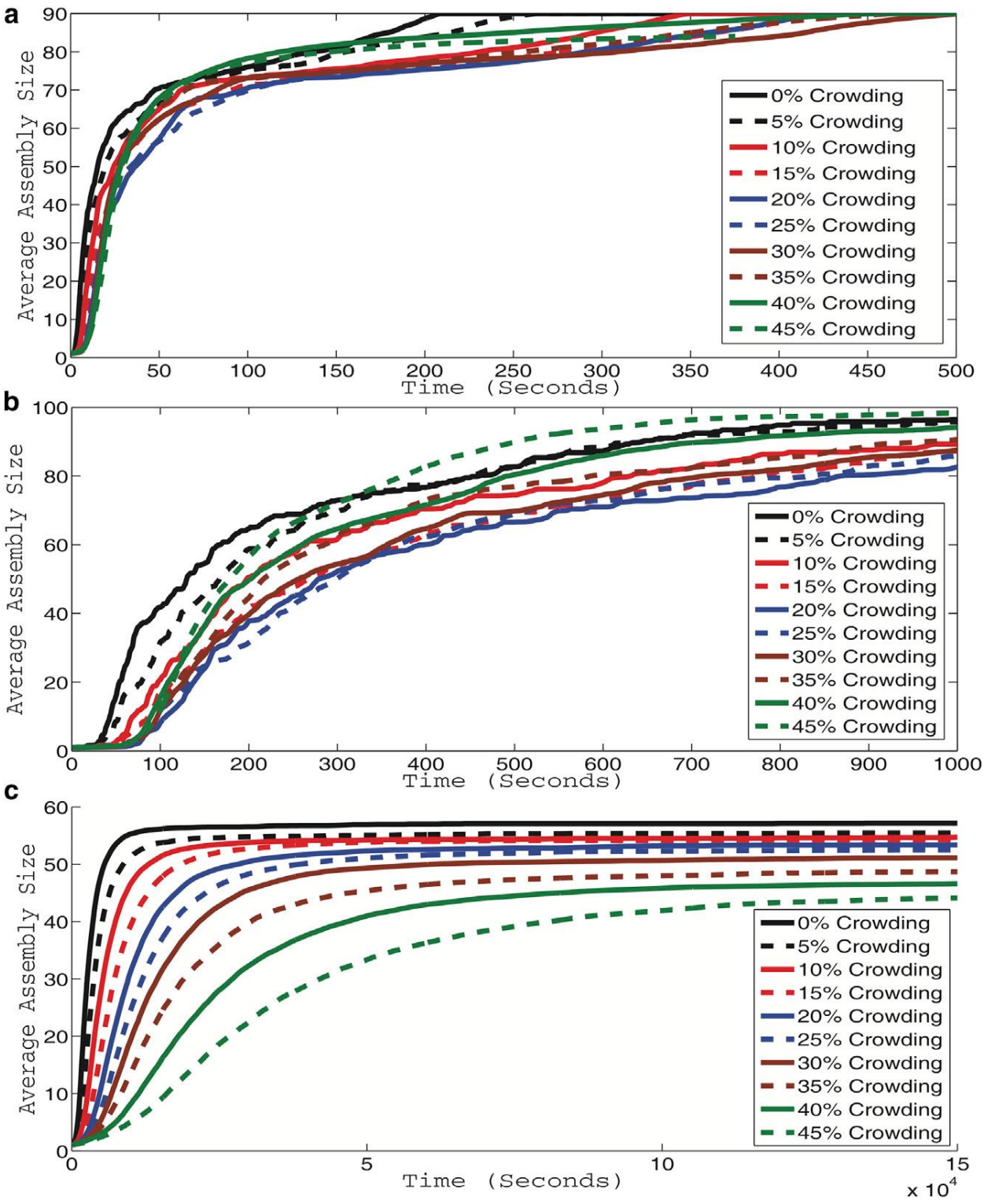

Figure 4.1. Simulated light scattering curves for CCMV (a), HBV (b), and HPV (c). Each curve represents an average simulated light scattering over 100 simulation trajectories. Curves are shown for levels of non-specific crowding agents from 0% to 45% of simulation solution volume in increments of 5%.

Figure 4.1(a) shows results for CCMV. At low crowding levels, the figure shows a pattern of decreasing speed and yield of assembly, with the assembly rate reaching a minimum at 25% crowding. The effect reverses at higher crowding levels, with the assembly rate at 35% crowding approaching that of the uncrowded system, and with 40% and 45% crowding yielding faster assembly at intermediate time points of the simulation. All trajectories go to equivalent levels of completion eventually, although with varying kinetics.

Figure 4.1(b) shows curves for HBV, which show qualitatively similar behavior to those for CCMV. HBV also shows a pattern of decreasing speed and quantity of assembly at low crowding levels, again reaching a minimum at 25% crowding, but increased assembly with respect to both speed and yield as crowding levels continue to increase. Crowding levels above 30% begin to approach the assembly rate of the 0% crowding state. 45% crowding yields higher assembly rates than 0% crowding levels in the later stages of assembly. HBV yields a higher apparent variance in the final yield of completed capsids than does CCMV. With HBV, assembly yield initially drops along with assembly rate as crowding is introduced, with yields at 10–35% crowding well below those of the uncrowded case. Yield approaches that of the uncrowded system by 40% crowding and surpasses it at 45% crowding.

Figure 4.1(c) shows curves for HPV, which show strikingly different behavior than the CCMV or HBV simulations. HPV shows a monotonic decrease in both rate and yield of assembly with increasing crowding rates. The curves also show a much lower variance than did the HBV or CCMV curves, with the effects of increasing crowding clearly distinguishable from the noise in the individual averaged simulated light scattering curves.

I next looked in greater detail at the kinetic effects of crowding on the different viruses' assemblies by comparing the average time required for each virus to reach 50% completion and

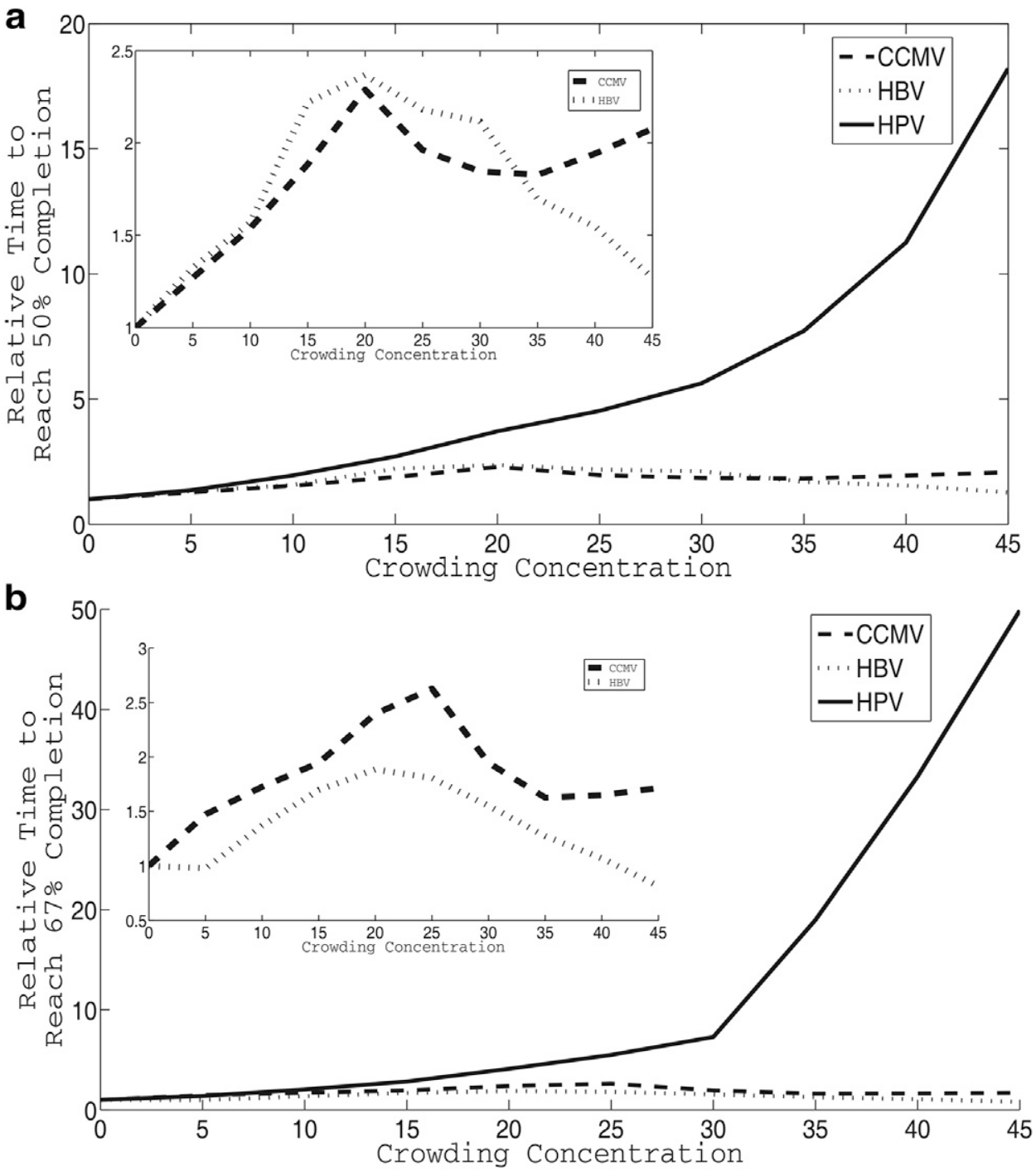

Figure 4.2. Average times required to reach 50% (a) and 67% (b) completed capsids for HPV, HBV, and CCMV models. All graphs are scaled to yield a starting time of 1 at 0% crowding. Subsequent crowding levels show the multiple of that initial time required to reach a given level of completion at a given crowding level. The inset in each figure represents a graph of just HBV and CCMV in order to highlight variations that are obscured by the large change in HPV completion times as a function of crowding level.

67% completion of the capsids in simulation. These measures are used to bypass a problem of mass averaging across assembly sizes in light scattering curves, namely that it tends to conflate productive assembly leading to complete capsids with unproductive off-pathway assembly leading to kinetically trapped intermediates. Figure 4.2 plots these times to partial completion for all crowding levels examined for the three viruses. Figures 4.2, (a) and (b), show a dramatic difference again between how the HPV model reacts to increased crowding compared to the CCMV and HBV models, with the CCMV and HBV times largely stable across crowding levels but the HPV times increasing greatly with higher crowding. Because the large change in HPV times makes it difficult to appreciate variations in CCMV or HBV times on a common scale, insets are added to the figures showing results solely for CCMV and HBV. The insets show an increase in time required for both viruses at low crowding levels, peaking between 20% and 25% crowding. Both viruses show about a 125% increase in time to 50% completion and a 75% increase for HBV in time and a 150% increase for CCMV in time to 67% completion. At higher crowding levels, CCMV shows a reduction in relative time to a local minimum at 35% crowding for both the 50% (a) and 67% (b) completion graphs. The average times then slightly increase again at the highest crowding levels. The highest crowding levels for CCMV are slower to reach 50% and 67% completion than the 0% level, although the higher levels eventually yield faster kinetics to reach 80% completion. For HBV, there is no local minimum at 35% as the relative time decreases at all crowding levels above 20%. 45% crowding approaches the 0% crowding case in 50% completion time (Figure 4.2(a)) and there is a reduction in relative time for the highest crowding level for the 67% completion time (Figure 4.2(b)). The figure implies that crowding acts earlier in HBV than CCMV assembly to accelerate the process.

4.3.2 Crowding effects on individual assembly trajectories

I next examined individual assembly trajectories to gain insight into the mechanisms by which crowding enhances bulk assembly of the HBV and CCMV models and suppresses bulk assembly of the HPV model. To this end, I produced plots of mass fractions of each assembly size over time for each virus at each crowding level. Although space would not allow me to show such figures for all simulation trajectories, I can provide an illustrative sample in Figure 4.3. Figures 4.3, (a), (d), and (g), show mass fraction plots for the three viruses with negligible crowding. The short peaks present in the HBV and CCMV plots (Figures 4.3, (a) and (d)), correspond to nucleation events followed by the rapid production of finished capsids, a feature absent from the HPV plot (Figure 4.3(g)). Both HBV and CCMV also maintain pools of trimers-of-dimers, which previous studies found in uncrowded simulations to be an important participant in the assembly pathways for these viruses (96). No large pools of assemblies are consistently present for either HBV or CCMV, aside from monomers, complete capsids, and this trimer-of-dimers intermediate. HPV, by contrast, shows a growing pool of partially assembled structures of varying sizes that form and then persist throughout the simulation. This latter pattern is indicative of kinetic trapping, in which many partial capsids form simultaneously and deplete free monomers to such a degree, that none can assemble to completion. This kinetic trapping has been previously observed in many capsid assembly models (33,96,107-112).

In Figures 4.3, (b), (e), and (h), I examine the changes in individual simulation trajectories induced by moderate (20%) crowding levels. Figure 4.3(b), a trajectory of CCMV with 20% crowding, shows a qualitatively similar picture to that of CCMV without crowding: a growth process with clearly defined nucleation events each touching off growth of a single capsid. Quantitatively, however, the process is slowed in both the nucleation and growth phases.

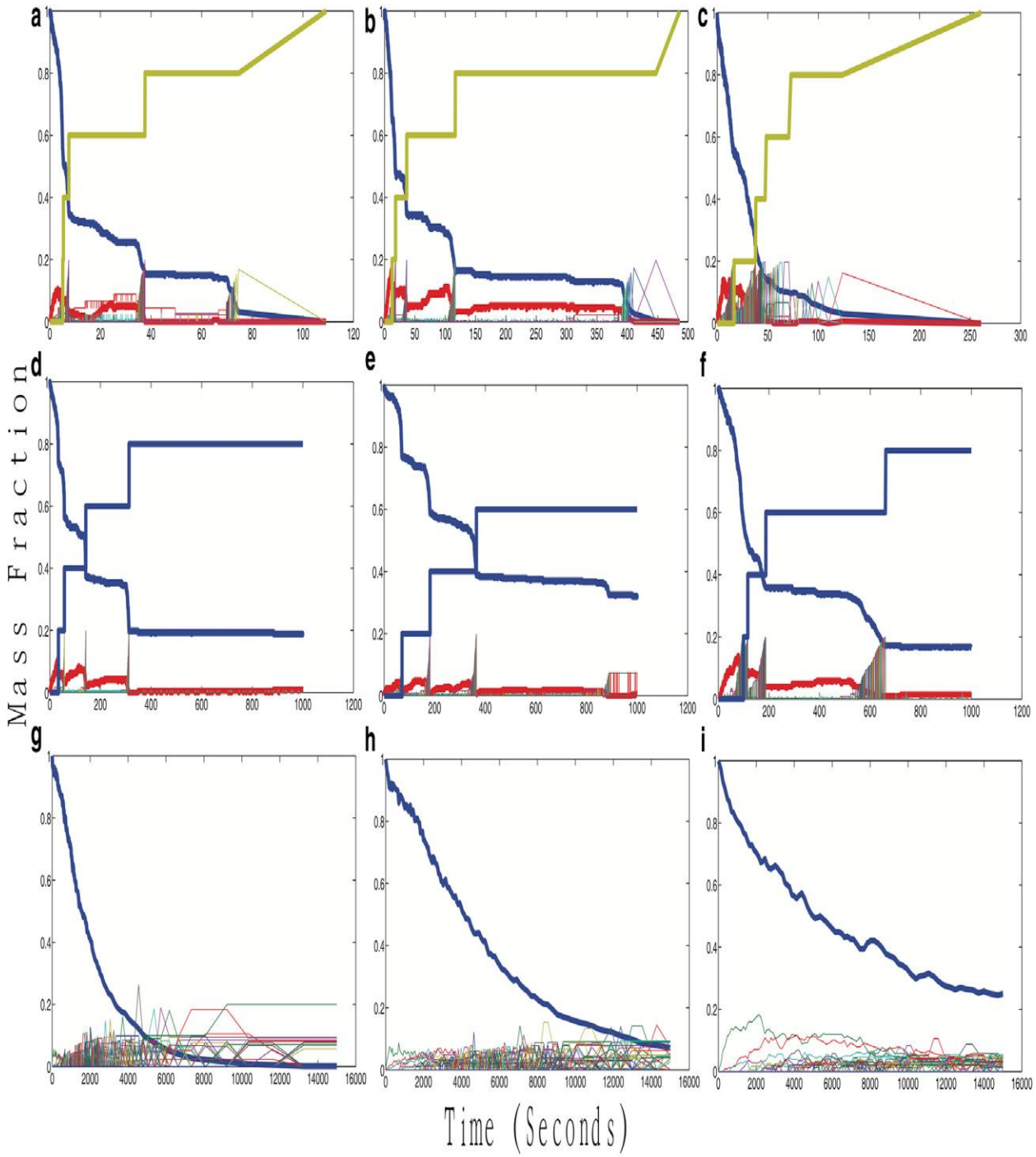

Figure 4.3. Mass Fraction Plots for CCMV at crowding levels of 0% (a), 20% (b) and 40% (c), HBV at crowding levels 0% (d), 20% (e) and 40% (f), and HPV at crowding levels of 0% (g), 20% (h) and 40% (i). Each plot shows a mass fraction for each size of intermediate versus time. The large number of intermediates makes it impractical to provide a full color key and necessitates repeating colors, however the most frequently observed intermediates are colored as follows: Monomers are blue, trimers are red, pentamers are purple. The plateauing yellow line for CCMV and blue line for HBV represent completed capsids.

Nucleation events are more widely spaced and the peaks corresponding to growth after nucleation are noticeably widened relative to the uncrowded case. The introduction of crowding also seems to produce a visible pool of 10-mers of dimers only evident for brief time periods in the uncrowded case; subsequent analysis showed this pool to be present for the majority of simulation runs at crowding levels between 10% and 25%. For levels lower than 10% and above 25%, 10-mers of dimers are still produced, although they are then added to larger assemblies at fast enough rates that no consistent pools are formed, except toward the end of simulation runs when assembly slows down due to fewer unbound capsid proteins available to take part in assembly reactions. Figure 4.3(e) shows that, for HBV, 20% crowding yields similar appearances for individual trajectories save for an increase in time to assemble and a lower likelihood of completing as many capsids by the end of the simulation time. There is no pool of 10 mers-of-dimers at 20% crowding for HBV, unlike CCMV. Figure 4.3(h) shows a highly distinct effect of simulated crowding on the HPV model. Kinetic trapping is visible both with and without crowding, with simulated capsomers largely absorbed into partially formed structures rather than complete capsids at both crowding levels. There is, however, a noticeable shift in the crowding simulations toward smaller partial intermediates.

Figures 4.3, (c), (f), and (i), show the effects on assembly trajectories as crowding is increased to 40%. CCMV shows a greatly increased assembly rate for the majority of potential capsids, to the point that the process no longer appears nucleation limited. Although the nucleation-like peaks are still present, they now overlap, indicating multiple capsids in their elongation phases simultaneously. The second, third, and fourth capsids assemble nearly simultaneously. The fifth, however, takes far longer than the original four to assemble, as the effects of reduced free subunits greatly slow bond formation. I note that this slow growth yields

an opportunity to observe step-by-step addition of each subassembly to the growing capsid. A combination of single dimer, dimer-of-dimers, and trimer-of-dimers additions all occur during the assembly process. A similar pattern can be seen with regard to HBV in Figure 4.3(f). At 40% crowding, assembly is greatly accelerated relative to lower crowding levels. Peaks corresponding to nucleation and subsequent growth of individual capsids remain clearly defined, unlike in the CCMV case, but nonetheless begin to run together. One feature of note in the 40% crowding case is the presence of a small pool of pentamers-of-dimers early in the assembly process before the rapid nucleation and assembly of the first four capsids. In each case, only four of the five potential capsids are produced. It is also notable that after the last capsid is formed, there are no assemblies present other than persistent populations of completed capsids, single dimers, dimers-of-dimers, and trimers-of-dimers, and the transient appearance of occasional tetramers-of-dimers. There is no construction of pentamers-of-dimers, which are often seen forming shortly before a nucleation step in HBV trajectories. For HPV assembly, the sample trajectory for 40% crowding in Figure 4.3(i) shows a qualitatively similar picture to that for 20% crowding in Figure 4.3(h). The system is unable to produce completed capsids, instead yielding a spectrum of partially built, kinetically trapped forms. There is a further noticeable shift toward smaller trapped species relative to the uncrowded case. Meaningful differences between 20% and 40% crowding are difficult to discern from single trajectories, although Figure 4.1(c) implies that the shift toward smaller trapped species continues as crowding levels increase.

To investigate individual assembly pathways more deeply, I constructed movies of single simulation trajectories for the non-zero crowding cases shown in Figure 4.3. These can be compared to the zero crowding movies described in Chapter Two. As in previous chapters, I have included figures showing individual important frames during a single assembly pathway

from monomer to completed capsid or in some cases the largest intermediate formed. Each figure will show two frames. In the case of CCMV and HBV, the frames will represent the moment of nucleation and the final completed capsid structure. In the case of HPV, the frames will highlight the less ordered nature of the intermediates and then the subsequent final structure assembled.

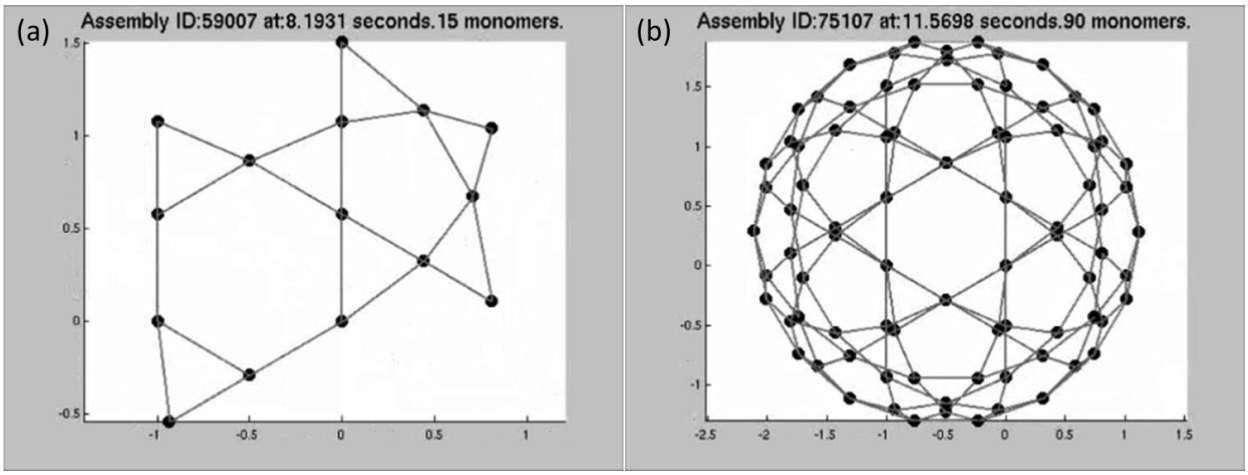

Figure 4.4. Individual frames of a movie following the same assembly trajectory of a CCMV capsid under 20% crowding as shown in Figure 4.3(b). Assembly information for each frame is listed in the header for each sub-image.

Figure 4.4 follows the assembly of a single CCMV capsid under 20 percent macromolecular crowding, as seen in Figure 4.3(b). While it is difficult to use the trajectory of a single capsid to describe the overall trend towards slower initial nucleation, I can examine changes in elongation phase following the nucleation event. In this case, a stable hexagon is formed at 8.19 seconds, after which a completed structure is assembled by 11.56 seconds. This is an increase in time required for elongation compared to the no crowding case described in Chapter 2 as well as a slight increase in the percentage of overall assembly time devoted to elongation: 29% compared to 24.8%. Overall, a similar usage of pathways is seen in both the no crowding and 20% crowding cases.

Figure 4.5 follows the assembly of a single CCMV capsid under 40% macromolecular crowding as seen in Figure 4.3(c). Here the effects of slowed on rates and increased kinetic trapping are very apparent in the assembly time devoted to the elongation phase, now over 53% of the assembly time. I will further note that the nucleus listed here is far smaller than most described in this work. In this simulation trajectory, a hexagon is formed in an intermediate with only nine subunits present. Very quickly following this, assembly begins to increase in pace, albeit somewhat diminished because of the slower rate of bond forming events. This hexagon structure can break apart during assembly simulations, so it is imperative that upon forming this structure, it survives long enough so that larger intermediates can be formed with it as a basis. This either requires fast on rates or slow off rates to either increase speed of assembly or increase stability of intermediates.

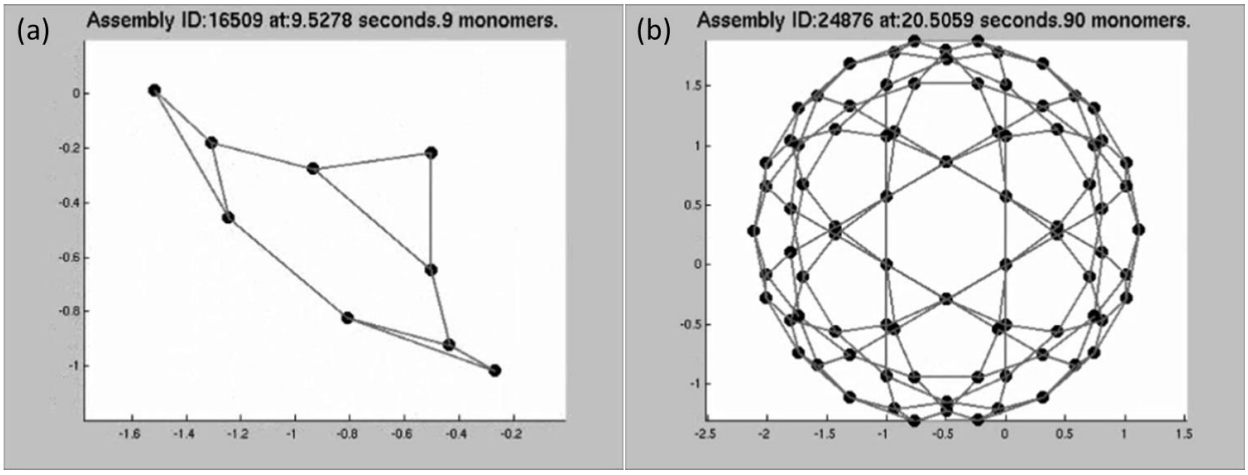

Figure 4.5. Individual frames of a movie following the same assembly trajectory of a CCMV capsid under 40% crowding as shown in Figure 4.3(c). Assembly information for each frame is listed in the header for each sub-image.

Figure 4.6 follows the assembly of a single HBV capsid under 20% macromolecular crowding as seen in Figure 4.3(e). HBV has similar proclivities in nucleation step; in this case, when two 10mers with stable pentagons bind together to produce a hexagon wedged between the pentagons, as seen in Figure 4.6(a). The time required for the elongation phase has again

increased to over 14 seconds compared 2 seconds in the assembly seen in Chapter 2. This is a change in percentage of assembly time from 5.3% to over 16%. While pathway selection is similar to the uncrowded case, there are fewer simulation events overall and rate of assembly has slowed down.

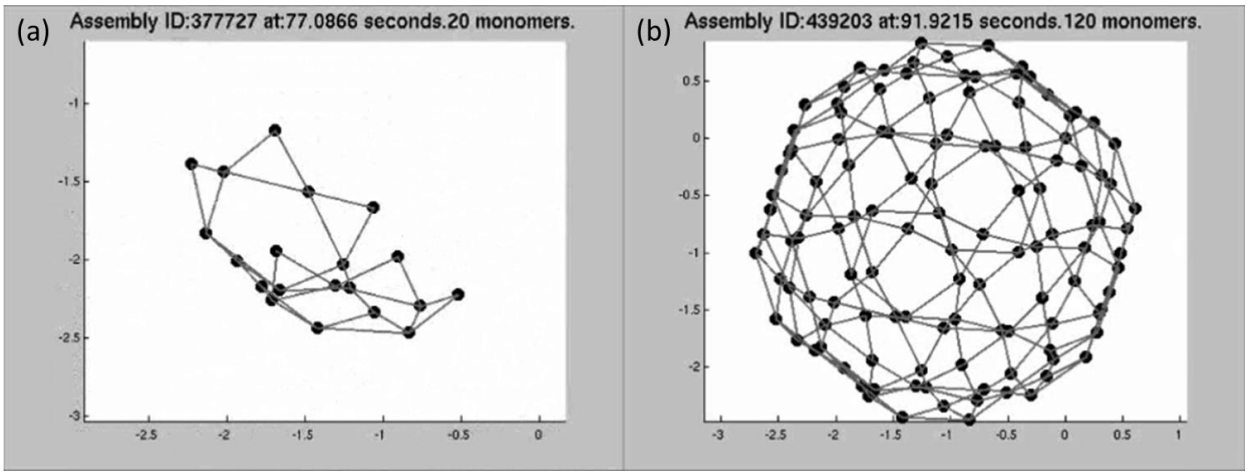

Figure 4.6. Individual frames of a movie following the same assembly trajectory of a HBV capsid under 20% crowding as shown in Figure 4.3(e). Assembly information for each frame is listed in the header for each sub-image.

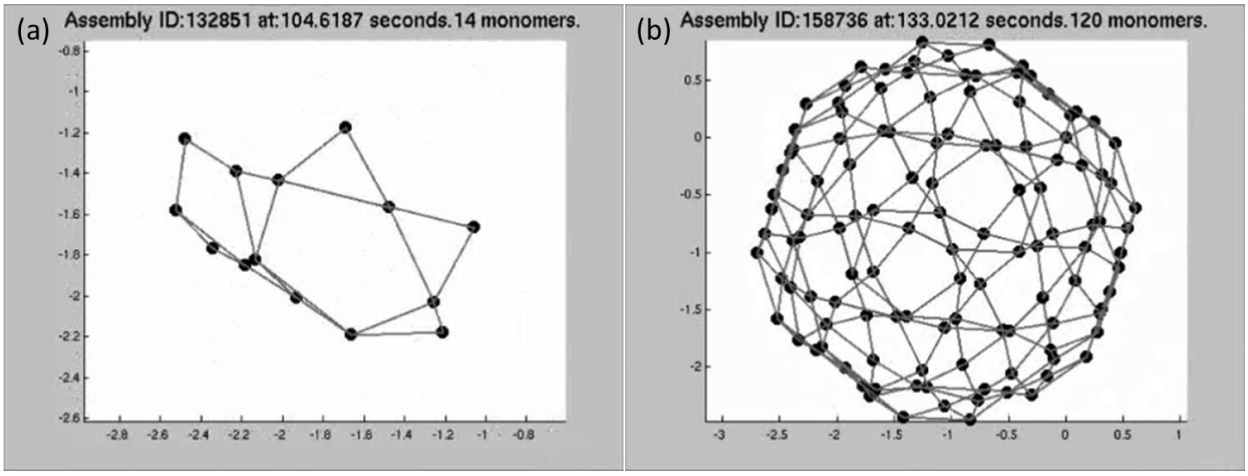

Figure 4.7. Individual frames of a movie following the same assembly trajectory of a HBV capsid under 40% crowding as shown in Figure 4.3(f). Assembly information for each frame is listed in the header for each sub-image.

Figure 4.7 then follows the assembly of a single HBV capsid under 40% macromolecular crowding as seen in Figure 4.3(f). Once again, the percent of assembly time devoted to the elongation phase increases to over 21%. Overall more completed capsids are assembled faster in

the 40% case compared to the 20% case. This suggests that despite the increase in time for individual elongation phases, nucleation must be sped up over all.

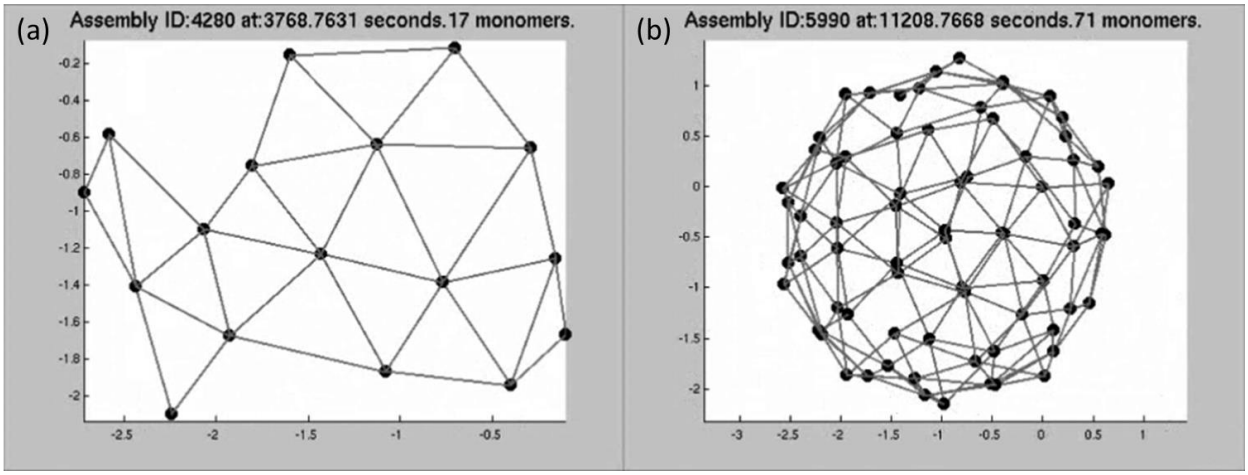

Figure 4.8. Individual frames of a movie following the same assembly trajectory of a HPV capsid under 20% crowding as shown in Figure 4.3(h). Assembly information for each frame is listed in the header for each sub-image.

Figure 4.8 follows attempted assembly of a single HPV capsid under 20% crowding, as shown in Figure 4.3(h). Despite enough subunits to potentially produce 10 completed capsid structures, the largest intermediate formed in this simulation was only a 71mer, whereas a full HPV capsid would have contained 72 subunits. There is no specific order in the pathway by which the larger intermediate is formed. Individual subunits add on to a growing lattice of triangles as the overall structure is built. Even at low or no crowding levels, the time to complete a capsid following the formation of an intermediate comparable to the size of a CCMV or HBV nucleation step is already over 50% of the overall assembly time.

This is only exacerbated in Figure 4.9 which follows the attempted assembly of a single HPV capsid under 40% crowding, as shown in Figure 4.3(i). Here the largest intermediate formed is only a 69mer and in Figure 4.9(a), it is even more evident the sporadic nature of the addition of individual subunits. Because of the decreased dissociation times, assembled structures are very stable and kinetic trapping is rampant. Because of this, 90% of the assembly

time is devoted to building from 34 subunits to 69 subunits, an incredibly slow process in contrast to nucleation-limited CCMV and HBV.

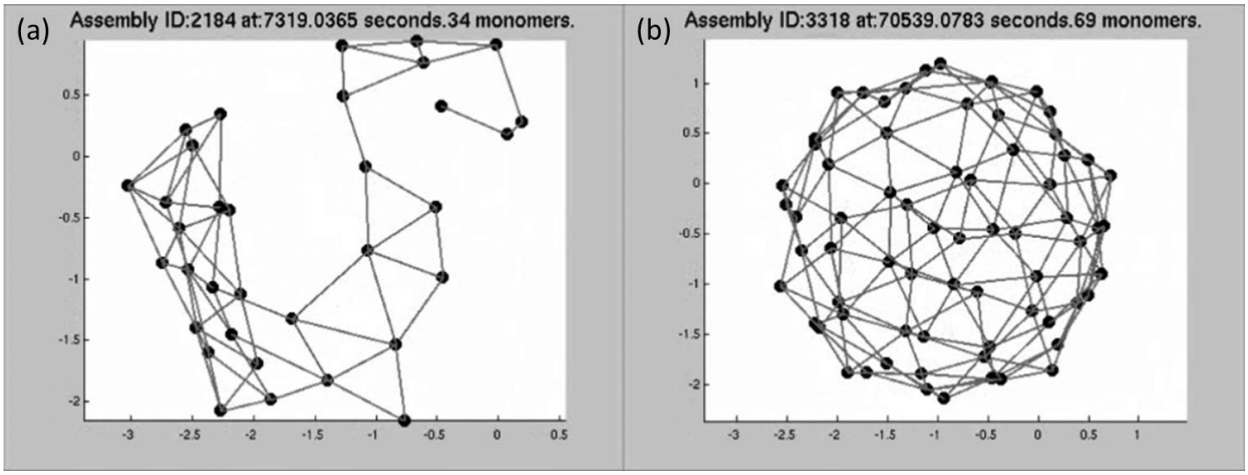

Figure 4.9. Individual frames of a movie following the same assembly trajectory of a HPV capsid under 40% crowding as shown in Figure 4.3(h). Assembly information for each frame is listed in the header for each sub-image.

4.3.3 Measuring average pathway usage across trajectories.

Another way to examine the changes seen in the assembly pathways for these viruses is to measure frequencies with which individual bonds are observed, averaged over many trajectories. Figure 4.10 shows a visualization of these bond frequency tables, showing relative frequencies with which different assembly sizes are used as reactants in producing any given larger assembly size (e.g., the frequency with which a dimer is a reactant to a reaction producing a pentamer). Figures 4.10, (a) and (b), show the progression in binding frequency tables for CCMV at negligible and 40% crowding, in each case averaged over 100 simulation trajectories. Both crowding levels yield similar bond usage patterns, with essentially the same combinations of reaction steps. In each case, assembly proceeds by addition of either subunit monomers (coat dimers), subunit trimers (coat trimers-of-dimers), or, in some early steps, subunit pentamers (coat pentamers-of-dimers). Most steps favor monomer addition with a low frequency of trimer addition, aside from a few conserved steps with high trimer frequency early in assembly.

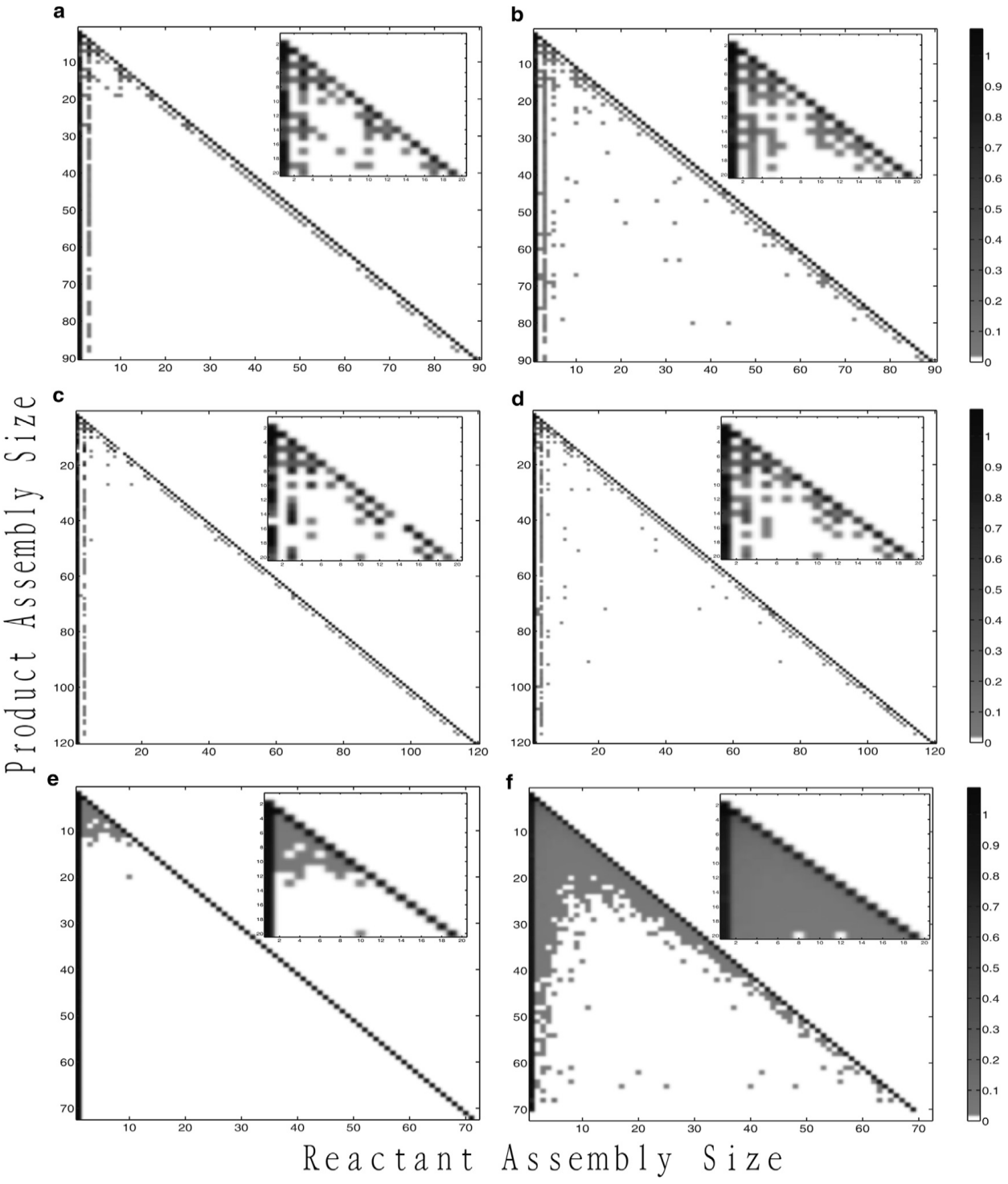

Figure 4.10. Binding frequency tables for CCMV at crowding levels of 0% (a) and 40% (b), HBV at crowding levels of 0% (c) and 40% (d), and HPV at crowding levels of 0% (e) and 40% (f). Each pixel shows the frequency with which a particular reactant size (x-axis) is used to produce a particular product size (y-axis). A scale bar relating shading to frequency appears on the right.

However, two changes are noted in the 40% crowding case. First, there is an overall trend toward an increased use of single dimers versus trimers-of-dimer or pentamers-of-dimer at 40% crowding. Second, there is an increased frequency of reactions involving larger intermediates in the 40% crowding case, potentially out of necessity as binding reactions become less kinetically favorable.

Figures 4.10, (c) and (d), show comparable plots for the HBV model. For HBV, as for CCMV, pathways are characterized predominantly by monomer addition, with a lesser rate of trimer addition. There are, however, a few conserved steps at which trimer addition is used preferentially. HBV shows only rare pentamer additions at a few early steps. As with CCMV, pathways are qualitatively similar across crowding levels, but again a slight trend occurs toward increased preference for assembly by single dimers versus other small oligomers and more frequent appearances of interactions between larger assemblies at 40% crowding. Figures 4.10, (e) and (f), again show a very different portrait for HPV. As found in previous work (30,32), HPV assembly is driven almost entirely by single capsomer additions. However, as crowding levels increase, the frequency of other binding events occurring also begins to increase. By the 40% crowding level every binding combination of sizes is observed at least infrequently among the lower assembly sizes. Thus, although crowding reduces the frequency of usage of minor assembly pathways for CCMV and HBV, it increases their use for HPV. I note that the final two rows of the HPV binding frequency table for 40% crowding are all white, showing that no binding events result in assemblies of >70 capsomers (HPV consists of 72 capsomers in the simulation model) and thus that the HPV model was unable to assemble any complete capsids at the 40% crowding level.

4.4 Discussion

This work is intended to take one step toward transitioning capsid assembly simulation away from models of *in vitro* environments toward more realistic representations of *in vivo* capsid assembly, with a specific focus on modeling the possible effects of macromolecular crowding on assembly. By using a previously developed approach for inferring the influence of crowding on association and dissociation rates of assembly reactions, I adjusted models of capsid assembly learned from *in vitro* data to better reflect how the coat subunits studied might behave in a crowded environment such as a living cell. Although the exact effects of crowding are notoriously hard to predict with accuracy, one can infer general trends by scanning across a range of possible simulated crowding levels. The resulting computational models made it possible to analyze detailed simulated assembly trajectories for three different icosahedral viruses—CCMV, HBV, and HPV—under levels of crowding between 0% and 45%, a range that should span the range from typical *in vitro* to plausible *in vivo* crowding levels.

The study revealed an important distinction between the effects of crowding on nucleation-limited capsid assembly and non-nucleation-limited capsid assembly. In particular, growth in the absence of a defined nucleation step generally was impeded by crowding, as the net acceleration of coat-coat binding led primarily to increased kinetic trapping and thus to a loss of productive assembly. Nucleation-limited growth, previously known to provide protection against kinetic trapping (30,33), appeared to provide a similar buffer against kinetic trapping that would otherwise be induced by crowding, allowing crowding to accelerate rather than inhibit productive assembly. One can speculate that the reason this effect can occur in the detailed assembly models used is that nucleation is not a single step; crowding-induced changes in both forward and reverse rates of intermediate steps leading to nucleation can simultaneously slow

each individual assembly step and yet accelerate the overall assembly rate. Furthermore, nucleation-limited assembly buffers against off-pathway growth, allowing this acceleration without an increase in off-pathway assembly over at least a broad range of crowding levels.

This enhancement of assembly through crowding operates over a limited range of crowding levels as the buffer effect will eventually break down when nucleation events occur in too rapid succession, which begins to be seen for CCMV at the highest crowding levels. Nucleation-limited growth allows crowding to work to promote effective growth, but a sufficiently strong crowding effect will itself block nucleation-limited growth. These effects are evident in a shift of both of the nucleation-limited capsids toward a single monomer-accretion pathway, evident in Figure 4.10, as well as in the overlapping nucleation/elongation peaks seen for these viruses in 40% crowding trajectories in Figure 4.3. These results suggest that crowding is neither inherently a benefit nor a disadvantage to capsids, but rather a complicated effect variable across biologically plausible parameter domains. One can conjecture that viruses would evolve to function effectively in the domains in which they operate in nature. Understanding these tradeoffs may be helpful in better designing *in vitro* systems for capsids or other complex self-assemblies as well as in the design of novel self-assembling nanotechnology that might take advantage of similar environmental factors.

A key question for this work is to understand the degree to which one can trust that results from models, whether computational or *in vitro*, accurately reflect what happens in the cell. These simulations suggest a mixed answer. The results suggest that, at least for nucleation-limited assembly processes, assembly mechanisms and pathways are well conserved over a broad range of crowding values likely to cover both *in vitro* and *in vivo* levels of crowding. This observation would suggest that one can, indeed, use conclusions from the *in vitro* system to

make predictions about pathways *in vivo*. This is a question that may have important practical consequences, such as for strategies for developing capsid assembly targeted antivirals (54–56). On the other hand, these models suggest significant quantitative differences in assembly rate and yield can be induced by crowding and that one thus cannot make accurate quantitative predictions about yield of a capsid system without adequately accounting for crowding in that system. This observation, too, may have important practical consequences. For example, one might expect that changes in usage of minor assembly pathways could substantially alter a virus's ability to resist an assembly-targeted drug.

The work presented here is intended to make one step forward in understanding how conditions in the cell might alter pathways for complex self-assemblies such as a viral capsid, with focus on the specific factor of molecular crowding, but it still falls far short of a real representation of viral capsid assembly *in vivo*. Designing computational models that better represent the conditions in which viruses normally assemble is essential to developing biologically accurate and predictive models of virus assembly *in vivo*. Furthermore, crowding effects themselves are complex and difficult to predict. I cannot claim that any particular parameter choices in the model will accurately measure the effect of crowding on the specific viral systems examined here. By exploring a range of values of crowding effects and the general trends across that range, one can, however, make general observations likely to prove useful despite imprecision in predicting crowding effects precisely. Nonetheless, improved crowding models or better empirical evidence from which to parameterize them for these specific systems would be valuable in making more specific and confident predictions. Furthermore, numerous other factors interact with growing capsids in the cell, including a diverse array of binding partners, chaperones, cytoskeletal structures, nucleic acid to be packaged or encapsidated in the

final capsid product, and likely other actors not yet known to us. Determining which of these are actually relevant to modulating assembly pathways and how they act, singly and in combination, will require extensive work, both experimentally and computationally.

Chapter 5: Analyzing Assembly Pathways under Combined RNA and Crowding Effects

5.1 Introduction

Learning how individual changes in assembly environment affect the pathways and rate by which capsids form is an important step; however, this alone does not fully model the *in vivo* capsid assembly process. What is truly needed is to examine how these individual effects, when combined, work to either inhibit or improve assembly. To date, I have solely examined two such environmental changes: macromolecular crowding and nucleic acid. While there are numerous other potential modifications that can be considered, seeing how combining these two modifications with known effect on capsid assembly is an important next step to developing biologically accurate capsid assembly models.

To accomplish this, I once again examine the combined RNA effects case detailed in Chapter 3 which models the assembly of a single CCMV capsid around the 3171 nucleotide strand of CCMV RNA1. I then apply the same crowding regression model described in Chapter 4 to calculate a series of kinetic rate modifications for crowding levels between 0 and 45 percent to assembly simulations of both a single hollow CCMV capsid and a single CCMV capsid under the combined RNA effects. The results described in this chapter are preliminary and quite surprising and suggest that, despite macromolecular crowding having a potentially positive effect on larger nucleation-limited assembly systems, when conducting simulations with only enough subunits to produce a single capsid, crowding has a decidedly negative effect on both the hollow and combined RNA effects cases. To attempt to reconstruct a positive crowding effect for the combined RNA effects case, I ran further simulations this time on the same monomer counts as the original crowding simulation work. This change in simulation size did reconstruct some of the previous complexity in crowding effect seen in Chapter 4.

5.2 Methods

5.2.1 Capsid simulation method

The Schwartz Lab has previously developed a rules-based discrete event stochastic simulator called Discrete Event Simulator of Self-Assembly (DESSA) (33) to model the process of capsid assembly from individual subunit building blocks through individual association and dissociation events into completed capsids. Simulated assembly is governed by simple biochemical rule sets specifying the geometries of the subunits, three-dimensional positioning of binding sites, and the specificities and on- and off-rates of binding events between binding sites. DESSA samples among all possible bond formation (association) and breaking (dissociation) events at each step in the simulation using a variant of the stochastic simulation algorithm (49,50). More details of the DESSA simulator and its application are provided in Section 1.2.3. Original best-fit rate parameters to *in vitro* light scattering data were found using the parameter estimation method described in more detail in Sections 2.2.1 and 2.2.2 and represent the uncrowded hollow capsid case of this study.

5.2.2 Applying the RNA Effects

I devised theoretical models for modifications to rate parameters dependent upon four individual effects of RNA on the free energy of the assembly system. Two of these were negative effects on assembly rate: RNA-RNA interaction and RNA compression, and two were positive effects: RNA-capsid protein interaction and increased capsid protein concentration. I then calculated rate modifications for any combination of these four effects, details of which are provided in Section 3.2.2. The hollow capsid case represents no RNA effects applied to the best-fit rate parameters for CCMV learned in Chapter 2. The combined effects case represents all four rate modifications being applied to the parameters for the hollow capsid case.

5.2.3 Applying Macromolecular Crowding Effects

To apply crowding effects to the hollow capsid and combined RNA effects cases, I once again used the regression model developed to predict equilibrium constant modifications under increasing levels of macromolecular crowding and based upon the 3DSOLM crowding simulator developed in the lab (85,86). Applying crowding effects was more complicated than just applying the previously learned corrections. The regression model is dependent upon the variable *C*, which is not just the concentration of crowding particles in the simulation, but instead the concentration of all binding and crowding molecules in the simulation space, represented by the percentage of volume in simulation space composed of capsid subunits and crowding agents. Therefore, in the far more concentrated capsid protein case examined in the combined RNA effects case, the values for C had to be increased by the still relatively small, but significantly different, value of .871 percent of simulation volume. In the previous crowding experiments, the capsid subunits only comprised .03 percent of simulation volume.

To add further complexity to the crowding simulations, I applied the results of Blum et al. (113) which estimated changes in diffusion coefficient based upon differences in crowding agents. This work calculated scaling factors that could be applied to diffusion rates based upon changes in shape of crowding particles as well as differently-sized lattices of crowding particles for a range of crowding levels that encompasses the levels discussed in this work. Technically, for each potential size, shape or style of crowding particle, the study calculated the ratio of diffusion for a spherical binding agent through a volume including the crowding particle of interest compared to diffusion for the same binding agent through water. For this study, I compared changes in diffusion for a spherical crowding particle as is seen in the crowding simulations done in the Schwartz Lab with a more complicated lattice of beams with square

cross-section. Modifications to rate of diffusion can be applied directly to the diffusion coefficient *D* in the crowding regression model. Changes to equilibrium constant calculated by the regression model are applied in the same manner as described in Section 4.2.2. The resulting on and off rate modifications are detailed in Table 5.1.

Table 5.1. Crowding-corrections for equibilibrium constants, on rates, and off rates for CCMV. The corrections calculated here are applied to the best-fit *in vitro* parameters of the capsid assembly simulator to reflect increasingly crowded assembly conditions.

	Crowding Level	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%
CCMV	$K_{eq}(10^{-24}mol^{-1}m^2)$	1	1.10	1.13	1.22	1.40	1.85	2.70	4.19	8.40	24.0
	k_+ (10^{-16} $mol^{-1}m^3s^{-1}$)	1	.596	.386	.263	.183	.133	.097	.067	.042	.024
	k_- (10^{-8} ms^{-1})	1	.542	.343	.215	.131	.072	.036	.016	.005	.001

5.2.4 Simulation Experiments and Data Analysis

I conducted 200 simulations at each crowding level for both the single hollow CCMV capsid case and the single CCMV capsid under combined RNA effects case. Only 10 simulation replicates were done for each crowding level for multiple CCMV capsids under combined RNA effects because of time limitations, though those simulations were more a proof of concept than complete analysis. For each CCMV model under each crowding level, I produced a simulated light scattering curve to measure changes in assembly rate.

5.3 Results

Figure 5.1 shows two panels of simulated light scattering curves of a single CCMV capsid assembling under increasing levels of macromolecular crowding. Figure 5.1(a) shows a single hollow capsid while Figure 5.1(b) shows a single CCMV capsid under combined RNA effects.

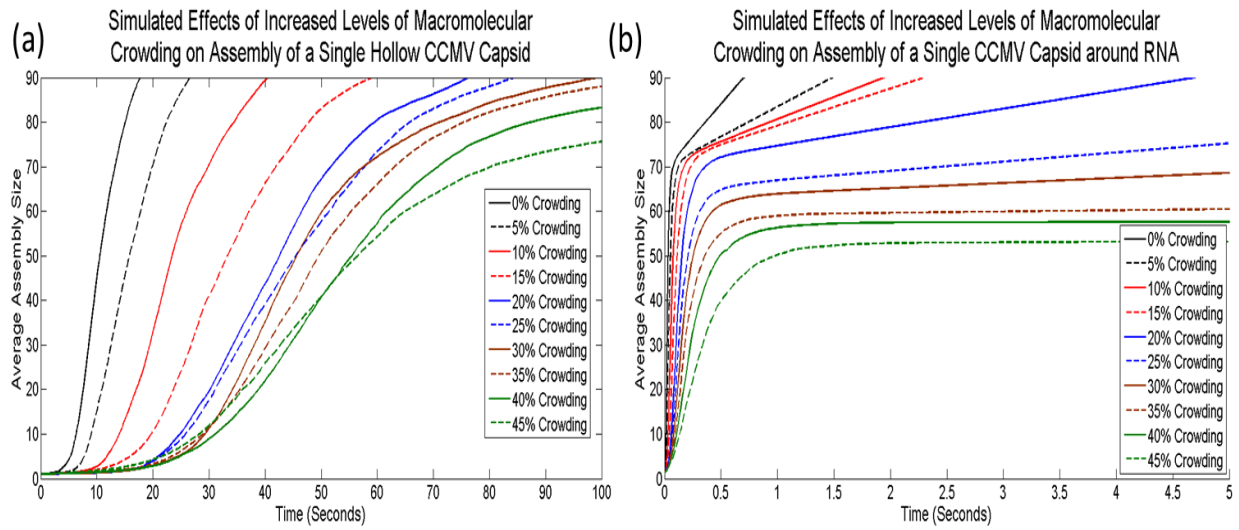

Figure 5.1. Simulated light scattering curves for (a) a single hollow CCMV capsid under increasing levels of macromolecular crowding and (b) a single CCMV capsid under combined RNA effects under increasing levels of macromolecular crowding.

Unlike the crowding experiments done in Chapter 4, the effects of increasing levels of macromolecular crowding are decidedly negative in both instances. Because high levels of crowding can have positive effects for CCMV under the exact same concentration conditions, one possible explanation for why crowding is so detrimental in both figures is because there are only enough capsid proteins present to form a single CCMV capsid in each simulation. While crowding can be beneficial in larger CCMV assembly simulations, kinetic trapping is too prevalent when dealing with such a small number of subunits where intermediates are exceedingly stable and the lack of excess capsid proteins makes assembly completion much more difficult. To test if this theory was correct, I ran the same RNA + crowding simulations with 450 subunits, the same number of subunits as the crowding experiments in Chapter 4. The results are seen in Figure 5.2 and, while still not showing as beneficial of an effect at high crowding levels, there is a return to a more complex relationship between increasing levels of crowding and assembly rate and yield. Here, any amount of crowding provided a decrease in total assembly yield compared to the 0% crowding case; however, low and medium crowding

levels do not show much distinction in assembly rate. Assembly yield does begin to decline significantly at 40% crowding, however.

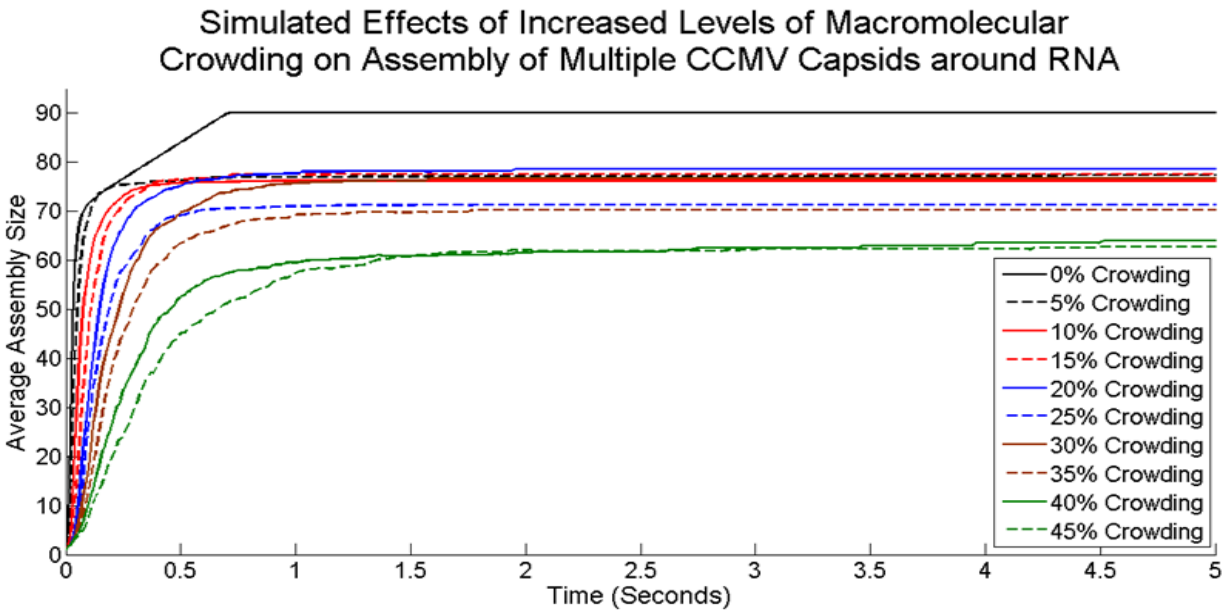

Figure 5.2. Simulated light scattering curves for multiple CCMV capsids under combined RNA effects under increasing levels of macromolecular crowding

5.4 Discussion

This is only a very preliminary attempt at understanding how these two important contributions to the *in vivo* assembly environment can further alter capsid assembly when combined. In fact, combining RNA and crowding effects raises the question of how prevalent macromolecular crowding would be in the extreme case examined in Chapter 3 where all capsid proteins are pre-localized to the direct vicinity of an RNA strand. Alternatively, in order to accurately model crowding and RNA by discrete event simulation, it might be necessary to conduct simulations with larger initial monomer counts and finding a new method to account for the rapid localization of capsid proteins to specific target sites representing the presence of RNA as a recruitment tool. Any simulations of that nature would have to account for spatial location of particles more than the current DESSA simulator.

Regardless, the complexities of this interplay between RNA and crowding effects so far provide more questions than answers. It is entirely plausible that combining the faster overall assembly rate of the combined RNA effects and the more stable intermediates of the highest crowding cases lead to a much higher risk of kinetic trapping. Nucleation-limited growth depends upon a buffer of available subunits to be present prior to nucleation to be used in the resulting rapid assembly of the capsid. If conditions of the simulation preemptively deplete this buffer, assembly cannot be completed. This can be seen when crowding is applied to single capsid assembly simulations and potentially in cases with applied RNA and high crowding level effects.

What is also interesting is that the mid-range crowding levels studied are now no more of a deterrent than the lowest of nonzero crowding levels seen in Figure 5.2. Previous estimations of *in vivo* crowding suggest that 40 to 45% would be a very extreme level seen inside of a cell whereas 15 to 20% crowding would be more realistic. It is interesting, then, that in this niche of examining two characteristics of the *in vivo* assembly environment, that more realistic crowding levels are favored compared to applying crowding corrections purely to best-fit *in vitro* parameters.

Chapter 6: Conclusions

6.1 Summary of Thesis Work

This thesis has made several critical strides in the areas of data fitting algorithms, viral pathway analysis and understanding the thermodynamic implications of transitioning from an *in vitro* to an *in vivo* assembly environment, specifically with regards to macromolecular crowding and capsid assembly around nucleic acid. This work has provided unprecedented insight into the mechanisms by which multiple icosahedral virus capsids assemble and how this assembly varies over a variety of changes to kinetic rates, capsomer concentration and assembly environmental conditions. Furthermore, detailed analysis of simulation results provides novel insight into the critical elements necessary in certain capsid assembly systems, information that would otherwise be unattainable from experimental data, theoretical models or from alternative simulation approaches, which are more coarse-grained in the assembly intermediates considered.

Modifying the parameter estimation algorithm previously designed in the lab to be applicable to any virus self-assembly system as well as to fit multiple curves of time-dependent assembly data simultaneously produced biologically realistic rate parameter values that fall in line well with previous free energy experiments in related virus systems. Furthermore, this work corroborated well the HPV data fitting results from the previous work in both kinetic rate values and resultant assembly pathways learned. What was surprising, however, was the ability of this modified parameter estimation code to learn rate parameters that show more complex assembly mechanisms than the simple monomer-based addition model for HPV. What was found then was a stark contrast between the assembly model discovered for HPV and the models discovered for CCMV and HBV. Capsid protein interactions in HPV assembly are far stronger with respect to reaction free energy than HBV or CCMV binding interactions, leading to far less trial-and-error

in the assembly process. This is tightly correlated to serious problems with kinetic trapping that plague HPV capsid assembly *in vitro* and *in silico* where stable intermediates form that cannot be broken down and no further assembly reactions are possible to continue assembly from the intermediate to a completed structure. Furthermore, HPV assembly is non-nucleation-limited in the simulations run and the vast majority of all binding reactions consist of single monomer-based addition to growing intermediates.

In comparison, the binding interactions for both CCMV and HBV capsomers are relatively much weaker, leading to far more trial-and-error in the assembly process despite an overall much faster rate of assembly. This in part leads to two striking differences in assembly mechanisms. First, both CCMV and HBV capsid assembly are examples of nucleation-limited growth models where a steady equilibrium of small oligomer concentrations is maintained until the formation of an assembly intermediate labeled the nucleus which in turn causes rapid completion of the capsid assembly process. Second, kinetic trapping is far less pronounced compared to HPV. Another critical distinction is the variety of assembly pathways utilized in both CCMV and HBV capsid assembly. Whereas HPV assembly was almost entirely monomer-based, both CCMV and HBV show two distinct well used pathways of monomer-based and trimer-based assembly which are used interchangeably at random as assembly progresses. Because of this variability, examining individual essential reactions is made more difficult when more than one pathway can be utilized by a virus to assemble and, indeed, the number of potential combinations of assembly reactions utilized to transition from individual monomers to completed capsid is astronomical. What can be examined instead is determining what nucleation event helps spark rapid completion of capsid assembly. In both HBV and CCMV, this nucleation event appears to be the first formation of a hexagon structure during assembly. This hexagon

does not appear in one specific size of intermediate; however, upon its formation whether it is in a 9mer, 17mer or 20mer, the rate of assembly dramatically increases until the final capsid structure is formed.

Drawing pathway conclusions about assembly simulations fit to *in vitro* data, and thus an *in vitro* assembly environment, is only the first step of my work. Indeed, what is most critical is to understand the nature of virus capsid assembly *in vivo* and my work has shown that changes in assembly environment can play a dramatic role in modifying assembly rate and assembly pathways. To accomplish this, I considered two separate important distinctions between the two assembly environments: nucleic acid and macromolecular crowding. While there are numerous changes in assembly environment that can be considered, the lack of knowledge of how each effect would, alone or in consort, modify overall assembly rate and pathways suggests an approach where each important environmental modification is considered individually and then combined as necessary when individual effects are found to be important. Previous experimental and simulation evidence had suggested that both crowding and nucleic acid could play major roles in modifying assembly.

To account for the effects of nucleic acid on capsid assembly, I examined the relatively simple system of CCMV capsid assembly about a single viral RNA strand. I designed a theoretical model based upon Flory theory to account for individual modifications to rate parameters based upon specific elements of the RNA effect on capsid assembly, including negative effects such as RNA-RNA interaction and compression of the RNA strand within the capsid, and positive effects such as RNA-capsid protein interaction and increased capsomer concentration in the near vicinity of the RNA strand at the beginning of assembly. What I found was an interesting combination of expected and surprising results.

As expected, individual negative effects would indeed slow down overall assembly rate and individual positive effects would speed up the rate of assembly. However, as more complex combinations of positive and negative effects were examined, it became clear that it was the combined RNA effects case which produced the most efficient overall assembly. Surprisingly, simply considering positive effects on assembly rate yielded simulations unable to produce completed capsid structures primarily because of issues with kinetic trapping. Alternatively, solely examining negative effects failed to yield any larger intermediates at all. This suggests a viral assembly system that is well evolved to the environmental conditions most commonly seen during the CCMV life cycle, which although this is mere speculation, does make sense that viruses would be adapted to the environment in which they naturally assemble. Myriad additions of simulation complexity and experimental data would be necessary to truly corroborate this speculation, but it is an interesting result in this simplistic system nonetheless. More detailed analysis of CCMV assembly pathways under each potential combination of RNA effects showed an increased diversity in binding reactions, especially in the combined effects case where a pentamer-based addition pathway was also quite common. Furthermore, these simulations showed the robustness of the previously defined nucleus in CCMV capsid assembly simulations under a lot of kinetic rate perturbations. While there were extreme cases where larger intermediates were not assembled and where bond association rates became too fast yielding kinetically trapped intermediates unable to form full capsids, in the remainder of the cases examined, the nucleus of CCMV capsid assembly was still the formation of the same hexagon structure.

To address the role of macromolecular crowding on capsid assembly, I applied a regression model previously designed in the lab to predict changes in equilibrium constant based

upon modifications in percentage of assembly environment composed of crowding agents. The details of this regression model were learned by fitting a regression function to the results of a three-dimensional crowding simulator based upon Green's function reaction dynamics that analyzed changes in association and dissociation rates based upon varying levels of macromolecular crowding.

Because crowding levels at specific capsid assembly locations are not known, I calculated changes in rate parameters for a wide range of crowding levels, from 0 to 45%, to analyze the global trends of how crowding affects assembly. I applied these rate parameter corrections to the best-fit parameters for all three virus systems studied and once again found surprising contrast between HPV and the other two viruses, CCMV and HBV. At low levels of macromolecular crowding, there was a decrease in assembly rate of all viruses studied; however, as crowding levels continued to rise, assembly rate and yield for CCMV and HBV both began to increase again and in certain circumstances surpass the uncrowded curves. In contrast, HPV assembly rate and yield continued to decline as crowding levels increased. I deduced that this distinction is in large part because HBV and CCMV assembly are both examples of nucleation-limited growth while HPV is not. Despite the nature of macromolecular crowding to slow down association rates, it also serves to stabilize intermediates by slowing down dissociation rates as well. In the case of nucleation-limited assembly, further stabilization of the nucleus can then ensure overall more efficient assembly. Without a nucleus or any specific assembly to stabilize, increasing crowding simply serves to further slow the rate of HPV assembly. Once again, with respect to both HBV and CCMV, the formation of a hexagon served as the nucleation step for rapid elongation and completion of the capsid regardless of crowding level. This once again

shows remarkable robustness of this essential piece in the assembly pathways of both HBV and CCMV.

These insights derived from simulations of virus capsid assembly emphasize the ability of simulation techniques to investigate biological problems that, to date, cannot be addressed solely through experimental means. Furthermore, the results attained by the techniques described here offer biologically plausible and relevant information about the nature of virus assembly to a level of detail that could not be addressed previously. These results are of specific relevance finding a critical intermediate structure in the assembly pathways of both CCMV and HBV which, while this structure cannot necessarily be described by a single size of intermediate, the geometry of the structure that is essential does not change and could potentially be targeted for antiviral therapy. This work also has direct applications to a variety of other fields, whether it is analyzing other self-assembly systems in nature or in engineering, understanding the role of macromolecular crowding on cellular processes as well as gaining further insight into the detailed interactions between proteins and nucleic acid, a topic that is crucial in numerous subcategories of biological research. In many ways, this work is a perfect example of the major trends in modern research, where the relatively simple question of how individual subunits form together to create a completed icosahedral capsid structure requires an incredibly varied interdisciplinary approach combining biological experimental results, computational simulation algorithms, theoretical physics and chemistry, as well as deep virology understanding to properly frame both the questions raised and the answers found. As individual scientific fields become further intermingled, this form of interdisciplinary research will become all the more necessary to address the biological questions that will dominate the twenty first century.

6.2 Future Directions

The work described in this thesis can be progressed further on a number of different fronts ranging from improving parameter estimation techniques, acquiring potentially better data sources, developing more complicated models for the impact of nucleic acid and macromolecular crowding on virus assembly, tackling new virus systems to search for further novel assembly pathways, as well as addressing further complexity in distinctions between *in vitro* and *in vivo* assembly environment. Beyond direct applications of the current work, there is great room for more ancillary studies to attempt to corroborate or explain the findings reported here, both experimentally and computationally.

6.2.1 Producing Improved Fits to In Vitro Data

There are many potential approaches to improving the parameter estimation technique used in Chapter 2. Current work submitted for publication in the lab has applied derivative-free optimization (DFO) methods as an alternative approach to parameter estimation for the incredibly noisy capsid assembly simulations, specifically examining two popular DFO algorithms Multilevel Coordinate Search (MCS) (114) and Stable Noisy Optimization by Branch and FIT (SNOBFIT) (115). This work has suggested that DFO methods might be more capable of handling the noisy objective functions seen during parameter estimation over viral assembly systems.

Another potential avenue for producing improved parameter estimation is by utilizing alternative forms of experimental data, such as small angle x-ray scattering (SAXS), dynamic light scattering (DLS) and non-covalent mass spectrometry (NCMS), which can offer more informative data sets and theoretically produce better fits. To date, the lab has only examined synthetic data in the case of NCMS and is working on producing synthetic data to represent

SAXS and DLS experiments and the lab is actively pursuing real experimental data on icosahedral virus systems in each case. The results found so far suggest that fitting parameters based upon NCMS data produces better overall fits to the ground truth of the synthetic data and hopefully once real experimental data is acquired, this can be examined further. Another potentially useful source of experimental data is Surface Plasmon Resonance (SPR) (116) which can study molecular interactions in real-time without the presence of labeling agents and can potentially determine kinetic rate constants experimentally. While this technology could not necessarily explain all of the complex kinetics of capsid assembly, especially *in vivo*, it could provide a reference point for individual subunit interactions.

6.2.2 Corroborating Results with New Virus Systems

To date, the assembly rate and pathways of three icosahedral virus systems have been studied in the lab. Two of these viruses share a similar nucleation-limited growth model while HPV assembly is strikingly different. This is a very small sample of viruses that have been tested using this simulation and parameter estimation approach. It will be important to determine if simulations on other virus systems share similar assembly pathways to either CCMV and HBV, or HPV, or alternatively if something else entirely is discovered. Currently, both CCMV and HBV inferred assembly models are very similar despite their differences in capsid geometry and kinetic rates. They share similar assembly pathways and even a desired nucleation step. The question remains if the simulator can discover an alternative nucleus for another nucleation-limited virus assembly pathway or if this hexagon structure is pervasive across all nucleation-limited assembly simulations, which is doubtful. Another interesting question would be if there is a case of fast capsid assembly on the same time scale as HBV or CCMV that is not nucleation-limited or if nonnucleation-limited growth models are always relegated to exceedingly slow

assembly pathways. Some of this work can be done theoretically by examining parameter spaces for standard icosahedral geometries, although theoretically discovering novel pathways is less enlightening than novel pathways discovered while fitting real data.

6.2.3 Delving Deeper into Understanding Pathway Selection in Each Virus

This thesis described in detail important pathways and intermediates during assembly simulations. What is essential now is to understand why certain pathways and certain intermediates are so favored during virus capsid assembly. Structural studies both experimentally and computationally would be interesting to understand the stability of this hexagon nucleus for both HBV and CCMV as well as what makes this structure so able to spark rapid assembly completion. Molecular dynamics (MD) studies may also play an important role in understanding individual capsid protein interactions on small time scales to measure the free energy of a simulated capsid protein interaction and compare with the results of fitting experimental data.

One can also ask how can these current assembly pathways be altered, something that could potentially be very important when finding vulnerable steps in assembly for targeting by antivirals (92-94). An important question would be what changes to assembly environment, kinetic rates or addition of competitive binding agents to simulation studies alter the nucleation-limited growth models seen in HBV and CCMV. If capsid assembly could be slowed dramatically *in vivo* during an infection such that the yield of new virus particles was 10-fold or 100-fold lower, that could have a dramatic effect fighting disease. Another question would be how potential mutations in coat proteins might alter assembly pathways and if changes in DESSA simulation results corroborate well with results of MD studies showing binding of similarly mutated coat proteins.

6.2.4 Adding Further Complexity to Nucleic Acid – Capsid Protein Effects

In deriving the model in Chapter 3 for the effects of nucleic acid on capsid assembly, I sought to design a series of theoretically sound rate parameter corrections that were based on as much relevant experimental evidence as possible. In certain situations, however, estimations had to be made where there was no experimental evidence for a specific virus studied. To estimate the thickness of the RNA inside the CCMV capsid, I used experimental cryo-electron micrograph data showing the location of pre-genomic RNA inside the HBV capsid surface of a certain thickness (99) and then extrapolated to CCMV based upon differences in genome length, and thus volume filled, and capsid radius. For more accuracy of these RNA models, it will be essential to gain the experimental evidence necessary to fill these gaps and reduce the need for estimation based upon data from other viral systems. In addressing RNA-capsid protein interaction, this could be another good use of MD simulations. Currently, the state of the art supercomputers designed specifically to run MD simulations of large molecular assemblies have proven capable of all-atom simulations of entire viral capsids on a microsecond timescale; however, they have to date not included any form of nucleic acid within the capsid structure. Constructing even short MD simulations of viruses with RNA contained within the capsid might provide more detailed information about the free energy of this system.

Acquiring experimental time-dependent evidence of viral capsids assembling around RNA would be exceedingly helpful in validating the results of this thesis. Virus capsids have been assembled *in vitro* around a variety of charged polymers, including virus-specific RNA strands (67,68). Generating light scattering or NCMS data of the assembly progression of capsids forming around their own RNA or different charged polymers would be of great assistance. In fact, a great validation test would be varying the size of charged polymers around which the

capsids must assemble and comparing the simulation results to the theoretical predictions from my RNA models.

6.2.5 Improvement of Macromolecular Crowding Modeling

There are a number of facets of the modeling of macromolecular crowding that can be improved upon to more accurately reflect the complex nature of crowding particles *in vivo*. Currently the three-dimensional GFRD crowding simulations only consider a single size and shape of crowding agent, whereas *in vivo*, crowding agents will not be a single uniform particle. Theoretical models do offer some ability to extrapolate from a single crowding particle size to a more complex crowded environment (113). It is always a tradeoff, however, between model complexity, computational efficiency, and learnability of model parameters. As computing power, algorithms, and experimental methods for monitoring capsid assembly improve, it should become possible to move to more detailed and realistic models.

Another example of this tradeoff between further complexity and computational efficiency is that crowding effects are applied uniformly to all bond association and dissociation reactions regardless of whether they are monomer-monomer interactions or 10mer-20mer interactions which potentially could see differences in diffusion capability, especially as crowding level increases. Developing a method of applying crowding effects relative to the size of assembly in the association or dissociation event could potentially provide more realism.

In the last few years, there has been much greater interest in the effects of macromolecular crowding on cellular processes. Because of this interest, a number of experiments have been done to begin to examine crowding in various novel manners. Recent research has developed a tool to measure levels of macromolecular crowding in living cells (117). This could be exceedingly helpful in resolving correct levels of crowding at the specific

site of capsid assembly, provided this crowding level experiment was paired with a localization experiment of capsid proteins inside the cell. Another very useful direction for experimental research would be to measure viral capsid assembly in a crowded environment with one of the experimental techniques previously discussed providing assembly progression over time. This would be very beneficial for validation of the crowding model described in Chapter 4 as well as in fitting more realistic rate parameters directly from experimental evidence.

6.2.6 Developing More Accurate Depictions of the In Vivo Assembly Environment

This thesis has addressed two specific changes between *in vitro* and *in vivo* assembly environments; however, there are far more distinctions that could produce a significant change in viral assembly. The lack of an important chaperone protein in the *in vitro* HPV experiments could very well explain the much slower nonnucleation-limited assembly process learned. Understanding the nature of the impact of this protein on the formation of HPV capsids would be essential in developing a model for HPV assembly *in vivo*. Developing a realistic model for HBV assembly *in vivo* also produces challenges not present with CCMV. HBV assembles around its pre-genomic RNA as well as a polymerase protein which will later be necessary for reverse transcribing the RNA into a final DNA form. While the polymerase protein is relatively small in comparison to the RNA strand, it is not known how these two interact inside the capsid and how this interaction might affect the free energy of the total system. Structural or simulation studies of this interaction would be crucial to developing an accurate model of HBV *in vivo*. Furthermore, as I touched on in Chapter 5, it will be imperative to combine different modifications to assembly environment to produce a more accurate picture of real capsid assembly pathways. This is not as trivial as simply taking individual rate corrections and merging them together, however. Each environmental change will likely have an effect on other

environmental changes included in the model. Much more thought and theoretical development will have to be included to produce as accurate a recreation of the viral capsid assembly environment as possible.

It will be imperative as experimental and computational techniques continue to improve to continually push the boundaries of what can be measured and analyzed and what levels of complexity can be included in a model for virus capsid assembly. While it can always be said that a model can be made more complex or a better form of data or computational algorithm could be designed or applied, the work described here not only shows the potential for computational algorithms to address virus capsid assembly, but rather illustrates the unprecedented impact computational algorithms have already achieved in unraveling the mysteries of this critical point in the viral life cycle.

References

1. Goff, S. P. and Berg, P. (1976). Construction of hybrid viruses containing SV40 and lambda phage DNA segments and their propagation in cultured monkey cells. *Cell,* 9:695-705.

2. Matsuzaki S., Rashel, M., …, and Imai, S. (2005). Bacteriophage therapy: a revitalized therapy against bacterial infectious diseases. *J. Infect. Chemother.,* 11(5):211–219.

3. Blattner, W.A. (1999). Human retroviruses: their role in cancer. *Proc Assoc Am Physicians,*111(6):563-72.

4. Bartsch, S.M, Gorham, K., and Lee, B. Y. (2015). The cost of an Ebola case. *Pathogens and Global Health*, DOI: 10.1179/2047773214Y.0000000169.

5. Goodenow, M., Huet T., …, and Wain-Hobson, S. (1989). HIV-1 isolates are rapidly evolving quasispecies: evidence for viral mixtures and preferred nucleotride substitutions. *J. Acq. Imm. Def.,* 2(4):344-352.

6. Moutouh, L., Corbeil, J., and Richman, D. D. (1996). Recombination leads to the rapid emergence of HIV-1 dually resistant mutants under selective drug pressure. *Proc. Natl. Acad. Sci. USA*, 93:6106-6111.

7. Schweiger, B., Zadow, I., and Heckler, R. (2002). Antigenic drift and variability of influenza viruses. *Med. Microbiol. Immun.,* 191:133-138.

8. Lou, Z., Sun, Y., and Rao, Z. (2014). Current progress in antiviral strategies. *Trends in Pharmacological Sciences*, 35(2): 86-102.

9. Hanna, G.J., Lalezari, J., …, and Grasela, D. M.. (2012). Antiviral activity, pharmacokinetics, and safety of BMS-488043, a novel oral small-molecule HIV-1 attachment inhibitor, in HIV-1-infected subjects. *Antimicrob. Agents Chemother.* 55, 722–728.

10. Yamashita, M. (2010). Laninamivir and its prodrug, CS-8958: longacting neuraminidase inhibitors for the treatment of influenza. *Antivir. Chem. Chemother.* 21, 71–84.

11. Desai, T.M., Marin, M., …, Melikyan, G.B. (2014). IFITM3 Restricts Influenza A Virus Entry by Blocking the Formation of Fusion Pores following Virus-Endosome Hemifusion. *PLOS Pathogens,* 10(4):e1004048.

12. Chandran, K., Sullivan, N. J., …, Cunningham, J.M. (2005). Endosomal Proteolysis of the Ebola Virus glycoprotein is necessary for infection. *Science,* 308:1643-45.

13. Kilby, J.M., Hopkins, S., …, and Saag, M. S. (1998) Potent suppression of HIV-1 replication in humans by T-20, a peptide inhibitor of gp41-mediated virus entry. *Nat. Med.,* 4, 1302–1307.

14. Andrei, G., De Clercq, E., and Snoeck, R. (2009). Drug targets in cytomegalovirus infection. *Infect. Disord. Drug Targets*, 9, 201–222.

15. Palumbo, E. (2008). New drugs for chronic hepatitis B: a review. *Am. J. Ther.*, 15, 167–172

16. Wegzyn, C.M. and Wyles, D.L. (2012). Antiviral drug advances in the treatment of human immunodeficiency virus (HIV) and chronic hepatitis C virus (HCV). *Curr. Opin. Pharmacol*., 12, 556–561.

17. Teschke, C.M., King, J., and Prevelige, P. E. Jr. (1993). Inhibition of viral capsid assembly by 1,1'-Bi(4-anilinonaphthalene-5-sulfonic acid). *Biochemistry*, 32:10658-65.

18. Eastman, S.W. and Linial, M. L. (2001). Identification of a conserved residue of foamy virus gag required for intracellular capsid assembly. *Journal of Virology*, 75(15): 6857-64.

19. Kanamoto, T., Kashiwada, Y., …, Nakashima, H. (2001). Anti-human immunodeficiency virus activity of YK-FH312 (a betulinic acid derivative), a novel compound blocking viral maturation. *Antimicrobial Agents and Chemotherapy*, 45(4):1225-30.

20. Ivanov, A. I. (2014). Pharmacological inhibitors of exocytosis and endocytosis: novel bullets for old targets. *Methods in Molecular Biology*, 1174:3-18.

21. Whitesides, G. M., Mathias, J. P., and Seto, C. T. (1991). Molecular selfassembly and nanochemistry: a chemical strategy for the synthesis of nanostructures. *Science*, 254:1312–1319.

22. Whitesides, G. M., and Grzybowski, B. (2002). Self-assembly at all scales. *Science*, 295:2418–2421.

23. Zlotnick, A. 1994. To build a virus capsid. An equilibrium model of the self assembly of polyhedral protein complexes. *J. Mol. Biol.*, 241:59–67.

24. Berger, B., Shor, P. W.,…, and King, J. (1994). Local rule-based theory of virus shell assembly. *Proc. Natl. Acad. Sci. USA*, 91:7732–7736.

25. Schwartz, R., Shor, P. W.,…, and Berger, B. (1998). Local rules simulation of the kinetics of virus capsid self-assembly. *Biophys. J.*, 75:2626–2636.

26. Rapaport, D., Johnson, J., and Skolnick, J. (1999). Supramolecular selfassembly: molecular dynamics modeling of polyhedral shell formation. *Comput. Phys. Commun.*, 122:231–235.

27. Hagan, M. F., and Chandler, D. (2006). Dynamic pathways for viral capsid assembly. *Biophys. J.*, 91:42–54.

28. Zandi, R., van der Schoot, P,…, and Reiss, H. (2006). Classical nucleation theory of virus capsids. *Biophys. J.*, 90:1939–1948.

29. Nguyen, H. D., Reddy, V. S., and Brooks, C. L. (2007). Deciphering the kinetic mechanism of spontaneous self-assembly of icosahedral capsids. *Nano Lett.*, 7:338–344.

30. Zhang, T. and Schwartz, R. (2006). Simulation study of the contribution of oligomer/oligomer binding to capsid assembly kinetics. *Biophys. J.*, 90:57–64.

31. Keef, T., Micheletti, C., and Twarock, R. (2006). Master equation approach to the assembly of viral capsids. *J. Theor. Biol.*, 242:713–721.

32. Misra, M., Lees, D.,…, and Schwartz, R. (2008). Pathway complexity of model virus capsid assembly systems. *Comp. Math. Meth. Medicine.*, 9:277–293.

33. Sweeney, B., Zhang, T., and Schwartz, R. (2008). Exploring the parameter space of complex self-assembly through virus capsid models. *Biophys. J.*, 94:772–783.

34. Ceres, P. and Zlotnick, A. (2002). Weak protein-protein interactions are sufficient to drive assembly of hepatitis B virus capsids. *Biochemistry*, 41:11525–11531.

35. Parent, K. N., Zlotnick, A., and Teschke, C. M. (2006). Quantitative analysis of multi-component spherical virus assembly: scaffolding protein contributes to the global stability of phage P22 procapsids. *J. Mol. Biol.*, 359:1097–1106.

36. Zlotnick, A., Johnson, J. M.,…, and Endres, D. (1999). A theoretical model successfully identifies features of hepatitis B virus capsid assembly. *Biochemistry*, 38:14644–14652.

37. Toropova, K., Basnak, G.,…, and Ranson, N. A. (2008). The three-dimensional structure of genomic RNA in bacteriophage MS2: implications for assembly. *J. Mol. Biol.*, 375:824–836.

38. Reddy, V. S., Giesing, H. A.,…, and Johnson, J. E.. (1998). Energetics of quasiequivalence: computational analysis of protein-protein interactions in icosahedral viruses. *Biophys. J.* 74:546–558.

39. Hemberg, M., Yaliraki, S. N., and Barahona, M. (2006). Stochastic kinetics of viral capsid assembly based on detailed protein structures. *Biophys. J.*, 90:3029–3042.

40. Bourne, C.R., Katen, S.P.,…, and Zlotnick, A. (2009). A mutant hepatitis B virus core protein mimics inhibitors of icosahedral capsid self-assembly. *Biochemistry*, 48(8):1736-1742.

41. Lavelle, L., Michel, J-P. and Gingery, M. (2007). The disassembly, reassembly and stability of CCMV protein capsids. *Journal of Virological Methods*, 146:311-316.

42. Modis, Y., Trus, B.L. and Harrison, S.C. (2002). Atomic model of the papillomavirus capsid. *The EMBO Journal*, 21(18):4754-4762.

43. Casini, G. L., Graham, D, and Wu, D. T. (2004). In vitro papillomavirus capsid assembly analyzed by light scattering. *Virology*, 325: 320–327.

44. Finnen, R. L., Erickson, K. D., and Garcea, R. L. (2003). Interactions between papillomavirus L1 and L2 capsid proteins. *J. Virol.*, 77: 4818–4826.

45. Zhou, S., and Standring, D. N. (1992). Hepatitis B virus capsid particles are assembled from core-protein dimer precursors. *Proc. Natl. Acad. Sci. USA*, 89: 10046–10050.

46. Kenney, J. M., von Bonsdorff, C. M., and Fuller, S. D. (1995). Evolutionary conservation in the hepatitis B virus core structure: comparison of human and duck cores. *Structure*, 3: 1009–1019.

47. Zlotnick, A., Aldrich, R., and Young, M. J. (2000). Mechanism of capsid assembly for an icosahedral plant virus. *Virology*, 277: 450–456.

48. Speir, J. A., Munshi, S., and Johnson, J. E. (1995). Structures of the native and swollen forms of cowpea chlorotic mottle virus determined by x-ray crystallography and cryo-electron microscopy. *Structure*, 3: 63–78.

49. Gillespie, D.T. (1977). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81:2340-2361.

50. Jamalyaria, F., Rohlfs, R., and Schwartz, R. (2005). Queue-based method for efficient simulation of biological self-assembly systems. *Journal of Computational Physics*, 204: 100-120.

51. Endres, D. and Zlotnick, A. (2002). Model-based analysis of assembly kinetics for virus capsids or other spherical polymers. *Biophys. J.*, 83:1217–1230.

52. Zhang, T., Rohlfs, R., and Schwartz, R. (2005). Implementation of a discrete event simulator for biological self-assembly systems. *In Proceedings of the 37th Winter Simulation Conference, Orlando, FL.* 2223–2231.

53. Kumar, M. S. and Schwartz, R. (2010). A parameter estimation technique for stochastic self-assembly systems and its application to human papillomavirus self-assembly. *Phys. Biol.*, 7:045005.

54. Box, G. E. P. and Wilson, K. B. (1951). On the experimental attainment of optimum conditions (with discussion). *J. Roy. Stat. Soc. B.*, 13:1–45.

55. Ermoliev, Y. 1983. Stochastic quasigradient methods and their application to systems optimization. *Stochastics*, 4:1–37.

56. del Álamo, M., Rivas, G. and Mateu, M.G. (2005). Effect of Macromolecular Crowding Agents on Human Immunodeficiency Virus Type 1 Capsid Protein Assembly In Vitro. *Journal of Virology*, 79(22):14271-14281.

57. Lin, B. Y., Makhov, A. M.,…, and Chow, L T. (2002). Chaperone proteins abrogate inhibition of the human papillomavirus (HPV) E1 replicative helicase by the HPV E2 protein. *Mol. Cell. Biol.*, 22:6592–6604.

58. Verchot, J. 2012. Cellular chaperones and folding enzymes are vital contributors to membrane bound replication and movement complexes during plant RNA virus infection. *Front Plant Sci.*, 3:275-1–275-12.

59. Ivanyi-Nagy, R., Davidovic, L,…, and Darlix, J.-L. (2005). Disordered RNA chaperone proteins: from functions to disease. *Cell. mol.Life Sci.*, 62:1409-1417.

60. Lanman, J., Crum, J.,…, and Johnson, J.E. (2008). Visualizing flock house virus infection in Drosophila cells with correlated fluorescence and electron microscopy. *Journal of Structural Biology*, 161:439-446.

61. Kegel, W. K. and van der Schoot, P. (2004). Competing hydrophobic and screened-coulomb interactions in hepatitis B virus capsid assembly. *Biophys. J.*, 86:3905–3913.

62. Venter, P. A., Krishna, N. K., and Schneemann, A. (2005). Capsid protein synthesis from replicating RNA directs specifics packaging of the genome of a multipartite, positive-strand RNA virus. *Virology*, 79:6239–6248.

63. Dykeman, E. C., Stockley, P. G., and Twarock, R. (2013). Building a viral capsid in the presence of genomic RNA. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 87:022717-1–022717-12.

64. Annamalai, P., Apte, S.,…, and Rao, A.L.N. (2005). Deletion of Highly Conserved Arginine-Rich RNA Binding Motif in Cowpea Chlorotic Mottle Virus Capsid Protein Results in Virion Structural Alterations and RNA Packaging Constraints. *Journal of Virology*, 79(6):3277-3288.

65. Stockley, P. G., Twarock, R. and Tuma, R. (2013). Packaging signals in single-stranded RNA viruses: nature's alternative to a purely electrostatic assembly mechanism. *J. Biol. Phys.*, 39:277-287.

66. Morales, L., Mateos-Gomez, P. A. and Sola, I. (2013). Transmissible gastroenteritis coronoavirus genome packaging signal is located at the 5' end of the genome and promotes viral RNA incorporation into virions in a replication-independent process. *J. Virol.*, 87(21):11579-90.

67. Comas-Garcia, M., Cadena-Nava, R. D. and Gelbart, W. M. (2012) In vitro quantification of the relative packaging efficiencies of single-stranded RNA molecules by viral capsid protein. *J. Virol.*, 86:12271-12282.

68. Cadena-Nava, R. D., Comas-Garcia, M. and Gelbart, W. M. (2012). Self-assembly of viral capsid protein and RNA molecules of different sizes: requirement for specific high protein/RNA mass ratio. *J. Virol.*, 86:3318-3326.

69. Sun, J., Dufort, C.,…, and Dragnea, B. **(**2007). Core-controlled polymorphism in virus-like particles. *Proc. Natl. Acad. Sci. USA*, 104:1354 –1359.

70. Verduin, B. J. M. and Bancroft, J. B. **(**1969). The infectivity of tobacco mosaic virus RNA in coat proteins from spherical viruses. *Virology*, 37:501–506.

71. van der Schoot, P. and Zandi, R. (2013). Impact of the topology of viral RNAs on their encapsulation by virus coat proteins. *J. Biol. Phys.*, 39:289-299.

72. Elrad, O. M. and Hagan, M. F. (2010). Encapsulation of a polymer by an icosahedral virus. *Phys. Biol.*, 7(4):045003.

73. Perlmutter, J. D., Perkett, M. R. and Hagan, M. F. (2014) Pathways for virus assembly around nucleic acids. *J. Mol. Biol.*, 426(18):3148-3165.

74. Zeng, Y., Larson, S. B.,…, and Harvey, S. C. (2012) A model for the structure of satellite tobacco mosaic virus. *J. Struct. Biol.*, 180(1):110-116.

75. Dykeman, E. C., Stockley, P. G. and Twarock, R. (2014) Solving a Levinthal's paradox for virus assembly identifies a unique antiviral strategy. *P. Natl. Acad. Sci. USA.*, 111(14):5361-6.

76. Minton, A. P. (2001). The influence of macromolecular crowding and macromolecular confinement on biochemical reactions in physiological media. *J. Biol. Chem.*, 276:10577–10580.

77. Zimmerman, S. B., and Minton, A. P. (1993). Macromolecular crowding: biochemical, biophysical, and physiological consequences. *Annu. Rev. Biophys. Biomol. Struct.*, 22:27–65.

78. LeDuc, P. R., and Schwartz, R. (2007). Computational models of molecular self-organization in cellular environments. *Cell Biochem. Biophys.*, 48:16–31.

79. Ellis, R. J. (2001). Macromolecular crowding: obvious but underappreciated. *Trends Biochem. Sci.*, 26:597–604.

80. Rinco´n, V., Bocanegra, R.,…, Mateu, M. G. (2011). Effects of macromolecular crowding on the inhibition of virus assembly and virus-cell receptor recognition. *Biophys. J.*, 100:738–746.

81. Johnston, I. G., Louis, A. A., and Doye, J. P. (2010). Modelling the selfassembly of virus capsids. *J. Phys. Condens. Matter*, 22:104101–104110.

82. McGuffee, S. R. and Elcock, A. H. (2010). Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. *PLOS Comput. Biol.*, 6:e1000694.

83. Balbo, J., Mereghetti, P.,…, and Wade, R. C. (2013). The shape of protein crowders is a major determinant of protein diffusion. *Biophys. J.*, 104:1576–1584.

84. Mittal, J., and Best, R. B. (2010). Dependence of protein folding stability and dynamics on the density and composition of macromolecular crowders. *Biophys. J.*, 98:315–320.

85. Lee, B., Leduc, P. R., and Schwartz, R. (2008). Stochastic off-lattice modeling of molecular self-assembly in crowded environments by Green's function reaction dynamics. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 78:031911-1–031911-9.

86. Lee, B., LeDuc, P. R., and Schwartz, R. (2012). Three-dimensional stochastic off-lattice model of binding chemistry in crowded environments. *PLoS ONE*, 7:e30131.

87. Flory, P. J. (1942). Thermodynamics of high polymer solutions. *J. Chem. Phys.*, 10:51-61.

88. Rubinstein, M. and Colby, R. H. (2003). *Polymer Physics* (Oxford: Oxford University Press).

89. Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quart. Appl. Math.*, 2:164–168.

90. Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. *SIAM J. Appl. Math.*, 11:431–441.

91. Conway, M. J., and Meyers, C. (2009). Replication and assembly of human papillomaviruses. *J. Dent. Res.*, 88:307–317.

92. Ternois, F., Sticht, J.,…, and Rey, F. A. (2005). The HIV-1 capsid protein C-terminal domain in complex with a virus assembly inhibitor. *Nat. Struct. Mol. Biol.* 12:678–682.

93. Stray, S. J., Johnson, J. M.,…, and Zlotnick, A. (2006). An in vitro fluorescence screen to identify antivirals that disrupt hepatitis B virus capsid assembly. *Nat. Biotechnol.*, 34:358–362.

94. Zlotnick, A., Lee, A.,…, Stray, S. J. (2007). In vitro screening for molecules that affect virus capsid assembly (and other protein association reactions). *Nat. Protoc.*, 2:490–498.

95. Minton, A. P. (2006). How can biochemical reactions within cells differ from those in test tubes? *J. Cell Sci.*, 119:2863–2869.

96. Xie, L., Smith, G. R.,…, and Schwartz, R. (2012). Surveying capsid assembly pathways through simulation-based data fitting. *Biophys. J.* 103:1545–1554.

97. Strobl, G. R. (2007). *The Physics of Polymers: Concepts for Understanding their Structures and Behavior 3rd Edition* (Berlin: Springer-Verlag).

98. Chen, H., Meisburger, S. P.,…, and Pollack, L. (2012) Ionic strength-dependent persistence lengths of single-stranded RNA and DNA. *P. Natl. Acad. Sci. USA*, 109(3):799-804.

99. Wang, J. C-Y., Dhason, M. S. and Zlotnick, A. (2012). Structural organization of pregenomic RNA and the carboxy-terminal domain of the capsid protein of hepatitis b virus. *PloS Pathog*, 8(9):e1002919.

100. Zandi, R, and van der Schoot, P. (2009). Size regulation of ss-RNA Viruses. *Biophys. J.*, 96:9-20.

101. Sakaue, T. and Raphael, E. (2006). Polymer chains in confined spaces and flow-injection problems: some remarks. *Macromolecules*, 39:2621-28.

102. Zlotnick, A. (2003). Are weak protein-protein interactions the general rule in capsid assembly? *Virology*, 315:269-274.

103. Prevelige Jr, P. E., Thomas, D. and King, J. (1993). Nucleation and growth phases in the polymerization of coat and scaffolding subunits into icosahedral procapsid shells. *Biophys. J.*, 64:824-835.

104. van Zon, J. S. and ten Wolde, P. R. (2005). Green's function reaction dynamics: A particle based approach for simulating biochemical networks in time and space. *J. Chem. Phys.*, 123: 234910.

105. Krissinel, E. and K. Henrick. (2007). Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.*, 372: 774-797.

106. Voss, N.R., Gerstein, M.,…, and Moore, P. B. (2006). The geometry of the ribosomal polypeptide exit tunnel. *J Mol Biol.*, 360:893-906.

107. Hagan, M. F. and Elrad, O. M. (2010). Understanding the concentration dependence of viral capsid assembly kinetics—the origin of the lag time and identifying the critical nucleus size. *Biophys. J.*, 98:1065–1074.

108. Morozov, A. Y., Bruinsma, R. F. and Rudnick, J. (2009). Assembly of viruses and the pseudo-law of mass action. *J. Chem. Phys.*, 131:155101-1–155101-17.

109. Katen, S. P., Chirapu, S. R.,…, Zlotnick, A. (2010). Trapping of hepatitis B virus capsid assembly intermediates by phenylpropenamide assembly accelerators. *ACS Chem. Biol.*, 5:1125–1136.

110. Tuma, R., Tsuruta, H.,…, and Prevelige, P. E. (2008). Detection of intermediates and kinetic control during assembly of bacteriophage P22 procapsid. *J. Mol. Biol.*, 381:1395–1406.

111. Johnson, J. M., Tang, J.,…, and Zlotnick, A. (2005). Regulating self-assembly of spherical oligomers. *Nano Lett.*, 5:765–770.

112. Hagan, M. F., Elrad, O. M. and Jack, R. L. (2011). Mechanisms of kinetic trapping in self-assembly and phase transformation. *J. Chem. Phys.*, 135:104115-1–104115-13.

113. Blum, J.J., Lawler, G.,…, and Shin, I. (1989). Effect of cytoskeletal geometry on intracellular diffusion. *Biophysical Journal*, 56:995-1005.

114. Neumaier, A.: MCS: Global Optimization by Multilevel Coordinate Search. http://www.mat.univie.ac.at/neum/software/mcs/

115. Huyer, W. and Neumaier, A. (2008). SNOBFIT—Stable noisy optimization by branch and fit. *ACM Trans. Math. Softw.*, 35, 1–25.

116. Wang, S., Boussaad, S. and Tao, N. (2003). Surface Plasmon Resonance Spectroscopy: Applications in Protein Adsorption and Electrochemistry. *Surfactant Science Series*, 111:213-251.

117. Gnutt, D., Gao, M.,…, and Ebbinghaus, S. (2015). Excluded-volume effects in living cells. *Angew. Chem. Int. Ed.*, 54:2548-2551.