Models of Molecular Self-Assembly for RNA Viruses and Synthetic DNA Cages

Nicholas Edwin Grayson

Thesis submitted for the degree of PhD

University of York

Department of Biology

July 2012

Abstract

A significant number of RNA viruses assemble their protein containers and genomic material simultaneously. Here the implications of this protein-RNA co-assembly are investigated using an extended version of a model first proposed by Adam Zlotnick in 1994 (Zlotnick, 1994). The inspirations for this extended model are the cases of bacteriophage MS2 and the STMV virus, viruses that have been well characterised experimentally. Example pathways of RNA virus assembly have been enumerated and kinetic simulations have been run on these networks. The results show the most likely pathways of virus assembly and the concentrations of the intermediates. This work will also demonstrate how kinetic traps may be avoided when proteins are able to bind RNA during assembly. Additionally modelled are DNA cages, which are three-dimensional shapes made from double-helical DNA molecules. Such cages have been seen within viruses but may also be constructed artificially. This model has been used to produce energetically optimised designs for icosidodecahedron-shaped DNA cages.

Contents

List of Figures

List of Tables

Acknowledgements

First I would like to thank my supervisor Prof. Reidun Twarock for all her help and enthusiastic discussions. I would also like to thank the members of my thesis advisory panel, Dr. Leo Caves and Dr. James Cussens for all their guidance. Furthermore I would like to thank my collaborators in Leeds, Prof. Peter Stockley, Dr. Katerina Toropova, Dr. Neil Ranson and colleagues at The Astbury Centre for Structural Molecular Biology whose research has greatly influenced this thesis and also, in Durham, Dr. Anne Taromina. I would also like to thank my friends and colleagues in YCCSA, especially Adam, Alastair, Eric, David, Jenny, Phil and Tom. Finally I would like to thank my family and friends for all their support, particularly Jo, Anne, Jenny M and most importantly Lizzie.

Declaration

The research presented in this submission is entirely my own work except where otherwise indicated in the text. Research in chapter 6 is based on “Self-Assembly of Viral Capsids via a Hamiltonian Paths Approach: The Case of Bacteriophage MS2”, Foundations of Nanoscience, N. E. Grayson, T. Keef. S. Severini and R. Twarock (Grayson et al., 2007). Later work in chapter 6 has been published in, “Simple Rules for Efficient Assembly Predict the Layout of a Packaged Viral RNA”, Journal of Molecular Biology, E.C. Dykeman, N.E. Grayson, K. Toropova, N.A. Ranson , P.G. Stockley and R. Twarock (Dykeman et al., 2011). Work in chapter 7 has previously been published in, “DNA duplex cage structures with icosahedral symmetry”, Theoretical Computer Science, N.E. Grayson, A. Taormina and R. Twarock (Grayson et al., 2009).

Chapter 1

Introduction

Understanding the assembly of the bacteriophage MS2 has been the main inspiration for the construction of models that take the co-operative roles of the genomic RNA and viral protein into account. The bacteriophage MS2 is a well studied model organism in the field of virus research. It is also the subject of an ongoing and productive collaboration with Prof. Peter Stockley and colleagues at The Astbury Centre for Structural Molecular Biology in Leeds, whose research has greatly influenced this thesis. After introducing the known assembly and structural aspects of the MS2 bacteriophage, it is shown how the constraints imposed by the RNA during capsid assembly can be modelled via Hamiltonian paths, a mathematical concept from graph theory. This Hamiltonian path model has consequences for the ensemble of assembly intermediates and the kinetics of the virus assembly

A further virus, Satellite tobacco mosaic virus (STMV), will also be introduced to demonstrate how the Hamiltonian path model applies to different assembly scenarios. Finally the possible virus modelling methods in the literature will be discussed.

1.1 Introduction To Bacteriophage MS2

MS2 is a bacteriophage that infects *E. coli* (see figure 1.1). In 1976 the RNA of this virus became the first genome ever to be sequenced (Fiers et al., 1976). Since this time many more extensive biochemical and structural studies have become available on MS2. This makes MS2 an ideal test system on which to base a model of RNA virus formation and it is the inspiration for the resulting model.

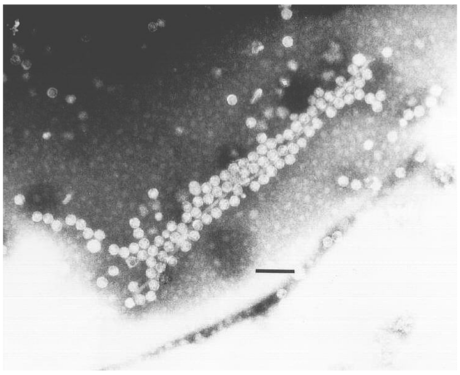

Figure 1.1: MS2 virus particles infecting an E. coli bacterium (Ackermann, 2006), the black bar represents 100 nm.

The MS2 virus has the same symmetry as an icosahedron, allowing its capsid proteins to fit into quasi-equivalent positions around the three-fold and five-fold axes of symmetry. In the Caspar-Klug classification of viral capsids, the MS2 virus is a T=3 virus (Casper and Klug, 1962) i.e. it has 180 capsid proteins. T=3 viruses are relatively simple; for example MS2 contains only 4 genes. The wild-type MS2 virion consists of a single-stranded RNA genome of 3,569 nucleotides surrounded by a protein capsid of coat protein and a single maturation protein that is important for infection. The coat protein first forms dimers of which there are two main conformations, an A/B and a C/C conformation as shown in figure 1.2. The crystal structure of the virus is shown in figure 1.3(a), in comparison with the tiling representing its surface structure in figure 1.3(b). The B monomer of the A/B dimer conformation has a flipped FG-loop allowing five copies of the dimers to meet around the five-fold axes without steric clashes. As a result, A/B dimers form pentamers of dimers around the five-fold axes of the MS2 capsid, whilst the C/C conformers

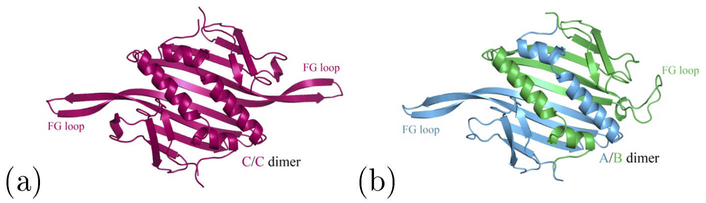

Figure 1.2: (a) A dimer in the symmetric C/C conformation and (b) in the asymmetric A/B conformation. Note the flipped FG loop in the asymmetric A/B monomer is the main source of asymmetry. This more compact FG loops allows the dimers to bind together around a five-fold axes. (Reproduced from (Toropova et al., 2008)).

only form part of the hexamers around the three-fold axes of the virus capsid. In the final capsid there are 60 dimers in the A/B conformation and 30 in the C/C conformation.

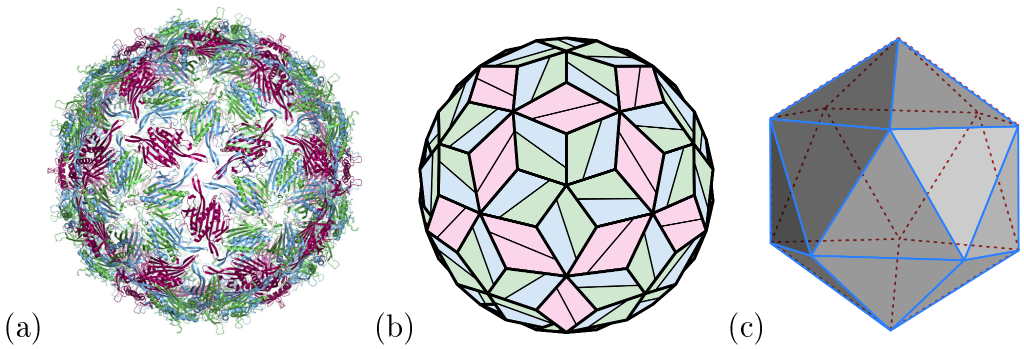

Figure 1.3: Crystal structure showing the relative positions of the coat protein dimers (reproduced from (Toropova et al., 2008)). Red C/C dimers sit on the two-fold axes of the virus and blue/green A/B dimers form clusters around five-fold axes and, interspersed with C/C dimers, in clusters of six around three-fold axes. (b) A schematic representation of the layout of the capsid. (c) An icosahedron that has the same symmetry as the capsid.

The coat protein dimer switches from a C/C into an A/B conformer by the binding of an RNA stem-loop via an allosteric effect (Stockley et al., 2007), (Dykeman and Twarock, 2010), (Morton et al., 2010). In the biological experiments described in (Stockley et al., 2007) RNA filled capsids are observed within 10 minutes, in comparison to days without RNA being present, which indicates the high efficiency of assembly in the presence of RNA.

The coat protein dimer shows different binding energies to different RNA stem-loops (Lago et al., 2001). In particular there is a high-affinity 19 nucleotide stem-loop at almost the exact centre of the MS2 genomic RNA, referred to as the translational repressor or TR sequence. This TR coat protein binding site is located at the ribosome binding site of the replicase MS2 gene. The binding of coat protein to this site is important in regulating the replicase translation. It is also thought that the TR site is the first in the genome to bind a coat protein dimer and therefore act as a nucleation site on which to grow the rest of the capsid (Stockley et al., 2007).

First experiments probing the cooperative roles of genomic RNA during capsid assembly only use multiple copies of the TR RNA sequence, instead of the full-length genome (Dykeman and Twarock, 2010) (Knapman et al., 2010). This allowed for the conduction of experiments that would not be possible with the full length genome such as detailed mass spectrometry of intermediates

(Knapman et al., 2010). Interestingly, assembly of the MS2 capsid in the presence of the full length genome progresses more slowly than that based on copies of TR alone (Rolfsson et al., 2008). This suggests that it takes longer to pack the greater amount of RNA present in the full genome. As there are 60 A/B dimers in the capsid, there must be 60 stem-loops of RNA binding to them to impart the correct conformation change. Since there are only a few RNA stem loops that bind the A/B dimers with high affinity (Lago et al., 2001), it may be concluded that most stem-loops bind only weakly with the dimers.

1.2 Visualisations of the MS2 RNA

It is possible to visualise the location of the MS2 RNA using cryo-electron microscopy (cryo-EM) (Toropova et al., 2008) (Van Den Worm et al., 2006), a technique that involves imaging the capsid using an electron microscope after freezing to 22 K. These low-resolution images can then be combined computationally and a 3D image of resolution typically around 9 Å produced (Toropova et al., 2008), see figure 1.4 for an example of the raw images used. The resulting RNA images show long and short segments of RNA density beneath the protein capsid, organised in a polyhedral shell arrangement, and further density within the virus making up a second shell of RNA. Two views of the reconstructions are shown in figure 1.5 (Toropova et al., 2008), (Koning et al., 2003). From these images the possible places where the RNA is located can be mapped onto the virus tiling in figure 1.3(b). Figure 1.6 shows the RNA locations in the outer shell of RNA on an icosahedral net of the virus.

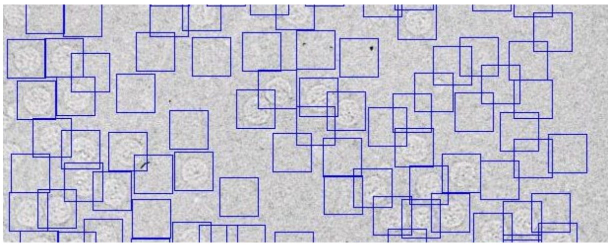

Figure 1.4: Cryo-electron microscopy images of MS2 (Van Den Worm et al., 2006).

The cryo-EM reconstructions shown in figure 1.5 from (Toropova et al., 2008) and (Koning et al., 2003) are icosahedrally averaged. This is because the only reference to align the virus images are the symmetry axes and different

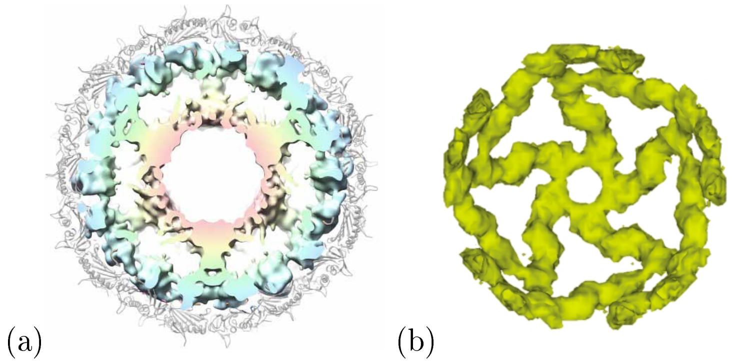

Figure 1.5: Cryo-electron microscopy reconstructions of bacteriophage MS2. (a) showing a slab viewed along a three-fold axis (Toropova et al., 2008) and (b) the outer shell of RNA viewed along a five-fold axis (Koning et al., 2003).

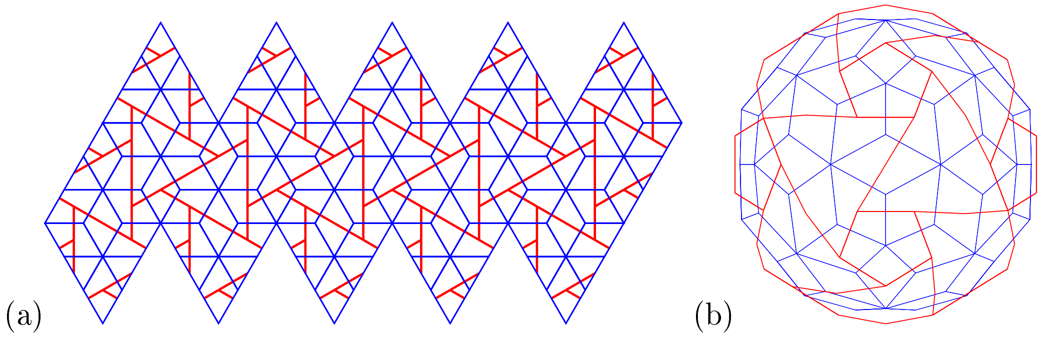

Figure 1.6: Planar representations of the MS2 capsid with locations of RNA density shown in red. (a) The virus represented as a net with dimeric building blocks shown as rhombs. (b) A view along a two-fold axis of symmetry.

symmetry axes cannot be distinguished. This results in the information on the actual organisation of the RNA in a single particle being lost. The RNA layout in any particular virus is also likely to be different, which will add to the averaging effect. It is estimated that 90 % of the virus RNA has corresponding density in the cryo-EM reconstructions and that the RNA paths in the outer shell of the virus are likely to be single-stranded (Toropova et al., 2008).

1.3 Hamiltonian path model for RNA virus formation

The biological data for the MS2 virus contains many constraints for a model of RNA virus formation to take into account. The major constraint is that a stem-loop from the RNA is required in each A/B dimer location, in order to trigger the required allosteric conformer switch in the dimer. An average of the possible pathways between these A/B dimers, that the single-stranded RNA may take, has also been shown by experiments and corresponds to the red polyhedral cage shown in figure 1.6(b), which is a diagrammatic representation of the data in figure 1.5. Further to this, the genomic RNA must be packaged within the MS2 virus such that there are no knots formed. This is because any knot formation would likely impact on the disassembly of the virion and certainly any transcription of the RNA. An RNA path that meets these constraints would be one that reaches every vertex, and therefore A/B dimer, of the red cage shown in figure 1.6(b) and 1.7(a). This 3D polyhedron of RNA may also be shown as the flattened Schlegel diagram in figure 1.7(b). This Schlegel diagram in mathematical terms may be thought of as an undirected graph. It is possible to find connected paths on this graph that reach every vertex precisely once. These correspond to "RNA paths" that reach each A/B dimer and are single-stranded along the edges. Such paths on this graph are called Hamiltonian paths. This Hamiltonian path requirement is here termed the "dimer switching model of capsid assembly" (Dykeman and Twarock, 2010) (Grayson et al., 2007). This is due to the allosteric switching of the RNA, which is required for the A/B dimers to bind the RNA. The later kinetic modelling depends on this assumption that A/B dimers have to bind the RNA in order to acquire the correct conformation to bind within the growing capsid.

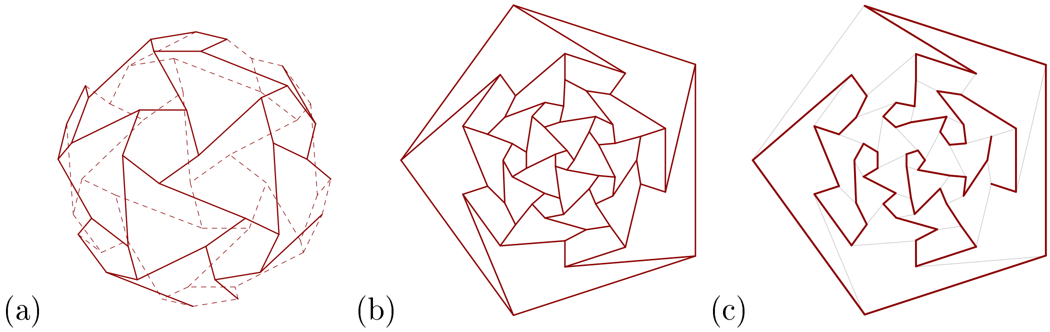

Figure 1.7: (a) A 3D polygon representing the possible RNA pathways from the cryo-EM data of the MS2 virus. (b) The Schlegel representation of the 3D polygon. (c) A Hamiltonian path on the graph of the possible RNA pathways.

Hamiltonian paths have also been used to model the RNA cryo-EM density of pariacoto virus (Rudnick and Bruinsma, 2005), shown in figure 1.8 (Tang et al., 2001). Reconstructions of pariacoto virus show a dodecahedral layout of double stranded RNA density. To replicate this structure Rudnick suggested

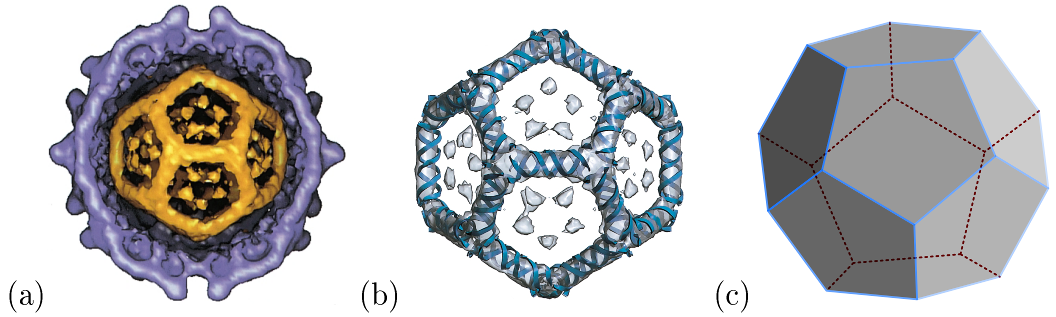

Figure 1.8: (a) Three Dimensional reconstruction of pariacoto virus from (Tang et al., 2001). (b) The inner double stranded RNA layout also from (Tang et al., 2001). (c) The data shows that the outer shell of the RNA density has the shape of a dodecahedron.

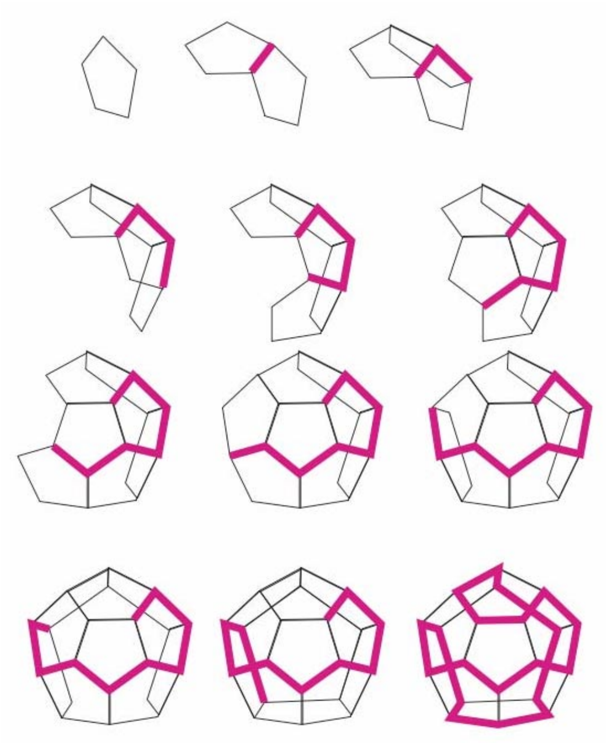

Figure 1.9: The assembly of a dodecahedral shell of RNA following the Hamiltonian path idea that would result, when icosahedrally averaged, in similar density to pariacoto (Rudnick and Bruinsma, 2005).

that assembly pathways follow Hamiltonian paths such as the one shown in figure 1.9. It has also been shown that encapsidation of RNA by pariacoto virus is not dependant on the RNA sequence (Johnson et al., 2004). However, the cognate genome might result in more efficient and faster assembly. For the MS2 virus it has been shown that the assembly depends closely on the RNA sequence. In a paper by Horn *et al.* (Horn et al., 2006) it was shown that the MS2 virus is able to discriminate between its own RNA and the genetically very similar $Q\beta$ bacteriophage RNA. This indicates the importance of the RNA in the assembly of the MS2 virus. This reason, and because so much is known about MS2 assembly, is why the MS2 virus is the basis for the subsequent RNA virus assembly model.

1.3.1 RNA connectivity

There is a slight complication in the Hamiltonian path model in that the Hamiltonian path must be contiguous on the outer shell of the cryo-EM RNA density. However from the density we can see that there are double stranded transitions between this outer shell and the inner shell at the 5-fold axes (see figure 1.10) (Toropova et al., 2008). The figures 1.10(b) and (c), reproduced from (Toropova et al., 2008), offer two possible explanations for the density. The first explanation is that the single stranded RNA dips to the inner shell and returns back to the same 5-fold axis. Alternatively, (see 1.10(c)) the RNA may return to the outer shell at a different 5-fold axis, base-pairing as it does so. An efficiency argument due to the speed of capsid assembly suggests that the RNA returns back to the same five-fold axis, because otherwise finding the correct axis to return at would be a slow process (Toropova et al., 2008). Assuming continuity in the Hamiltonian path is therefore a good representation of the process.

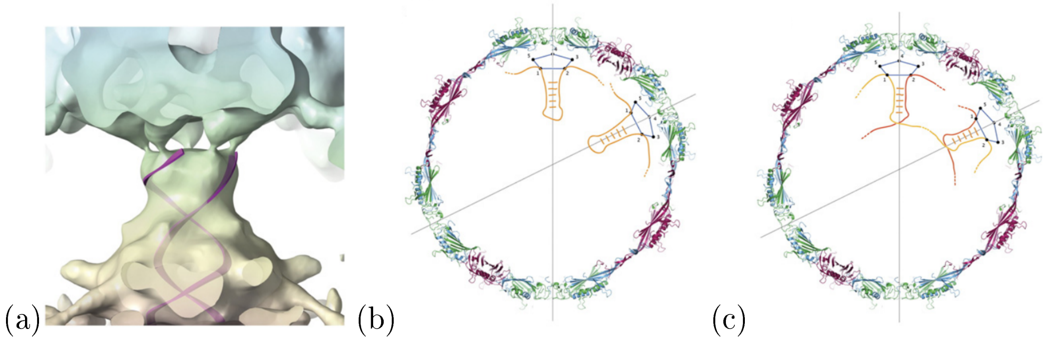

Figure 1.10: (a) A close up of the MS2 cryo-EM density shown in figure 1.5. This shows the double stranded RNA transitions that occur between the two shells of RNA. (b) and (c) show two possible explanations for this RNA density, (a) shows the RNA transitioning to the inner shell and returning at the same axis, (b) shows the RNA returning at a different five-fold axis.

1.4 Introduction to STMV

Satellite tobacco mosaic virus (STMV) exhibits similar assembly properties to the MS2 virus, such as having dimers that must bind the RNA in order to assemble (Larson and McPherson, 2001). STMV is also a very well known model virus and its crystal structure has the highest resolution of any virus (Larson et al., 1998). The reason for introducing STMV is that it is a much smaller T=1 (Casper and Klug, 1962) virus, consisting of only 30 dimeric building blocks. This smaller system is later much easier to model than the full MS2 capsid of 90 dimers. A possible layout of the STMV RNA on a net of the virus is shown in figure 1.11 (Larson and McPherson, 2001).

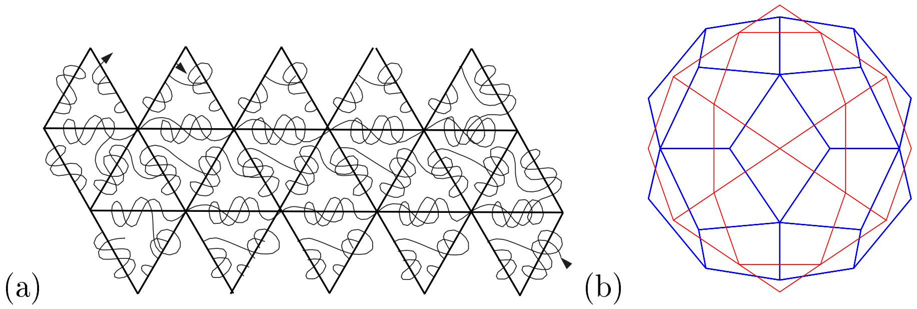

Figure 1.11: (a)A possible RNA path shown superimposed on an icosahedral surface representing STMV (from (Larson and McPherson, 2001)). (b) The 3D tiling to show the location of the building blocks. The STMV virus capsomeres must bind the RNA in order for the virus to assemble.

1.5 Modelling the self-assembly of viral capsids

Viruses spontaneously self-assemble within their host cells and do so with high fidelity. Molecular self-assembly processes are usually described as nano-scale components coming together to form larger structures with a higher degree of order. To form these higher order structures, the process is usually required to be reversible, in order to correct mistakes by removing building blocks. To achieve this reversibility the interactions between individual building blocks are required to be weak, this usually means non-covalent interactions. When large amounts of backward reactions are possible, the self-assembling system is normally at or near equilibrium. The assembly process would then be driven by only a relatively small reduction in the free energy of the final structure.

Many varying models of virus assembly have been proposed and developed to answer both viral and nano-technology assembly questions. The basis for these models may be broadly separated into models that use more statistical approaches and models that use coarse-grained molecular dynamics (MD). The prominent examples for both approaches are now introduced and discussed. The comparisons will show the reasons for using the Zlotnick model of virus assembly (Zlotnick, 1994) in the later chapters.

1.5.1 Molecular dynamics approaches

In molecular dynamics approaches, building blocks are simulated dynamically in 3D space. The rules placed on how these building blocks may bind to each other determines the intermediates formed and the assembly pathways. There is also a built-in spatial and time dependence in any simulation due to the simple rules describing the building blocks and environment. However, the computation of large numbers of viral proteins moving and colliding in 3D space is very computationally intensive. This problem requires significant amounts of coarse graining to simplify the amount of calculations that have to be performed.

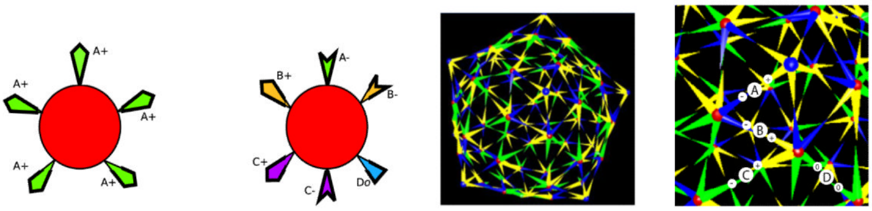

Figure 1.12: The building blocks and resultant capsids that assemble using the local rules approach. Reproduced from Kumar *et al.*(Kumar and Schwartz, 2010).

Early viral MD simulations were conducted in 1998 by (Schwartz et al., 1998) and continued with (Zhang and Schwartz, 2006) and (Kumar and Schwartz, 2010). These MD simulations are based on local rules. These local rules define specific distances and relative angles that the building blocks require before the are able to bind one another. The building blocks and resulting structure from (Kumar and Schwartz, 2010) are shown in figure 1.12. The building blocks may be thought of as spheres with sticky arms, which are able to bind other spheres if the correct arms match up. From these very local interactions large capsid-like structures assemble readily. However, due to the rules in place, there is little option for any imperfect assembly. The advantage of using the local rules to govern the interactions between building blocks is that they greatly reduce the computational requirements. This also allows for more computation to be used to simulate the movement of the building blocks, in the case of (Schwartz et al., 1998) relatively complex Brownian motion was simulated.

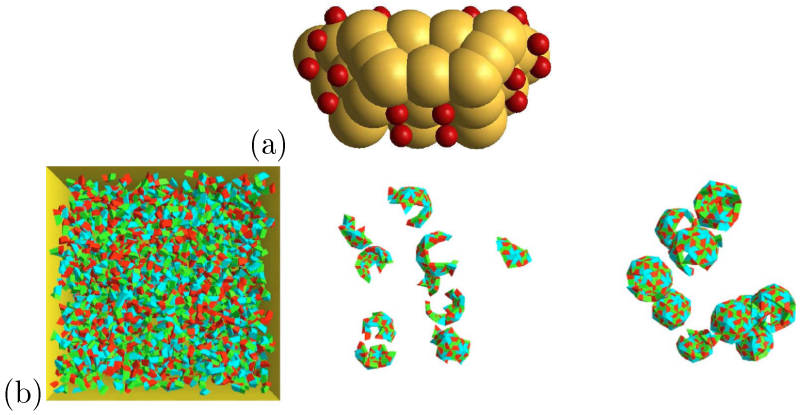

Figure 1.13: (a) A building block that forms a T=1 capsid from (Rapaport, 2012). (b) Three time points of a simulation from (Rapaport, 2004), first there is only free capsomere, then partial intermediates and finally complete capsid.

An interesting progression of MD simulations that have less restrictive rules governing subunit interactions are by Rapaport, these include (Rapaport, 2012) and (Rapaport, 2004). In these papers various trapezoidal shaped building blocks assemble into their respective capsids. An example of a building block that forms a T=1 capsid is shown in figure 1.13(a) (Rapaport, 2004). Figure 1.13(b) shows a simulation from (Rapaport, 2004) at 3 time points, where first there is only free capsomere, then partial intermediates and finally complete capsid. Typically, these simulations contain about 1000 building blocks at the start of the simulation. This is due to the computational limitations of the MD approach. Further coarse-graining was also required for computational tractability. An example of the level of coarse graining required in Rapaport's 2004 paper (Rapaport, 2004) is that the viruses were modelled in a vacuum with random ballistic movements. This more simplistic motion compared to (Schwartz et al., 1998) is to compensate for the more complex subunit interactions. A further simplification was to model the binding reactions as irreversible. To counter this irreversibility, the partially built and malformed capsids were arbitrarily broken up after a certain time period.

Rapaport's later paper (Rapaport, 2012) included a solvent in the simulation and larger building blocks that undergo reversible reactions. These additions resulted in improved sigmoidal kinetics but came at a large cost in computational tractability, requiring 20 times more computational power than the (Rapaport, 2004) paper and the use of more powerful computers with co-processors. Even using this larger computing capability each simulation only formed on the order of 36 viral capsids (Rapaport, 2012).

Different coarse-graining techniques have also been applied to the movements of more complex building blocks. This includes the use of determining the movements stochastically such as in (Johnston et al., 2010) and the use of Newtonian dynamics such as in (Hagan and Chandler, 2006). However these techniques, like all the MD approaches, are also limited to similarly low numbers of viral particles.

So far, only assembly models that consider self-assembling virus protein capsomeres have been discussed. There have also been attempts to model to model RNA using molecular dynamics. Initially the RNA mediated assembly was modelled as protein capsomeres assembling around a charged sphere (Elrad and Hagan, 2008). However, to model MS2 assembly more details regarding the RNA structure need to be taken into account. This is in order to allow the affects of the RNA path to impact on the assembly in accordance with the Hamiltonian path model. Such a model should allow for the genome to spontaneously form a Hamiltonian path as a result of the assembly rules. Two MD papers that model flexible polymer encapsidation are by Michael Hagan's group; (Kivenson and Hagan, 2010) and (Elrad and Hagan, 2010). These papers are primarily concerned with effects of polymer length. The first paper models the assembly of cube shaped capsids around a theoretical polymer and discusses the effects of nucleation rates and polymer length. A typical assembly pathway is shown in figure 1.14 (Kivenson and Hagan, 2010). The cube capsids formed in this paper have no limits imposed on their size, unlike the icosahedral geometry of an actual virus. The second paper, (Elrad and Hagan, 2010), models a much more realistic situation and that has also been inspired by the MS2 virus. The building blocks of this simulation are shown in figure 1.15. The design of these building blocks allows them to bind to each other as well as to an RNA polymer. In a fully assembled capsid, 20 of the building blocks will form an icosahedral shape. However, again, this paper is more concerned with polymer length and interaction energy, considering virus formation, or lack of formation, as more of a binary condition. This limitation is again due to only being able to simulate low numbers of building blocks. A phase diagram of capsomere-polymer affinity against polymer length reproduced from (Elrad and Hagan, 2010) is shown in figure 1.16. This diagram shows the range of values in which successful capsid formation occurs within the observed time.

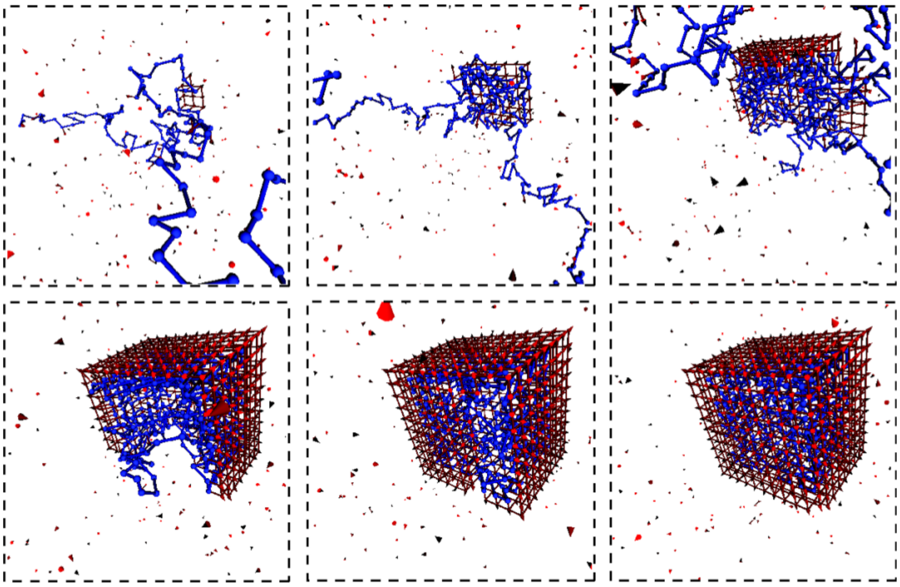

Figure 1.14: Six time points of polymer encapsidation by building blocks that are able to form cube shaped capsids, reproduced from (Kivenson and Hagan, 2010).

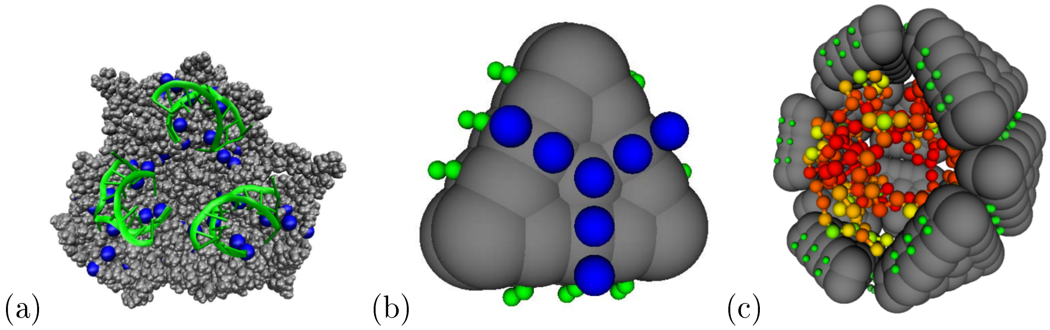

Figure 1.15: (a) shows a trimer of MS2 dimer proteins that inspired the building block design shown in (b) (Elrad and Hagan, 2010). This trimer of dimers configuration is from a crystal structure in (Valegård et al., 1997) that shows the C/C dimer binding a TR stem-loop. Later results have shown that these C/C dimers do not bind TR stem loops during efficient assembly (Morton et al., 2010), (Knapman et al., 2010). (c) A view of a complete capsid with half the capsomeres removed to show the internal RNA (Elrad and Hagan, 2010).

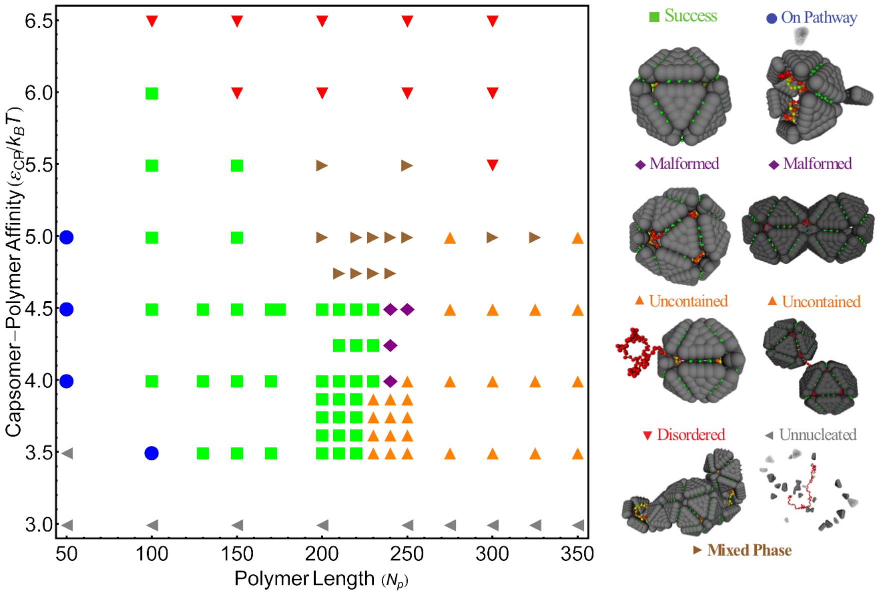

Figure 1.16: A phase diagram showing the typical assembly product as a consequence of polymer length and polymer-capsomere contact energy. Reproduced from (Elrad and Hagan, 2010).

1.5.2 Statistics based approaches

Molecular dynamics attempts to directly replicate the real world physics during virus assembly computationally. An alternative are more statistical approaches that represent the viral components more abstractly. The advantage of this is that the following statistical techniques are able to capture much more of the parameter space of viral assembly. This is achieved by taking into account many more viral capsids and building blocks than would be possible in an MD calculation. However in achieving this the statistical models tend to have much simpler representations of the building blocks and physics.

A popular statistical technique is the use of potential energy surfaces pioneered by (Wales, 2005). In (Wales, 2005) and (Fejer et al., 2009), Wales investigates all the possible capsomere orientations for virus models constructed from pentagonal subunits. The energies of these capsid configurations were then measured to create potential energy surfaces in parameter space. An example of a potential energy surface is shown in figure 1.17(a). This shows a funnel of local minimum energies to the minimum energy at the bottom. Figure 1.17(b) shows the same potential energy surface represented as a disconnectivity graph (Becker and Karplus, 1997), where transition points between minimal energies are shown as branches in the graph. The virus capsid potential energy surface investigated by Wales follows this funnel pattern and is shown in

1.17(c). This funnel energy surface is similar to those seen in models of protein folding (Chiti and Dobson, 2009). The use of the largely thermodynamic considerations in creating the potential energy surfaces results in convincing virus assembly pathways that follow intermediates with high numbers of bonds. However kinetic effects such as competition for building blocks between different pathways are ignored. The introduction of RNA into this model would also be very complicated and would require a large number of extra parameters to model the RNA polymer shape.

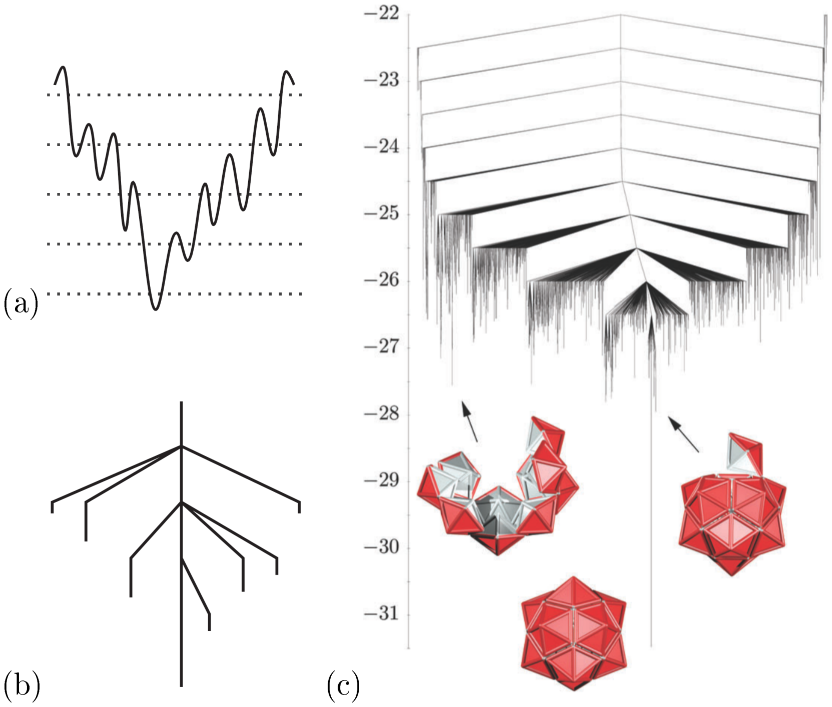

Figure 1.17: (a) A one dimensional potential energy surface and (b) its corresponding disconnectivity graph. (c) The disconnectivity graph corresponding to a T=1 capsid, the global minimum energy is at the bottom of the graph. Reproduced from (Wales, 2005) and (Fejer et al., 2009).

Further statistical approaches to virus assembly are able to take advantage of the fact that self-assembly is usually at or close to equilibrium. At thermodynamic equilibrium, the concentrations of any intermediates in the building process will be related to their Gibbs free energy. The Gibbs free energy is made up of the enthalpy contribution from the bonds formed and the entropy term. By counting the number of bonds in an intermediate and the number of ways to form an intermediate it is possible to model the Gibbs free energy using Boltzman statistics (Endres et al., 2005). Once the possible assembly intermediates and the reactions between them have been determined, the Boltzman statistics result in a probability of formation for each intermediate. Master equations are one way of using the Gibbs free energy to share out the initial concentration of the building block into the equilibrium concentrations of the other intermediates (Keef et al., 2005), (Keef et al., 2006). In the case of (Keef et al., 2006), equations describing the Gibbs free energy of each intermediate were recursively combined to determine each intermediate's probability and therefore equilibrium concentration. An example pathway of viral assembly from (Keef et al., 2006) is shown in figure 1.18(a).The construction of a master equation will be demonstrated in the next chapter for a 12-step pathway of virus formation. The advantages of master equations is that they are relatively easy to compute for small systems. However, as the number of possible intermediates increases, master equations are not solvable explicitly (Hemberg et al., 2006). Like the potential energy surface technique, it is also not possible to model the kinetics of assembly (outside of thermodynamic equilibrium) using master equations.

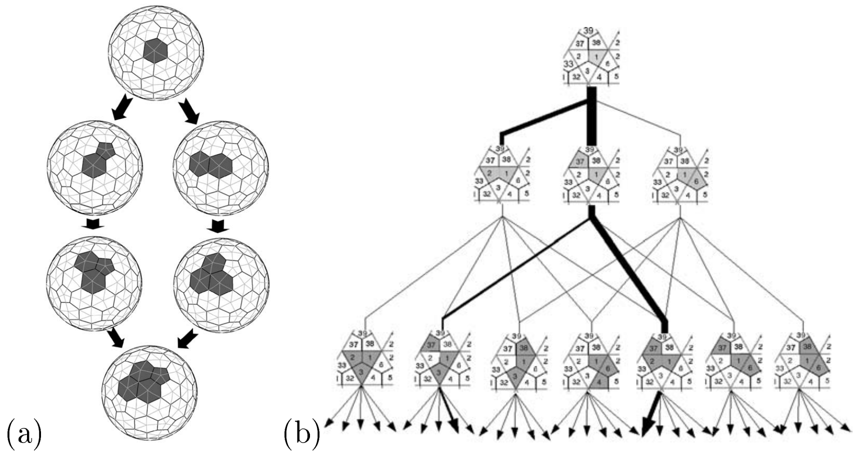

Figure 1.18: (a) An example of a viral assembly pathway reproduced from Keef *et al.*(Keef et al., 2006). (b) The beginning of the network representing a T=1 capsid reproduced from Hemberg *et al.*(Hemberg et al., 2006)

A similar statistical technique is to use a Gillespie algorithm (Gillespie, 1977) to investigate virus assembly. To use a Gillespie algorithm for virus assembly, first a network of the possible reactions between the intermediates is constructed, similar to the master equation approach (Keef et al., 2006). Then on this network, the reactions between the intermediates are modelled as discrete steps. The probability of a reaction happening depends on the bonds formed and the network topology. An extended Gillespie algorithm was used in (Hemberg et al., 2006) to model the assembly network shown in figure 1.18(b). This Gillespie algorithm (Hemberg et al., 2006) started with 1000 monomers and modelled the movement of each monomer through the network individually. If Gillespie algorithms use the same probabilities as the master equation approach, the two techniques should achieve the same distribution of material. However, Gillespie algorithms have advantages over master equations in that they are able to approximate the master equation using far less computation, albeit at a cost of accuracy.

Zlotnick's model of virus assembly uses the same equilibrium considerations as the master equation approach, but in addition is able to gain kinetic insights (Zlotnick, 1994). Again, the first step in this model is to construct a network of intermediates. Rate equations are then created for each reaction between intermediates in this network. These rate equations use putative forward rates for diffusion-limited protein binding and backward rates based on the Boltzman statistics. Since this model system was eventually chosen to simulate the MS2 assembly, a full account of Zlotnick's model is presented in the next chapter.

The most interesting feature of these variety of molecular dynamics and statistics based simulations is how much they have in common. For instance, several quite different simulations (Hagan and Chandler, 2006) (Johnston et al., 2010) (Kumar and Schwartz, 2010) (Rapaport, 2012) (Endres et al., 2005) (Hemberg et al., 2006) all show sigmoidal assembly kinetics. A further common theme across the self-assembly simulations is that the concentrations of partially built capsid intermediates are very low (Kumar and Schwartz, 2010), (Hemberg et al., 2006), (Rapaport, 2012), (Endres et al., 2005). Hysteresis is also a theme observed in the reactions building up to capsid (Kumar and Schwartz, 2010) (Rapaport, 2012). Finally, many of the simulations show kinetic trapping occurs when the building block concentration is diminished (Kumar and Schwartz, 2010), (Rapaport, 2012), (Endres et al., 2005). The reasons for these assembly behaviours will be discussed in the next chapter as Zlotnick's model is able to capture all these behaviours.

1.5.3 Choosing an assembly model

As we have seen, there are a number of choices of model frameworks that could be chosen to model MS2 virus assembly. The more advanced model frameworks exhibit the sigmoidal kinetics, paucity of intermediates, kinetic traps and hysteresis, seen during *in vivo* experiments. This still leaves a choice of whether to use a statistics based or molecular dynamics approach. The most obvious difference between these two approaches is that the statistics approaches use a precomputed network of assembly intermediates. This limits the building blocks to the positions they would be within the fully assembled capsid. This is because the computation of networks with other than the perfect geometry of the capsid would create unfeasibly large assembly networks.

In reality, the interactions in the surfaces of the virus proteins are quite large and complex. This complexity in the binding interfaces of the proteins helps ensure they bind in the correct orientation (ElSawy et al., 2010). Using small simple building blocks such as in early MD papers e.g. (Rapaport, 2004) and (Wales, 2005), see lots of malformed capsids. Increasing the size and complexity of the building blocks in both the later Rapaport (Rapaport, 2012) and Wales (Fejer et al., 2009) approaches had the effect of greatly reducing the malformed capsids. This result suggests that having the constraint that the geometry of the capsid proteins limits the assembly intermediates to those considered in the assembly networks is a valid assumption biologically. Since the triangular building blocks in Hagan's RNA model (Elrad and Hagan, 2010) are relatively small with simple interfaces, this could explain the large numbers of malformed capsids observed in the results.

The main disadvantage of using MD techniques is the computational power required. This practical constraint requires lots of assumptions in order to model the systems in a reasonable amount of time.

In general, by only modelling a small number of building blocks and assembling small numbers of virus particles all the molecular dynamic simulations have a problem in covering the parameter space of virus assembly. As a result of this many possible virus intermediates never occur over the time frame of the simulation. Zlotnick's approach is able to cover the full parameter space in that every intermediate in the network of assembly intermediates will have a concentration.

Using a precomputed network also has the further advantage that is very easy to characterise the intermediates, because they have defined configurations. The reactions in Zlotnick's approach are modelled continuously so that the conversion of smaller sized intermediates to larger ones is a continuous process. This exchange of material may also be easily characterised quantitatively and tracked, which is much harder to do in MD simulations given their discrete events.

A continuing debate (McPherson, 2005) in the RNA virus assembly field is whether all the capsomeres bind to positions on the RNA, and then the RNA folds and condenses to form capsid, or whether capsomeres bind one at a time to the RNA and growing capsid edge (see figure 1.19(a) and (b)). These two possibilities are likely to be protein concentration dependant, with high protein concentrations favouring the saturation of the RNA and low protein concentrations favouring cooperative single capsomere additions. Looking at single capsids assembling in the RNA polymer model (Kumar and Schwartz, 2010) based on a fixed protein concentration, subunit-polymer association energy and association rate were found to distinguish between the two assembly pathways (see figure 1.19(d) and (e)). The largest ranges of parameters in this RNA MD model favoured the sequential addition of capsomeres, while only a narrow range of parameters made the en masse association of capsomeres to the RNA more efficient. It has been suggested that the STMV virus starts binding capsomeres to its RNA genome as soon as the RNA is transcribed by the RNA replicase (Larson and McPherson, 2001) (see figure 1.19(c)). This immediate binding to the genome of STMV favours the sequential addition scenario. For these reasons the later assembly models of both the MS2 and STMV pathways are assumed to be through single, sequential capsomere additions.

1.6 Conclusions

In conclusion we have seen that there is a large amount of evidence for a Hamiltonian path model of MS2 virus assembly. This main evidence is that each A/B dimer must bind the RNA in accordance with the dimer switching model (Dykeman and Twarock, 2010) and that there are defined paths of RNA between these dimers on the inner surface of the capsid proteins (Toropova et al., 2008) (Van Den Worm et al., 2006). This is sufficient information from which to create a model of virus formation for the small single stranded RNA viruses. Further biological knowledge of MS2 and data on which to validate the model for this particular virus will be introduced in chapter 6.

Various model frameworks that could be used to simulate MS2 assembly have been discussed. Models that only investigate the thermodynamic equilibrium have been discounted in favour of models that show realistic kinetic behaviour. One set of the remaining models are the molecular dynamics simulations. These models, although very interesting, are computationally limited in the size and number of virus particles that can be simulated. The large 90-mer of the MS2 capsid, along with its RNA genome means that no current MD simulation could hope to characterise the full parameter space. The model chosen, pioneered by Zlotnick, has been shown to exhibit complex kinetic be-

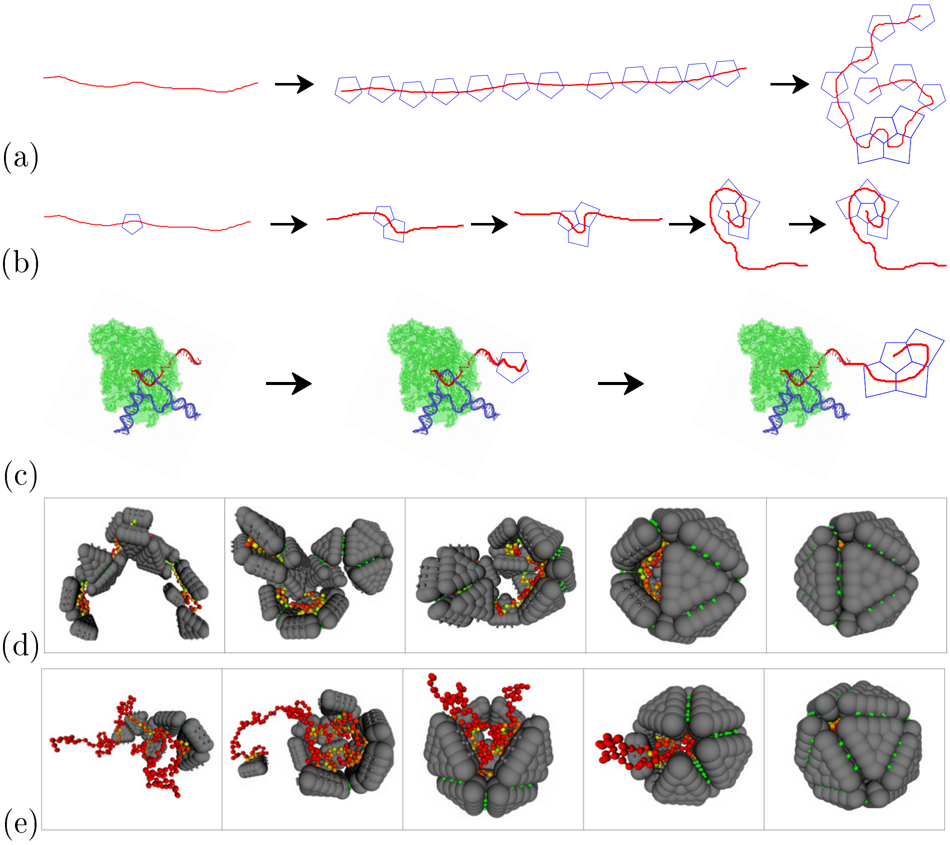

Figure 1.19: (a) An assembly pathway where all the capsomeres bind to the RNA, which then folds into the capsid shape. (b) The alternative scenario where capsomeres bind sequentially to the RNA and the previous capsomeres. (c) The proposed STMV assembly scenario where capsomeres bind to the RNA as it is transcribed from the replicase of the TMV virus. (Larson and McPherson, 2001). (d) and (e) are pathways from the RNA MD model by Hagan *et al.* that correspond to scenarios (a) and (b) respectively (Elrad and Hagan, 2010).

haviours for protein capsomeres in the absence of RNA (Endres et al., 2005). Additionally Zlotnick's model can consider all the assembly intermediates in a network and assign each one a particular concentration. Zlotnick's model is also able to consider virus assembly over large periods of time and at a very large range of parameters efficiently. Being able to investigate a large range of parameters, such as bond strength, is necessary to show all the possible assembly behaviour, not only *in vivo* but also at the more extreme conditions often used *in vitro*.

Zlotnick's model has also successfully been used to replicate the assembly behaviour with multiple copies of the 19 nucleotide TR RNA sequence (Morton et al., 2010). Finally Rudnick and Bruinsma, who used Hamiltonian paths to describe the RNA cryo-EM density of pariacoto virus (Rudnick and Bruinsma, 2005), suggest that limiting the assembly of the pentagonal building blocks using the Hamiltonian path would be, “A natural extension of the Zlotnick model”. A full account of Zlotnick’s model and its progression in complexity is given in the next chapter. Chapter 3 will show how this model is extended to follow the constraints imposed by the Hamiltonian path model of RNA virus assembly.

Chapter 2

The Zlotnick Virus Assembly Model

2.1 Original Zlotnick Equilibrium Model

The first description of Zlotnick's assembly model was in a 1994 paper entitled "To Build a Virus Capsid" (Zlotnick, 1994). It is a protein only model using a simple assembly scenario to illustrate equilibrium assembly behaviour. The assembly scenario used is that of a single pathway through 12 capsomeres to form a dodecahedron. Here a capsomere refers to a protein subunit that is the building block of the virus capsid. The 12 capsomere pathway is shown in figure 2.1. This network contains only the most energetically favourable intermediate for each size, i.e. the one with the most inter-capsomere contacts. With this sequence of assembly intermediates we have 11 forward reactions and 11 backward reactions. The forward reactions are second order, since they depend on the concentration of the previous intermediate and also that of the free capsomere (intermediate 1), while the backward reactions are first order as it is simply a large intermediate breaking apart. Zlotnick's equilibrium model assigns rates to the equations in this linear reaction scheme.

For the reaction to form a particular intermediate, denoted as (n), Zlotnick considers its growth from the previous intermediate $(n-1)$ and free capsomere (1) as shown in equation (2.1). To determine the concentration change for a particular intermediate, (n), equation (2.2) must be constructed, here the concentration of intermediate (n) is denoted by $[n]$. The first part of this equation $(k_f[n-1][1])$ is the forward reaction of the previous intermediate reacting with the free capsomere to increase the concentration of (n). The second part $(k_b[n+1])$ is increase in concentration of (n) due to the backward reaction of

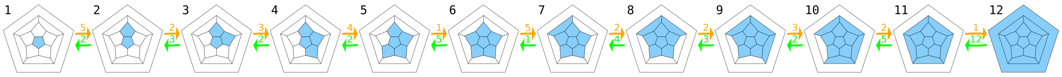

Figure 2.1: A single assembly pathway for assembly of a dodecahedral shape adapted from Zlotnick *et al.* (Zlotnick, 1994). This pathway contains the most energetically favourable intermediates, i.e. those with the largest numbers of capsomere-capsomere contacts at every step of capsomere addition. Each dodecahedron is represented as a Schlegel diagram to show the face connectivity and the pentagon at the back has also been expanded to show its presence. The numbers in orange are the “build up” symmetry factors and the green numbers “build down”.

the lager intermediate breaking apart to give (n) and free capsomere. Thirdly, $(k_f[n][1])$ is the forward rate of (n) gaining a free capsomere and becoming $(n+1)$. Finally, there is the backward rate of (n) itself breaking apart $(K_b[n])$.

$$(n-1)+(1) \rightleftharpoons (n) \tag{2.1}$$

$$\frac{d[n]}{dt} = k_f[n-1][1] + k_b[n+1] - k_f[n][1] - k_b[n] \tag{2.2}$$

In order to model the reaction kinetics, it is necessary to assign numbers to the forward (k_f) and backward rates (k_b) of equation 2.2. To do this, Zlotnick has based the model around the Arrhenius equation shown in equation (2.3). This formula describes the temperature dependence of the rate constant k in a reaction.

$$k = Ae^{(\frac{-E_a}{RT})} \tag{2.3}$$

In the Arrhenius equation A is the attempt frequency factor, $-E_a$ is the activation energy of the reaction, R is the gas constant ($8.314\ JK^{-1}mol^{-1}$) and T is the temperature at which the reaction takes place (set to 298 K). $(\frac{-E_a}{RT})$ gives the percentage of reactants that have the required energy to complete a reaction, and the attempt frequency encodes how many of the reactions are attempted. Multiplied together, these give the number of reactions that actually occur per second. The Arrhenius equation is used to model both the 2^{nd} order forward reaction and the 1^{st} order backward reactions.

The activation energy in the Arrhenius equation is difficult to estimate, but the bond strengths in a particular intermediate can be estimated relatively easily. By using two combined Arrhenius equations, it is possible to derive an equation relating the difference in bond strength to the reaction rate. By ignoring the activation energies at this point we are left with an equilibrium model. To derive the corresponding equation using the difference in bond strengths, first the Arrhenius equations for the forward reaction rate k_f and backward reaction rate k_b are established:

$$k_f = A_1 e^{(\frac{-F_a}{RT})} \tag{2.4}$$

$$k_b = A_2 e^{(\frac{-B_a}{RT})} \tag{2.5}$$

Now let F_a be the activation energy of the forward reaction and B_a the activation energy for the backward reaction. Then the Arrhenius equations can be combined to give the quotient $\frac{k_f}{k_b}$, and then be rearranged, as follows:

$$\begin{aligned}
\frac{k_f}{k_b} &= \frac{A_1 e^{(\frac{-F_a}{RT})}}{A_2 e^{(\frac{-B_a}{RT})}} \\
\frac{k_f}{k_b} &= \frac{A_1}{A_2} e^{(\frac{-F_a}{RT})} e^{(\frac{B_a}{RT})} \\
\frac{A_2}{A_1} k_f &= k_b e^{(\frac{-F_a+B_a}{RT})} \\
\frac{\frac{A_2}{A_1} k_f}{e^{(\frac{-F_a+B_a}{RT})}} &= k_b
\end{aligned}$$

This yields the following expression for the backward rate:

$$k_b = \frac{A_2}{A_1} k_f e^{(\frac{F_a - B_a}{RT})} \tag{2.6}$$

With reference to the energy diagram for this reaction in figure 2.2, $F_a - B_a$ in (2.6) corresponds to the difference in contact energy C_e (bond energies) of the two intermediates: $-C_e := F_a - B_a$. The forward and backward attempt frequencies (A_2 and A_1) are assumed to be the same and therefore cancel out. This leaves only the multipliers to the attempt frequencies (S_2 and S_1) that come from the symmetry of the intermediates and are described next.

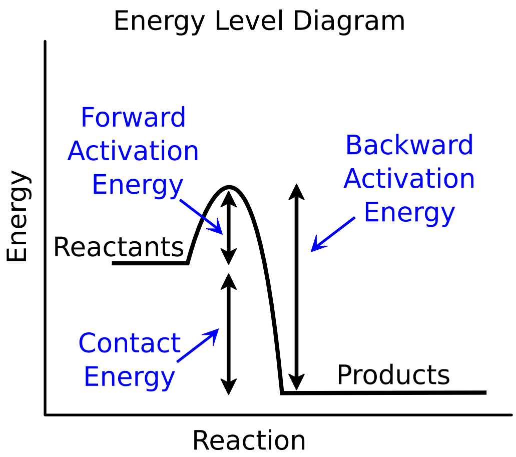

Figure 2.2: Energy diagram for intermediate capsomere addition.

2.1.1 Symmetry factors

The symmetry factors in this model arise from the symmetry of the intermediate and, for forward reactions, the symmetry of the incoming capsomere. Since the incoming capsomere for the dodecahedron is the shape of a pentagon, its symmetry is always 5. The symmetry of the intermediate for a particular reaction may be more easily thought of as the number of ways a capsomere can bind or break off to give the product intermediate. For example, in figure 2.1 between intermediates 2 and 3 there are two ways to add a capsomere to form intermediate 3, and three ways to remove a capsomere to form intermediate 2, hence $S_1 = 2$ and $S_2 = 3$. These symmetry factors may also be thought of as adding to the entropy term of the Gibbs free energy of an intermediate.

The final form of the equation is:

$$k_b = \frac{S_2}{S_1} k_f e^{(\frac{-C_e}{RT})} \tag{2.7}$$

This equation relates C_e to k_f and k_b. Therefore, it permits to choose a forward rate and have the appropriate backward rate determined by the number of capsomere contacts. Zlotnick follows this procedure and chooses a k_f of 10^8 $\mathrm{M}^{-1}\mathrm{s}^{-1}$ for a single protein binding event because it is, “a value that is close to the diffusion limited association of two proteins” (Zlotnick, 1994). Choosing a forward rate that is diffusion limited is convenient in that it applies to every possible forward reaction equally (modulo the symmetry factors). It is justified by the assumption that the coming together of an intermediate and free capsomere is likely to be the rate-limiting step in their binding. When the symmetry factor is 1, e.g. the reaction of 5 to 6 in figure 2.1, there is only one place an incoming capsomere may bind. This single symmetry factor, along with the symmetry factor of 5 for the incoming capsomere, and the k_f of 1 $\times 10^8$ M^{-1}s^{-1}, results in a forward rate of 50 $\times 10^8$ M^{-1}s^{-1}. When the intermediate has a symmetry factor of five the rate is as high as 250 $\times 10^8$ M^{-1}s^{-1}. Although this is an extremely high rate it comes from the multiplicity in the number of possible capsomere reactions rather than one particularly quick protein binding event.

Following is a worked example to find the backward rates between intermediates 2 and 3 using a bond strength (ΔG_c) of -11.4 kJmol^{-1}. Since there are two capsomere contacts formed, C_e follows from the equation; -11.4 kJmol^{-1} $\times$ 2 $\times$ 1000, where the $\times$1000 is to convert to Joules. The forward rate of this equation is defined to be 1 $\times 10^8$ M^{-1}s^{-1}. Inserting these values into equation 2.7 with the numbers for this reaction is shown in equation 2.8. Using equation 2.8 the backward rate for this reaction is 3025 M^{-1}s^{-1}.

$$k_b = \frac{3}{2 \times 5} 1 \times 10^8 M^{-1} s^{-1} e^{(\frac{-11.4kJmol^{-1} \times 2 \times 1000}{8.314 \times 298})} = 3025 M^{-1} s^{-1} \tag{2.8}$$

2.1.2 Master Equation method

At equilibrium the concentration of the initial capsomere concentration is spread across all intermediates in the reaction scheme, i.e. across the network of assembly intermediates, proportionally to the number of inter-capsomere bonds in each intermediate and the symmetry factors. To work out the intermediate concentration of intermediate 2 in figure, 2.1 equation 2.9 may be used. This equation uses the fact that the equilibrium constant ($k_{equ.}$) is simply the ratio of k_f and k_b. With a set value of 0.88 $\times$ 10^{-6} M for the free capsomere concentration, the concentration for intermediate 2, [2] may be worked out using equation 2.10. This concentration for intermediate 2 may then be substituted into equation 2.9 to find a concentration for intermediate 3. With iterative substitution the concentration of capsid can be determined. The full table of substitutions, reproduced from (Zlotnick, 1994), is shown in table 2.1. The equation resulting from the series of substitutions is a master equation. Master equations such as this are useful where the final concentrations are dependant on the probability of occurrence of the intermediates and not time i.e. they give information on thermodynamic equilibrium, but cannot be used to compute assembly kinetics. Master equations have been used to determine the statistically dominant pathways through the reaction networks (Keef et al., 2005), (Keef et al., 2008).

The series of substitution equations relating the capsid concentration to the concentration of free capsomere may be solved to find a capsomere and capsid concentration that are equal in thermodynamic equilibrium. The initial free capsomere value of 0.88×10^{-6} M, in figure 2.1 is the determined concentration using this method. For the value of 0.88×10^{-6} M the equilibrium concentrations show that anything other than the free capsomere or capsid has a very small concentration. This is because it is only the capsid that has sufficient bonds to be stabilised. At an initial capsomere concentration of 0.44×10^{-6} M the capsid concentration is dramatically reduced. This is due to the fact that the forward reactions pushing the equilibrium towards capsid are reduced, because there are only limiting amounts of capsomere present to react. Equally, at higher concentrations such as 1.8×10^{-6} M free capsomere, there is far more capsid present at equilibrium. The initial capsomere concentration that results in capsid having the same concentration at equilibrium has been used later in the results chapters as an interesting starting point.

$$k_{equ.} = \frac{k_f}{k_b} = \frac{[n]}{[n-1][1]} \tag{2.9}$$

$$\frac{k_f}{k_b} = \frac{[2]}{[1][1]} \tag{2.10}$$

2.2 Zlotnick's Initial Kinetic Simulations

Zlotnick's first kinetic simulations based on this model also formed part of his seminal 1994 paper (Zlotnick, 1994). These kinetic simulations used the forward and backward reaction equations for the 11 assembly reactions in figure 2.1, totalling 22 (2 × 11) simultaneous equations. These simultaneous equations were then numerically integrated with respect to time to give the assembly kinetics of the linear dodecahedral pathway. Using the same bond strength of -11.4 kJmol^{-1} these simulations were conducted for initial capsomere concentrations of 13 μM, 50 μM and 500 μM. Zlotnick found that at the concentrations of 13 μM and 50 μM capsid formed swiftly with 90 % of the capsid equilibrium value being reached after 10 milliseconds. However, at concentrations of 500 μM it took 40 milliseconds to reach 90 % of the equilibrium value. This increase in the required time is due to the scarcity of the free capsomere building block, which stems from the fact that much of the

Intermediate	Concentration (M)		
1	0.44×10^{6}	0.88×10^{6}	1.8×10^{6}
2	2×10^{10}	1×10^{9}	5×10^{9}
3	4×10^{12}	3×10^{11}	2×10^{10}
4	1×10^{13}	2×10^{12}	3×10^{11}
5	5×10^{15}	2×10^{13}	5×10^{12}
6	2×10^{15}	1×10^{13}	1×10^{11}
7	2×10^{16}	3×10^{14}	4×10^{12}
8	2×10^{16}	7×10^{14}	2×10^{11}
9	6×10^{16}	2×10^{13}	1×10^{10}
10	1×10^{15}	1×10^{12}	1×10^{9}
11	1×10^{13}	2×10^{10}	5×10^{7}
12	2×10^{10}	0.88×10^{6}	3.6×10^{3}

Table 2.1: Reproduced from (Zlotnick, 1994), this table shows the intermediate equilibrium concentrations for three different initial capsomere concentrations at a ΔG_c of -11.4 kJmol^{-1}.

free capsomere assembles into smaller intermediates. With no free capsomere available to grow these intermediates to capsid, any further capsid formation is dependant on intermediates breaking apart, which is a slow process. This kinetic trapping of free capsomere in smaller intermediates is a recurring theme in this thesis, later discussed are its effects on virus assembly efficiency. We will see later how this kinetic trapping becomes more important when larger and longer networks of viral assembly are modelled, and how kinetic traps are related to the bond strength. Since Zlotnick's protein only models are later recreated in order to compare these to assembly in the presence of the genomic RNA, the fine details of the assembly kinetics will be discussed later.

2.3 Model Assumptions

The assumptions underlying Zlotnick's model are as follows: Firstly, it is an equilibrium model and therefore is more appropriate when concentrations are close to equilibrium. This is because the actual forward and backward activation energies and attempt frequencies of reactions are not taken into account. However, the resulting kinetic model does give the equilibrium concentrations expected from the Boltzman statistics. A related assumption is that there is a free choice of basic on-rate, which is taken to be the same for each reaction and is diffusion limited. This is unlikely to be the case, but it is a suitable simplification for certain regions of concentrations of free capsomere concentrations. Note that the assumption of being diffusion limited results in the maximum reaction rate possible. Moreover the forward reaction rates are likely to be much larger than the backward rates, and this ratio is maintained for a large range of basic on rates. This means that the kinetics is mostly driven by the forward rates which have been found in biological experiments.

Finally, only reactions that are described in the assembly network are permitted. This excludes the possibility of aggregates forming, misshapen capsids and rearrangements of intermediates to more favourable layouts, but comprises of only the reactions corresponding to the perfect geometry of the virus shape. Each reaction in the described network only involves addition of one free capsomere to the previous intermediate. This leaves out the possibility of intermediates being constructed of multiple capsomeres binding together. Certainly for larger viruses, this reaction may be less likely as it would rely on the associating intermediates having the correct geometry in order to fit together and it is therefore a good assumption. Likewise intermediates may only break up one capsomere at a time.

With these assumptions Zlotnick finds congruence with biological experiments, which further corroborates the validity of these choices. Zlotnick finds qualitative agreement with brome mosaic virus assembly (Cuillel et al., 1983) and also trypsin treated virus particles (Cuillel et al., 1981). Later, Zlotnick is also able to find strong correlation to experiments conducted in his own lab, see the papers titled; “A Theoretical Model Successfully Identifies Features of Hepatitis B Virus Capsid Assembly” (Zlotnick et al., 1999) and “Observed hysteresis of virus capsid disassembly is implicit in kinetic models of assembly” (Singh and Zlotnick, 2003).

2.4 Extensions Of The Model

The first significant expansion of the equilibrium model was in 2002 (Endres and Zlotnick, 2002). Here, to model the putative nucleated assembly of the viruses CCMV and HBV a nucleation step was added to the model. The nucleation step was introduced by using a rate of either 100 $\mathrm{M}^{-1}\mathrm{s}^{-1}$, 1000 $\mathrm{M}^{-1}\mathrm{s}^{-1}$ or 10000 $\mathrm{M}^{-1}\mathrm{s}^{-1}$ for the initial reaction in the linear assembly network, while the elongation rate was chosen to be $1{\times}10^{6}$ $\mathrm{M}^{-1}\mathrm{s}^{-1}$. There were four reaction networks used in this paper, the first two relate to a dodecahedron with a

12 intermediate linear pathway as shown in figure 2.1, corresponding to the most stable intermediates, and a further 12 intermediate linear pathway of intermediates that have an average number of inter-capsomere contacts. The remaining two reaction networks are similar, but contain a 30 intermediate pathway relating to the assembly of a 30-mer T=1 capsid, (see figure 2.3), in 2005 (Endres et al., 2005) Zlotnick used the model to investigate the first complete network of interactions for the dodecahedron; this network is shown in figure 2.4. This full dodecahedral network shows all 73 geometrically different intermediates and the 263 links between them. Also in this paper the full network for the 20 faced icosahedron was constructed. In this network there are 2649 intermediates and 17,241 reactions between them. These intermediates correspond to all the combinatorial ways of combining the proteins as first introduced by Wales (Wales, 1987). Not included in this 2005 paper was the nucleation step introduced in 2002 (Endres and Zlotnick, 2002). Instead, a factor μ was introduced that acts to reduce the probability of intermediates that only form one capsomere contact. This is in addition to the relatively high backward rate such an intermediate would have. The effect of this is similar to the nucleation step in that the number of intermediates during the kinetic simulation has been reduced, although arguably using the μ factor is more artificial.

The current complexity of Zlotnick's model was published in 2011 in (Moisant et al., 2010), where the complete network for the more computationally intensive 30 monomer polygon (figure 2.3) has been determined. This network consists of 2,423,212 intermediates and 26,823,095[*] bi-directional reactions between them. Interestingly, Zlotnick finds only 97,741 of these intermediates have a continuous surface of capsomeres i.e. a surface without holes. However, due to the computational cost of numerical integration only networks of up to 1124 were used. The selection of these intermediates was based on their stability and probability. The kinetics was modelled over these networks with μ and also a nucleation reaction. In order to significantly avoid kinetic traps, the nucleation reaction was set to be as low as 80 $\mathrm{M}^{-1}\mathrm{s}^{-1}$, although this was still not enough to completely avoid the kinetic traps.

[*] This figure was determined using algorithms described in the next chapter.

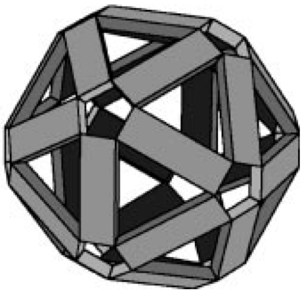

Figure 2.3: The 30 monomer polygon described in Moisant *et al.* (Moisant et al., 2010) originally taken from Keef *et al.* (Keef et al., 2005).

All of Zlotnick *et al.*'s virus assembly papers follow broadly the following similar steps:

1. Determining the combinatorially possible virus capsid assembly intermediates.
2. Placing these intermediates into a network with the edges representing the chemical reactions between them.
3. Running a kinetic simulation on this network to determine the kinetics of assembly and the concentrations of assembly intermediates.

In the next chapter the same steps will be followed to generate a co-assembly networks, i.e. networks of assembly intermediates that interact with genomic RNA.

2.5 Assembly kinetics versus thermodynamics

The concentration of any particular intermediate at any given point in time during the simulation will be a combination of the kinetics leading to the intermediate and the intermediate's equilibrium concentration. Zlotnick was the first to note this in the full dodecahedral network in 2005 (Endres et al., 2005). The extreme example of this interplay would be the free capsomere which, with strong capsomere contacts, may have an initial concentration of 8 μM and an equilibrium concentration of effectively zero. With the forward reaction rates fixed, using less negative contact energies leads to relatively quick backward reactions, and at more negative bond contact energies the backward reactions are relatively slow. For choices of contact energies close to zero, any larger intermediates formed immediately break apart and only the

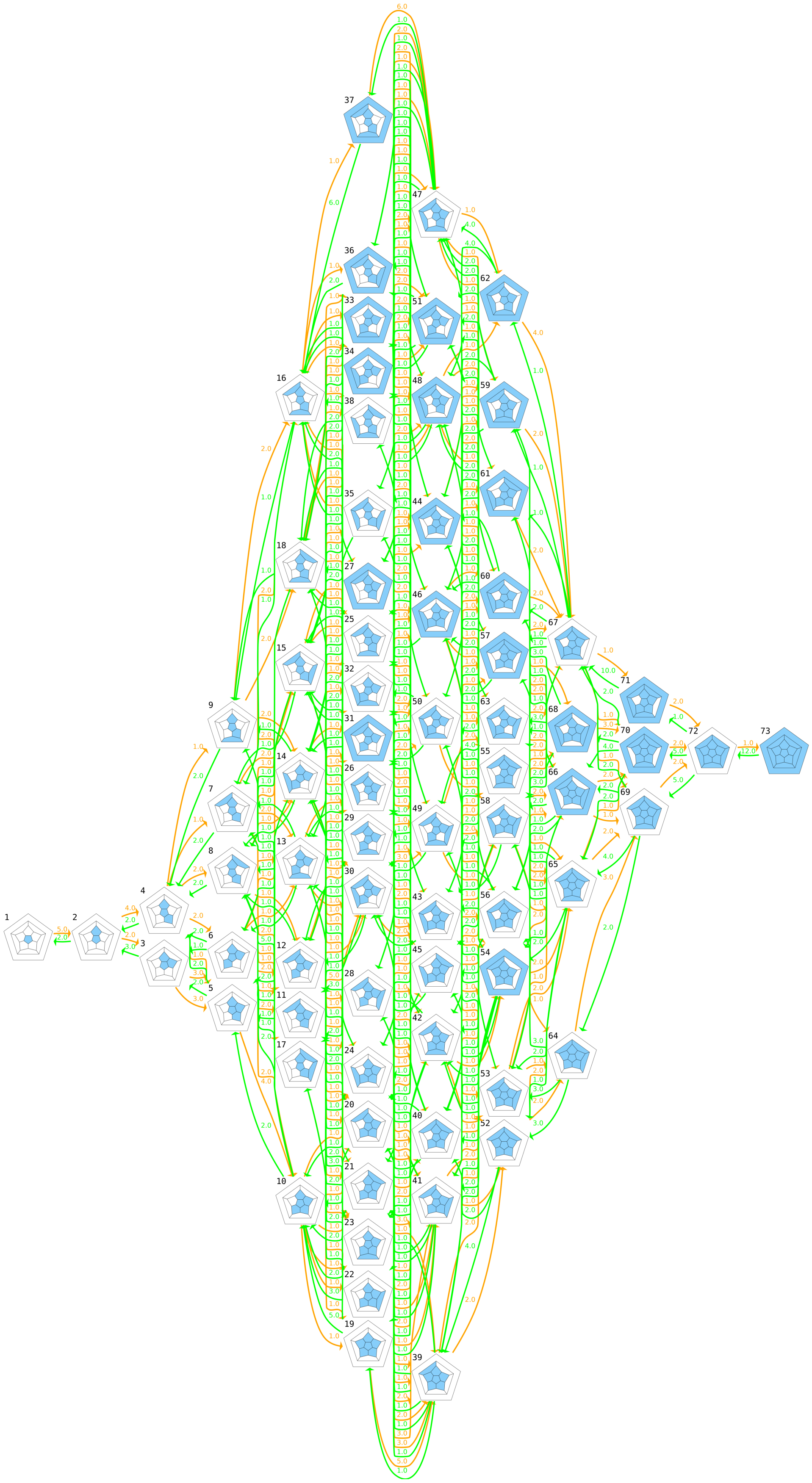

Figure 2.4: The full dodecahedral assembly network of 73 intermediates and 263 reactions between them, modelled according to Endres, Zlotnick*et al.* (Endres et al., 2005).

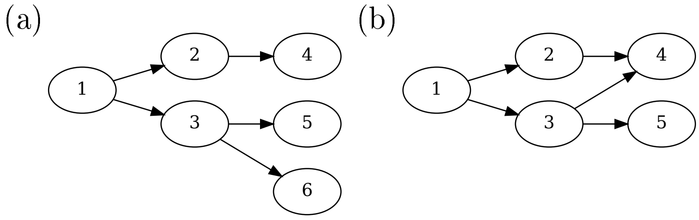

Figure 2.5: An example of path branching (a) and recombination (b).

free capsomere will have a significant concentration. At very negative contact energies, the backward reaction rates are so small that there effectively will be no backward reactions taking place. With effectively no backward reactions, the pathways that protein material takes through the network depend only on the branching within the network. Eventually, even with very small backward reactions, the model system would equilibrate to thermodynamic equilibrium. However, the simulated time period required may be on the order of years.

The kinetic factors that affect the concentration of intermediates are not only the reaction rates, but also factors coming from properties of the network of assembly intermediates. To emphasise how network branching affects the intermediate concentrations, two example networks are described. The first network (a) in figure 2.5 shows how the path to intermediates 5 and 6 branches twice, while only branching once to intermediate 4. If, in the kinetic simulation of this network, we set a large negative bond strength and choose a short time period, where the backward rate would be insignificant, intermediate 4 would have a higher concentration than 5 and 6. The concentrations between 4, 5 and 6 would be split $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{4}$, respectively. In Network (b) of figure 2.5, showing paths recombining, intermediate 4 would have $\frac{3}{4}$ of the total concentration and intermediate 5 would have $\frac{1}{4}$ in the case of a large negative bond strength.

2.6 Conclusion

Zlotnick's assembly model has been used in several papers ((Zlotnick, 1994), (Zlotnick et al., 1999), (Endres and Zlotnick, 2002), (Zlotnick and Stray, 2003), (Singh and Zlotnick, 2003), (Zlotnick, 2003), (Endres et al., 2005), (Zlotnick, 2005a), (Zlotnick, 2005b), (Zlotnick, 2007),(Katen and Zlotnick, 2009), (Moisant et al., 2010), (Zlotnick and Mukhopadhyay, 2011)) and found to be a useful and interesting model of viral assembly. Even the early more simple versions of the model have been found to describe the assembly kinetics of DNA viruses such as Hepatitis B (Zlotnick et al., 1999).

The model is certainly a lot quicker than any molecular dynamics simulation due to the limited number of differential equations in comparison. However, as larger viruses are considered the number of intermediates grows to vast numbers, corresponding to all the possible combinatorial states the virus capsomeres may be in. The scale of the numbers is similar to the number of possible states in protein folding. Protein folding is a well known problem and Lenvinthal's paradox (Levinthal, 1969) tells us that in the protein can not sample all possible states and yet proteins fold correctly. By analogy, not all the possible virus intermediates need to be sampled for virus formation to occur. As Zlotnick has shown (Moisant et al., 2010), only subsets of the network are necessary to capture most of the overall assembly behaviour.

The model shows a great deal of interesting behaviour. For example assembly at equilibrium may be a very quick process where intermediates would be almost undetectable in biological experiments. The model also produces complex emergent behaviour such as the formation of kinetic traps. It is the need for removing these kinetic traps that lead to later modifications to the model. The first modification was to introduce the μ factor when considering larger networks. This μ factor, which down-weights less stable intermediates, acts to reduce the amount of capsomere in partially built capsids. This increases the amount of free capsomere available to allow kinetically trapped intermediates to build up to capsid. A criticism of the model would be that the instability of the intermediates over time was not enough to reduce their concentration sufficiently without the μ factor. However, the μ factor may be justified because in an *in-vitro* experiment there would likely be a cumulative disadvantage for binding additional capsomeres with only single bonds. Forming single bonds in this way would create long, thin, more flexible intermediates that are more likely to break apart. There would also likely be a cumulative favouring of particularly compact and stable intermediates due to longer-range stabilisation across the capsid.

The introduced nucleation rate also acts to reduce the concentrations of the intermediates to leave more free capsomere. Each reduction of the nucleation rate below the standard 1×10^{8} $\mathrm{M}^{-1}\mathrm{s}^{-1}$ would reduce the kinetic trapping, allowing the optimal capsid concentration to be reached more quickly. Further reduction of the nucleation rate from the optimal one would increase the time to form capsid due to the slow rate. A nucleation rate would also decrease the equilibrium capsid concentration amount, although for most values this would be insignificant.

In this thesis it has been possible to reproduce the graphs in the 1994 paper (Zlotnick, 1994) and also the reaction networks, as well as all 73 intermediates and the 263 edges between them (for the dodecahedral network) and the 2,423,212 for the 30-mer intermediates[†]. In the next chapter it will be shown how this model may be extended to include the Hamiltonian path model constraints in order to model the cooperative roles of genomic RNA during capsid assembly.

[†] Note that the numbers for the 30-mer given in (Moisant et al., 2010) are slightly different at 2,423,313 and 2,423,323. Since other numbers in the paper do match up to subsets of the intermediates calculated here, the difference of the final number is presumed to be due to typographical errors.

Chapter 3

Incorporating RNA into the protein assembly model

3.1 Introduction

Building on the work by Zlotnick *et al.* it will be shown how the RNA may be incorporated into the assembly model. Following the steps of the model, this chapter is concerned with analysing the intermediates of assembly and placing these intermediates into reaction networks. These networks will show the orders in which successive building-blocks may be attached to form capsid. In later chapters we will see the kinetic simulation of these networks. To demonstrate the procedure, a simple cube shape will be first used as an example. Using such a small shape it is possible to show entire assembly networks. These networks are used to illustrate the assumptions made in the RNA assembly model. Kinetics of the RNA model will also be shown for the dodecahedron, which is the polyhedral shape discussed in Zlotnick's earlier papers. The larger viruses of STMV and MS2 will be discussed later in chapter 6.

3.2 Theoretical "Cube Virus" Assembly

To model RNA virus formation for a hypothetical cube virus, each face may be thought of as representing a capsomere subunit. As there are only 6 capsomeres/faces in the cube the model of assembly is easily computable. First, we can look at the assembly network of the cube without RNA as shown in figure 3.1, henceforth referred to as the "protein-only" scenario. This protein-only model has been formulated in the same manner as Zlotnick's dodecahedron

(Endres et al., 2005). Here we can see eight intermediates and all possible assembly pathways between them. The first intermediate is the "free" capsomere which will then bind a further "free" capsomere to become the second intermediate. Assembly then proceeds through intermediates with three, four and five capsomeres to the complete cube capsid.

The STMV virus has been used as the inspiration for the first network of RNA assembly intermediates. Since STMV is thought to assemble by binding capsomeres along its RNA starting at the 5 prime end (Larson and McPherson, 2001), the model has been limited to this assumption. Following the Hamiltonian path theory of virus assembly, a complete cube capsid must correspond to a Hamiltonian path on the graph shown in figure 3.2 (b). There are 40 such paths that sequentially connect each face of the cube starting at face 1. Removing the initial four-fold symmetry gives the ten shown in table 3.1. In addition to these ten Hamiltonian paths, dead-end paths that do not lead to capsid and an initial RNA binding step are used to create the first RNA assembly network, shown in figure 3.3. The resultant network contains 36 intermediates including the free capsomere (35), free RNA (36) and the ten final capsids. The central cube in each of the cube Schlegel diagrams of the network in 3.3 is always face 1 of the cube shown in figure 3.2(a).

The first difference in the assembly network of the RNA to the protein assembly network (compare figure 3.1 and figure 3.3) is the greater number of intermediates. This is because the additional RNA structure breaks the symmetry of the intermediates, resulting in multiple different RNA layouts for a single protein capsomere configuration. As an example of the assembly behaviour we can look at the growth of the two capsomere intermediate labelled 2 in figure 3.3. Intermediate 2, like all other intermediates, has the 5' end of the RNA bound to face one of the cube. This leaves the trailing end of the RNA then bound to face two. Since (using the numbering convention for cube faces given in figure 3.2) the adjacent faces are three, five and six, the next assembly intermediates numbered 3, 4 and 5 respectively correspond to each possible capsomere binding event to the 3' end of the RNA. Disassembly of an intermediate may equally only occur through capsomeres at the 3' end of the RNA. This is based on the assumption that disassembly is not allowed to occur where there is a special high affinity binding site on the RNA such as the 5' end of the RNA for the STMV model or later for the TR position in MS2. This is due to the high affinity of such sites which makes dissociation unlikely.

From a protein layout perspective we can see that the intermediate 5 in the RNA network (figure 3.3) has the same protein layout as intermediate 4 in the protein-only network 3.1. However, in order to create intermediate 4 in the protein-only network there are two capsomeres that may bind to intermediate 2, whereas in the RNA network there is only one possibility of capsomere binding. This illustrates an important feature of the RNA model, which is that, like the protein-only model, it is protein diffusion driven. In the protein-only model capsomeres may diffuse to any position adjacent to an already present capsomere and bind. However, in the RNA model the capsomere will only bind if the end of the RNA is also adjacent. The RNA is required to allow the capsomere to form the correct conformation in order to bind. The justification for the protein diffusion limited reactions is the same as in the protein-only model, in that the rate-limiting step in the reaction will be due to the protein capsomeres’ diffusion rather than any subsequent binding events.

No RNA contact energies are present in this RNA model. Only the number of capsomere contacts are used to work out the energy of the intermediate for the kinetic analysis. The reason for this simplification comes from the fact that in many viruses only a few high affinity RNA binding sites are known, while the vast majority of stem loops bind relatively weakly compared to capsomere association energies. Omitting this RNA binding energy reduces the number of parameters in the model and therefore provides a simple and transparent testing ground to investigate how the change in the network of assembly pathways in the presence of RNA impacts on the assembly kinetics. However, there is an initial RNA-capsomere binding event in each RNA network. To account for this, because the initial RNA binding sites are assumed to have high affinities, resulting in a diffusion-limited forward reaction with no backward reaction. This therefore never allows a capsomere bound to the single high-affinity position to dissociate.

Another interesting feature of the RNA assembly network is the existence of dead-end species such as intermediates 22 and 23, that have no direct path to the completed capsid. These intermediates are termed dead-ends as the only pathway to capsid is for, in this case, two capsomeres to fall off, creating intermediate 5, which in turn does have a possible pathway to capsid. As will become clear, it is dead-ends such as these that have a major influence in the later kinetic simulations. It is possible to reduce the complexity of the network by factoring out mirror symmetry. The procedure is illustrated in figure 3.4 which has been obtained by removing all intermediates with mirror symmetry. By way of example the intermediates 3 and 4 in the original figure 3.3 are mirror images and are therefore combined, resulting in intermediate 3 in the new network (figure 3.4). To correct for this combination the symmetry factors are modified. In this example the symmetry factors from intermediate 2 to intermediates 3 and 4 were both 1 in the original network, therefore the symmetry factor in the new network is the sum, which is 2.

This new network is henceforth referred to as the UniRNA network of the cube because assembly precedes in a single direction along the RNA. With the correct symmetry factors, removing the mirror image intermediates in any of the investigated networks does not affect the later kinetic simulations. Except that for each intermediate that has had a mirror image removed the actual concentration would need to be divided equally in order to yield that of each individual in the original pair of intermediates.

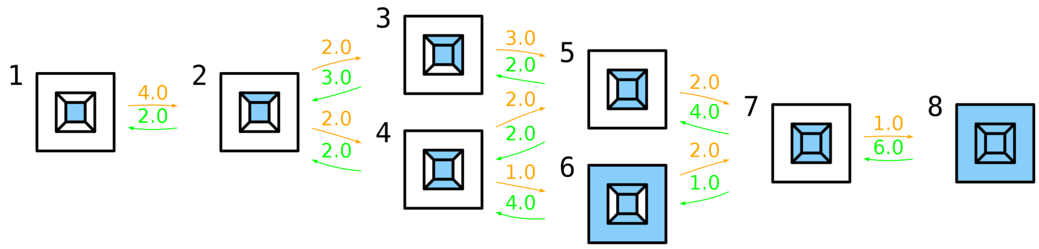

Figure 3.1: Assembly of a cube without RNA, showing the 8 possible intermediates as flattened Schlegel diagrams with an extended back square. Positions occupied by capsomeres are shaded in blue, and intermediates are numbered 1-8 in the upper left corner.

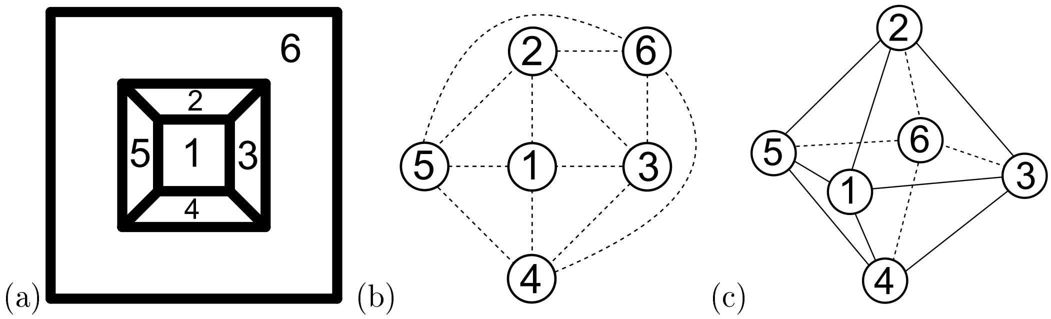

Figure 3.2: (a) The Schlegel diagram of the cube and (b) the corresponding RNA connectivity graph. This connectivity graph is called an octahedral graph as it corresponds to the vertices and edges of an octahedron (c).

The MS2 capsid has the TR position in the centre of the genome which is believed to be the first position to bind a capsomere and hence initiate

Forwards	Backwards
123456	123456
123465	125634
123645	123645
123654	125463
125436	125436
125463	123654
125634	123465
125643	125643
126345	126345
126543	126543

Table 3.1: The Hamiltonian paths corresponding to the capsids in the cube UniRNA network. There are 40 Hamiltonian paths for the octahedral graph that start at a single point, which is labelled 1 in figure3.2(c). The path number may be divided by 4 to give only those that then proceed to point 2, this gives the 10 shown. Eight of the Hamiltonian paths have the same geometry forwards and backwards, shown by the blue links. The remaining four paths are each other backwards. When this directionality is not required the Hamiltonian paths may be described by only 8 of the 10 shown.

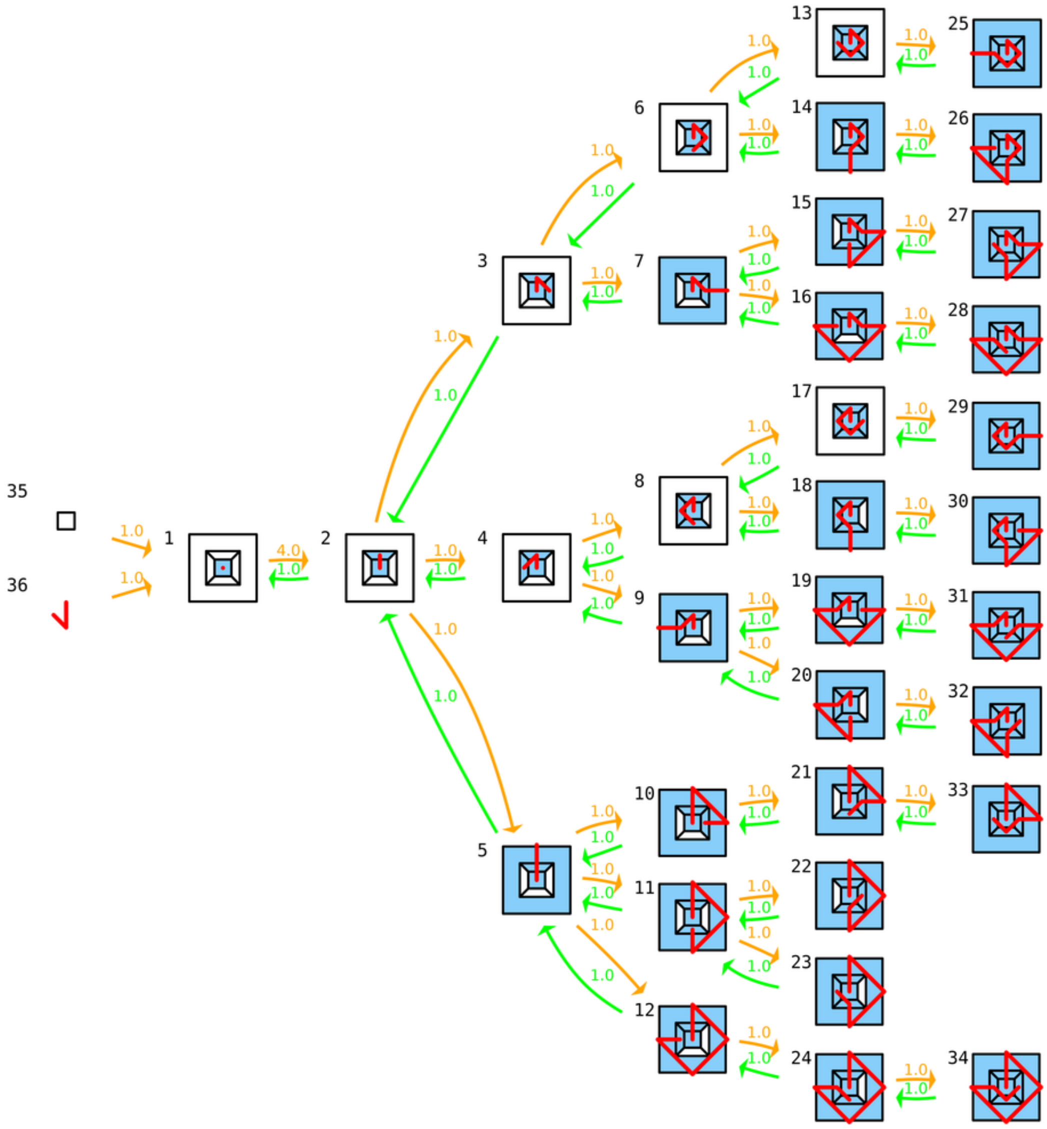

Figure 3.3: Assembly of a cube with RNA where assembly proceeds from the 5’ end, showing the 36 intermediates including the 10 final capsids with their RNA configurations as flattened Schlegel diagrams. Intermediates are numbered 1-36 in the upper left of each intermediate. The RNA binding to the back face, number 6, has been shown by drawing a line to the center of one of its edges. To show the RNA connectivity proceeding from the back face, edge center points have been connected when appropriate. The red line numbered 36 represents the free floating RNA. The 10 final capsids shown each correspond to a distinct Hamiltonian path.

assembly. To model this scenario using the cube, a reaction network with the first capsomere binding in the middle of the RNA has been constructed (see figure 3.5). The network has been simplified by making no distinction between whether the second capsomere is bound by the 5’ or 3’ prime end

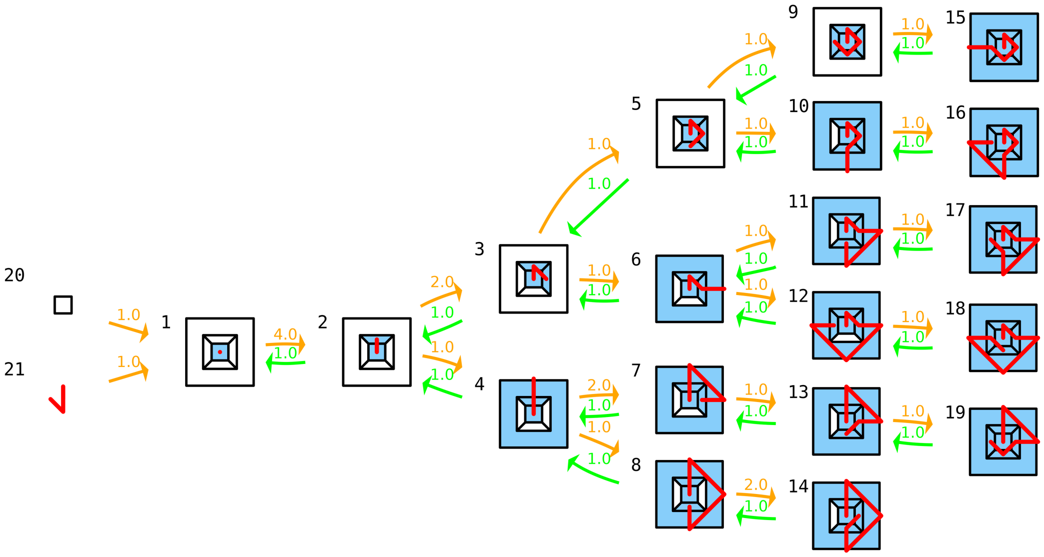

Figure 3.4: The UniRNA network, this network is the same as figure 3.3 but with the mirror images removed and the symmetry factors updated.

of the RNA. This leaves 30 final capsids rather than the 60 which would be the case if the direction was distinct. This number of 60 is consistent with the fact that there are 6 possible distinct TR positions for each of the 10 Hamiltonian paths detailed for the cube above. Again this assumption is justified as the rate-limiting step in the reaction is assumed to be due to the protein capsomeres’ diffusion. This means that having two ends of the RNA available to bind a single capsomere would not affect the speed of the reaction. In the kinetic simulations each final capsid may be thought of as arising form either a pathway on which the second capsomere binds at the 5’ or at the 3’ end, and therefore its concentration should be halved into these two possibilities. An implicit assumption in this network is that the 5’ and 3’ strands of the RNA, from the TR position, are individually long enough to complete the capsid. In future work changing this assumption and limiting the number of capsomeres able to bind to each side will be investigated.

Since the assumptions of the model do not presume significant RNA binding energies there is no consequence to combining the 5’ and 3’ directions in this way. If RNA binding energies were introduced, in order to not separately model the 5’ and 3’ binding the RNA would have to be assumed or designed to be palindromic around the TR position. A future potential model could take different RNA binding energies into account, at which point the 5’ and 3’ RNA directions will be modelled separately.

In comparison to the first RNA cube assembly network there are again many more intermediates. This is because intermediates are now distinguished by the TR position. Thus, even if the RNA and the protein layouts look the same a configuration can represent different assembly scenarios. An example of this is shown in figure 3.6. Another difference to the first cube network is that, when starting in the middle of the RNA, both ends are available to bind capsomeres. The result of this is that there are no longer any dead-end intermediates. This is likely to be an advantage kinetically as no material will become trapped in these dead-ends.

To further illustrate the protein driven nature of these networks, the symmetry factors relating to the addition of the third capsomere will be explained. A larger view of this portion of the network from figure 3.5 is shown in figure 3.7. There are four places a protein may diffuse to bind intermediate 2 and therefore the forward symmetry factors (shown in orange) add up to four. This is the same number as the protein-only network, indicating that binding along both directions of RNA allows all the forward capsomere reactions. If a capsomere were to diffuse to face four of the cube (see figure 3.2) only one end of the RNA is adjacent and may bind to form intermediate 7. This therefore has a symmetry factor of 1, because only one intermediate may be formed. Likewise, only one end of the RNA may bind a capsomere diffusing into the position of face six of the cube. This reaction is also therefore given a forward symmetry factor of 1.

For capsomeres diffusing to the positions corresponding to faces three and five of the cube, either end of the RNA may bind. For a single capsomere being at face three half the time the RNA will bind and form intermediate 3, and the other half of the time intermediate 6 will form. Each forward symmetry factor is therefore 0.5, because the two future intermediates formed must share the single protein addition. However, the next intermediate 6 may also be formed with a 0.5 symmetry factor if a capsomere diffuses to cube face five and has a combined symmetry factor of 1. There are two capsomere additions that result in intermediate 6, because the direction is not taken into account. The backward symmetry factors depend on the number of single capsomere disassembly reactions that would recreate the previous intermediate.

To simplify the network, again, the mirror image intermediates have been removed to create the network that will be used in the kinetics. This network, shown in figure 3.8, is termed the cube TrRNA network due to the TR position in the network being unique.

	Protein Only	UniRNA			TrRNA	BiRNA
Intermediate Size	No.	No.	Dead-Ends	On D.E. Path	No.	No.
Free Capsomere/RNA	1	2			2	2
1	NA	1			1	1
2	1	1			1	1
3	1	2			4	2
4	2	4		1	8	3
5	2	6	1		16	4
6	1	5			15	4
Total	8	21	1	1	47	17

Table 3.2: The intermediate numbers for the 4 different cube scenarios. Also shown separately are the number of dead-ends and the number of intermediates that are only on dead-end pathways, relevant for the UniRNA network.

In addition to the previous two RNA networks (UniRNA and TrRNA) a third, termed the BiRNA network, has been created. The BiRNA network may be thought of as representing capsomeres binding to a uniform circular strand of RNA with no unique positions. This results in the much smaller and simpler network shown in figure 3.9. The network pictured in figure 3.9 does not discriminate between 5' and 3' RNA directions and the mirror images have already been removed. There are only four final capsids in this BiRNA network, rather than the five in the UniRNA network, because two of the final capsids in the UniRNA network actually have the same Hamiltonian path layout, as one forwards corresponds to is the other one backwards. This is shown in table 3.1.

For simplicity, once a capsomere has bound to the RNA, there must always remain a capsomere bound to the RNA, although the original capsomere is allowed to fall off. This removes the need for a reaction back to free RNA, which would require an RNA binding energy. Although there are many fewer RNA layouts in this BiRNA network due to the TR position not being distinguished, in terms of the protein assembly this network is very similar to the TrRNA network. The only difference, is that there are a few additional backward reactions. These backward reactions are those that would otherwise require the TrRNA bound capsomere in the TrRNA network to dissociate.

The different RNA binding network assumptions are summarised in figure 3.10, showing the assembly and disassembly reactions on the RNA. A summery of the intermediate numbers is given in table 3.2.

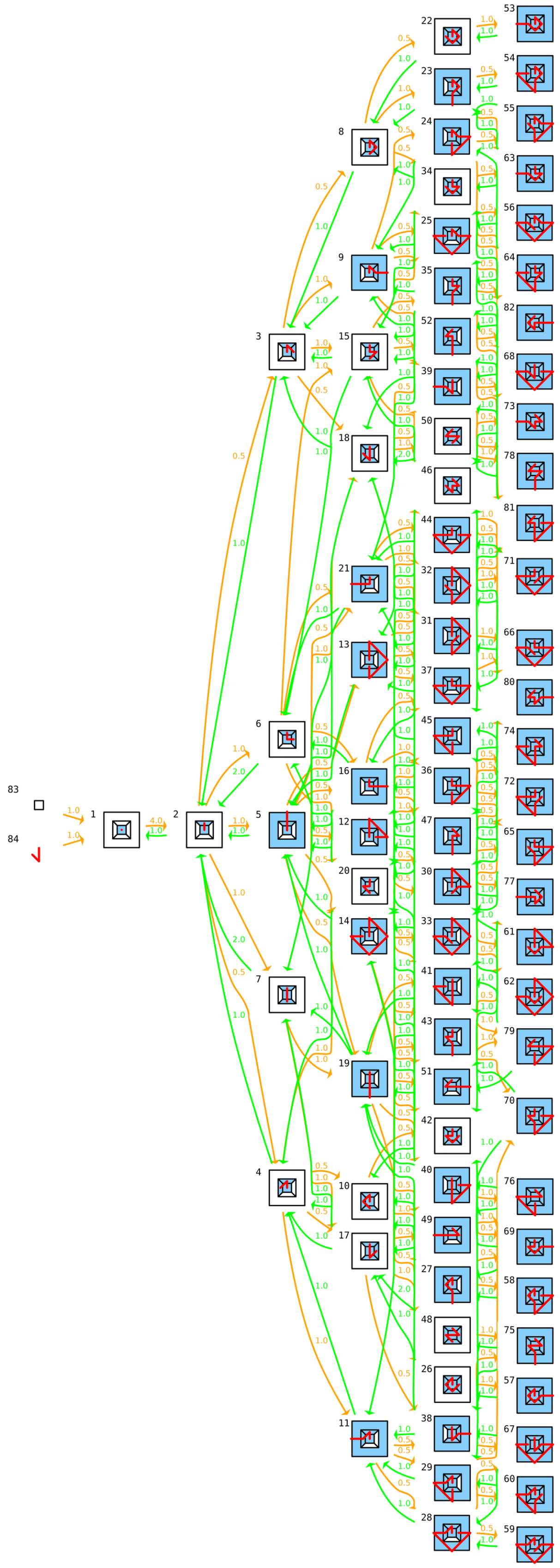

Figure 3.5: The network with the initial RNA-capsomere binding in the middle of the RNA. It should also be noted that the orientations of the Schlegel diagrams may not be maintained through the binding steps. This is due to drawing the minimal binding pattern for each intermediate which is explained later.

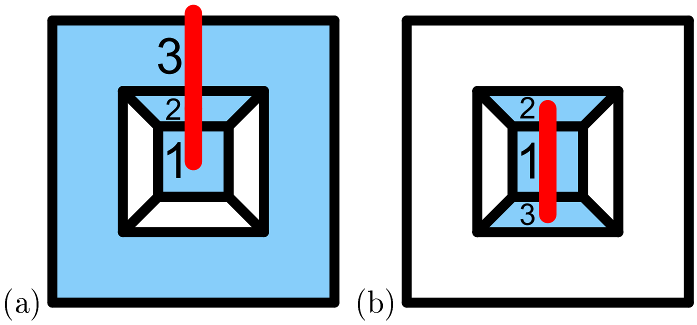

Figure 3.6: Two intermediates showing the same protein and RNA layouts, however the unique TR position labelled 1 is in geometrically different positions.

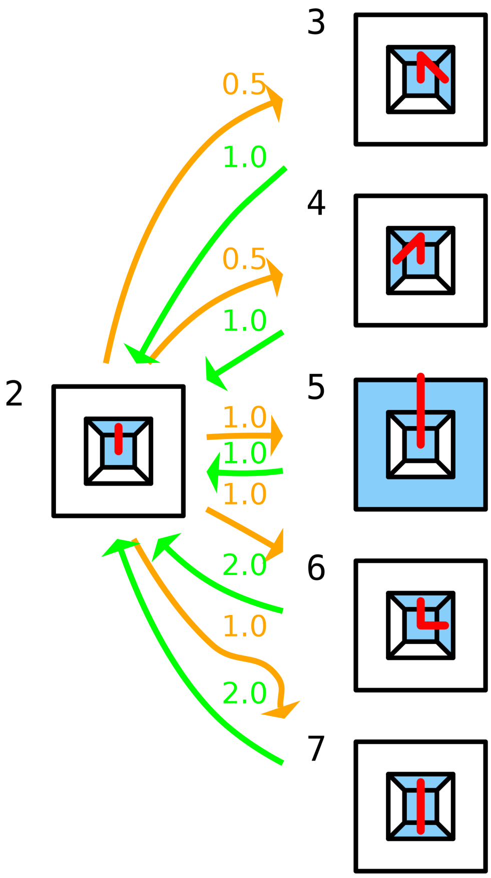

Figure 3.7: The start of the network shown in figure 3.5 to emphasise the symmetry factors.

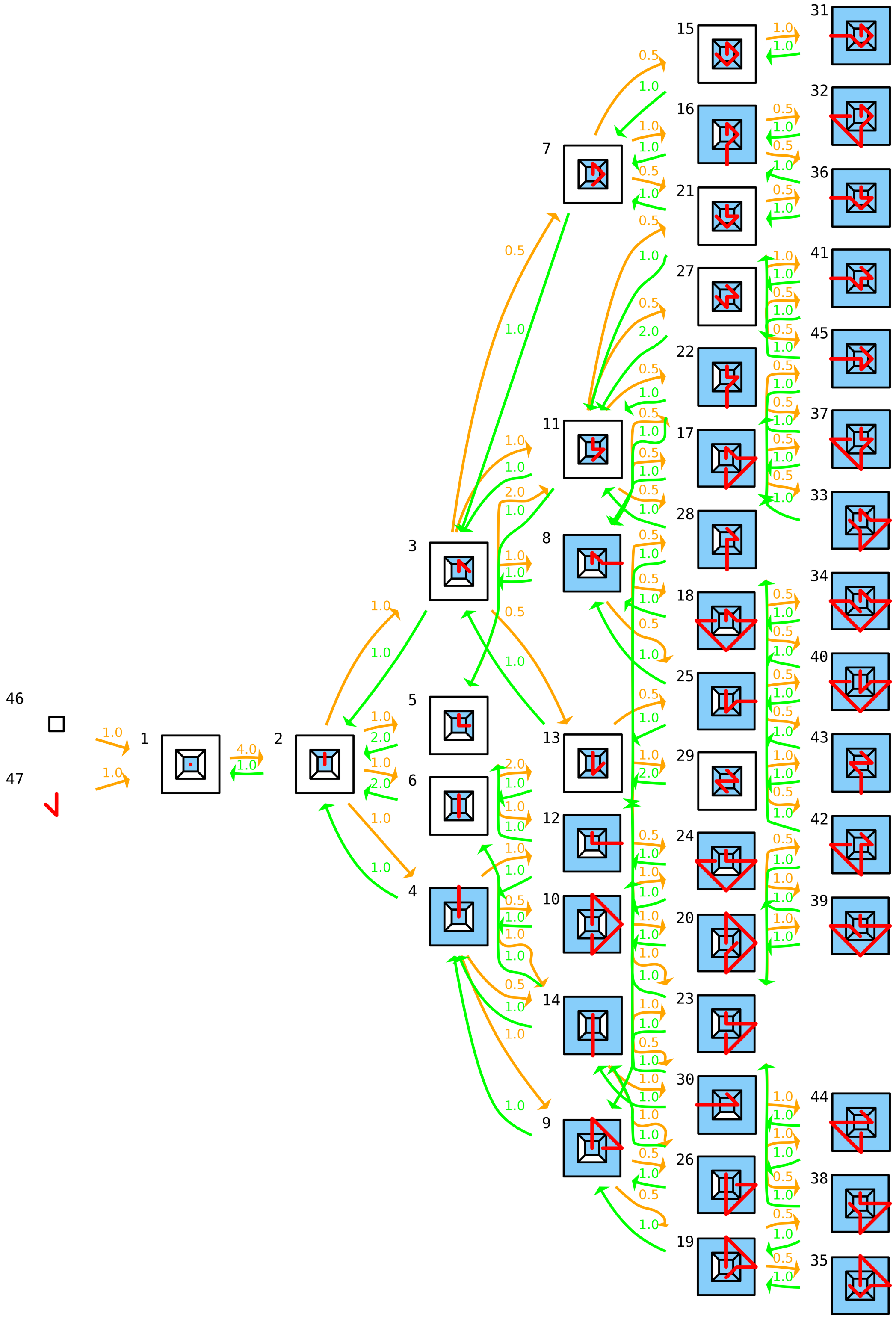

Figure 3.8: The TrRNA network, this network is the same as figure 3.5 but with the mirror images removed.

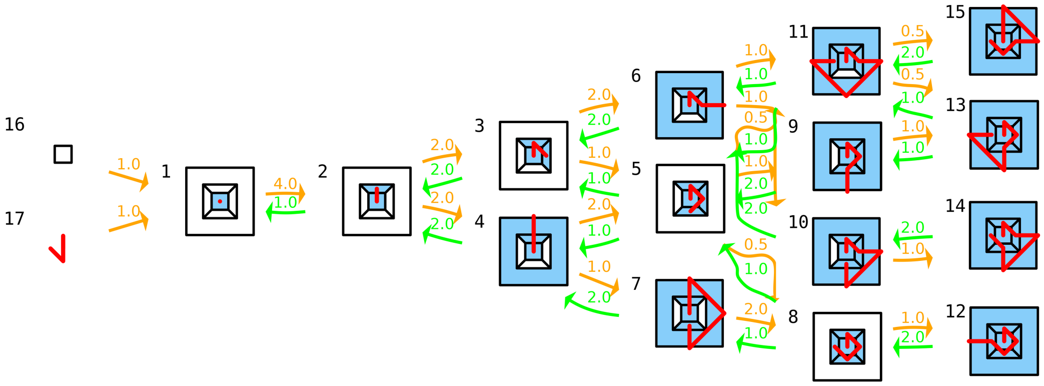

Figure 3.9: The BiRNA network.

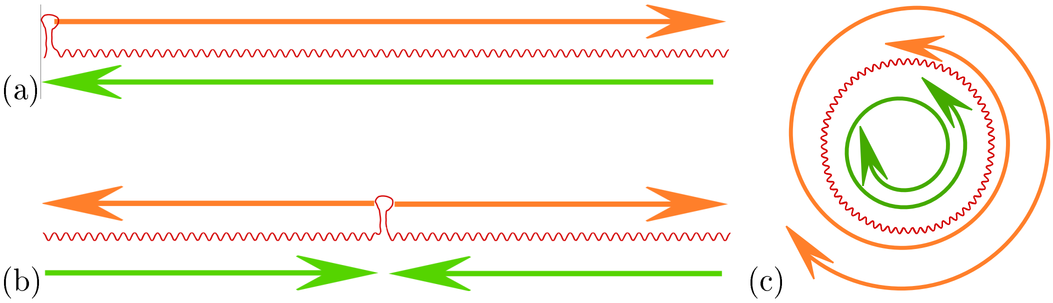

Figure 3.10: A summery of the different UniRNA, TrRNA and BiRNA RNA binding scenarios. The orange and green arrows show assembly and disassembly directions, respectively. The UniRNA network may only build up from one end of the RNA while the TrRNA network may build up from the middle TR point. The BiRNA network is not restricted to keeping a TR capsomere bound, and disassembly may happen across the original binding site.

3.3 Dodecahedral Reaction Networks

Reaction networks for the dodecahedron have also been created for the UniRNA, TrRNA and BiRNA scenarios. The start of each network is shown respectively in figures 3.12, 3.13 and 3.14, and the corresponding intermediate numbers are shown in table 3.3. There are 1264 Hamiltonian paths for the corresponding icosahedral graph shown in figure 3.11. Removing the mirror images halves this number and gives 632 final capsids in the dodecahedral UniRNA network. For the BiRNA network the final number of 3792 is $\frac{632\times12}{2}$ because there are now 12 positions for TR, given that direction is not taken into account. The BiRNA network only has 340 final capsids after mirror removal. Before mirror removal, there are 680 consisting of 96 paths that are the same forwards and backwards and 584 paths that are not. From this the 1264 Hamiltonian paths used in the UniRNA network are obtained as $(584 \times 2) + 96$. This demonstrates how the Hamiltonian paths may be combined to reduce the complexity of the network. There are a great many more paths for the dodecahedral networks because the intermediate number grows almost combinatorially with the capsid size.

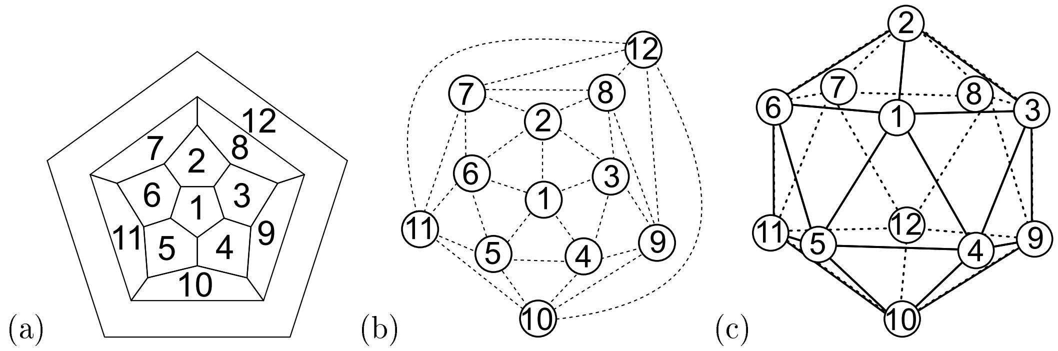

Figure 3.11: (a) The Schlegel diagram of the dodecahedron and (b) the corresponding RNA connectivity graph. This connectivity graph is an icosahedral graph and in three dimensions is shaped like the icosahedron (c).

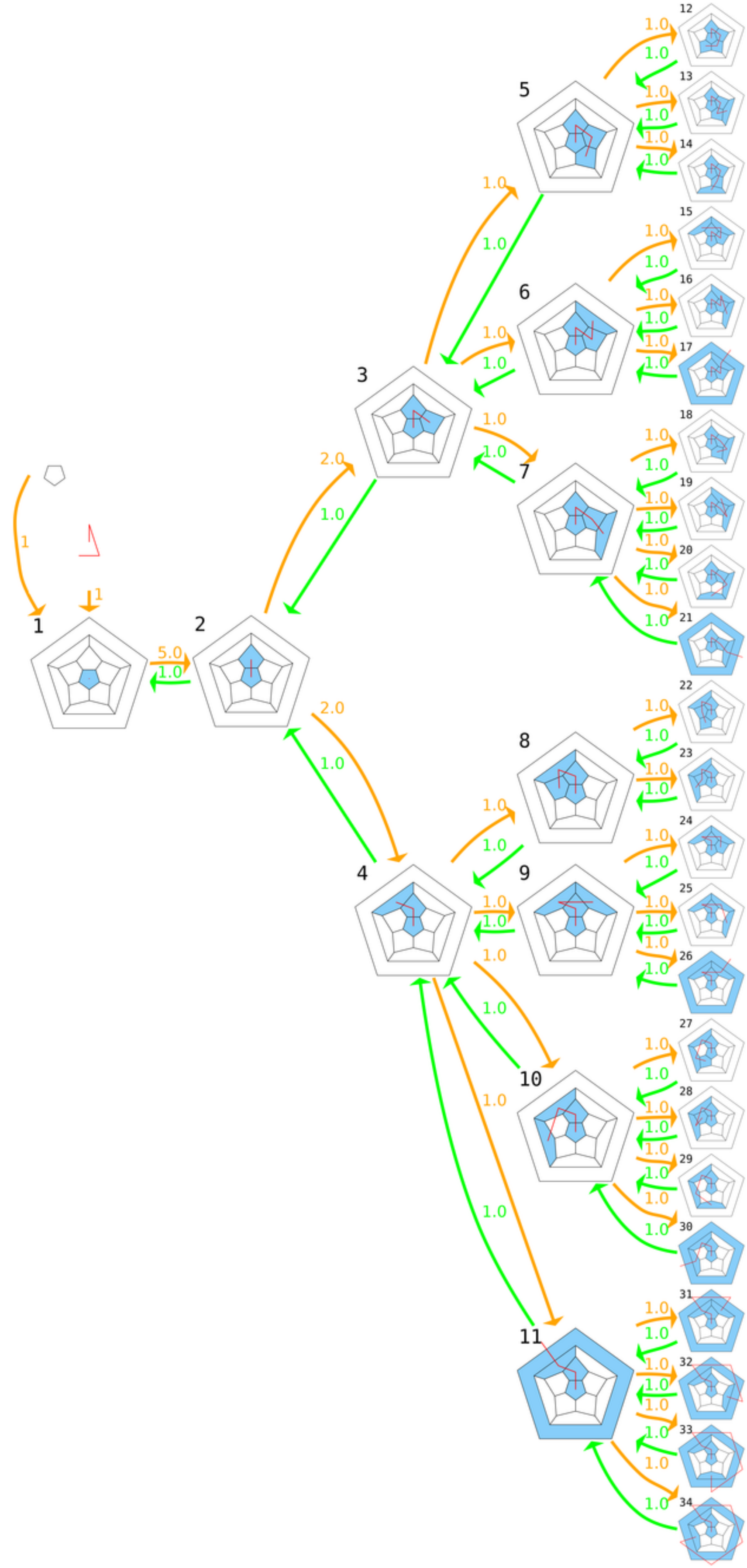

Figure 3.12: Beginning of the dodecahedral UniRNA network.

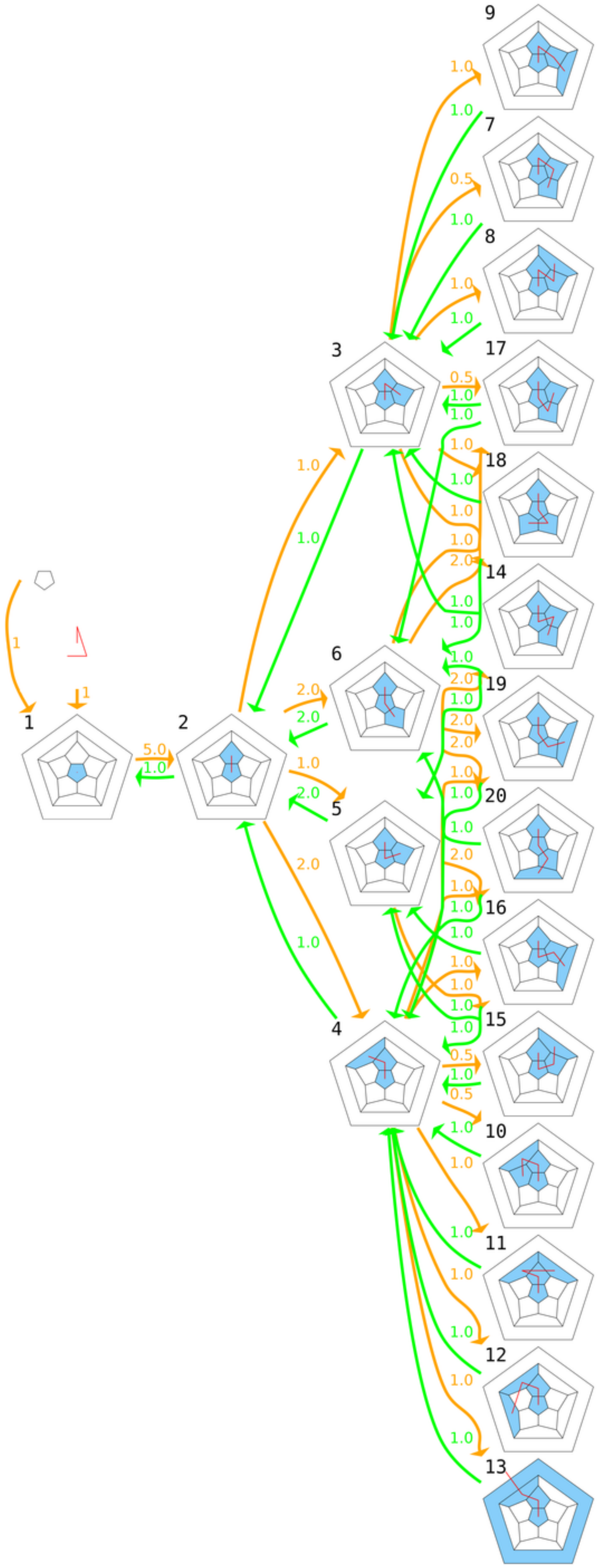

Figure 3.13: Beginning of the dodecahedral TrRNA network.

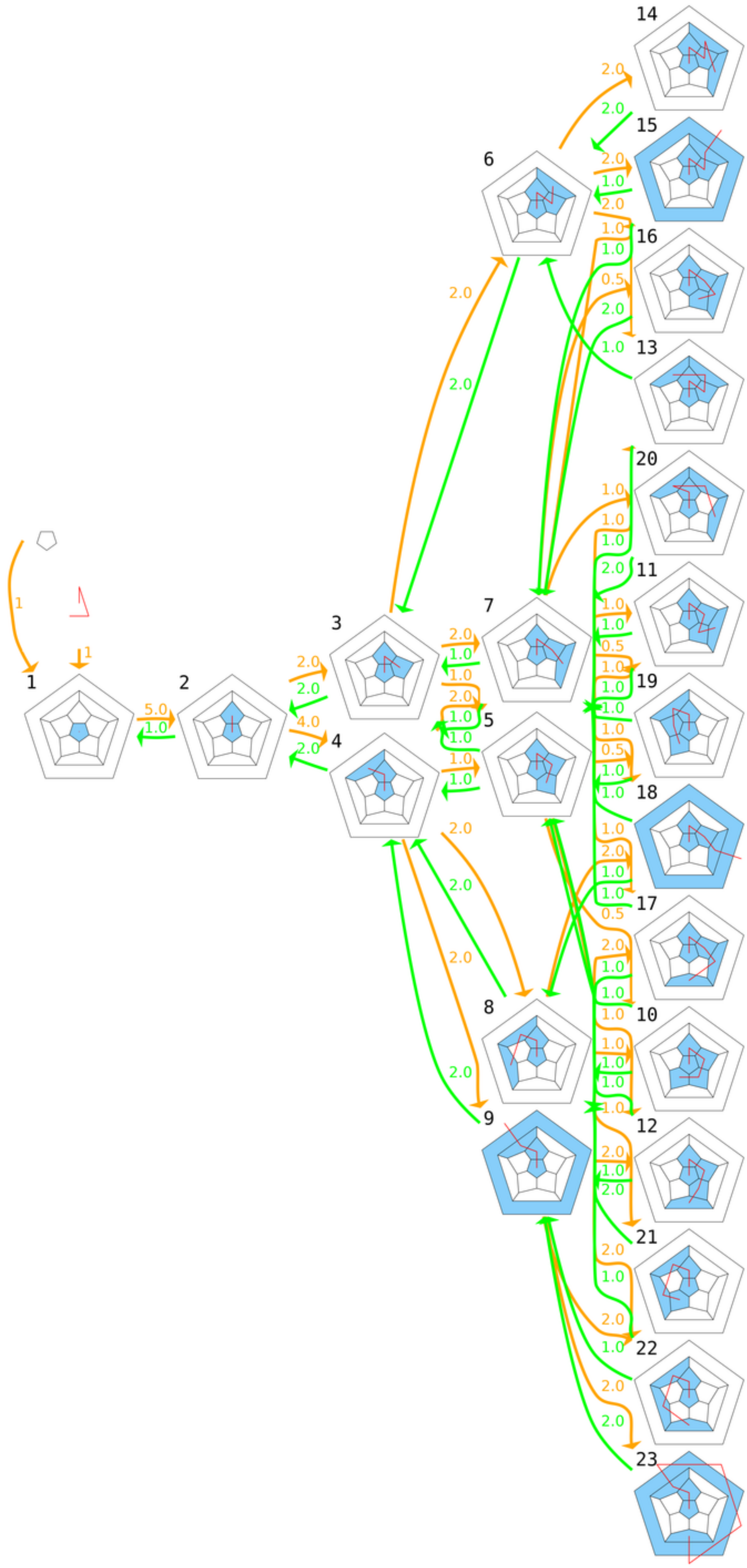

Figure 3.14: Beginning of the dodecahedral BiRNA network.

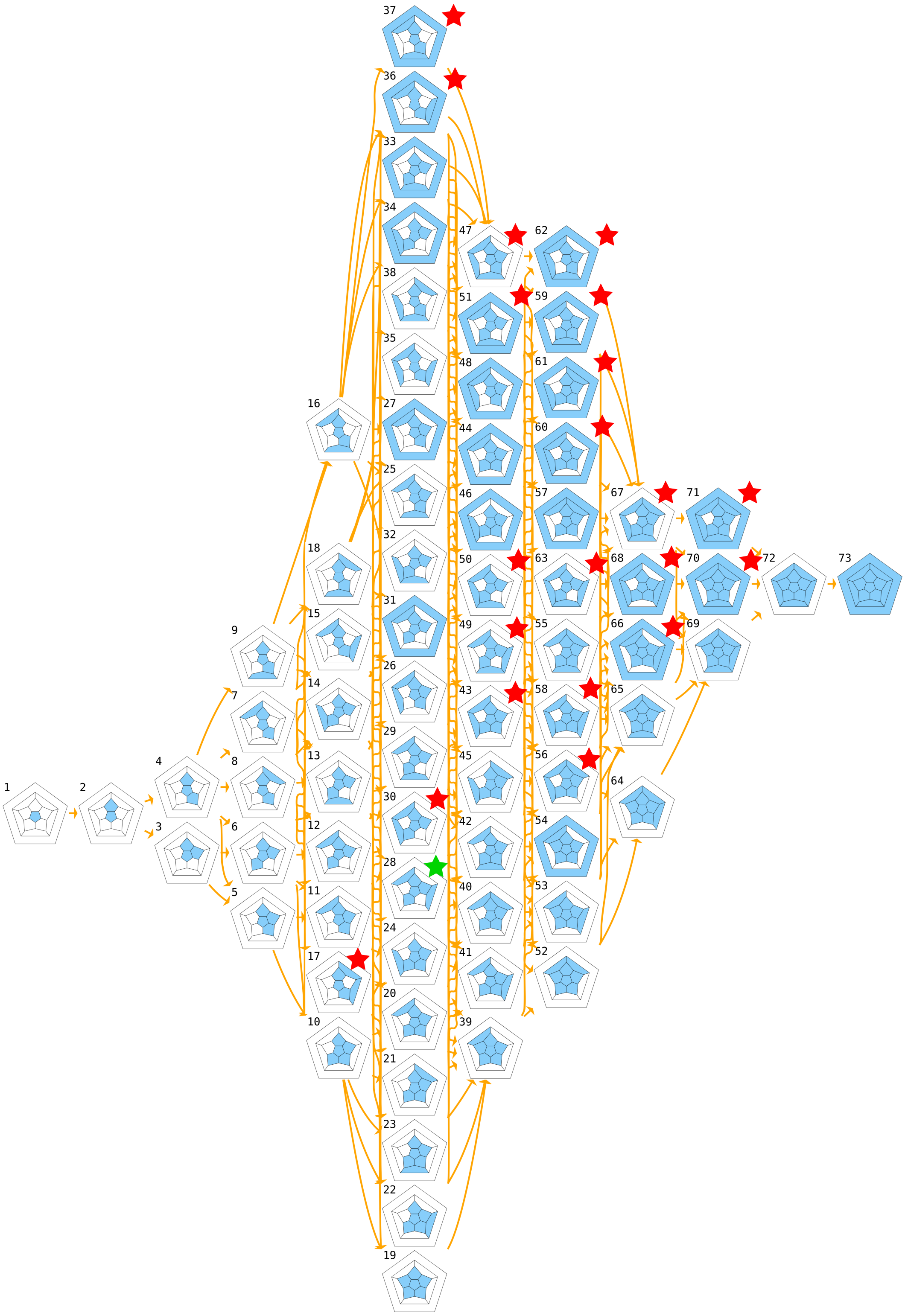

Figure 3.15: Protein only network for the dodecahedron. The red starred intermediates are those which would be kinetically trapped in the UniRNA network and the green stared intermediate is the only protein configuration not realisable in any of the RNA networks.

	Protein Only	UniRNA			TrRNA			BiRNA		
Intermediate Size	No.	No.	Dead-Ends	On D.E. Path	No.	Dead-Ends	On D.E. Path	No.	Dead-Ends	On D.E. Path
Free Capsomere/RNA	1	2			2			2		
1	NA	1			1			1		
2	1	1			1			1		
3	1	2			4			2		
4	2	7			14			5		
5	5	23		1	60			14		
6	9	71	1	12	213			40		
7	20	198	6	75	697		11	103		2
8	13	474	29	242	1896	8	164	249	1	22
9	12	916	112	492	4125	54	762	461	6	86
10	5	1336	340	555	6680	355	1390	691	38	139
11	3	1300	668		7150	1606		650	146	
12	1	632			3792			340		0
Total	73	4963	1156	1377	24635	2023	2327	2559	191	249

Table 3.3: The intermediate numbers for the 4 different dodecahedral scenarios. With the dead-end numbers and the number of intermediates that are only on dead-end pathways.

With the dodecahedron it is now possible for both ends of the RNA to be trapped in a dead-end. Although the percentage of dead-ends to capsid is still much less in the TrRNA and BiRNA networks than in the UniRNA dodecahedral network. The lengths of these dead-end pathways are listed in table 3.4. The length of a dead-end pathway is defined to be the number of backward reactions required until an intermediate on a pathway to capsid is reached. This table shows that the length of the dead-end pathways decreases as more reactions are allowed in the TrRNA and then BiRNA networks. Four examples of dead-end intermediates are shown in figure 3.16. Intermediate number 4120 in 3.16 is in the UniRNA network. This is the extreme example where seven capsomeres would need to dissociate from the RNA to obtain an on-pathway intermediate. This would then allow the RNA to bind face six, from which point the resulting intermediate is on pathway to capsid. The length of this dead-end pathway is due to capsomere 1 being bound to the end of the RNA, so there is no RNA available to bind face six. When there is RNA available to bind capsomeres either side of the TR position, the maximum length of the dead-end pathway is 5. An example of an intermediate in the TrRNA reaction requiring five backward reactions is number 14402 in figure 3.16. This intermediate requires the capsomeres on faces 11, 10, 4, 5 and 6 to fall off one end of the RNA. These are then bound by the other side of the RNA, while capsomere 8 is bound by the RNA that was bound to face 6. It is using both sides of the RNA like this that reduces the dead-end pathway length. In the TrRNA network we assume that capsomere 1 is always bound to the RNA because of the packaging signal. Therefore, the shorter way of correcting the dead-end, removing intermediates 1 and 2, is not possible hence making it more difficult to resolve the dead-end. Two examples of shorter pathways for the BiRNA network, where any capsomere may fall off, are intermediates 1025 and 2039 in figure 3.16. The shortest path to an intermediate that could lead to capsid would require removing capsomeres 6, 1 and 2 for intermediate 1025 and 1, 2 and 3 for intermediate 2039.

Dead End Path Length	UniRNA	TrRNA	BiRNA
1	0	885	90
2	434	831	85
3	271	232	16
4	215	60	
5	151	15	
6	57		
7	28		

Table 3.4: The lengths of the off-pathway portions that only lead to dead-ends for the three dodecahedral RNA reaction networks. The path length includes the dead-end intermediate and is the same number as the number of backward reactions that are required to arrive at an on-pathway intermediate.

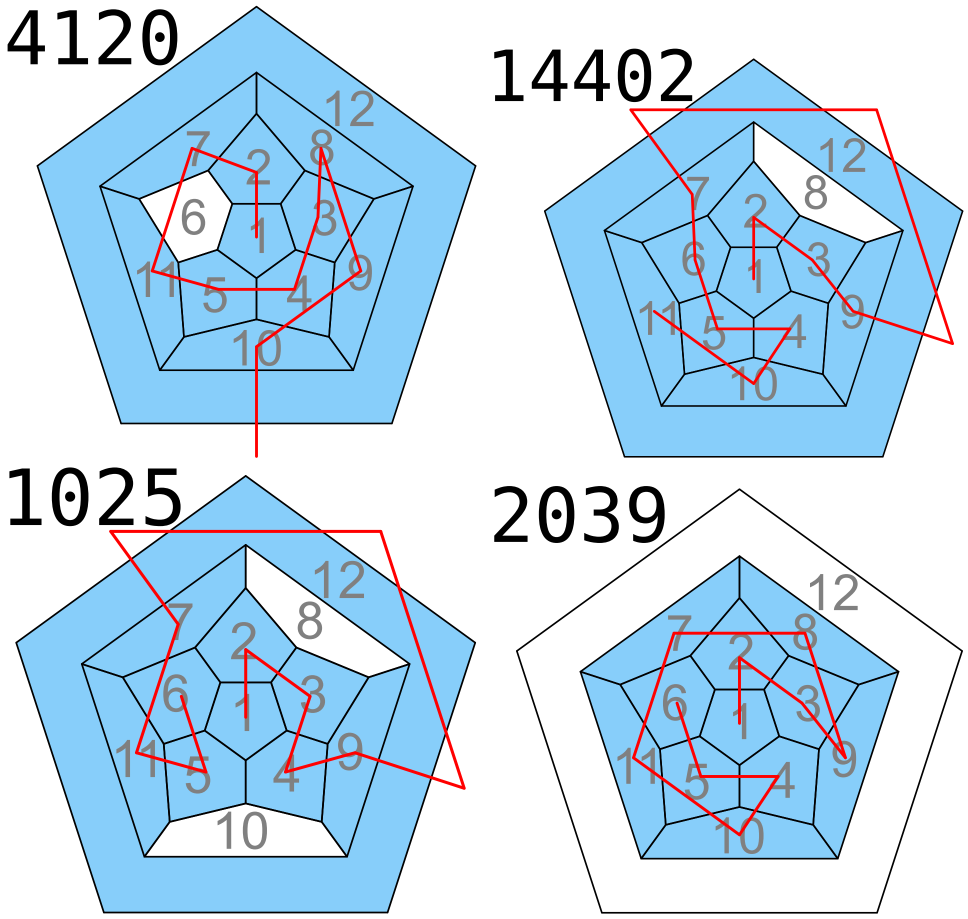

Figure 3.16: Examples of dead-end intermediates. 4120 is in the UniRNA network, 14402 in the TrRNA network and, 1025 and 2039 are in the BiRNA network.

3.4 Methods

3.4.1 Computation of Hamiltonian paths

The algorithms used for generating the assembly networks shown in this chapter have gone through several iterations. The improvements allow the generation of large reaction networks efficiently. Generating all the Hamiltonian paths for a mathematical graph is a well known NP-complete problem in computer science and is similar to the travelling salesman problem. The NP-complete nature of this problem is that although it is simple to test whether a particular path is a Hamiltonian path or not, determining a Hamiltonian path in the first place can only be done by combinatorially trying all the possibilities. The usual algorithm for finding Hamiltonian paths is to us a simple recursive algorithm that visits each edge in turn until either a Hamiltonian path or a dead-end is reached. The algorithm would then back-track and try all the other combinations of edges, saving the result as it proceeds. For the cube and dodecahedral networks which are relatively small and have low connectivity, the paths may be determined within minutes. However, for the MS2 virus finding all the Hamiltonian paths in its corresponding graph takes three weeks. This is simply due to the higher number of possible edge combinations.

The algorithm used for generating protein-only networks is very similar to the Hamiltonian path recursive algorithm, in that every combinatorially possible addition of a protein to a previous intermediate must be constructed to find all the intermediates. Constructing the intermediates is a relatively quick process. It is removing the duplicates that takes orders of magnitude more time.

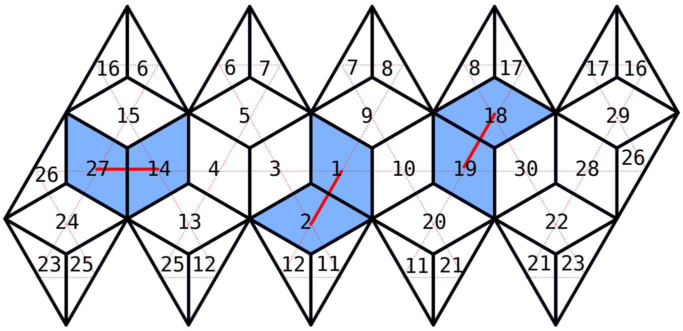

Figure 3.17: The net of the 30-mer with labelled face numbers showing 3 of the possible 60 dimer encodings.

3.4.2 Duplicate intermediate removal

The problem of the duplicate problem is illustrated for the 30-mer in figure 3.17, which shows three possible ways to encode two capsomers into pairs of face numbers. The encodings shown are, [1,2], [19, 18] and [14, 27], in fact there are 60 different ways of representing this protein layout due to the icosahedral symmetry of the capsid. As an example of the scale of the duplicate problem consider all intermediates of size 18 on the pathway to the STMV 30-mer. There are 4,403,010 intermediates generated of which only 455,307 are unique. Naively, to compare the first intermediate of size 18 generated to each of the 60 representations of the remaining 4403009 would require over a quarter of a billion comparisons. In total, up to 6×10^{14} comparisons would be required which would clearly take a disproportionate amount of time.

Previously in (Moisant et al., 2010) the intermediates to be deduplicated are first sorted by size, then number of capsomere contacts, perimeter path, and finally by the number of holes in the intermediate. This separation into groups to be compared is in order to reduce the overall number of comparisons. Zlotnick's use of the perimeter path is a clever way of avoiding the 60 separate symmetric encodings of a particular set of proteins. This is because the perimeter is the same, whether two capsomere intermediate is described by proteins 2 and 3 or 4 and 5. A perimeter path in (Moisant et al., 2010) is shown in figure 3.18. However, since the starting point of the perimeter is undefined the perimeter from every possible starting position must be compared. This increases the number of comparisons again, especially when larger intermediates are present. In his 2010 paper, Zlotnick was able to generate the unique set of STMV protein intermediates in 150 days of CPU core time.

3.4.3 New Algorithms

The sorting algorithm

To avoid the disproportionately large amounts of time required for the deduplication of generated intermediates a new algorithm has been developed. The key to this algorithm is to use a different way of representing the intermediates, to make the deduplication much faster. Using this new algorithm all of the 2,423,212 intermediates for STMV may be generated with forward symmetry factors in only 4 hours. This is a great improvement on the state of the art, especially considering this run was conducted on an average desktop computer using a single 2.4 GHz processor core. The inspiration for this algorithm are the similar hash algorithms often used for duplicating data. A hash algorithm is able to convert a large amount of data down to a much smaller and unique identifier that may easily be compared.

The problem then becomes how to generate a small and unique identifier for each intermediate. As we have seen there are 60 possible representations for each particular layout of building blocks in the STMV capsid. It is possible to pick a single representation from the list of 60 to act as the unique identifier and this representation is termed the "minimal binding pattern".

To determine the minimal binding pattern for a particular intermediate in the protein-only scenario first all 60 symmetry operations are applied. This creates a list of all the identical protein layouts with each one using different protein building-block numbers. Each set of protein numbers in this list of 60 is then individually sorted. The sorting may be conducted because the order of the proteins in the protein-only assembly does not matter. The list of 60 encodings is then itself sorted. Now the first item in the list of 60 is defined to be the unique identifier required. Crucially, which ever proteins were used to represent the intermediate originally the sorting conducted will always yield the representative encoding. The generation of this unique identifier scales linearly with the number of intermediates.

To demonstrate the generation of the unique identifier for an intermediate the cube will again be used as a simple example. Consider intermediate 3 in figure 3.1, which consists of three capsomeres that share a corner. The unique identifying representation of this intermediate is 1,2,3 corresponding to the faces pictured in the diagram. In the construction of the network the intermediate 3 was generated from intermediate 2 which has the binding pattern of 1,2. There was also another intermediate generated from intermediate 2 which had the representation 1,2,5. The duplicate finding algorithm was used as follows to correctly determine that these two generated intermediates are identical.

To convert this 1,2,5 representation to its unique identifier first all 24 possible symmetry operations for the cube (48 with mirror symmetries) must be found. Then each symmetry operation must be applied to create a list of, identical by shape, encodings. This first list of length 48 is summarised thus: [[1,2,5],[5,2,1],[1,3,2], ... ,[1,2,3], ... ,[2,5,6]]. The next steps are to sort each list of three and then the whole list. This brings the unique identifier/minimal binding pattern, of 1,2,3 to the front.

Slightly different processes are required to determine the minimal binding patterns for each of the RNA networks. For the UniRNA and TrRNA cube network only the eight symmetry operations (including mirrors) around face 1 need to be used, because this unique position has to stay in place, hence breaking the overall symmetry. There is also no need for sorting of each encoding in any of the RNA networks because this order represents the binding to the RNA which may be different even for the same protein layout. For the BiRNA network all 48 possible symmetry operations are used for the cube. Applying all the relevant symmetry operations in the TrRNA and BiRNA scenarios generates a list from which the minimal binding pattern may be chosen. However since the direction of the RNA is not being considered each encoding within this list must be copied, reversed and appended to the list. This ensures that when the list is sorted the minimal binding pattern is at the front, independent of the direction the proteins were bound in.

Computation of intermediate networks and symmetry factors

In generating the intermediates a "breadth-first" rather than recursive approach has been used, which makes storage of the network connectivity and calculation of the forward symmetry factors easier. To generate the next intermediates in the network for a protein-only intermediate, a new intermediate is generated for each of the adjacent unoccupied capsomere positions. A note is kept of which proteins were added to create which next intermediates. This is used to then compute the forward symmetry factors (build-up factors). For instance, suppose that two proteins may be added, and addition of each results in the same intermediate being formed. Then the symmetry factor would be two, provided that adding these same proteins did not form any other intermediates as can happen. When RNA is present there is the possibility that only one protein may be added, but due to differential RNA binding two different intermediates may be formed. Each intermediate would then have a build-up symmetry factor of 0.5. It is important to distinguish the RNA binding locations, because this determines where later incoming proteins are able to bind. The RNA intermediates are generated by adding protein adjacent to the ends of the RNA. It is also common for both ends of the RNA to be able to bind the same protein which must be taken into account when calculating the forward symmetry factors. The backward symmetry factors may be worked out by examining any intermediate and its connected edges. However, it is easier to use the forward algorithm but work backwards from the final capsids in each of the networks.

Further algorithms have been written to manage the geometry and symmetry of the virus shapes involved. These allow for automatically drawing an intermediate in two or three dimensions. These intermediate diagrams have then been laid out into the networks using an interface to the graphviz package (Gansner et al., 1993). These visualisations have proved invaluable in correctly determining the network layouts and in their explanation.

All the programs used to generated the intermediates, networks, visualisations and other data in this thesis may be found on the accompanying CD. The structure of the programs, instructions and dependencies are detailed in the read me file located in the root directory of the CD.

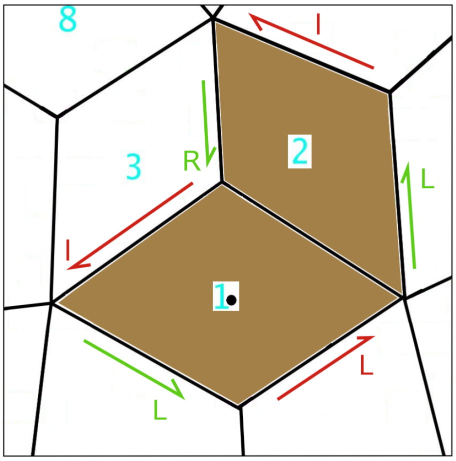

Figure 3.18: Intermediate representation reproduced from Moisant, Zlotnick *et al.*(Moisant et al., 2010). The above dimer is represented by the perimeter pathway:

IRILLL

3.4.4 Numeric Integration

To numerically integrate the assembly reaction equations the livermore solver for ordinary differential equations(LSODE) was used (Hindmarsh, 1983). The interface to this solver was via the python-scipy scientific programming software (Peterson, 2009). This solver is capable of solving stiff and non-stiff differential equations. Stiff equations are more numerically unstable than non-stiff equations in that very small errors build up more rapidly, i.e. exponentially rather than linearly. In the stiff case, backward differentiation formula methods are used and the non-stiff case uses Adams predictor-corrector methods to determine appropriate time steps. The solver was allowed to automatically determine the stiffness of the equations and the Jacobian matrix. Testing for conservation of mass established that the accuracy of the LSODE solver was significantly greater than Runge-Kutta 4th order solvers and that this increase in accuracy was necessary when integrating the larger reaction networks discussed. The correctness of the numerical integration is especially necessary when integrating at stronger, i.e. more negative, bond strengths, because the backward rates are very low compared to the forward rates. The time taken to integrate the cube networks is of the order of a few seconds. The dodecahedral networks take longer: the BiRNA dodecahedral network takes about half an hour, the UniRNA network about a day, and the TrRNA (due to the much larger number of intermediates) takes about a week. The main measure of the accuracy in the integration has been the conservation of mass of both the protein and RNA. Note that this conservation was not set as a constraint of the integration. This conservation of mass in all the kinetic simulations is accurate to at least seven significant figures. However, to achieve this at some of the longer times and at more negative capsomere contact energies a modification of the kinetic equations was required.

To compute the kinetics the (very small) 8×10^{-6}M protein concentration and the (very large) 1×10^{8} which are many orders of magnitude apart because these are the values derived from experiments. As a result, errors build up. In order to cope with this a protein concentration of 8M is used instead and a modified on-rate, after the calculation of k_{off}, to 1×10^{2}. This produces no change to the concentrations of the simulation other than that they are now all 1×10^{6} higher, which is compensated for by multiplying each concentration by 10×10^{-6}. In this setting the conservation of mass was found to be more accurate. In using the modified kinetic equations a ten-fold increase in the number of time steps were required to maintain the accuracy. This slowed down the kinetic runs, with the previously half-hour running BiRNA dodecahedral network now taking a day to complete on average. Running the unmodified kinetic equations with the greater number of time steps did not significantly increase the numerical accuracy.

3.5 Discussion

Based on Zlotnick's protein-only assembly model, models for capsid assembly in the presence of genomic RNA have been designed. These allow investigation into the pathways of RNA virus assembly. It has been shown that by determining the networks of RNA virus intermediates, new information can be found concerning the number of intermediates and pathways and also how these pathways are likely to affect the kinetics. Factors affecting the kinetics include the number of capsomere contacts formed in the pathways, and the path splitting to capsid as well as the occurrence of dead-end intermediates. The three different scenarios discussed above demonstrate the likely advantages, in terms of dead-end intermediates, of a scenario starting assembly in the middle of the RNA genome. As discussed for figure 3.16 fixing the TR position at the start makes self corrections close to the TR position more difficult.

In one simulated protein-only network for the 30-mer Zlotnick (Moisant et al., 2010) included only the lowest energy intermediates. By removing all other intermediates from the network this created dead-end pathways, where some of the lowest energy intermediates had no further forward path to capsid. In the kinetic simulations these intermediates formed a significant kinetic trap.

It is easy to predict that the dead-ends due in the RNA network assumptions will act as a similar kinetic trap. Due to the large number of dead-ends in the dodecahedral network and the length of the dead-end pathways it is likely that the dead-ends will severely inhibit the RNA virus formation. Even without the dead-ends the RNA networks are likely to assembly capsid more slowly as not all the protein assembly pathways are available. However, a potential kinetic advantage of the RNA networks is that the RNA effectively acts as a nucleating point for the protein following principles of heterogeneous nucleation. This limits the number of intermediates in a similar way to the nucleation step introduced for the STMV networks (Moisant et al., 2010). The nucleation step, introduced by Zlotnick, reduced the kinetic traps, avoiding the situation where much of the protein is stuck in smaller intermediates. This increased the formation rate of capsid and the RNA nucleation is likely to have a similar effect. The interplay of these various factors will be shown in the next chapter, using numeric integration to solve the rate equations for the cube to predict the assembly kinetics. Later the dodecahedral networks will be kinetically modelled to show the effects of the increase in scale.

In the generation of these networks there are a number of implicit and explicit assumptions. It is these assumptions that are to be carefully investigated and understood before moving onto larger viruses requiring further assumptions. Many of the assumptions correspond to those in the protein-only model of Zlotnick. Such as only reactions allowed in the network are able to occur and that it is an equilibrium based model.

An additional assumption in the presence of RNA is that the proteins may only bind to each other when RNA is available. Additionally the proteins may only bind and fall off at the ends of the RNA. It has also been assumed that the proteins in the RNA simulations diffuse to their binding sites similarly to the protein-only network. With the diffusion-limited protein binding step being the rate-limiting step, there has been no additional symmetry factors created for RNA binding, such as when both ends of the RNA are able to bind the same capsomere. Similar to this assumption, the different directions of the RNA have not been taken into account in the generation of these networks, because no RNA binding steps are present even though this does leave the first reaction of the RNA networks without a dissociation rate at all. In future work the first change to the assumptions would likely be to put in RNA binding energies. The multiple assumptions emphasise the importance of having a simple model for investigation of the qualitative features of the assembly process.

Finally, the algorithms developed for this project are demonstrably an improvement on what is already present in the literature. The increases in speed of the algorithms achieved here are essential when later calculating the much larger RNA networks for STMV, which contain hundreds of millions of different intermediates. The main duplicate intermediate finding algorithm is also trivially parallelisable, which results in a further speed increase. Useful improvements have also been found that increase the accuracy of the kinetic simulations. In future work it would be possible to investigate alternative models of virus formation using very similar networks and algorithms. The first alternative model could assume that proteins bind to the RNA first. A further model, potentially more relevant at high protein concentrations or high RNA binding energies, could have all the proteins bind to the RNA and then trigger refolding of the RNA for better packaging into capsid.

Chapter 4

Cube Results

4.1 Introduction

In this chapter the rate equations for all the cube assembly networks have been numerically integrated to determine the kinetics of assembly. The first simulations are conducted on the protein-only network of cube assembly in order to later contrast this with the RNA networks. It will be shown that the cube protein-only simulation has very similar kinetic behaviour to the dodecahedron in Zlotnick's 2005 paper (Endres et al., 2005). For simplicity the protein-only network has been modelled with no μ factor, which Zlotnick used to down weight unlikely intermediates. There is also no modelling of a nucleation step for the protein-only network. The UniRNA, TrRNA and BiRNA networks each have their own kinetic behaviour and these differences will be described for a full, range of capsomere contact values and time lengths. The interconnectedness of these time frame to capsomere contact energies will also be understood in terms of the viral capsid concentration. Finally an investigation in to the parameters of the kinetic simulations has been conducted using the easily understood cube model.

4.2 Protein-Only Simulations

The first experiment conducted was to determine an appropriate ΔG_c bond value at which to run the kinetic simulations. An interesting starting point is the bond contact energy that results in the free capsomere, also termed free monomer, having the same concentration as the final capsid. This has been shown previously with the dodecahedron in (Zlotnick, 1994). To find this value

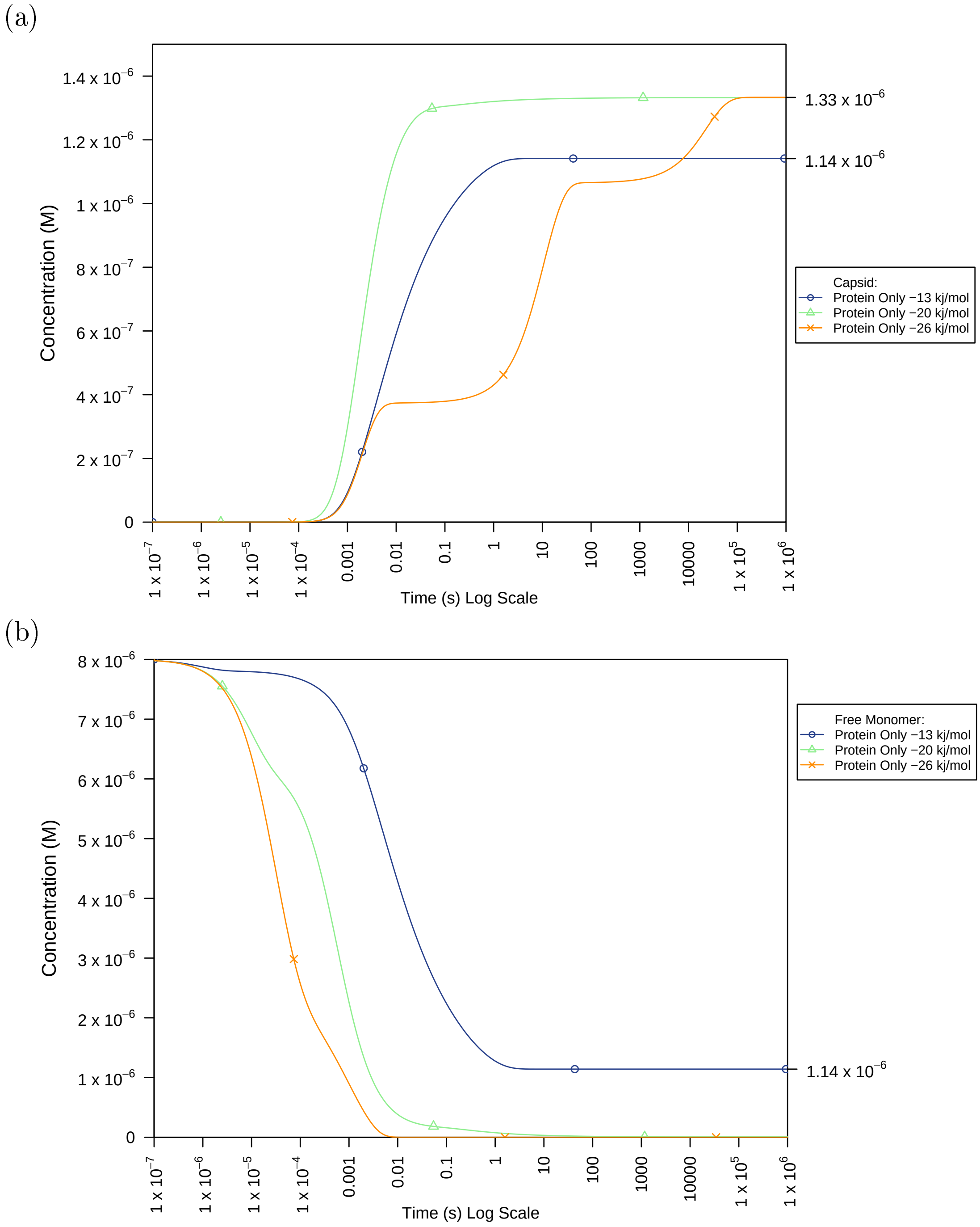

Figure 4.1: Protein-only simulation kinetics at ΔG_c values of -13, -20, and -26 kJmol^{-1}. Capsid concentration is shown in (a) and the corresponding free monomer concentration in (b).

Simulation	Equilibrium bond contact energy $kJmol^{-1}$	Equilibrium FM/Capsid amount (M)
Protein	-13.07	1.141 $\times10^{-6}$
UniRNA	-12.30	1.111 $\times10^{-6}$
BiRNA	-12.74	1.111 $\times10^{-6}$
TrRNA	-12.51	1.111 $\times10^{-6}$

Table 4.1: Bond strengths required for an equilibrium where the free monomer has the same concentration as the final capsid.

for the cube the differential equations for each simulation were numerically solved using iteratively chosen ΔG_c values. The final ΔG_c values were such to achieve accurate and equal concentrations for the free monomer and capsid. The results are summarised in table 4.1. The most negative ΔG_c value of -13.07 kJmol^{-1} is the protein-only experiment. This is a result of the protein-only network having the highest number of ways of disassembling the capsid across the network. Consistent with this, the least negative ΔG_c is required to stabilise the UniRNA capsid as there are relatively few ways the capsids can fall apart. The reasoning follows that the BiRNA experiment has a more negative value than the TrRNA experiment. This is because the pathways of the TrRNA network exclude the possibility of the TR bound dimer detaching from the RNA. This again leads to fewer disassembly pathways and acts to stabilise the capsid. This effect accounts for some of the subtleties in the later comparative graphs. Representative ΔG_c energies of -13, -20 and -26 kJmol^{-1} were chosen, relating to 1x, 1.5x and 2x the ΔG_c of free monomer/capsid equilibrium in the cube protein simulation.

A graph of the capsid concentration over time for the protein-only experiment at these representative ΔG_c values is shown in figure 4.1(a). The time scales here are relative and do not necessarily correspond to a real biological experiment. At the free monomer / capsid equilibrium ΔG_c energy the capsid builds up smoothly reaching 99 % of the equilibrium amount after 1.5 seconds. With the bond contact energy of -20 kJmol^{-1}, which would result in more energetically stable intermediates, faster virus formation takes place. The equilibrium is now in a position where (practically) all the protein is in complete capsids. However, against this pattern, increasing the ΔG_c further to -26 kJmol^{-1} slows the progression to capsid and produces plateaus along the way.

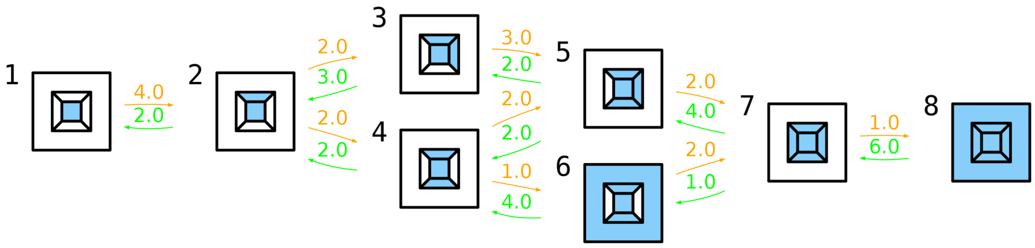

Figure 4.2: The protein-only cube network.

Since efficiently growing capsids need a ready supply of free monomer building blocks a corresponding graph showing the free monomer concentrations was produced for figure 4.1(b). Here we can see at -26 kJmol^{-1} there is no appreciable free monomer concentration after 0.001 seconds. This is the point at which the initially rapid capsid growth stops. There is still a significant ratio of free monomer to capsid at the -20 kJmol^{-1} bond value until all the protein is in capsid. Finally, by definition, the final concentration of free monomer at -13 kJmol^{-1} is 1.14 $\times 10^{-6}$ M, the same as the capsid.

Taking a closer look at what is going on in these initial simulations of -13 kJmol^{-1}, -20 kJmol^{-1}, -26 kJmol^{-1} we can look at the individual intermediate concentrations, figures 4.3, 4.4, 4.5 respectively. A reminder of the cube protein-only network layout is shown in figure 4.2.

At the free monomer / capsid equilibrium ΔG_c value of -13 kJmol^{-1} we see very low concentrations of the intermediate species due to their relatively low stability and likelihood to break apart. This was found also in the original Zlotnick paper (Zlotnick, 1994), where it was suggested that this graph could be mistaken for a reaction mechanism only between the free monomer and final capsid, without any stable intermediates. With the increase in ΔG_c to -20 kJmol^{-1} the intermediates have more significant concentrations and we see a quick succession from one size to the next. Notably the assembly is via the relatively stable intermediates 3 and 5 as opposed to 4 and 6.

At -26 kJmol^{-1}, like -20 kJmol^{-1} initially we see a quick build up of intermediates and to even higher levels as they are even more stable. Once the free monomer becomes scarce the forward reactions are reduced massively and with the backward reactions being slow due to the high bond strength this results in very little flux. Since there are so few reactions taking place the protein remains in the intermediates it was in when the free monomer became scarce.

The protein in these intermediates is said to be kinetically trapped. It is these kinetic traps that then affect the assembly as previously discussed and first described for the dodecahedron by Zlotnick (Zlotnick, 1994). It is not until the time scale is increased by orders of magnitude that we start to see these stable intermediates break apart, providing the free monomer for capsid formation. At even higher bond strengths it is possible to have more plateaus and even have the protein kinetically trapped in a two dimer intermediate.

Notably at -26 kJmol^{-1}, it is still the relatively stable intermediates 3, 5, and 7 that in turn contain the kinetically trapped protein. In contrast, the less-stable intermediates (4 and 6) have a much higher relative concentration than when using -20 kJmol^{-1}. This is because at the more negative ΔG_c value all the intermediates are reasonably stable, despite few numbers of bonds and it comes more down to the branching of the network. Due to the model setup the intermediates 3 and 4 both gain the same amount of material from intermediate 2 over the course of the simulation.

The steady states that appear in the -26 kJmol^{-1} simulation (e.g. between 0.001 seconds and 0.1) could be mistaken as the final equilibrium being reached. Therefore care has been taken to ensure equilibrium is always fully reached. The cube networks although very small sometimes take a whole simulated day to equilibrate. It is clear from this that larger and longer networks of interactions may never reach the true thermodynamic equilibrium, certainly over a time frame that could be numerically solved.

The assembly behaviour of the protein-only simulation over the full range of relevant ΔG_c energies is shown in figure 4.6. This graph shows the trade off between the ΔG_c value and the capsid assembly time. At the least negative ΔG_c values the reaction proceeds slowly due to the high dissociation rates. It therefore requires a long time for the stable capsid to form. For times longer that 1 second we can see that the final capsid concentration is able to equilibrate to its maximum value until -20 kJmol^{-1}. After this point, due to the more negative ΔG_c, the capsid concentration starts to require much longer to equilibrate. Although with longer time periods the capsid is able to equilibrate to its maximal value. To achieve the most capsid in the shortest amount of time, looking at the graph, a time of 0.1 seconds and a ΔG_c of -20 kJmol^{-1} would be a good choice. To achieve the optimal amount of capsid on a shorter time scale the ΔG_c energy should be less negative to coincide with the peak capsid amount. This reduction is to optimise against the kinetic traps that still take a little time to resolve.

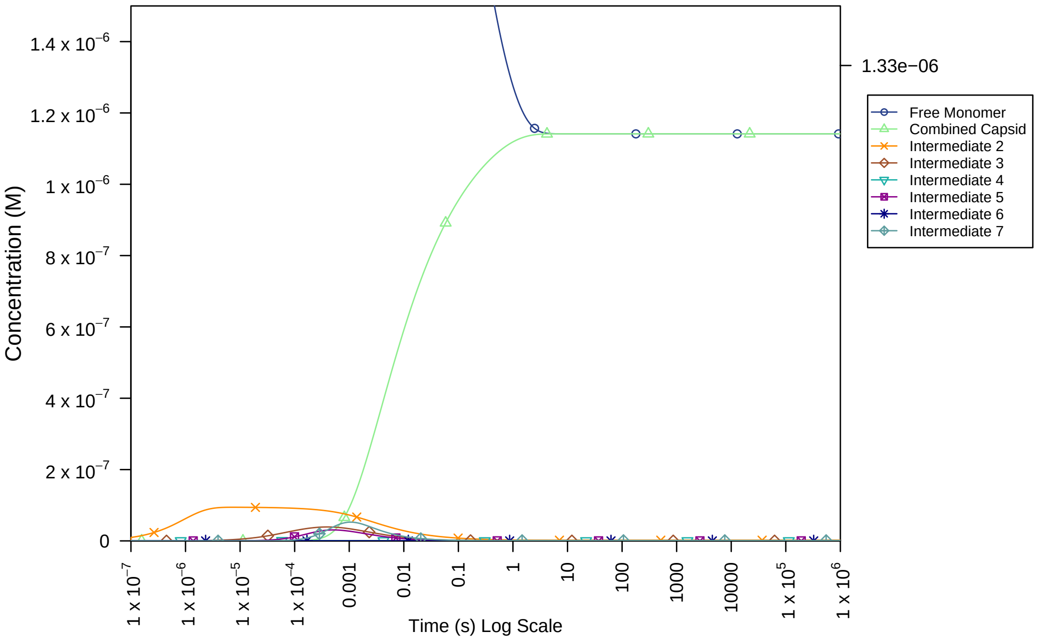

Figure 4.3: Intermediate concentrations in the protein-only cube capsid simulation at a ΔG_c of -13 kJmol^{-1}.

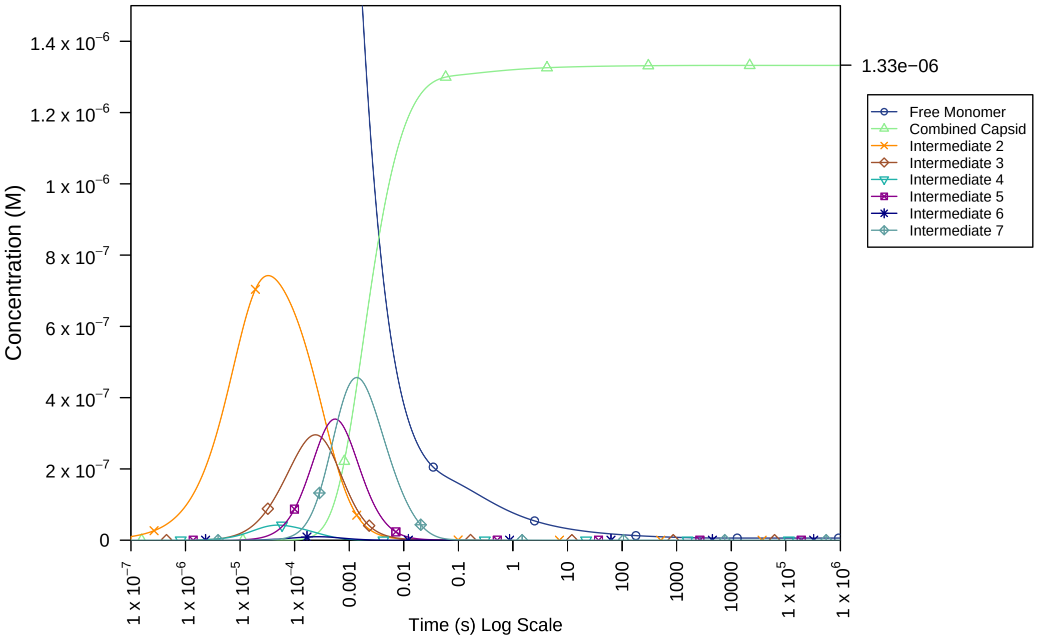

Figure 4.4: Intermediate concentrations in the protein-only cube capsid simulation at a ΔG_c of -20 kJmol^{-1}.

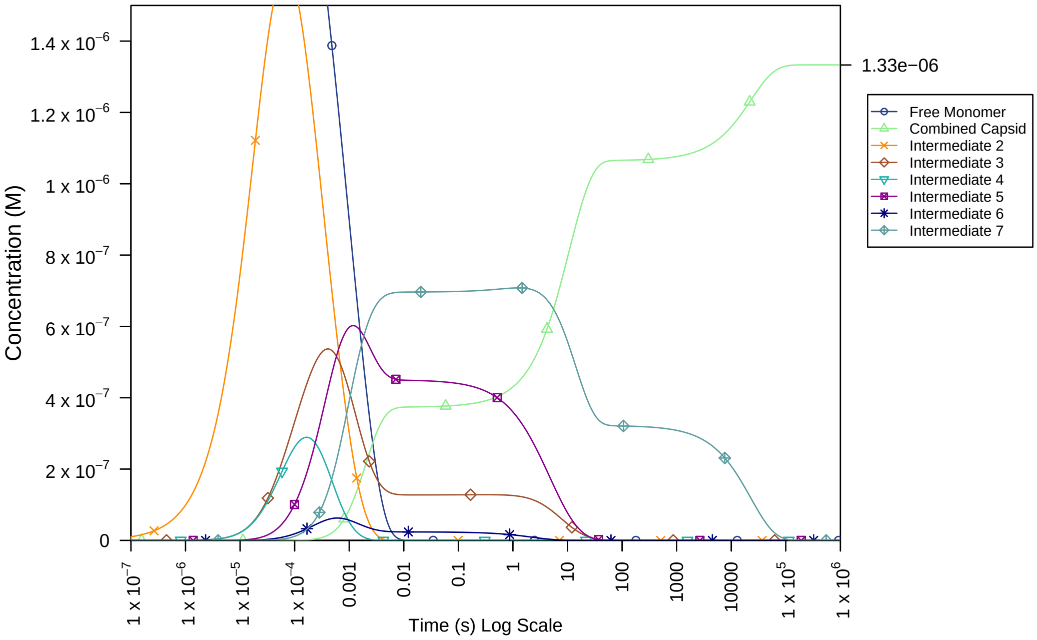

Figure 4.5: Intermediate concentrations in the protein-only cube capsid simulation at a ΔG_c of -26 kJmol^{-1}.

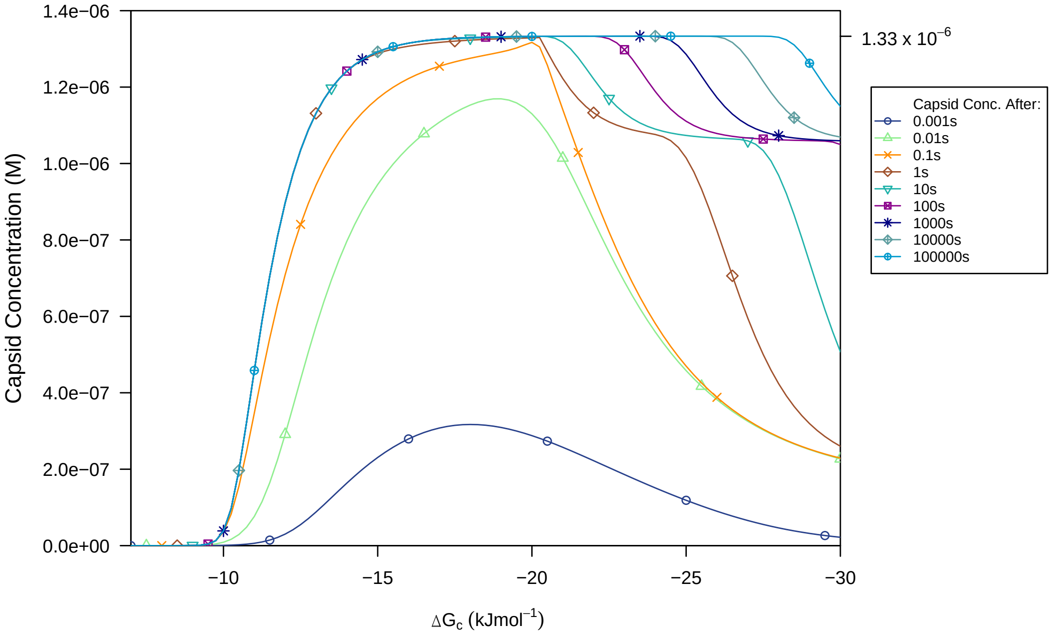

Figure 4.6: The capsid concentration after different time periods and a range of ΔG_c values. In the protein-only cube simulation.

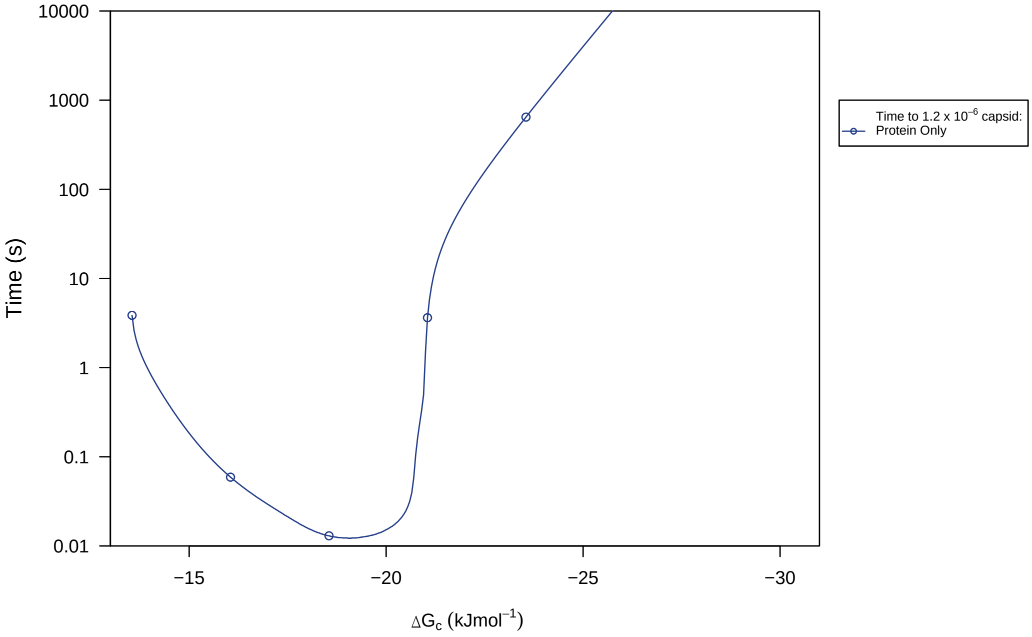

Figure 4.7: Time taken to reach 90 % of the maximum possible capsid concentration across a range of ΔG_c values in the protein-only simulation. The minimum is reached for a ΔG_c value of about -19 kJmol^{-1}.

The efficiency of capsid assembly over time is an interesting point to consider when simulating the networks. This is especially true when considering that even in the cube simulations it can still take several (simulated) hours to form large amounts of capsid. The absolute maximum possible capsid concentration is simply the initial protein monomer concentration divided by the number of proteins in the capsid. For the cube this capsid concentration works out to be a value of 1.33×10^{-6} M when using the initial monomer concentration of 8×10^{-6} M. To demonstrate the efficiency of capsid assembly a graph showing the time to 90 % of this maximum capsid value,1.33×10^{-6} M, was produced for figure 4.7. For the protein-only simulation there is quite a range of ΔG_c energies between -15.5 and -21 kJmol^{-1} where the time is less than 0.1 seconds. At more negative ΔG_c energies however the kinetic traps again come into play and dramatically increase the time taken.

4.3 RNA Simulations

The RNA simulations have been initially compared to the protein-only simulation in the following graphs; 4.8, 4.9 and 4.10. These graphs compare the capsid concentration over time at the three chosen representative ΔG_c values.

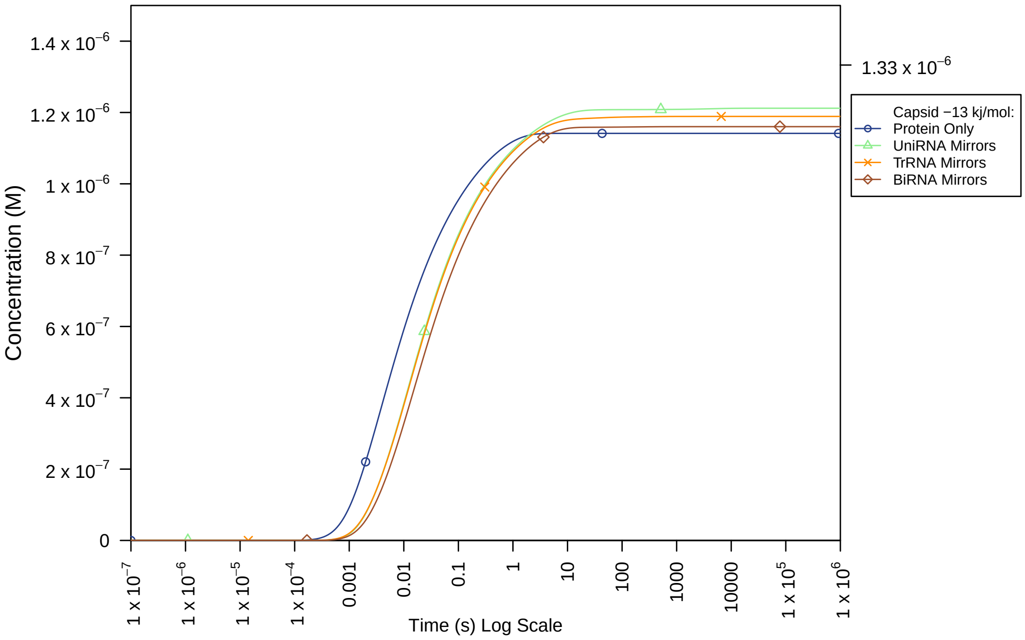

Figure 4.8: Formation of capsid over time in the four cube scenarios at a ΔG_c of -13 kJmol^{-1}

At the ΔG_c of -13 kJmol^{-1} the protein-only capsid appears the quickest. This is because there are more possible forward reactions and because the reaction network is slightly shorter. Later the final capsids of the four scenarios equilibrate in order of the number of ways to disassemble in their respective networks. This shows the RNA scenarios having a grater capsid concentration at equilibrium, as would be expected from the free monomer / capsid equilibrium ΔG_c values. For instance the UniRNA simulation has the least negative ΔG_c value at this free monomer / capsid equilibrium. This means that at any chosen ΔG_c value the equilibrium will be pushed more towards the capsid relative to the other simulations. A ΔG_c of -20 kJmol^{-1} pushes all the equilibriums further towards the now more stable capsid. The speed of assembly of the protein-only to RNA simulations is also now more similar. Finally we see the dead-end intermediate acts to reduce the UniRNA capsid concentration until this trap starts to disappear at the end of the time period. With the ΔG_c of -26 kJmol^{-1} we see the presence of the kinetic traps on the protein-only simulation take effect. This is in addition to a more pronounced kinetic trap in the UniRNA simulation.

A breakdown of the ΔG_c of -13 kJmol^{-1} for the three RNA scenarios are shown in; 4.11, 4.12, 4.13. In the graph of the UniRNA (4.11) we see that the

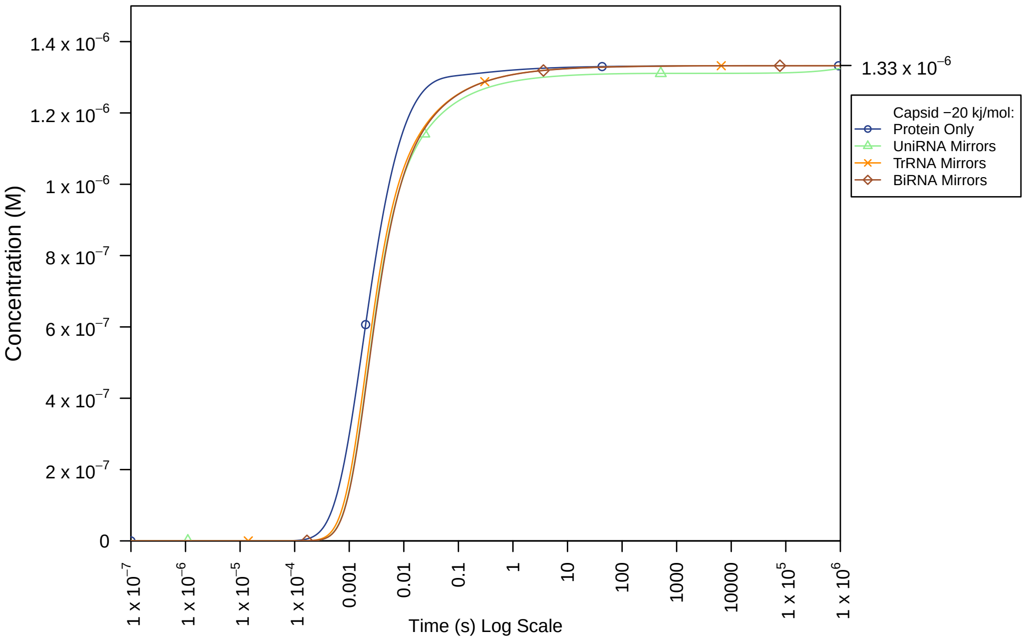

Figure 4.9: Formation of capsid over time in the four cube scenarios at a ΔG_c of -20 kJmol^{-1}

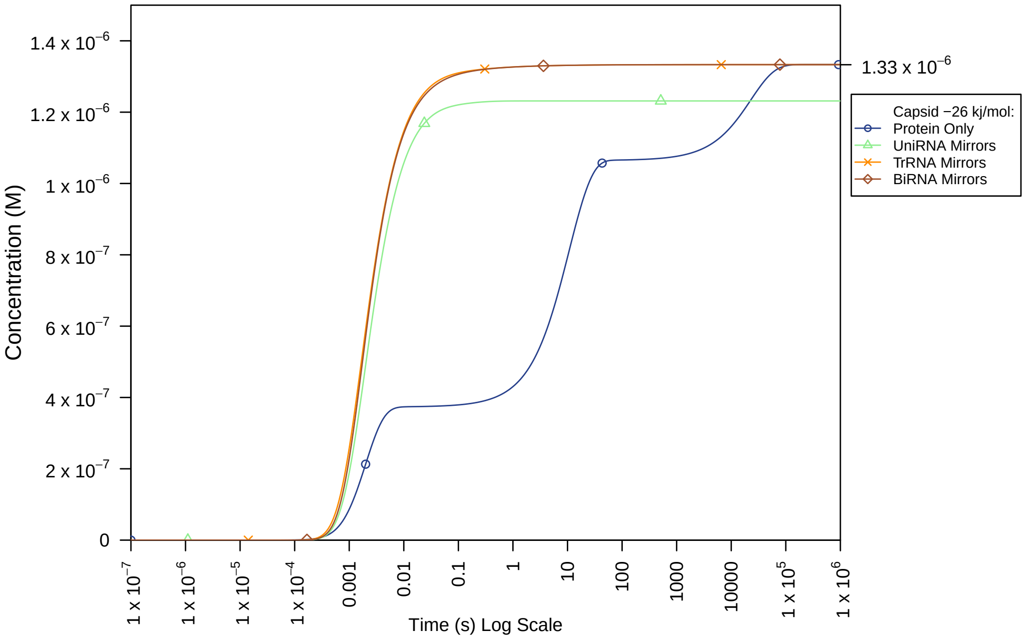

Figure 4.10: Formation of capsid over time in the four cube scenarios at a ΔG_c of -26 kJmol^{-1}

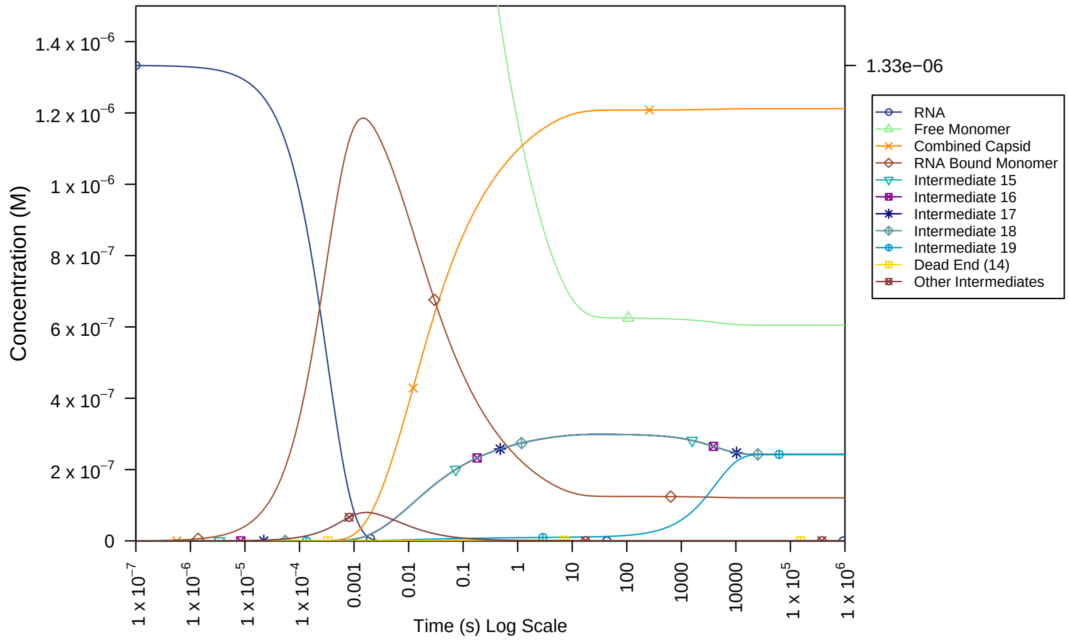

Figure 4.11: Intermediate concentrations in the UniRNA cube capsid simulation at a ΔG_c of -13 kJmol^{-1}.

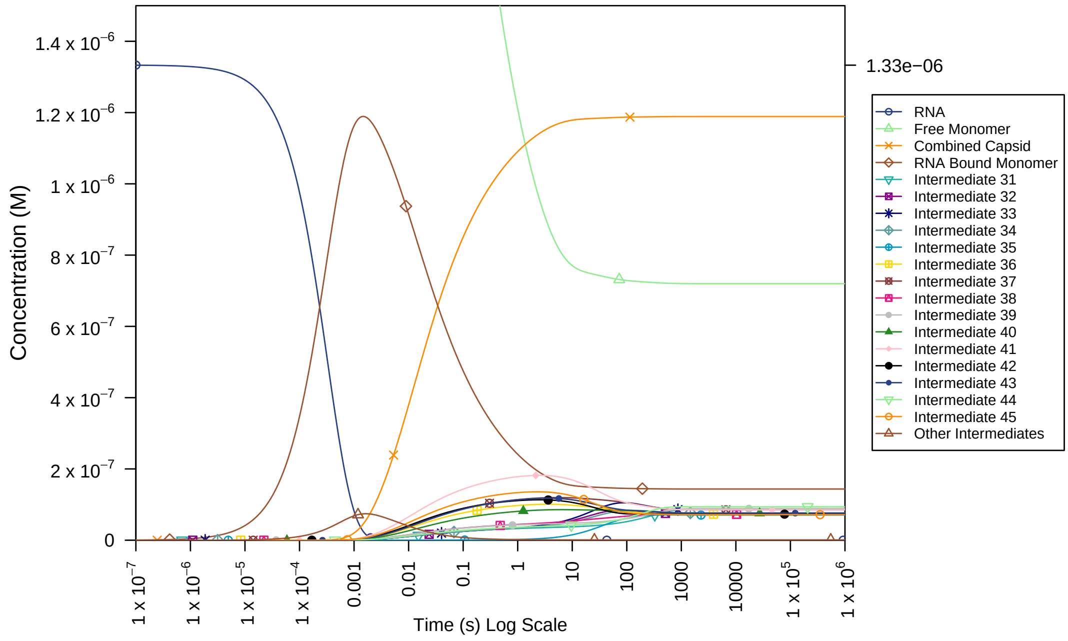

Figure 4.12: Intermediate concentrations in the TrRNA cube capsid simulation at a ΔG_c of -13 kJmol^{-1}.

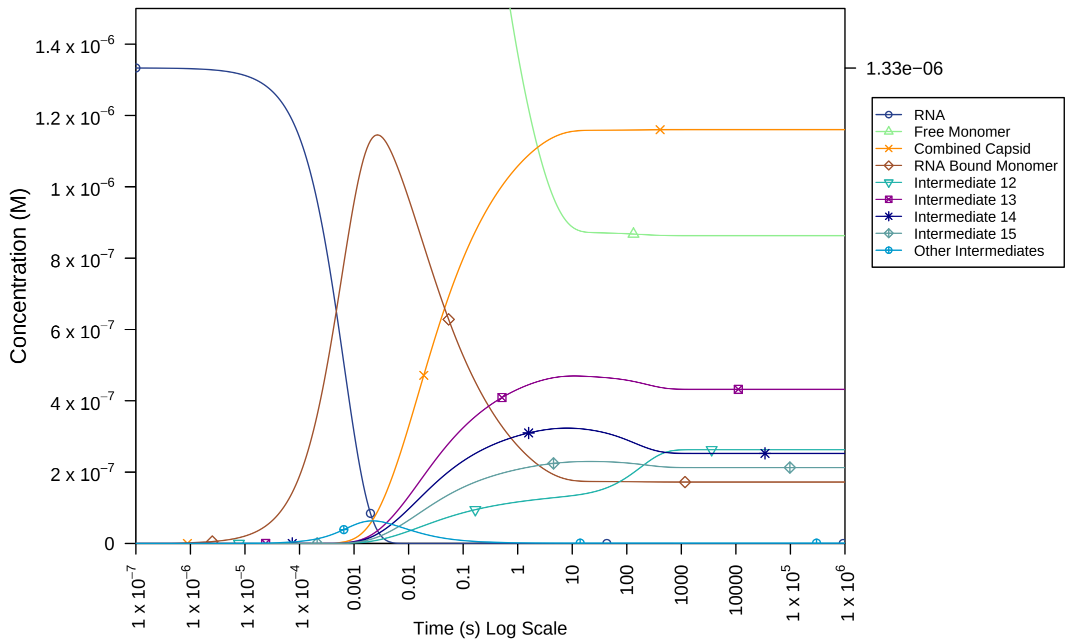

Figure 4.13: Intermediate concentrations in the BiRNA cube capsid simulation at a ΔG_c of -13 kJmol^{-1}.

final capsid intermediate 19 initially has a low concentration due to intermediate 4 on its pathway being relatively unstable. Eventually all five of the final capsids in the UniRNA network equilibrate to the same thermodynamic equilibrium concentration. This is because all the capsids have the same number of bonds and, in the UniRNA case, the symmetry factors to capsid are the same. As previously described this thermodynamic equality does not hold for the TrRNA or BiRNA simulations and these equilibrate to a few different possible capsid concentrations. For an example of kinetic versus thermodynamic properties we can consider the final capsid concentrations of the BiRNA at -13 kJmol^{-1}. Capsid 13 has the highest concentration. This is because its pathway contains relatively stable intermediates and also because it has two directly previous intermediates unlike the other capsids. The capsid with the lowest concentration initially is number 12, because intermediates 4 and 7 have relatively low numbers of bonds. Additional factors in capsid 12's concentration though is the symmetry between intermediates 7 and 8 and the additional pathway joining to intermediate 5. As the network equilibrates these extra factors push the capsid concentration of 12 from the lowest to the 2nd highest.

The corresponding breakdown for the ΔG_c of -20 kJmol^{-1} are shown in; 4.14, 4.15, 4.16. While the corresponding breakdown for the ΔG_c of -26

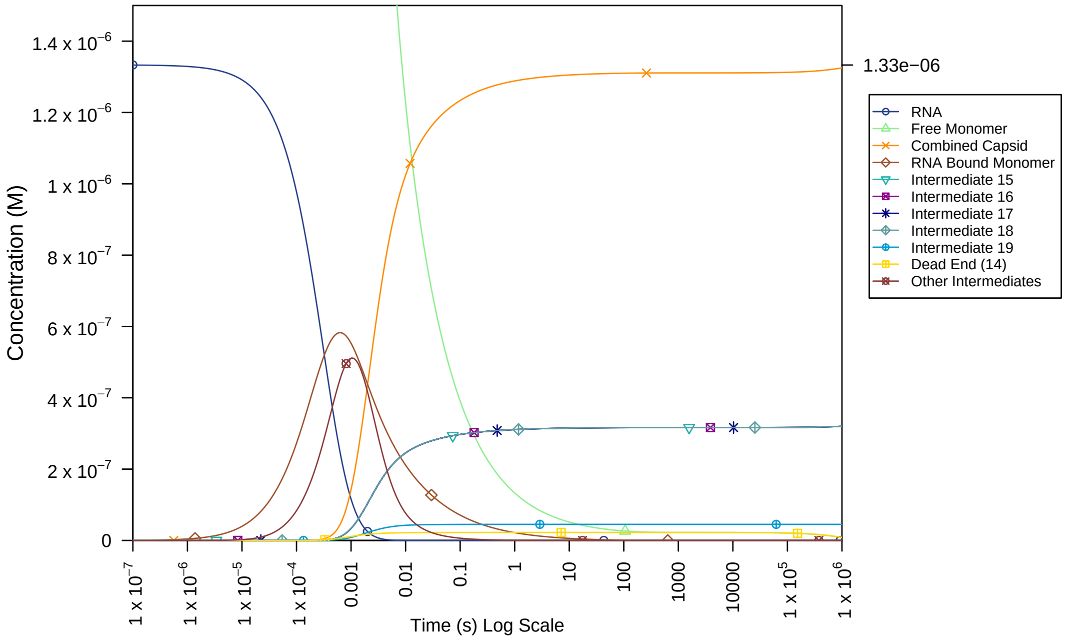

Figure 4.14: Intermediate concentrations in the UniRNA cube capsid simulation at a ΔG_c of -20 kJmol^{-1}.

kJmol^{-1} are shown in; 4.17, 4.18, 4.19.

At these more negative ΔG_c energies there is little equilibration in the networks due to all the material being in the final capsids. There is in fact very little disassembling of intermediates at all and the backward rate is effectively zero. Especially considering the number of backward reactions that would need to happen in a row for an RNA strand to form into a different capsid. When there is no effective backward rate it is only the symmetry factors that distinguish between the pathways in the network. Interestingly we still end up following the most energetically stable pathways. This is because these tend to be the more compact structures and have more symmetry in terms of adding proteins early on.

At the more negative ΔG_c of -26 kJmol^{-1} we can see a significant concentration of the dead-end in the UniRNA network 4.17. To investigate this further, graph 4.20 (similar to the protein-only simulation graph in figure 4.6) was produced. In figure 4.20 we see the kinetic trap having a larger effect after ΔG_c of -20 kJmol^{-1}. This level is similar for the kinetic traps in the protein-only simulation. Unlike the protein-only simulation this trap is not resolved in time. The reasons for this are as follows, if a protein was to detach from intermediate 14 (the kinetic trap) the only RNA it is likely to bind to

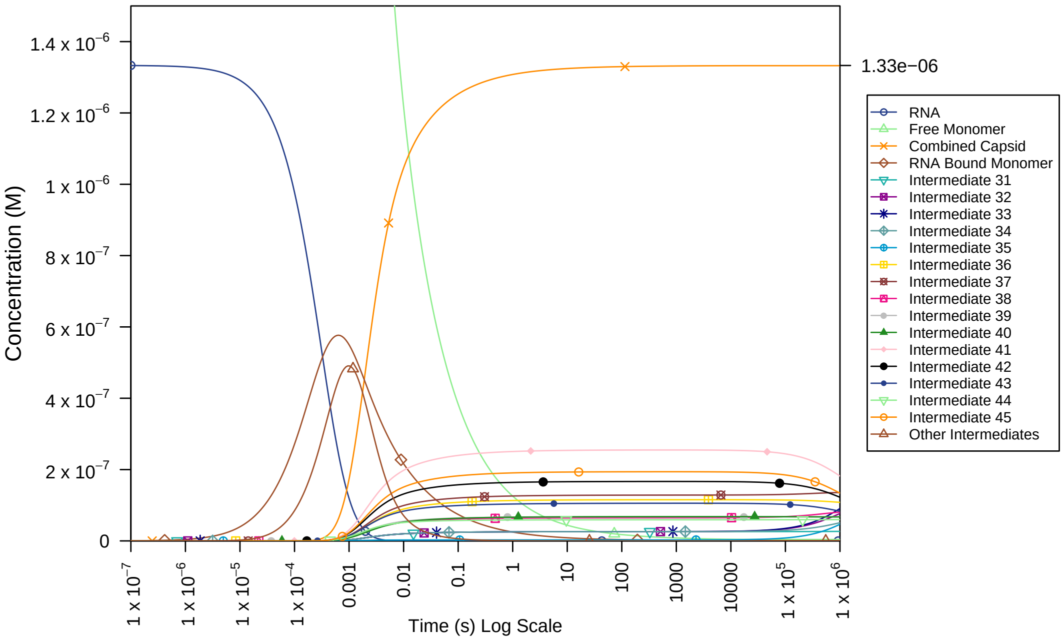

Figure 4.15: Intermediate concentrations in the TrRNA cube capsid simulation at a ΔG_c of -20 kJmol^{-1}.

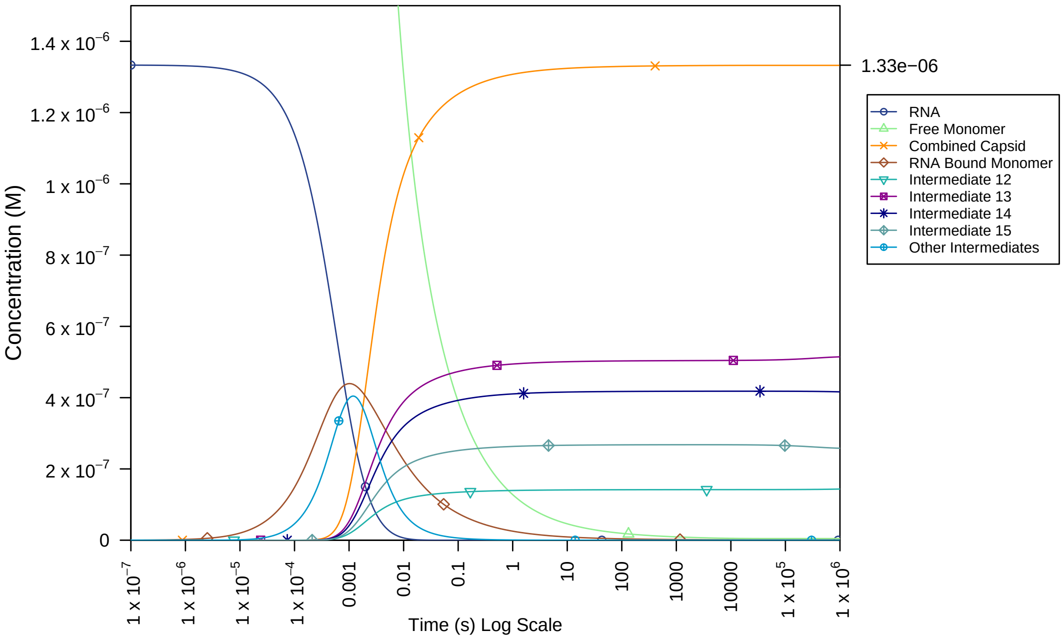

Figure 4.16: Intermediate concentrations in the BiRNA cube capsid simulation at a ΔG_c of -20 kJmol^{-1}.

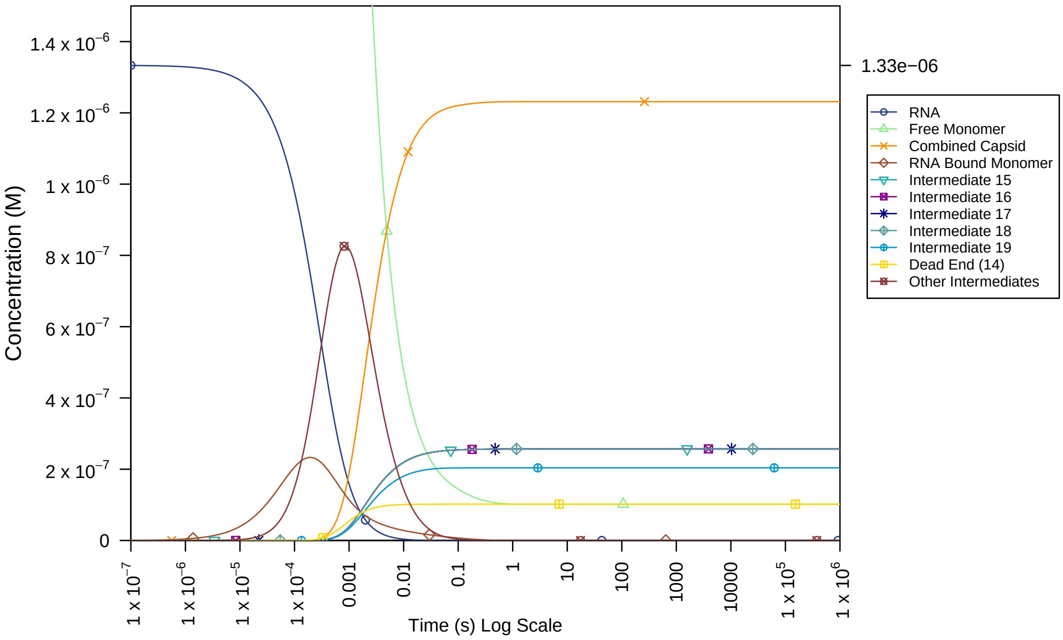

Figure 4.17: Intermediate concentrations in the UniRNA cube capsid simulation at a ΔG_c of -26 kJmol^{-1}.

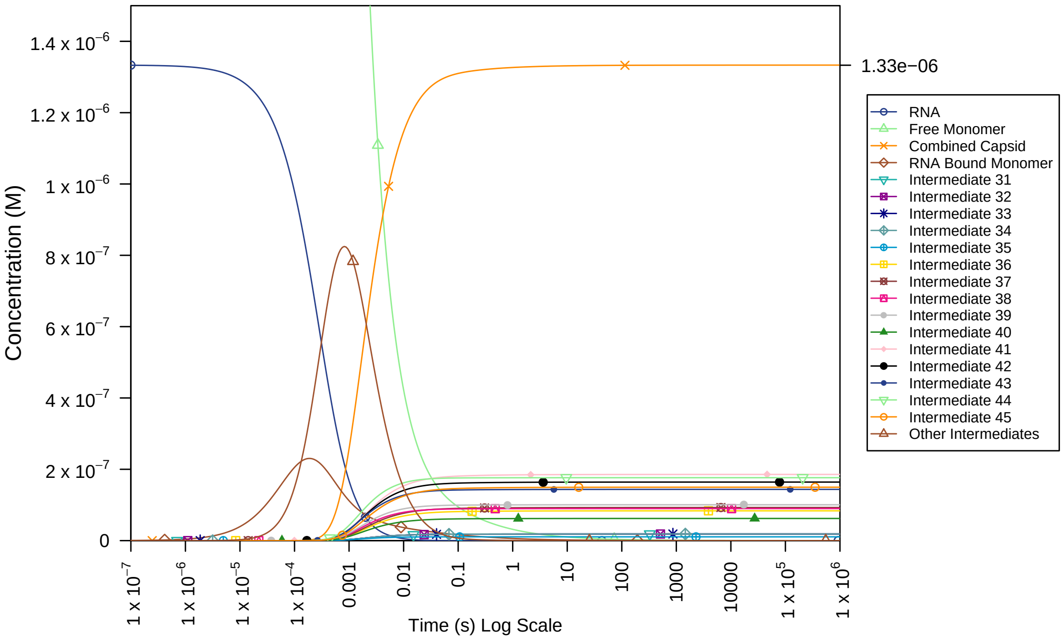

Figure 4.18: Intermediate concentrations in the TrRNA cube capsid simulation at a ΔG_c of -26 kJmol^{-1}.

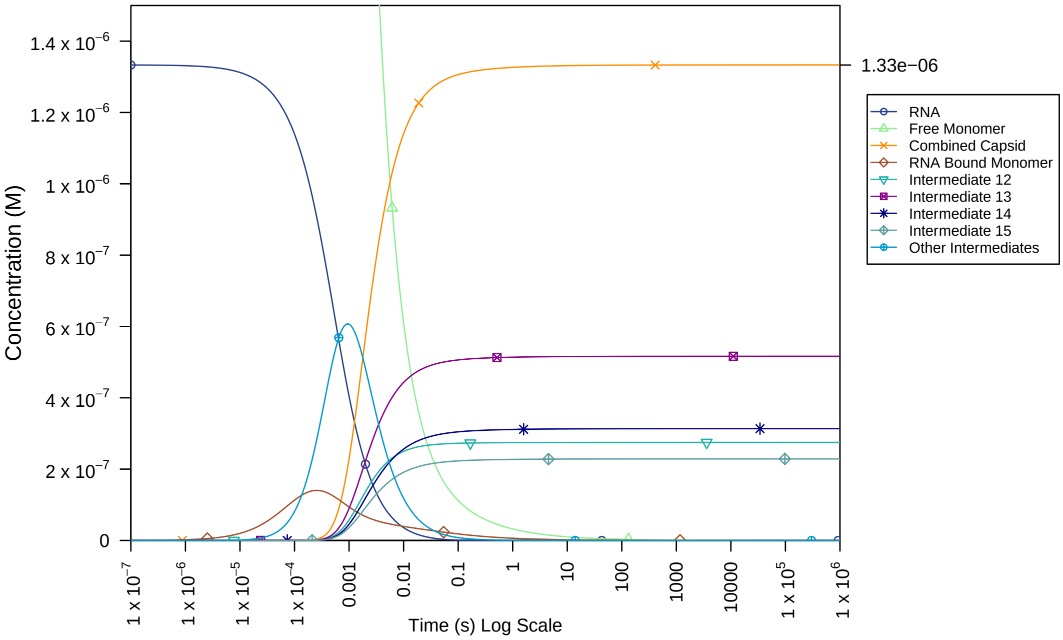

Figure 4.19: Intermediate concentrations in the BiRNA cube capsid simulation at a ΔG_c of -26 kJmol^{-1}.

is then available in intermediate 8. This recreates intermediate 14. In fact at least two proteins would need to fall off intermediate 14 to allow for a choice in pathway to a final capsid. Due to the small backward rates at the higher ΔG_c values, this is again extremely unlikely. The maximum time in this graph (4.20) of 100000s is briefly flat at the top, indicating this length of time allows the system to equilibrate to the maximum capsid concentration over a larger range of ΔG_c values. Increasing the time still further would increase the range of ΔG_c values the maximum capsid concentration is achieved. To complete the comparison figure 4.21 has been included to show the time lines for the TrRNA simulation. Since there are no dead-ends in this network the capsid concentration does not increase due to equilibration or decrease due to material being trapped in the dead-end. The corresponding BiRNA graph is practically identical to the shown TrRNA graph and has been omitted for this reason.

Finally to show the efficiency of assembly we add the RNA simulations to the graph showing the time until 90 % of the protein is in capsid, see figure 4.22. At ΔG_c values between about -15 kJmol^{-1} and -21 kJmol^{-1} the time taken is below one second across all the different networks with the protein-only experiment being the quickest. However once the kinetic traps form in

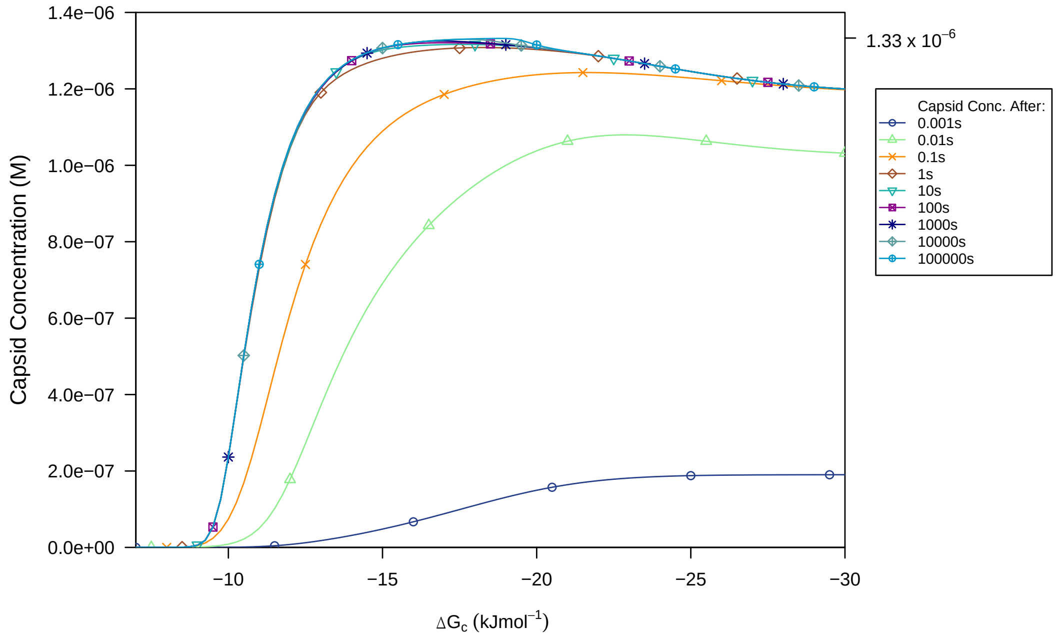

Figure 4.20: The capsid concentration after different time periods and a range of ΔG_c values in the UniRNA cube simulation.

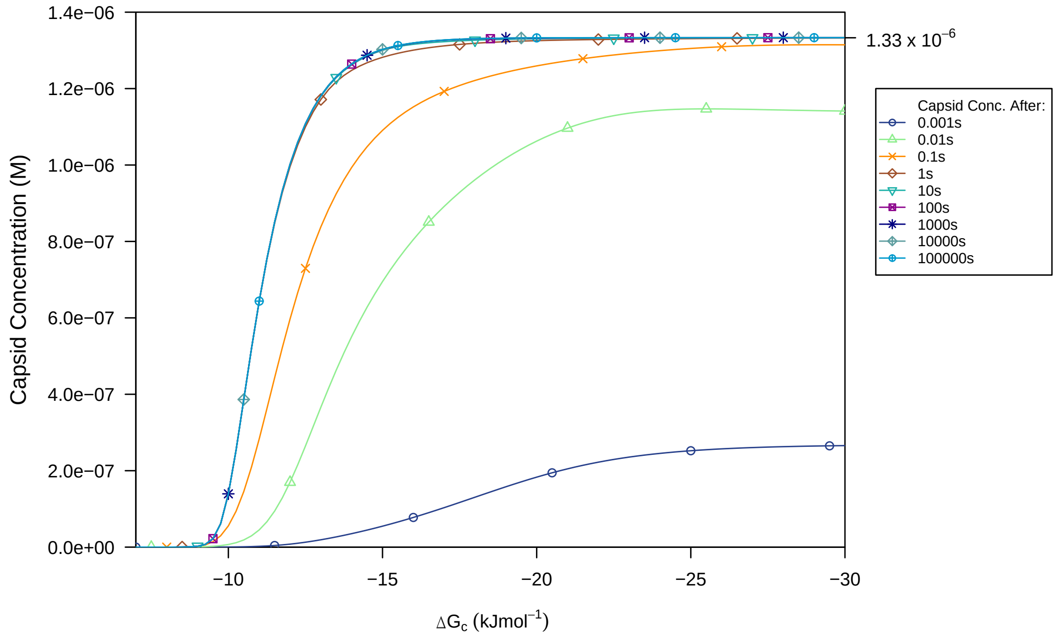

Figure 4.21: The capsid concentration after different time periods and a range of ΔG_c values in the TrRNA cube simulation.

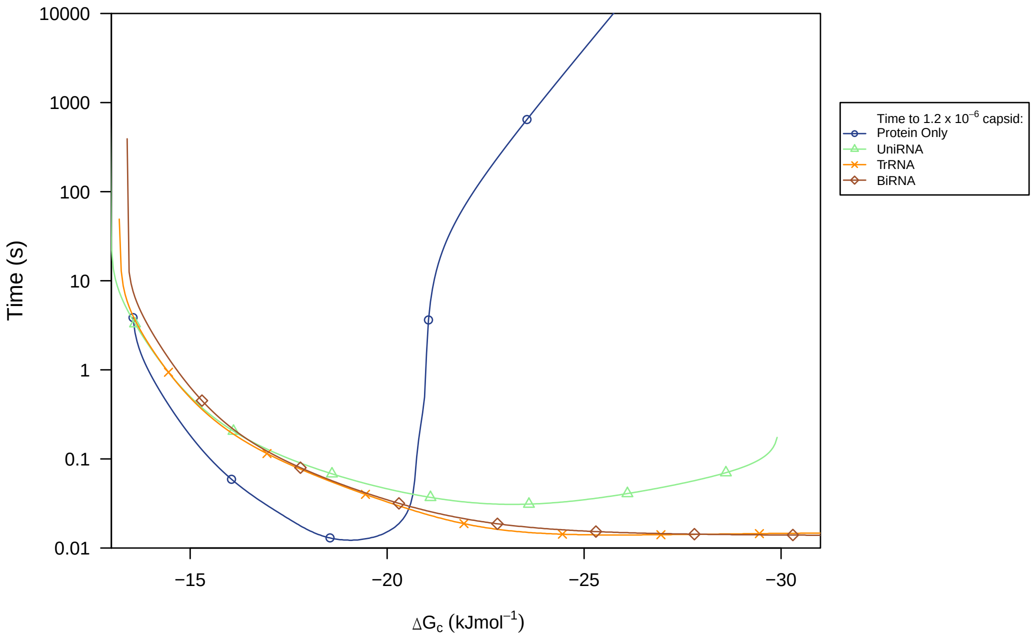

Figure 4.22: Time taken to reach 90 % of the maximum possible capsid concentration across a range of ΔG_c values comparing all the cube simulations.

the protein simulation the efficiency is dramatically reduced relative to the RNA simulations. The kinetic trap in the UniRNA network can also be seen manifesting as the time line tending upwards in the UniRNA simulation. The abrupt end of this UniRNA lane is the point at which 90 % capsid is just no longer reached. As can be seen in figure 4.20 increasing the time makes no difference. There is very little difference between the TrRNA and BiRNA, suggesting that the smaller BiRNA network would be a good model substitute for the more complicated TrRNA network.

4.4 Parameter Investigation

Using the cube assembly model we can take the opportunity to investigate some of the other parameters in the model. Changing the forward reaction rate just acts to speed up or slow down all the reactions proportionally. So all the graphs look exactly the same, however appear spread out in time. Theoretically predicting the on-rate for proteins of MS2 dimer size is almost impossibly complicated. For this reason on-rates in laboratory experiments are determined empirically due to the myriad of complex factors involved. One of the complex factors would be the possible change of shape in the proteins as

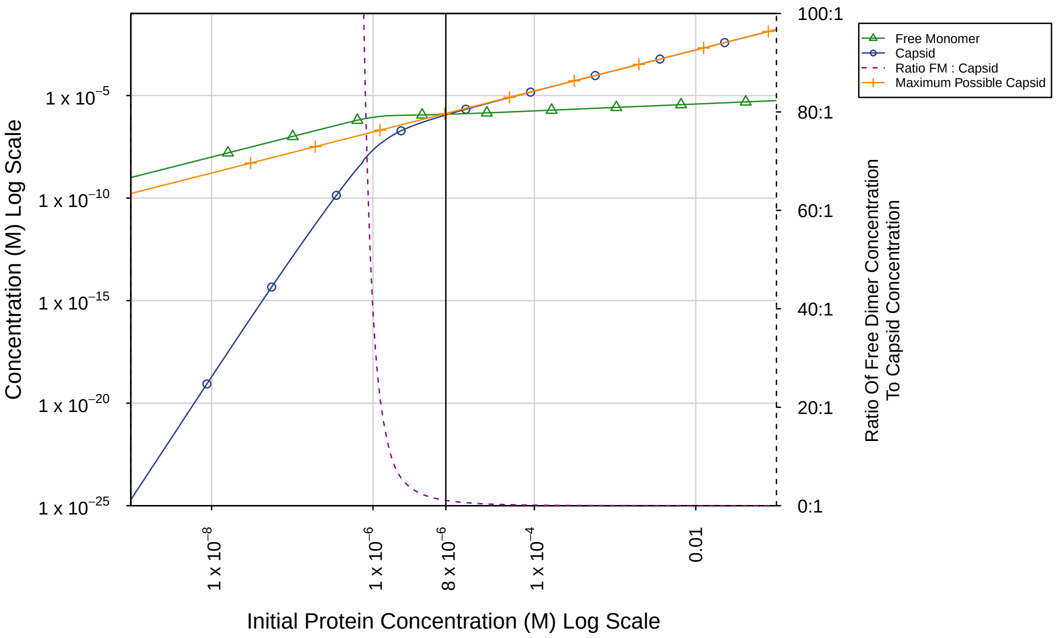

Figure 4.23: The free monomer concentration effects on the final capsid concentration. The protein-only network was simulated at a ΔG_c of -13 kJmol^{-1} and a time sufficient to reach thermodynamic equilibrium. The ratio shows the quickly diminishing free monomer up to the 8μM mark.

they begin to bind, another would be how the charges on the respective proteins interact over distance. Due to these factors the on-rate for two proteins binding in a particular experiment is likely to be different from the canonical value of 1×10^8 chosen. For instance in (Morton et al., 2010) a K_{on} rate of 1×10^5 was found to be more appropriate for this model in order to more closely match the biological reactions.

The protein concentration in the protein-only simulation has a major impact on the results, however increasing the protein concentration is very similar to having a more negative ΔG_c value and so in that regard the consequences are known. This is because the on-rate would be increased while the backward rate would be relatively smaller. A graph demonstrating the effects of protein concentration is shown in figure 4.23. In this graph we can see the change in gradient of the concentrations of capsid and free monomer around the initial protein concentration of 8μM. This change, of course, is from choosing the ΔG_c of -13 kJmol^{-1}. However this graph does explicitly show the change as the free monomer to capsid ratio is reduced.

The ratio of the protein to RNA in the RNA networks is also something we can investigate. So far we have only used the stoichiometric ratio of 6:1

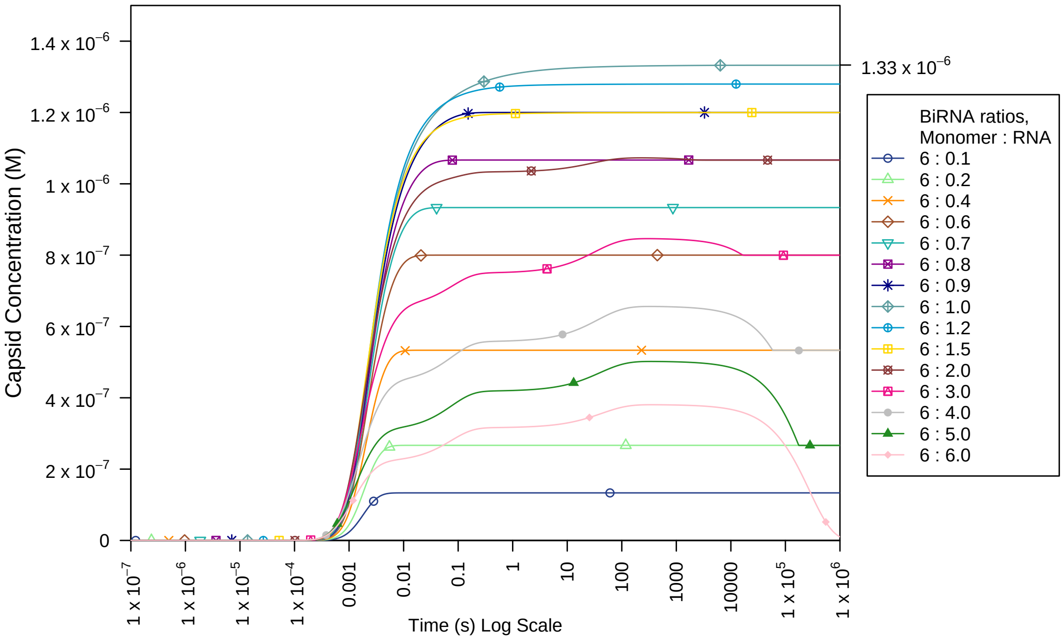

Figure 4.24: Different RNA : Protein ratios. -20 kJmol^{-1} was the ΔG_c value chosen to represent the is effect because it takes longer to reach equilibrium and as a result the effects are more obvious.

protein to RNA, respectively. Different ratios and their effects on the BiRNA network are shown in figure 4.24. Here we see that the best ratio to use at equilibrium is the stoichiometric ratio of 6:1, protein to RNA. Ratios with less RNA form correspondingly less capsid. While ratios with more RNA see continued reduction in the amount of capsid formed. This is until the extreme case of a ratio of 6:6 where the network equilibrates to each protein being bound to each RNA molecule and no concentration for capsid. Although since there is no off rate for a monomer leaving the RNA, this result is entirely predictable from the model assumptions. Although since we know from SELEX experiments (Shtatland et al., 2000) that the TR bound dimer is bound very tightly this may not be an unreasonable biological result. The equilibrium concentrations may be worked out mathematically; in the case of the 6:0.4 ratio the amount of capsid will be 0.4 x its maximum value of 1.33×10^{-6} M, giving 0.533×10^{-6} M. With a ratio of 6 : 4, the capsid concentration may be worked out by $\frac{8\times10^{-6}-(1.33\times10^{-6}\times4.0)}{5}$ to give 0.533×10^{-6} again.

4.5 Conclusions

The cube model investigations have shown a large amount of interesting and complex behaviour is able to occur even with such a small model system. When investigating the protein-only simulations we see all of the same patterns of behaviour described in the literature for the larger dodecahedron. The main result from this chapter is that at the higher bond strengths the presence of RNA greatly increases the final capsid concentration. This is due to the expected kinetic traps when only protein is present. The kinetic traps in the protein-only simulation have very much the same result, in that they collectively reduce the concentration of free monomer. Throughout all the simulations it is this existence of free building blocks that really determines the rate of capsid formation. To maintain the free monomer concentration and thus prevent the kinetic traps Zlotnick later introduced the nucleation step (Zlotnick et al., 1999) and elongation factors citeEndres2002. In the RNA simulations the proteins nucleate around the RNA and it is this that is the main cause of the increase in free monomer concentration. Unlike the protein nucleation step described in (Zlotnick et al., 1999), nucleation onto the RNA is not a slow step and therefore would reduce the time to form capsid. For the stoichiometric ratio of RNA to protein monomer, at no point in the virus assembly is the number of intermediates greater than the number of possible capsids. This maintains the highest possible free monomer concentration throughout the reaction. Although with very fine tuning of a protein-only nucleation step it is possible that the same result could be achieved, but at a likely cost of assembly speed.

The UniRNA simulation has a different type of kinetic trap and this single dead-end can have a large influence on the kinetics. The protein and RNA in this dead-end kinetic trap also takes a great deal of time to reach the thermodynamically favourable final capsids. The presence of free dimer still being available has no affect in this case. The final capsids in the RNA simulations, which are differentiated by their RNA layouts, can have large concentration differences. At thermodynamic equilibrium these differences reduce but up to this point the RNA capsids are strongly kinetically trapped. In laboratory experiments, as we have seen with MS2, it may be possible to detect these RNA layouts. From knowing the RNA layouts of the complete capsid, or at least the averaged RNA layout, it is then possible to infer the assembly pathways through the networks to those capsids.

The best RNA to protein monomer ratio at equilibrium is the stoichiometric one, however an increase in RNA concentration would speed up the first RNA-monomer binding reaction in the network. This slight speed increase can be seen in figure 4.24, where the ratios with more than the stoichiometric amount of RNA such as $6 : 1.2$ are very marginally quicker. If a ratio significantly different from this ideal stoichiometric ratio was discovered in a laboratory experiment, this model suggested that it would be an interesting phenomenon to investigate.

Being able to test all parameters in the model quickly and easily using the cube system has been a great advantage. It has been confirmed that in using the more compact reaction networks, that exclude mirror images or RNA direction, the final capsid concentration remains unchanged. Where pairs of intermediates and capsids have been combined, due to having a mirror image, the concentrations through the simulation have simply become double as a result. Equally every intermediate and capsid represents both possible RNA directions. For instance, an intermediate concentration in one of the RNA networks, that has a mirror image, should therefore be quadrupled to find the concentration of one of the four intermediates it represents. In choosing which graphs to show it has also become clear that the different parameters can not be considered in isolation and there is always a time or ΔG_c scale to consider.

The cube is still a small shape and not all of the conclusions drawn are likely to hold perfectly for larger viruses. This scale factor will be investigated in the next chapter with a dodecahedral shape.

Chapter 5

Dodecahedron Results

Following the same steps as in the cube chapter, the first simulations were conducted to find the ΔG_c energies required for the free monomer and total capsid to have the same concentration at equilibrium. These results, obtained via numeric integration of the network of reactions, are shown in table 5.1. The equilibrium concentrations of the final capsid and free monomer are 6.15×10^{-7} M. This concentration is one thirteenth of the starting 8μ M protein concentration used throughout these simulations. The reason for this is because at equilibrium only the free monomer or complete capsid/s have high concentrations. This leaves one thirteenth of the total protein in the free monomer and twelve thirteenths therefore must be in the final capsid/s for the equal concentrations.

5.1 Protein-Only Simulations

Using 1, 1.5 and 2 times the ΔG_c value of -11.74 kJmol^{-1} gives us an indication of the protein-only simulation over time, see figure 5.1. Here we can see the

Simulation	Equilibrium bond contact energy kJmol^{-1}	Equilibrium FM/Capsid amount (M)
Protein	-11.74	6.15 $\times 10^{-7}$
UniRNA	-11.1	6 $\times 10^{-7}$
BiRNA	-11.2	6 $\times 10^{-7}$
TrRNA	-11.3	6 $\times 10^{-7}$

Table 5.1: Bond strengths required for an equilibrium where the free monomer has the same concentration as the final capsid. Also shown are the concentrations that this occurs at.

plateaus indicative of the kinetic traps happening at the 1.5(-18 kJmol^{-1}) and 2(-23 kJmol^{-1}) ΔG_c bond energies. This is different from the cube, for which there were no similar plateaus at the more negative bond energies (ΔG_c of -20 kJmol^{-1}). The reasons why these kinetic traps occur at less negative bond strengths for the dodecahedron are due to the structure of the network of intermediates. Firstly, the network is longer in the case of the dodecahedron, which increases the time required for the free monomer to react to form capsid. If the free monomer then becomes scarce in a sufficiently short time period there will be more protein kinetically trapped within intermediates. Secondly the free monomer concentration is reduced more rapidly in the dodecahedron simulations than the cube simulations. This is because there are more possible reactions in the dodecahedron network, due to the increased connectivity of the network and the larger numbers of intermediates. This second reason acts to reduce the time to reach capsid, but this does not compensate for the increase in length of the network of interactions.

A breakdown of the concentrations by size in the protein-only simulation for the dodecahedron is shown in figure 5.2 for a bond strength of -18 kJmol^{-1}. This graph shows the quick succession of intermediates and which intermediates dominate when the free monomer becomes scarce.

5.2 RNA Simulations

For the dodecahedron, as for the cube before, first the RNA simulations are compared at each of the three ΔG_c values (figures; 5.3, 5.4 and 5.5). The results for a ΔG_c of -12 kJmol^{-1} are very similar to those for the cube. Again, we see the same ordering both in speed of formation and, as expected from the free monomer / capsid equilibrium values, the same relative equilibrium concentrations for fully assembled capsids. It is not until the higher ΔG_c of -18 kJmol^{-1} that large differences occur. At this ΔG_c we see effects of the dead-ends in the RNA networks in absorbing some of the protein. For example, the final capsid concentration for the UniRNA simulation has been reduced to about two thirds of the maximum possible value of 6.6×10^{-7} M. This concentration should not be confused with the thermodynamic equilibrium concentration, it is simply that material that will eventually form capsid (given enough time) is kinetically trapped for a significant amount of time in the dead-ends. The actual equilibrium capsid concentration for the UniRNA simulation would be the maximum of 6.6 ×10^{-7} M, like the protein-only simulation at this value of

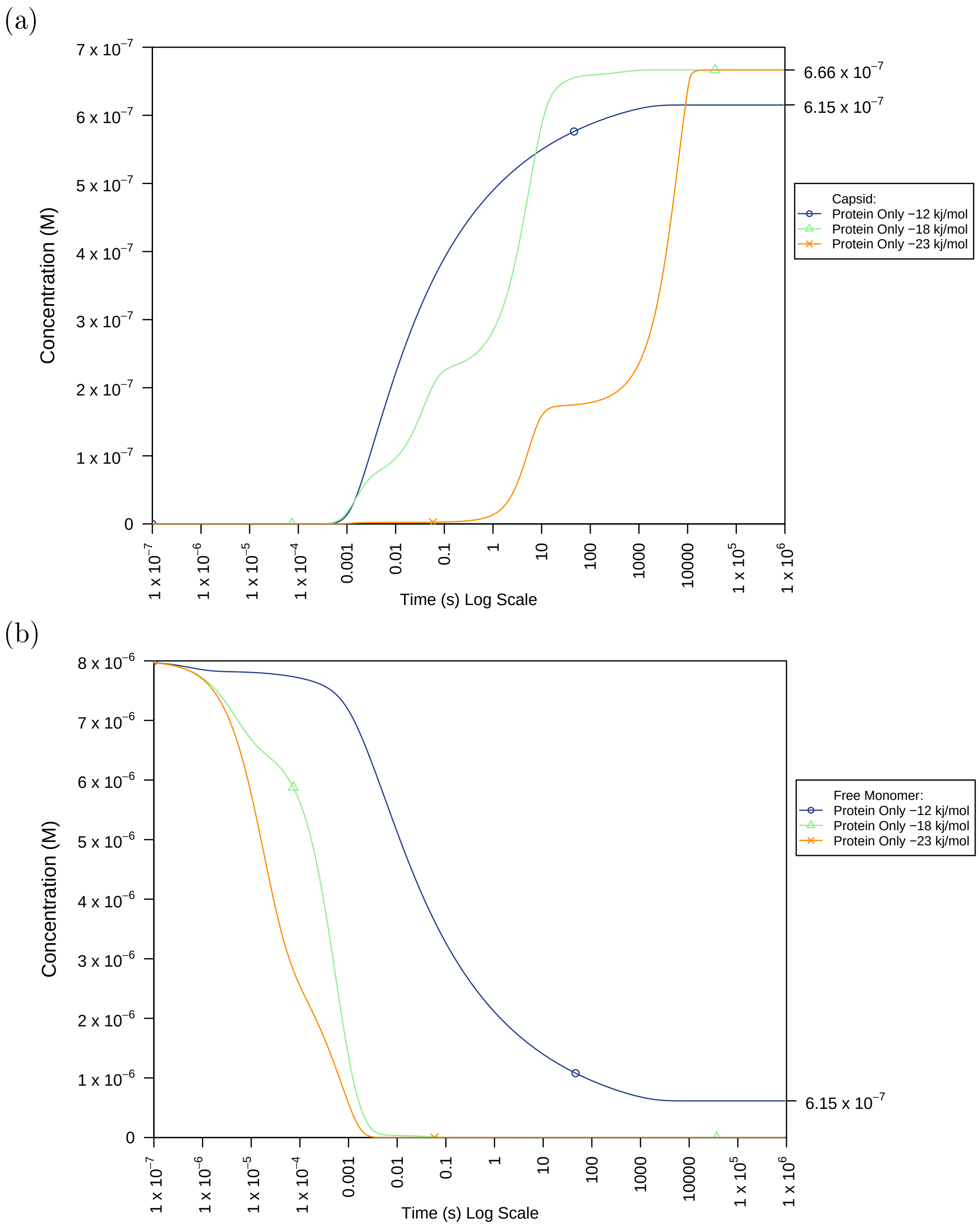

Figure 5.1: Protein-only simulation of assembly kinetics at ΔG_c values of -12, -18, and -23 kJmol^{-1}. Capsid concentration is shown in (a) and the corresponding free monomer concentration in (b).

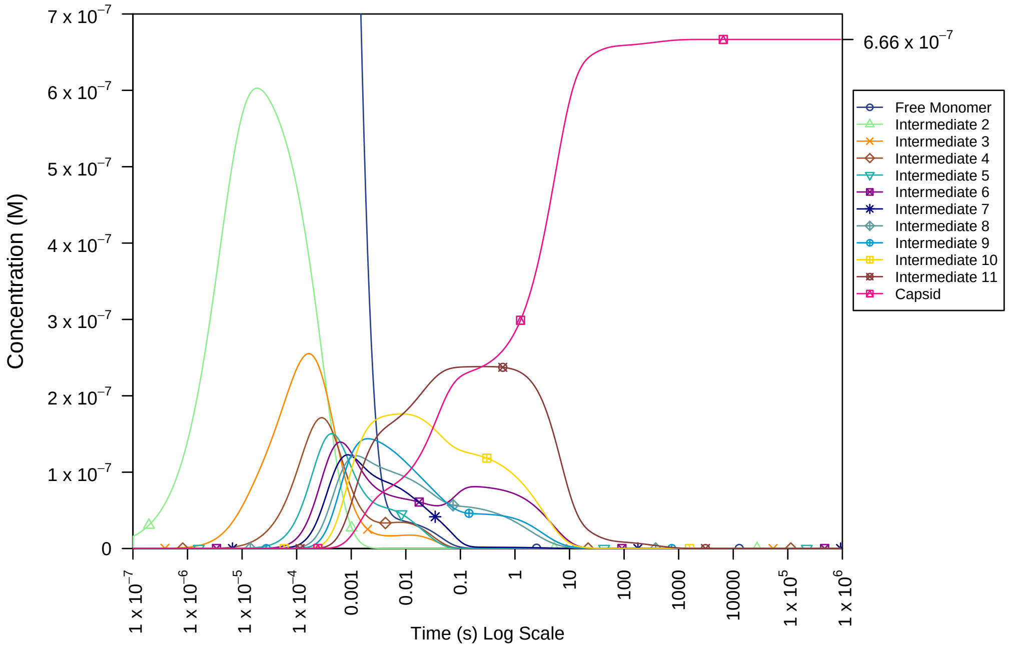

Figure 5.2: Intermediate concentrations in the protein-only cube capsid simulation at a ΔG_c of -18 kJmol^{-1}.

ΔG_c. The dead-end kinetic traps also manifest themselves in the TrRNA and BiRNA simulations, where they again act to reduce the capsid concentration. The reduction in final capsid concentration for the RNA simulations is even more pronounced at bond energies of -23 kJmol^{-1}. This implies that even with the low backward rates expected for a ΔG_c or -18 kJmol^{-1}, the backward reactions are still significant enough to avoid dead-end pathways and get out of dead-end traps.

The RNA kinetic trap dependence on the value of ΔG_c and time can be more easily seen in figures 5.6, 5.7 and 5.8. In figure 5.6 we can see that with a ΔG_c value of about -12 kJmol^{-1}, and given enough time, the capsid concentration is able to equilibrate to almost the full 6.6 $\times 10^{-7}$ M. This is because all the dead-end traps are able to disassemble to then allow for a path through the network to capsid to be taken. Note that this may require several successive backward reactions. At -21 kJmol^{-1} increasing the time period further than 0.1 seconds does not change the concentration profile of capsid. For an indication of how quickly kinetic traps may be resolved and avoided, observe the green line at 0.01 seconds reaching a peak and then decreasing as ΔG_c becomes more negative. At the peak level, significant backward reactions are occurring, favouring the formation of capsid. The pathways leading to

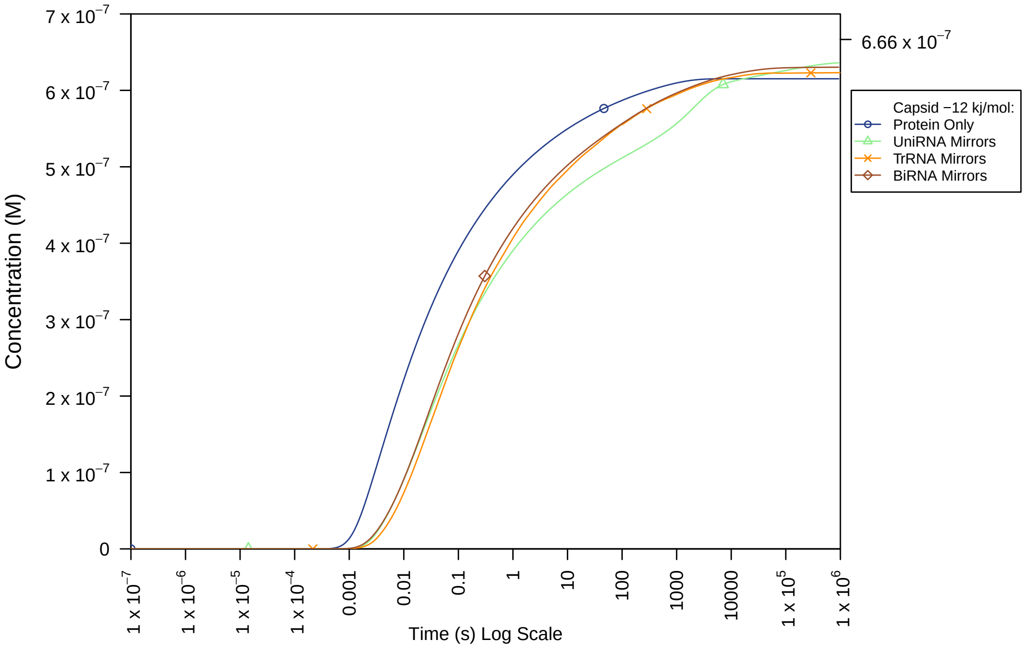

Figure 5.3: Time versus capsid concentration in all the dodecahedral scenarios at a ΔG_c of -12 kJmol^{-1}.

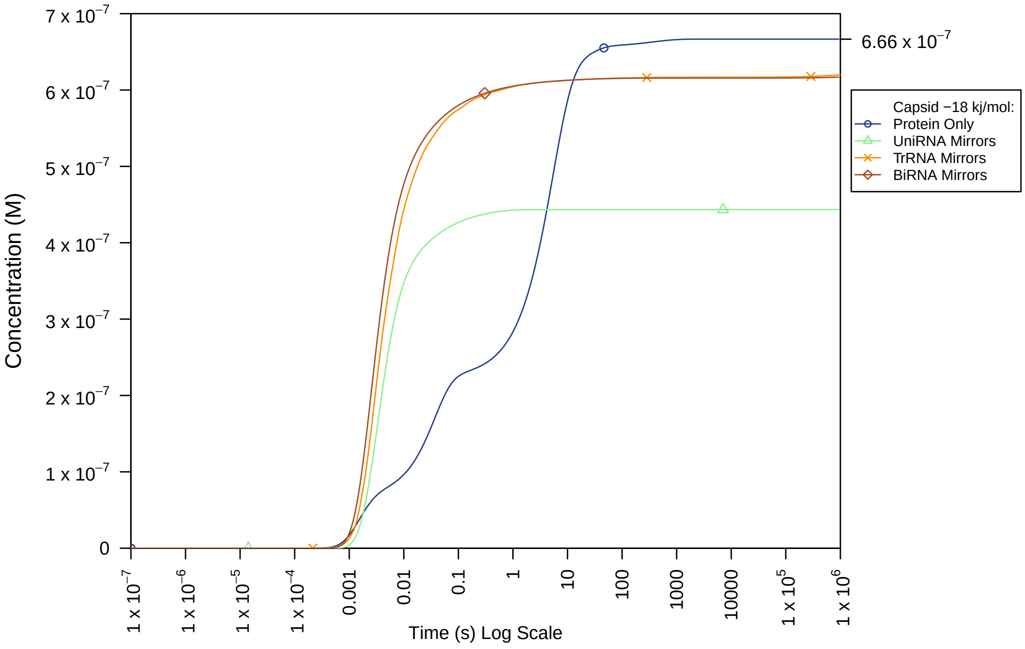

Figure 5.4: Time versus capsid concentration in all the dodecahedral scenarios at a ΔG_c of -18 kJmol^{-1}.

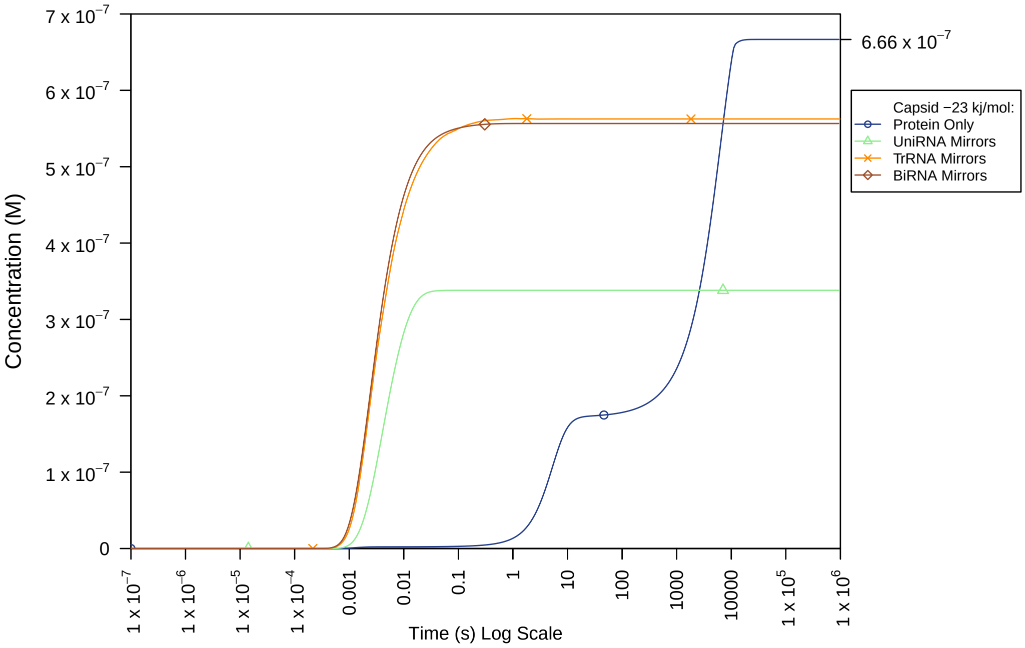

Figure 5.5: Time versus capsid concentration in all the dodecahedral scenarios at a ΔG_c of -23 kJmol^{-1}.

dead-ends are more likely to contain relatively unstable intermediates. This factor allows for quick corrections of dead-end pathways taken as intermediates on dead-end pathways break up readily, especially at the less negative ΔG_c values.

At -30 kJmol^{-1}, for times of 0.1 seconds or more, the capsid concentration reaches a level of 2.68 $\times 10^{-7}$ M and appears to be levelling out. At this ΔG_c level, there occur close to zero backward reactions. The result of this is that all the material that reaches a dead-end stays in that dead-end and likewise for material reaching the capsid. It is not necessary to numerically integrate the reactions to arrive at the correct capsid concentration when there are no effective backward rates because in this case capsid concentration can be determined as follows: First set a concentration of 8 $\times 10^{-6}$ M for the RNA-bound dimer at the start of the network. Then split up this concentration proportionally according to the symmetry factors to the next intermediates. Carrying this through to the final capsids gives them a combined concentration of 2.67 $\times 10^{-7}$ M. This procedure is a very quick way to find the minimum bound for capsid concentration. This method does presume that there has been enough time to allow all the protein to reach an endpoint in the network.

The graphs for the TrRNA and BiRNA simulations (5.7, 5.8) show similar

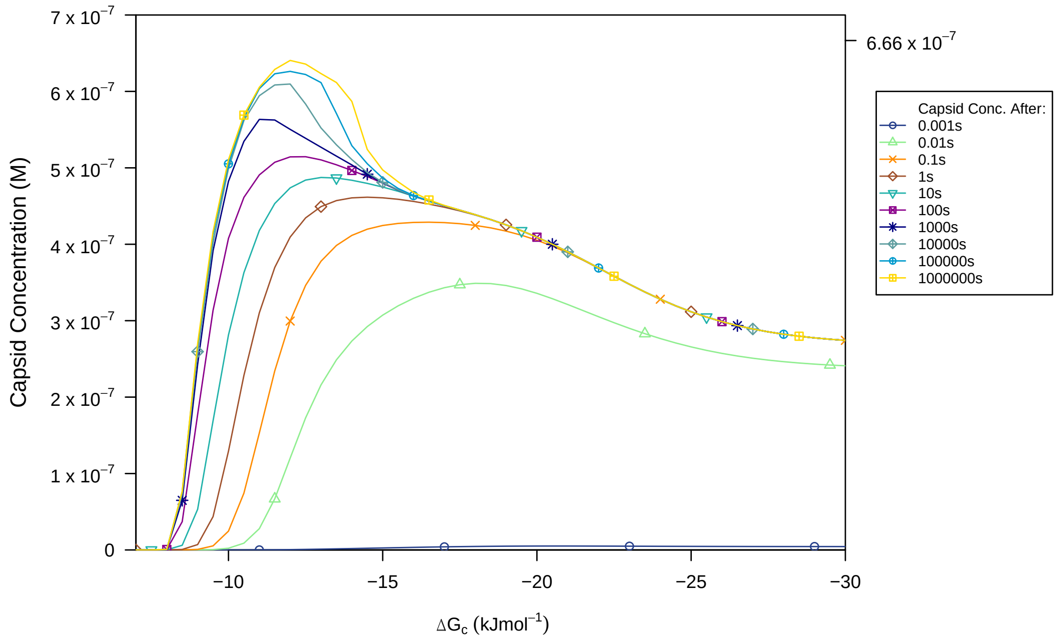

Figure 5.6: The capsid concentration after different time periods and a range of ΔG_c values in the UniRNA cube simulation.

behaviour to the UniRNA simulation and especially to each other, with final capsid concentrations of 5.02 $\times 10^{-7}$ M for the TrRNA simulation and of 5.00 $\times 10^{-7}$ M for the BiRNA simulation. Running the above mentioned splitting procedure on these networks gives exactly the same concentration value of 4.94 $\times 10^{-7}$ M for capsid in both networks. These figures show that it is only in the backward reactions that the BiRNA and TrRNA networks vary.

In order to investigate the efficiency of assembly, the time to reach 90 % capsid has been plotted (see figure 5.9). Like in the cube graph, we can see the BiRNA and TrRNA simulations become more efficient at producing capsid when the ΔG_c exceeds a critical value. In the dodecahedral case this happens after about -15.7 kJmol^{-1}, which is less negative than the -21 kJmol^{-1} for the cube BiRNA and TrRNA simulations. Note however the time taken to produce the 90 % capsid amount for the BiRNA and TrRNA simulations is not much less than a second at any ΔG_c. For the protein-only simulation it is clear from the graph that assembly efficiency is strongly dependant on the ΔG_c value, with -15 kJmol^{-1} being the optimum in quite a steep well. The UniRNA simulation barely reaches 90 % of the maximum possible capsid amount and it is not until we look at the efficiency to 66 % capsid that we see the UniRNA simulation become significantly more efficient than the protein-only simulation

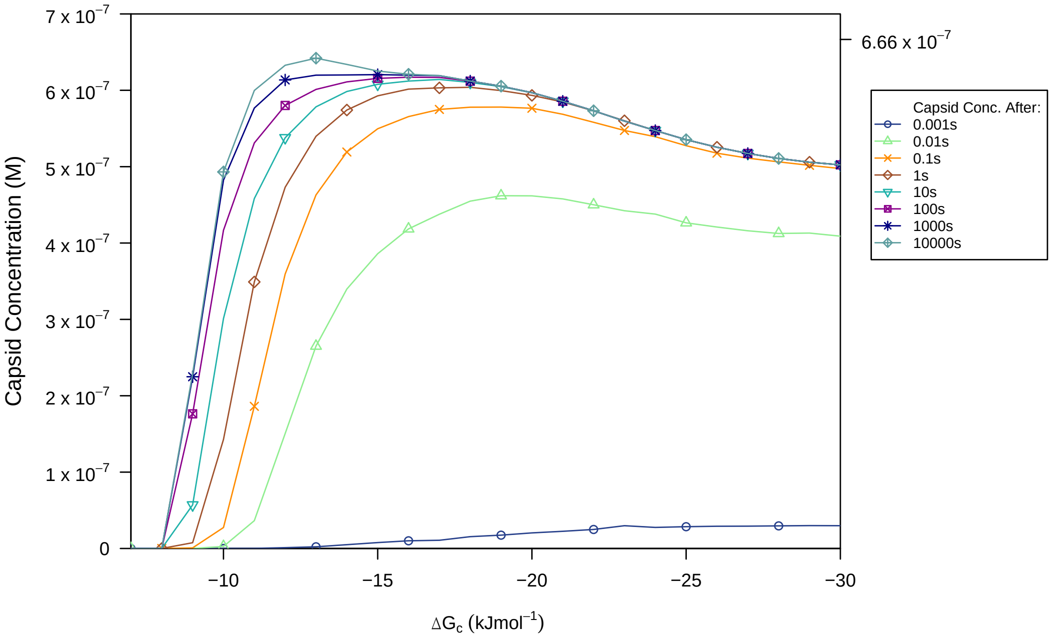

Figure 5.7: The capsid concentration after different time periods and a range of ΔG_c values in the TrRNA cube simulation.

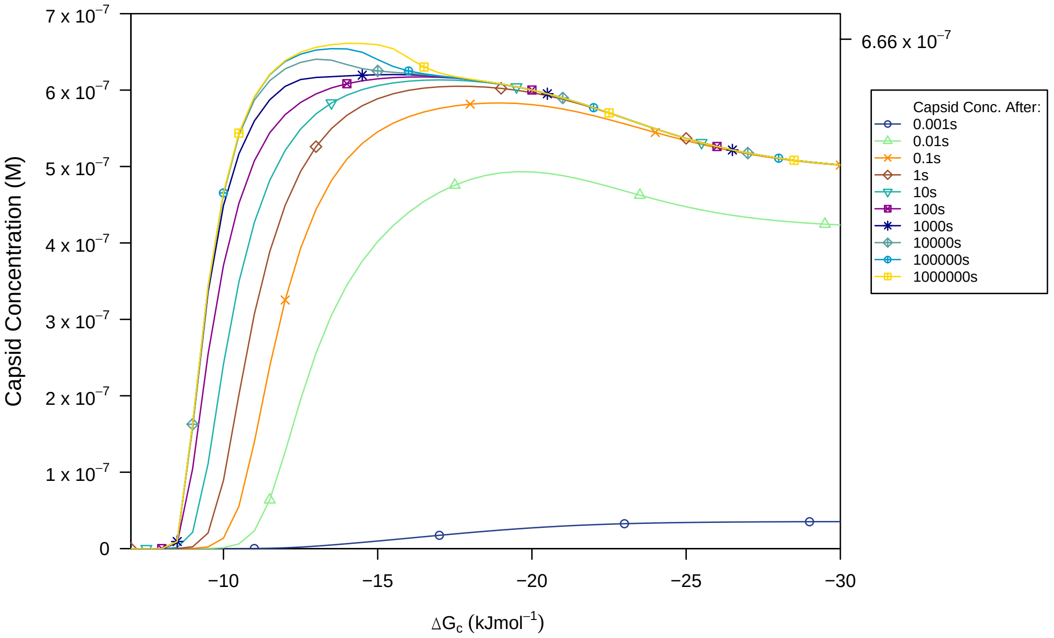

Figure 5.8: The capsid concentration after different time periods and a range of ΔG_c values in the BiRNA cube simulation.

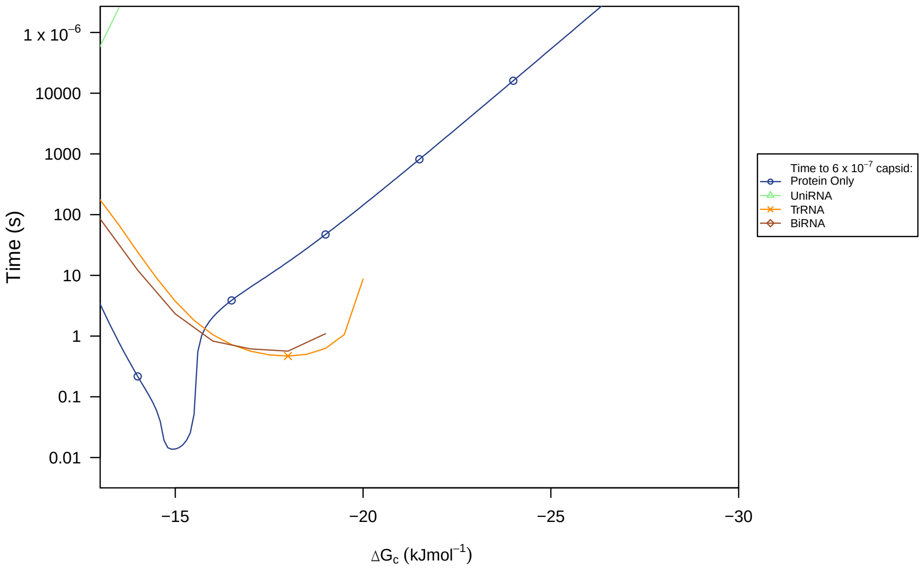

Figure 5.9: Time taken to reach 90 % of the maximum possible capsid concentration across a range of ΔG_c values comparing all the dodecahedron simulations.

(see figure 5.11). Note that again this is only for bond strengths of -15.7 kJmol^{-1} or more negative. Likewise the BiRNA and TrRNA also become more efficient at this ΔG_c. Additionally, the BiRNA and TrRNA simulations are much quicker in reaching the 66 % capsid level with a time close to 0.01 seconds, comparable to the protein-only simulation. To indicate the trend, the time taken to reach 75 % is shown in figure 5.10.

5.3 Conclusions

Overall, the behaviour of dodecahedron assembly and the conclusions drawn are very similar to those for the cube. It has been found that the increased scale of the network affects both the protein-only and the RNA simulations. Firstly, the protein-only simulation has significant kinetic traps at less negative ΔG_c values. This is due to the increased size of the network and additional possible reactions, i.e. it is very much scale related. As a result of the kinetic traps in the protein-only case the RNA simulations for the BiRNA and TrRNA simulations become more efficient for assembly at less negative ΔG_c values. The RNA simulations are dominated by the influence of the dead-ends, However the effect is not as large as might be expected from just looking at the network. The influence of kinetic traps is greatest in the UniRNA simulation.

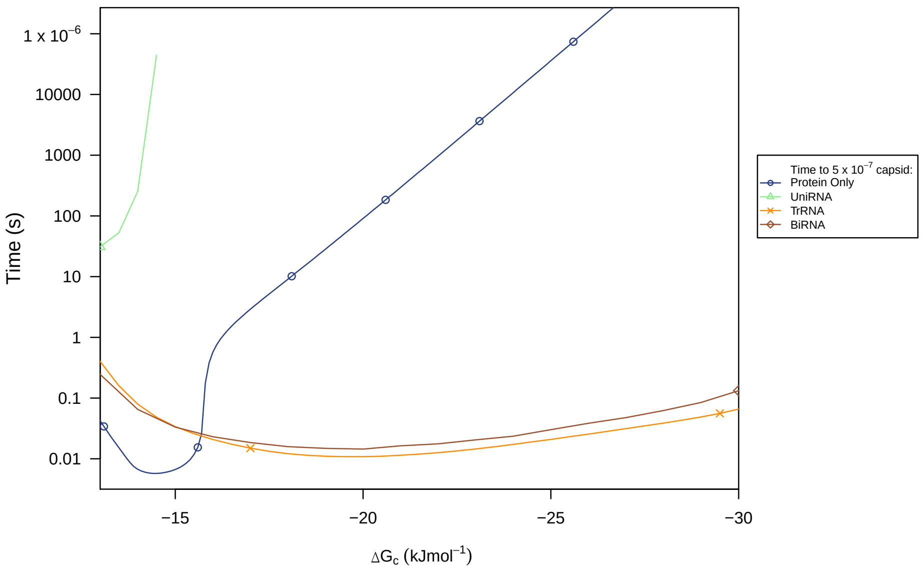

Figure 5.10: Time taken to reach 75 % of the maximum possible capsid concentration across a range of ΔG_c values comparing all the dodecahedron simulations.

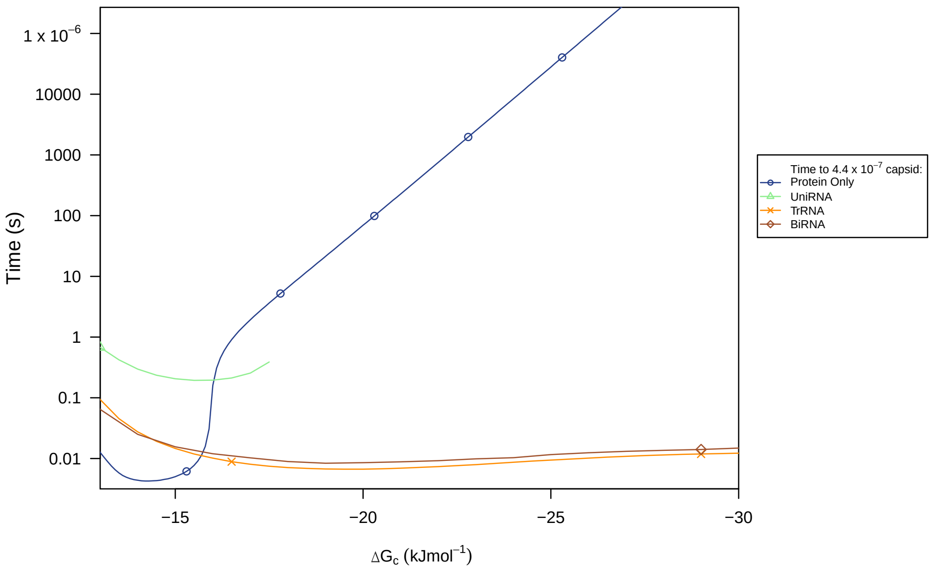

Figure 5.11: Time taken to reach 66 % of the maximum possible capsid concentration across a range of ΔG_c values comparing all the dodecahedron simulations.

This network has the highest number of dead-ends and also the dead-ends that are the most difficult to equilibrate out of, due to the length of the dead end pathways. However, considering that there are almost twice as many dead-ends in the UniRNA network as there are final capsids, the effect is not as large as might be expected. This is because the dead-ends tend to occur along pathways that are characterised by formation of relatively few bonds and smaller symmetry factors.

We have seen that these simple models can be very useful to illustrate the characteristics of the assembly kinetics resulting from specific sets of biological assumptions. In the next chapter, models for the STMV and MS2 viruses will be investigated to determine what may be learned regarding assembly in these large-scale assembly scenarios that are far less computationally tractable.

Chapter 6

STMV and MS2 Assembly

6.1 Introduction

In this chapter the larger viral capsids of the STMV 30-mer and the MS2 90-mer are considered. Previously we have seen how the number of intermediates grows combinatorially with the size of the capsid. This makes these much larger systems harder to model due to the computational intractability of analysing and simulating so many intermediates. However, simply by investigating the computed assembly networks it is possible to gain insights into the problem. Later in this chapter possible coarse-graining approaches are discussed and a successful published method is described.

6.2 STMV Reaction networks

The STMV virus is constructed from 30 dimers that take the shape of a rhombic triacontahedron. Joining the centres of the faces in this rhombic triacontahedron creates its dual polyhedron, the icosidodecahedron (see figure 6.1(a)). The planar Schlegel diagram of the icosidodecahedron gives us the graph on which we can construct the Hamiltonian paths taken by the RNA. The Hamiltonian paths corresponding to the protein-only, UniRNA, TrRNA and BiRNA are shown in table 6.1. It is easy to see how large the numbers of intermediates becomes with the increased size of the 30-mer. Even using the new algorithms described in chapter three, 82,000,000 intermediates of a single size is about the limit of what can practically be computed. This means that the TrRNA scenario is not able to be fully calculated. However, the BiRNA scenario may be fully computed, which as we have seen for the dodecahedron behaves similarly.

The the kinetic integrations of the previous chapter at most consider 24,635

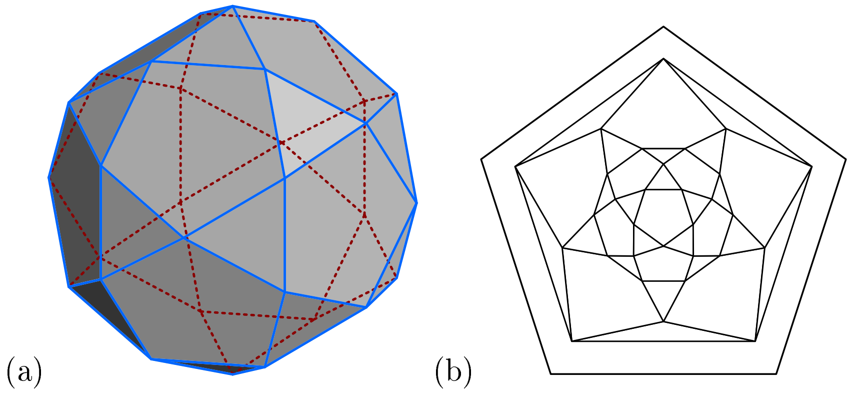

Figure 6.1: (a) The icosidodecahedron and (b) the corresponding planar Schlegel diagram.

intermediates, this was in the dodecahedral BiRNA scenario. Between these intermediates there are 44,284 edges. This leads to 88,567 equations that require integrating, because each edge has both a forwards and backwards reaction (except the very first reaction with the RNA). Since the integration of these 88,567 equations takes at least a week clearly the integration on the whole STMV network is unfeasible. Later, possible coarse-graining of the kinetics is discussed but first, what can be learnt from the network itself will be investigated.

6.2.1 Network Analysis

It is possible to gain a large amount of insight into what will happen in a kinetic simulation just by analysing the intermediate network of STMV. We have leaned from previous chapters that at lower bond values, the pathways to final capsid assembly will favour the more stable intermediates. While at more negative bond strengths the concentrations are determined by the network topology, i.e. the splitting. We have also seen the large influence of the dead-end intermediates in the RNA scenarios, especially at more negative bond strengths. The 12-mer, dodecahedron UniRNA scenario has 632 full capsids and, in total 1,156 dead-ends, which is almost twice as many. The 30-mer has 141,680 complete capsids and a total of 24,543,622 dead-ends, now 173 times as many. There is a similar story for the BiRNA scenario where the dead-end ratio to full capsid grows from 0.5:1 in the 12-mer to 71:1 in the 30-mer. It is presumed that these greater number of dead-ends in the STMV scenarios will have a very large effect on the virus assembly kinetics. This presumption is based on the dodecahedral model, where the dead-ends trapped material

	Protein Only	UniRNA		TrRNA		BiRNA	
Intermediate Size	No.	No.	Dead-Ends	No.	Dead-Ends	No.	Dead-Ends
Free Capsomere/RNA	1	2	0	2	0	2	0
1	NA	1	0	1	0	1	0
2	1	1	0	1	0	1	0
3	4	3	0	6	0	3	0
4	6	8	0	16	0	5	0
5	19	22	0	59	0	15	0
6	43	59	0	177	0	32	0
7	119	153	1	545	0	86	0
8	300	389	5	1 556	0	200	0
9	818	987	16	4 464	0	516	0
10	2 083	2 469	53	12 345	10	1 247	1
11	5 357	6 024	166	33 186	67	3 066	7
12	13 078	14 375	468	86 250	168	7 211	14
13	30 674	33 487	1 313	217 781	380	16 859	32
14	66 723	75 342	3 715	527 394	1 477	37 708	106
15	133 347	162 385	10 042	1 218 110	5 249	81 415	356
16	236 182	334 409	25 499	2 675 272	15 152	167 253	949
17	360 834	656 323	61 216	5 579 144	44 937	328 560	2 665
18	455 307	1 220 872	138 554	10 987 848	137 061	610 493	7 618
19	452 799	2 137 824	292 660	20 309 991	378 482	1 069 575	19 985
20	338 011	3 500 796	573 466	35 007 960	922 250	1 750 462	46 118
21	193 929	5 322 484	1 040 001	55 887 071	2 064 117	2 662 231	98 437
22	88 217	7 440 417	1 735 449	81 844 587		3 720 268	195 445
23	32 545	9 436 052	2 632 686			4 719 264	354 236
24	9 834	10 660 764	3 567 440			5 330 417	570 135
25	2 408	10 466 498	4 213 948			5 234 440	805 386
26	482	8 628 298	4 189 569			4 314 158	970 302
27	78	5 690 471	3 332 691			2 846 005	945 473
28	11	2 792 376	1 966 900			1 396 188	681 914
29	1	899 444	757 764			449 967	317 995
30	1	141 680				70 840	
Total	2 423 212	69 624 413	24 543 622	214 393 764	3 569 350	34 818 486	5 017 174

Table 6.1: The intermediate numbers for the 4 different 30-mer scenarios, With the dead-end numbers for each size. The original protein-only numbers were first determined in (Moisant et al., 2010).

and slowed the formation of capsid. Of course at less negative bond strengths and longer times these kinetic traps were resolved in the dodecahedron kinetic simulations.

The use of less negative bond strengths favours the formation of stable intermediates that have a large number of bonds between their intermediates. As a consequence of the larger number of bonds formed these intermediates also tend to be the most compact structures. The UniRNA networks are relatively easy to analyse because the topology of the network is only to branch and there is no recombination of pathways. As a consequence of this there is only one pathway to each of the final capsids, excluding any forward and then backward reactions on side branches.

Bond Formation and Branching Analysis

Every pathway to capsid forms the same number of bonds on capsid completion but the number of bonds formed at each point on the way can differ. One of the capsid pathways that forms the most number of bonds early on is shown in figure 6.2(a), a pathway that forms the least number of bonds early on is shown in figure 6.2(b). The bonds formed at each step in the assembly of these two paths are shown in table 6.2. From this table we see that the path that forms the most bonds early on in assembly does so at the earliest possibility and has more, or the same number of bonds, at all points. The largest differences in the bond number are at the start, this would favour paths that form the most bonds early on.

As well as the number of bonds that are formed we know from the kinetic simulations that the branching in the networks will also affect the pathways taken. To investigate this branching, the pathways to capsid in the UniRNA STMV network have been sorted and a pathway with the most splitting and least splitting are shown in figure 6.3. At a very negative bond strength there would effectively be no backward rates and the concentration of each capsid (and dead-end) in the network will only depend on this splitting in the network. In the previous chapter simply dividing the protein concentration at each branch point was successfully used to predict the capsid concentration, at such negative bond strengths. The same technique can be used for the two pathways to capsid with the most and least splitting. The result is that the pathway with the least splitting has 0.077 % of the initial protein concentration and the pathway with the most splitting has orders of magnitude less with only 0.0000037 %.

If we look at the number of bonds formed at early times for the path with the most splitting we find comparatively low numbers, see table 6.2. Similarly there is a relatively high amount of branching in the path that forms the most bonds (see figure 6.4), the opposite is also true. The reason for this is that compact structures form the most bonds, but also have the most branch points because there are many options to move away from compact structures. Forming a non-compact intermediate, with low numbers of bonds, constrains the Hamiltonian path because the ways to reach the remaining unbound dimers is reduced. Effectively this means that the branching of the network and the bonds formed on the paths favour opposite intermediate shapes and pathways. The detailed interplay of these factors could only be solved by some form of kinetic model. However, this does provide insight into the likely assembly pathways. This is because, not only are the pathways that continuously form the most bonds the most favourable due to the stability of the intermediates, they also avoid branching as much as possible. The branching is not avoided due to the lack of possible choices to branch but by forming the next most stable intermediate. Avoiding most of the branching in general will avoid the branches that lead to dead-ends. These dead-ends have already been associated

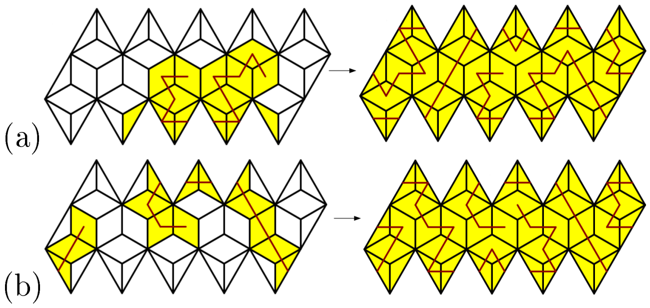

Figure 6.2: The halfway points and final capsids for the most number of bonds formed early on (a) and least number of bonds (b). These halfway points clearly show the difference in capsomere contacts formed.

Path	Capsomere Contacts Formed Along The Path																												
Most early bonds	1	3	4	6	7	9	11	13	14	16	18	20	22	24	26	28	29	32	35	37	38	41	43	46	48	51	53	56	60
Least early bonds	1	2	3	4	5	6	7	8	9	10	12	15	17	20	22	25	28	30	33	35	38	40	43	45	48	50	53	56	60
Most branching	1	3	4	6	7	9	10	12	13	15	16	19	22	24	25	28	30	32	34	35	38	41	43	45	48	51	53	56	60
Least branching	1	2	3	4	5	7	9	11	12	14	15	16	19	21	24	26	28	30	33	35	37	40	42	45	47	50	53	56	60

Table 6.2: How the capsomere contacts grow depending on the capsomere binding order of specific paths.

kinetically with intermediates that have relatively few bonds in the previous two chapters.

6.3 MS2 assembly pathways

The paths for the MS2 90-mer have also been calculated as far as the computational tractability allows (see table 6.4). The final UniRNA Hamiltonian path number of 40678 was first calculated by Simone Severini using the program "Gap" (GAP) (Grayson et al., 2007). This number was later confirmed using a simple backtracking algorithm, that was not able to save the intermediate steps, as discussed in chapter three. The protein-only intermediate numbers for the MS2 virus grow rapidly and combinatorially, this suggests that Zlotnick's estimation (Moisant et al., 2010) that there may be as many as 10^{18} in total could easily be correct. In the MS2 virus these numbers could be reduced

Path	Path Splits																													Percentage
Most early bonds	1	3	2	3	2	3	2	2	2	3	2	2	2	2	2	2	2	2	1	1	2	2	1	1	1	1	1	2	1	0.000019
Least early bonds	1	3	3	3	3	3	2	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.026
Most splitting	1	3	2	3	2	3	2	3	2	3	2	3	1	1	2	3	1	2	2	2	3	1	1	2	2	1	1	2	1	0.0000037
Least splitting	1	3	3	3	3	2	2	1	2	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.077

Table 6.3: STMV paths that show the variability in the number of branches along the path.

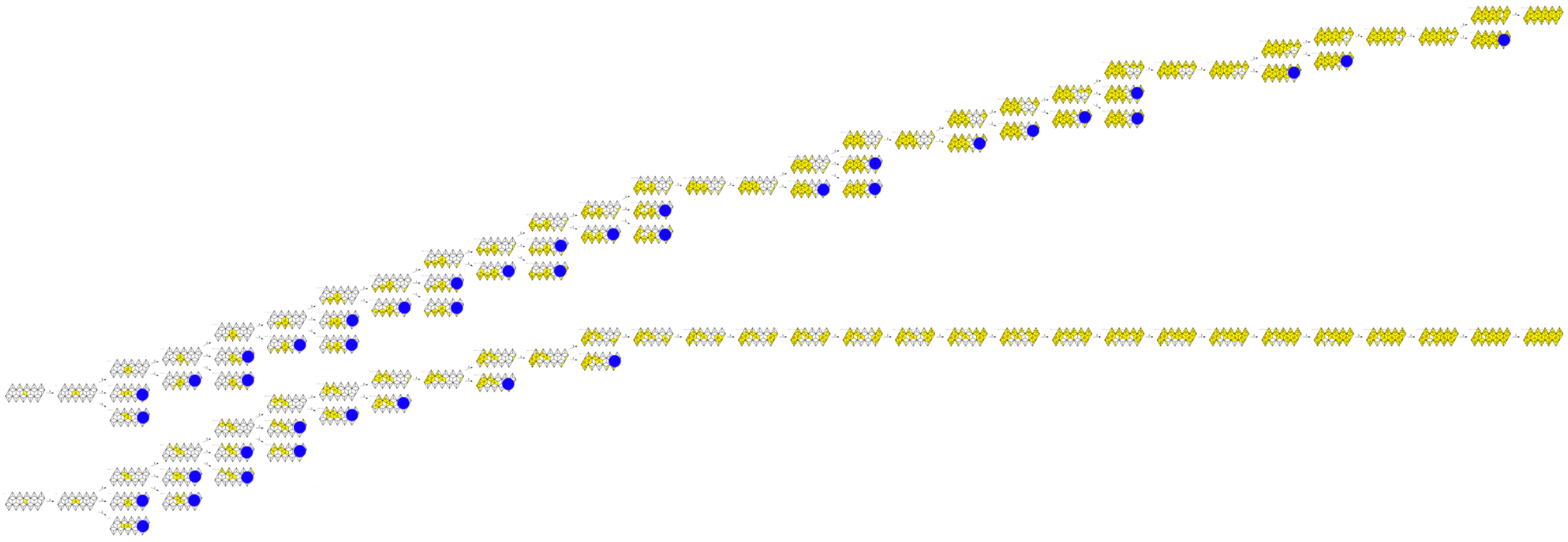

Figure 6.3: Two pathways to capsid in the UniRNA simulation, the top pathway shows the most splitting while the lower pathway has the least. Intermediates that are first on other pathways are also shown to emphasise the branching of the paths, these intermediates have been marked with a blue dot.

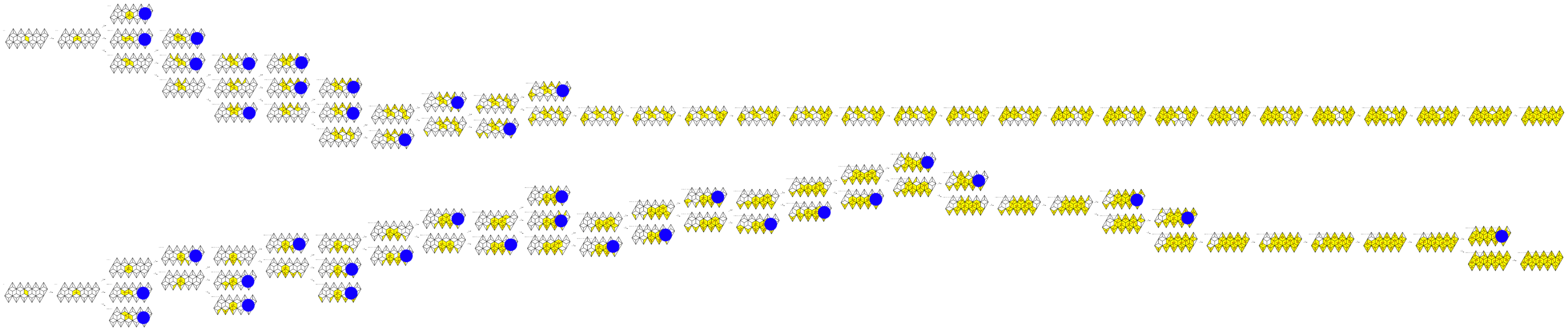

Figure 6.4: Two partial networks showing paths to STMV capsid formation. The upper pathway forms the least number of bonds early on, while the lower pathway forms the least number of bonds early in assembly. Intermediates that are first on other pathways to capsids are also shown to emphasise the branching of the paths, these intermediates have been marked with a blue dot.

by removing mirror image intermediates. Previously any mirror symmetries of the smaller capsids, such as in STMV, could be achieved by rotations.

To reduce the time for the computation of the intermediates in table 6.4, for the RNA scenarios, the intermediates are only constructed from A/B dimers, because these are defined by the Hamiltonian path. The maximum size of an intermediate is therefore 60, corresponding to the 60 A/B dimers. The C/C dimers may be added later. Interestingly at this size of virus, the number of RNA intermediates is less than the number of protein-only intermediates for the first time. This is due to the RNA path limiting the combinatorial positions that dimers may be placed or removed. Also the MS2 RNA intermediate numbers initially grow slower than for STMV, this is because the MS2 Hamiltonian paths are 3 coordinated at the junctions while the STMV Hamiltonian path possibilities are 4 coordinated. This higher connectivity of the faces in STMV allows for more choices in path and the resulting higher number of intermediates early on. For larger sizes of intermediate the increased size and length of Hamiltonian path results in larger intermediate numbers for MS2. The layout and helicity of the RNA connectivity in the MS2 virus is presumed to be due to differences in the bond strengths between different contact faces of the A/B and C/C dimers. This suggests the possibility that the virus has evolved to reduce the connectivity of the possible Hamiltonian paths, limiting the complexity.

The Hamiltonian paths on the MS2 capsid contain short steps, between A/B dimers on the same 5-fold axis and long steps between different 5-fold axes (see figure 6.5). These long steps pass underneath the C/C dimers. Since the C/C dimers do not bind the RNA in the same way as the A/B dimers, when to add C/C dimers to the growing capsid is currently undefined. If we assume there is any bonding interaction between the RNA and the C/C dimers that are above the RNA, however minimal, placing the C/C dimers when the RNA passes below them would be justifiable. This would also place the C/C dimer between the A/B dimer previously placed and the one about to be placed. Any Hamiltonian path on the RNA density can not actually use all the long or short edges on the graph of possibilities. This leaves some C/C dimers that do not have RNA underneath them. These remaining C/C dimers may be placed combinatorially with the rest of the dimers, at the cost of increasing the intermediate number by further orders of magnitude, or more simply these C/C dimers may be added when they form 2 or 3 contacts with the already present capsomeres.

A breakdown of the 40,678 MS2 Hamiltonian paths by long and short steps is shown in table 6.5. Even if only a very minimal amount of binding occurs between the C/C dimers and the RNA, this would result in the 212 with the most long paths having the highest thermodynamic equilibrium concentration. The likelihood of this equilibrium being reached in viral assembly though is low due to the large amounts of backward reactions that it would take to equilibrate between capsids, especially when the difference would be a very small free energy amount. We have already seen for the short cube network that equilibration between final capsids can take a long time even at less negative bond strengths. There are also paths in table 6.5 with lots of short steps, this corresponds to Hamiltonian paths that travel around the 5-fold axis as much as possible.

There are in-fact a multitude of interesting geometries in the Hamiltonian paths for the MS2 virus. Two further examples include what have been termed the spiral path, 6.6(a), and double spiral path, 6.6(b). Assembling along the spiral path, starting at one end of the RNA genome such as the UniRNA scenario, forms a relatively very high number of bonds early in the assembly. Whereas the assembly of the double path starting at one end forms a minimal amount of bonds at any point. However, if the assembly was allowed to proceed in both directions along the RNA, such as the TrRNA or BiRNA scenarios, this double spiral has the possibility of forming relatively large numbers of bonds early on by using both ends of the RNA. This emphasises the necessity to investigate assembly starting in the middle so that such interesting assembly possibilities are not artificially discounted.

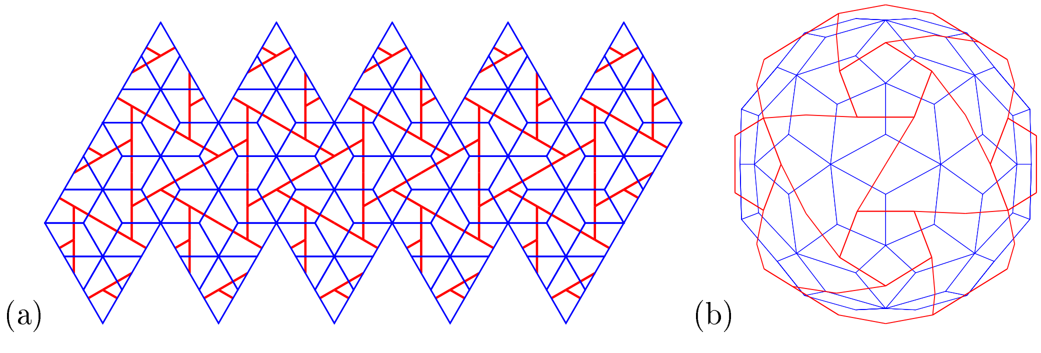

Figure 6.5: Planar representations of the MS2 capsid with locations of RNA density shown in red. (a) The virus represented as a net with dimeric building blocks shown as rhombs. (b) A view along a two-fold axis of symmetry.

Intermediate Size	Protein Only	UniRNA		TrRNA		BiRNA	
	No.	No.	Dead-Ends	No.	Dead-Ends	No.	Dead-Ends
Free Capsomere/RNA	1	2	0	2	0	2	0
1	NA	1	0	1	0	1	0
2	2	1	0	6	0	2	0
3	8	2	0	18	0	3	0
4	18	4	0	48	0	7	0
5	52	8	0	120	0	12	0
6	136	15	0	276	0	25	0
7	391	29	0	602	0	43	0
8	1 108	54	0	1 312	0	86	0
9	3 252	104	2	2 808	0	156	0
10	9 486	190	1	5 740	0	295	0
11	28 087	355	4	11 748	0	534	0
12	83 174	655	10	23 688	24	1 001	1
13	247 749	1 197	17	46 930	26	1 805	1
14	738 582	2 190	36	92 316	28	3 325	1
15	2 207 153	3 967	64	179 490	90	5 983	3
16	6 597 819	7 173	141	345 696	448	10 853	14
17	19 733 747	12 847	237	658 410	204	19 365	6
18		22 940	473	1 244 700	828	34 670	25
19		40 631	911	2 326 740	2 926	61 230	77
20		71 276	1 627	4 297 360	3 200	107 606	82
21		123 923	3 090	7 843 836	7 686	186 758	183
22		212 860	5 720	14 118 808	16 104	321 185	371
23		361 610	10 703	25 072 806		545 061	766
24		606 514	19 478			914 881	1 300
25		1 004 525	35 429			1 514 667	2 558
26		1 641 549	63 928			2 476 174	5 236
27		2 643 280	112 499			3 986 421	9 217
28		4 194 076	196 233			6 327 571	17 783
29		6 550 577	336 824			9 882 321	32 141
30		10 064 833	570 413			15 188 304	58 505
31		15 199 432	947 658			22 936 394	103 239
32		22 545 648	1 545 753			34 033 030	
33		32 822 588					
...							
90		40 678					

Table 6.4: The intermediate numbers for the 4 different MS2 scenarios. With the dead-end numbers for each size.

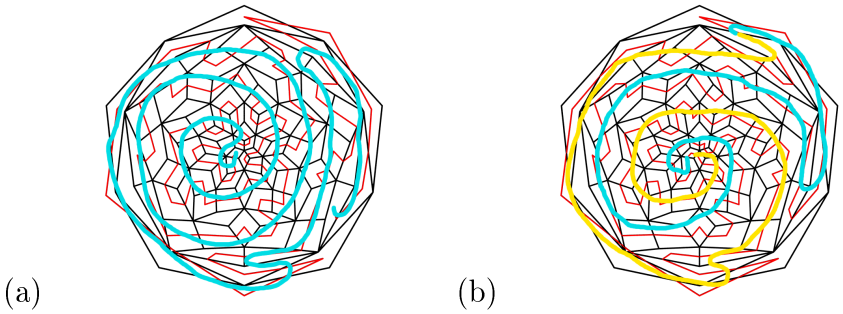

Figure 6.6: (a) A Hamiltonian path forming a spiral from center of a Schlegel representation of the MS2 capsid, (b) A “double” spiral Hamiltonian path, the qualitative layouts of the paths are highlighted in blue and yellow.

6.4 Reducing the complexity

To be able to further model the assembly of the MS2 virus it is necessary to reduce the large number of intermediates that need to be considered. In (Moisant et al., 2010) and (Endres et al., 2005) Zlotnick found that not all the protein-only intermediates were required in order to capture much of the assembly behaviour, in fact, as previously discussed only, 1,124 intermediates

Number of Paths	Short Steps	Long Steps
212	34	25
1132	35	24
2690	36	23
4116	37	22
5300	38	21
6204	39	20
6902	40	19
5868	41	18
4126	42	17
2684	43	16
1132	44	15
272	45	14
40	46	13

Table 6.5: The distribution of long and short steps in the 40,678 Hamiltonian paths of MS2.

were required of the 2,423,212 for STMV. However the kinetic traps that appear in the protein-only simulations are due to large numbers of partially built capsids forming. This could be the case no matter how many pathways are chosen. For instance a linear pathway with only 3 intermediates and 2 assembly steps to capsid could have large amounts of material trapped in the second intermediate at a negative enough bond strength.

The significant kinetic traps of the RNA scenarios depend on the network topology so reducing the size of the network will likely have more of an effect than in protein-only scenarios. For the MS2 UniRNA scenario there are 40,678 Hamiltonian paths, these paths have a further 40,677 branch points between them. Additionally many more branch points leading to dead-ends are present. The dead-end pathways could be reduced to a single dead-end with a slower backward rate, which would model the multiple backward steps required to equilibrate material out of the dead-end. More simply the forward rate to capsid intermediates, that have branches to dead-ends, could be reduced to model the time dependence of dead-end equilibration. This would lead to kinetically modelling about 80,000 intermediates which is closer to the 25,000 already successfully simulated. The network for this model could be created from the 40,678 Hamiltonian paths without the need to calculate all the other intermediates. This is because the final capsid pathways are already known and any other branching possibility for an intermediate must therefore lead to a dead-end.

It would also be possible to only consider intermediates that form large number of capsomere contacts in a coarse grained model. This is similar to how the smaller networks were constructed in (Moisant et al., 2010). With their large numbers of intermediates and dead-ends, the dodecahedral networks are expected to be a sufficiently complex system on which to test future coarse-grained kinetic implementations. It is likely that both the previous constructions of the cube and dodecahedral models will become invaluable in predicting the effects of coarse-graining on the larger viruses.

To coarse-grain the kinetic modelling itself, changing the way in which the free capsomere concentration is considered may help. This is because every forward and backward reaction updates the concentration of free dimer. As a result of this ensuring an accurate concentration of free capsomere requires very small time steps. Alternatively the free capsomere concentration could be set to a constant low amount. This scenario would reduce the integration complexity and perhaps also, be a more appropriate assumption for *in vivo* modelling where capsomeres are being produced concurrently with virus formation.

6.4.1 Further MS2 Biology

Maturation Protein

An alternative approach to simplifying the model is to use more assumptions based on the biological knowledge of the MS2 virus. It has long been known that the MS2 virus has both ends of its RNA bound to its single maturation protein (Shiba and Suzuki, 1981), which is important for infection. The maturation protein is situated at a 5-fold axis of the MS2 virus (Toropova et al., 2011). The implications of these biological results suggest that the ends of the RNA, in a Hamiltonian path, should start and finish at the same 5-fold axis. This leads to the two possibilities shown in figure 6.7, 6.7(a) is termed a cycle, 6.7(b) is termed a pseudo-cycle. The cycle possibility corresponds to a Hamiltonian path (Hamiltonian cycle) that starts and finishes on adjacent dimers, whereas the pseudo-cycle would start and finish on opposite dimers around the 5-fold axis. These cyclic constraints are assumed to limit the valid Hamiltonian paths, for instance only 1,456 of the original 44,678 Hamiltonian paths start and finish on adjacent A/B dimers. Completing the final link between these two dimers creates a Hamiltonian cycle, which reduces the 1,456 paths to 42 that are unique. This reduction occurs due to each Hamiltonian cycle having up to 60 starting points for a particular Hamiltonian path. The reason why this cycle number is not simply 1,456 divided by 60, is because some of the paths contain two or three repeated sequences of directions. The number of 42 cycles may be further reduced by not considering the direction of the path, similar to the BiRNA scenario, which leaves only 32 remaining unique paths.

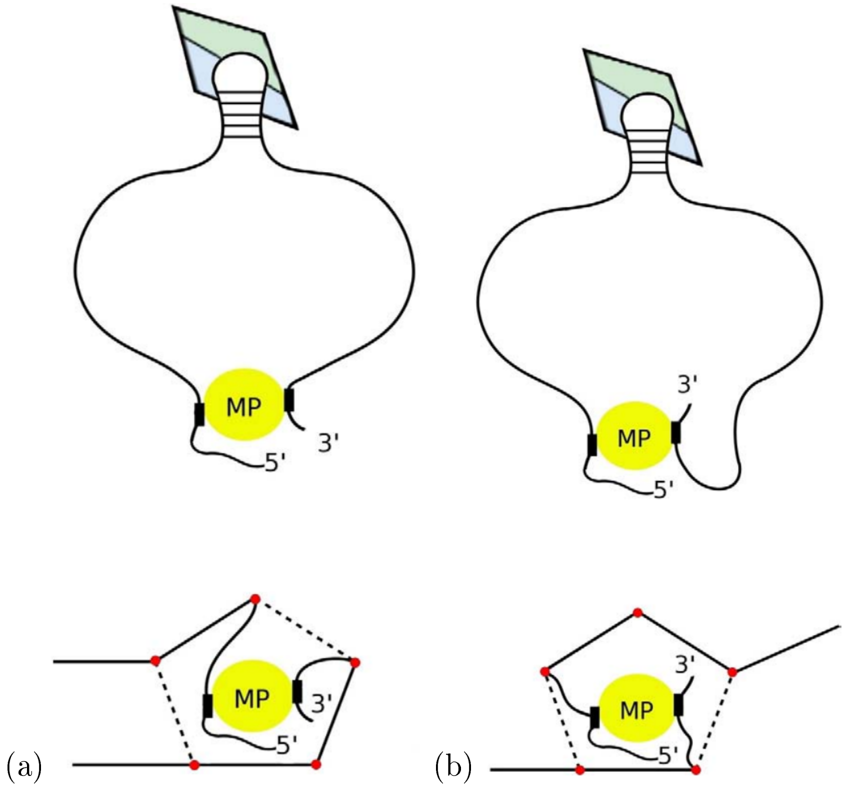

Figure 6.7: RNA Hamiltonian paths starting and finishing at the same 5-fold axis, (a) shows a Hamiltonian cycle possibility, (b) shows a pseudo-cycle Hamiltonian path

5-Fold Averaged Cryo-EM

A cryo-EM reconstruction of the MS2 bacteriophage bound to the F-pilus of bacteria has recently provided new insights into the internal RNA layout (Toropova et al., 2011). Binding the virus to F-pili determines a particular 5-fold axis of the 6 possibilities, therefore the resulting reconstructed image is only 5-fold averaged rather than icosahedrally averaged. The selection of particular virus particles in the cryo-EM micrograph images used only those in which the attachment site was visible, see figure 6.8. The reconstructed is shown in figure 6.9. The data from this 5-fold reconstruction shows different amounts of RNA density in the outer shell for different distances from the maturation protein. This information is later used to validate the selection of possible Hamiltonian paths.

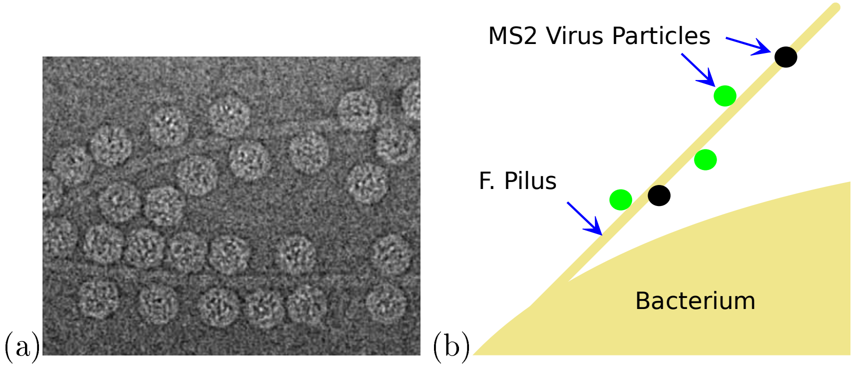

Figure 6.8: (a) A cryo-EM micrograph of MS2 particles bound to F-pili (Toropova et al., 2011). (b) A diagrammatic representation of example virus positions, where the attachment sites would be observable, are shown in green and example locations which are unsuitable for use in the reconstructions are shown in black.

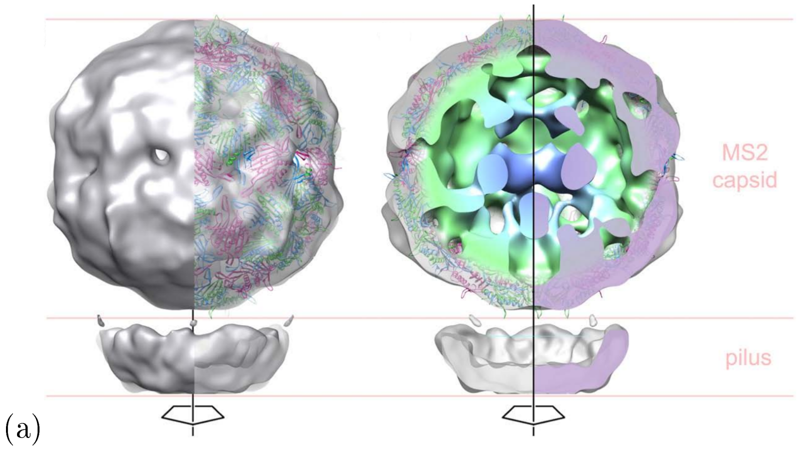

Figure 6.9: The left image shows the outer surface of the reconstruction and the image on the right shows only the back half (Toropova et al., 2011).

Mass Spectrometry Data

Mass spectrometry has been used to analyse the assembly of MS2 particles using only the TR stem loops (Knapman et al., 2010) (Morton et al., 2010). The results show that the assembly of hexamers, around a 3-fold axis and decamers around a 5-fold axis are most likely intermediates. These highly bonded intermediates were also predicted in a model of the mass spectrometry results (Dykeman and Twarock, 2010). It would make sense therefore to model the assembly profiles of MS2 that form the most hexamer and decamer configurations.

6.5 Simple Rules for Efficient Assembly

In the paper "Simple Rules for Efficient Assembly Predict the Layout of a Packaged Viral RNA" (Dykeman et al., 2011), a co-author, Dr. Eric Dykeman continued the analysis of the MS2 Hamiltonian paths, described above, to find only 66 that are consistent with binding the maturation protein at the 5-fold axis. Dr. Dykeman was then able to find 3 of these paths that are consistent with mass spectrometry data (Morton et al., 2010) and remarkably, validate this result against the 5-fold averaged cryo-EM data. The initial 66 paths are cycles and pseudo-cycles that have been reduced from the larger set of cycles and pseudo-cycles through removing symmetric and inversely identical paths. The paths were also filtered to remove those paths with only slight differences in the endings of the configurations. The paths were again reduced by considering only paths that lead to capsid assembly, following the Hamiltonian path model, through forming hexamers and decamers. The initial assembly of these paths was constructed such that the maximum number of capsomere contacts were formed at each step (see figure 6.10).

It was found that the RNA density of these three paths was indeed consistent with that of the 5-fold averaged cryo-EM data (Toropova et al., 2011). In order to align the Hamiltonian paths to the 5-fold averaged cryo-EM data first the Hamiltonian path was converted to an RNA density. The method used to do this was to measure the distances between the MS2 dimers in the crystal structure from (Valegård et al., 1997) and convert these numbers into the amount of RNA density expected for the long and short steps within the Hamiltonian paths. Secondly both the cryo-EM data in the outer shell and the Hamiltonian path densities were converted to 1D projections, these projections are compared in figure 6.11. Figure 6.11(a) shows the average of converting the 40,678 paths, 66 filtered paths and final 3 paths to an RNA density. Certainly the three final paths compare very favourably to the 1D projection of the outer shell of the 5-fold averaged cryo-EM data, also shown in figure 6.11(a). Interestingly paths with slightly less energetically favourable bonding did not match with the cryo-EM RNA density.

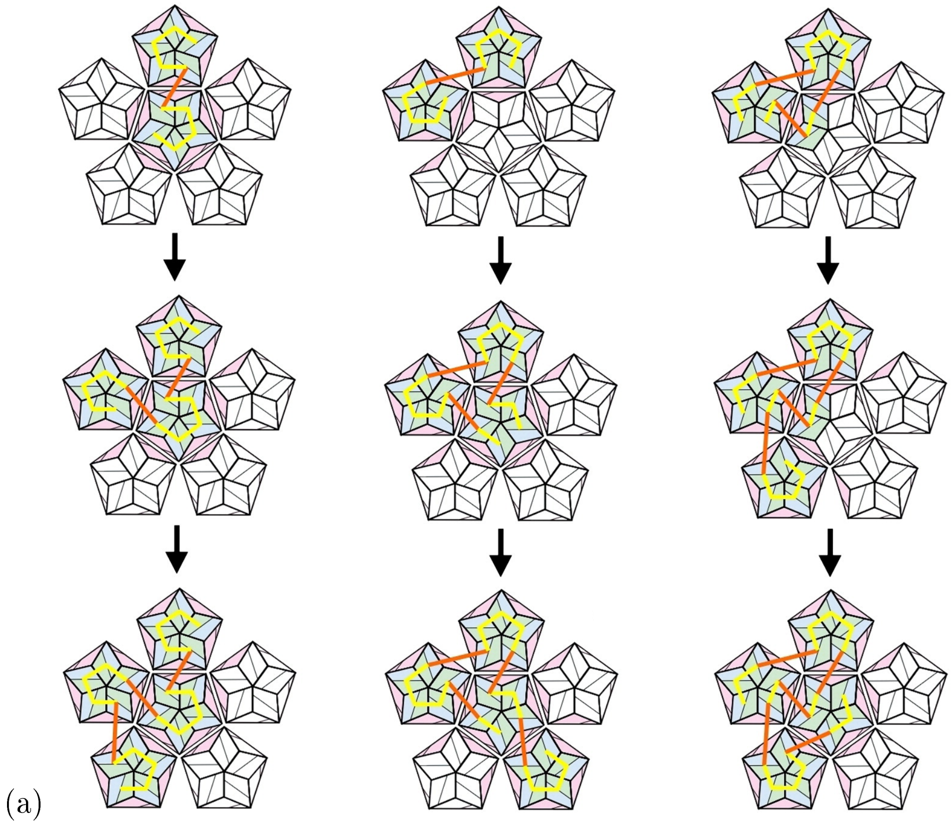

Figure 6.10: The three pathways that form the highest number of bonds at this early stage of assembly. The left most pathway has the highest number of bonds at each step whilst the centre and right pathways have the second highest number of bonds at each step. Reproduced from (Dykeman et al., 2011).

6.6 Conclusions

In this chapter we have seen how large the combinatorial possibilities of virus assembly become. Again this can be compared to the protein folding problem where there are also very large numbers of possible states (Levinthal, 1969). However proteins and viruses do successfully reach their final configurations quickly and accurately. In lieu of actually running kinetic simulations on the STMV and MS2 assembly networks it has been shown how the networks themselves may be analysed to gain insights to the likely assembly behaviour. This analysis has emphasised the interplay of the network branching and the bonds formed within intermediates. This network investigation shows that by forming energetically favourable intermediates the bulk of the pathways to kinetically trapped dead-end intermediates may be avoided.

The network analysis also shows the interesting geometry of the MS2 and

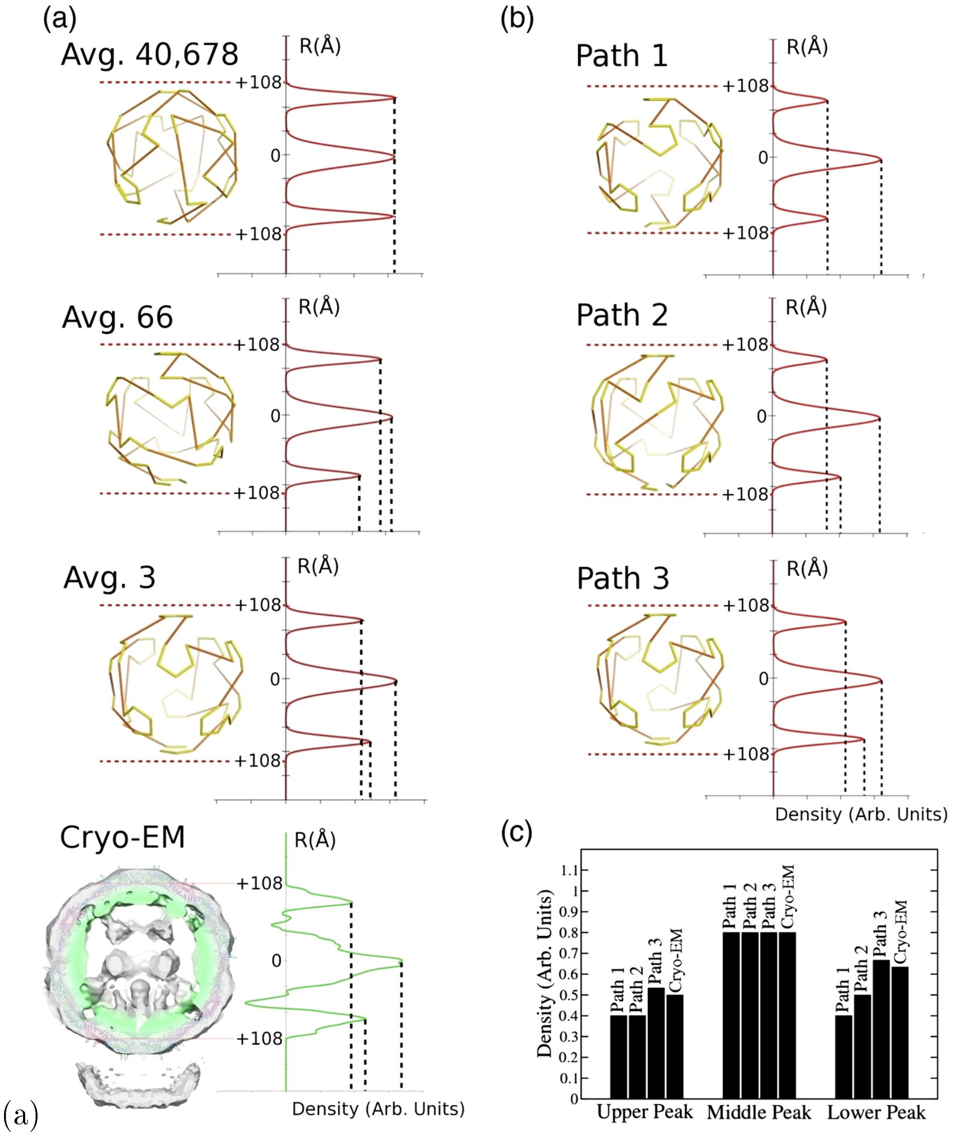

Figure 6.11: (a) shows the average of converting the 40,678 paths, 66 filtered paths and final 3 paths to an RNA density along with the 1D projection of the outer shell of the 5-fold averaged cryo-EM data. (b) The individual RNA densities of the 3 predicted paths and (c) a table of relative densities for the three distinct levels of RNA density found. Reproduced from (Dykeman et al., 2011).

STMV intermediates. Particularly interesting are the short and long steps of the walk in the MS2 capsid. These short and long steps correspond to changes in the length of the RNA between A/B dimers. This allows for the validation of paths with the cryo-EM data. The short and long steps will also determine where along the genome RNA stem loops are required to be in order to bind the A/B dimers. Since these distances along the RNA vary, it is possible to analyse the genome for stem-loops that match up to particular long and short steps. This is topic of soon to be submitted work in (Dykeman et al., 2012).

Different coarse-grainings of the kinetic model have been discussed in terms of finding a solution to the computational intractability of the large assembly networks. Coarse-grained models would be relatively easy to implement, especially using the existing computer program code base. These models could then be run on the STMV networks for which the UniRNA and BiRNA networks have already been fully computed. Running a coarse-grained model is relatively easy, it is understanding the effects, if any, that the coarse-graining produces on the results that is hard to determine. This is where the cube and dodecahedral kinetic simulations will become very informative, since any coarse-grained model could be first run on these networks and the effects on the results compared. The particular choice of coarse-graining is left to future work. In a break from the Zlotnick based model (Zlotnick, 1994), there is also the possibility of using Gillespie algorithms to analyse the networks already determined (Gillespie, 1977) (Hemberg et al., 2006). For larger, computationally intractable networks, a Gillespie algorithm could be implemented such that only a small number of intermediates was required at any point. These intermediates could be generated on the fly, removing the need to calculate the whole of the network in advance.

Additional biological information has also been considered in order to reduce the complexity of the analysis specific to the the MS2 bacteriophage. This has lead to finding three Hamiltonian paths that are constrained to binding the maturation protein, follow the energetically favourable pathways determined in mass spectrometry results and also happen to conform with the RNA density found in (Toropova et al., 2011). This is a remarkable result that, like the network analysis, suggests that the assembly pathways likely follow only the most stable intermediates.

The most stable intermediates would naturally occur at less negative bond strengths, where any intermediate forming only a few bonds would readily break apart. The high number of backward reactions required for this process favours a reaction system close to equilibrium. This is one of the hallmarks of a self-assembly. In the next chapter this theme of self-assembly continues with the topic of DNA cages.

Chapter 7

Designing an icosidodecahedral DNA cage

7.1 Introduction

A DNA cage is a three-dimensional shape made from double-helical DNA molecules. DNA and RNA cages have been seen in nature and may also be constructed artificially. A good example of an RNA cage in nature is the dodecahedral arrangement of the viral genome in pariacoto virus (figure 7.1) (Tang et al., 2001). This is the same virus that inspired the Hamiltonian path model of assembly in earlier chapters (Rudnick and Bruinsma, 2005). In this chapter we will look at constructing DNA cages using nanotechnology process as opposed to the self-assembly approaches used by viruses. Designed DNA cages may be constructed that are stable in solution without any further bonding from proteins or other molecules. In 1991 (Chen and Seeman, 1991) created a DNA cage structure in the shape of a cube. Further examples of DNA cages include a truncated octahedron (Zhang and Seeman, 1994), octahedron (Shih et al., 2004), tetrahedron (Goodman et al., 2005), dodecahedron (He et al., 2008) and buckyball (He et al., 2008). These example DNA cages are shown in figure 7.2. DNA is generally used to make these nanoscale structures because it is more stable than RNA. There is also no wobble base pairing in DNA which makes the sequences easier to design. Finally what makes DNA such a great nanoscale material is that many of the previously developed techniques and enzymes from Biology may be employed in its manipulation. DNA cages show promise in a wide variety of nanoscale applications. In medicine there are possibilities for drug delivery (Destito et al., 2007) and diagnostic purposes (Chhabra et al., 2010). There is also potential in molecular nanofabrication,

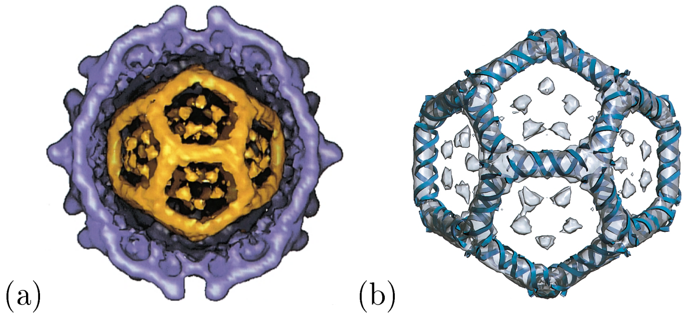

Figure 7.1: (a) The three dimensional reconstruction of pariacoto virus from (Tang et al., 2001). (b) The double-stranded RNA layout of the dodecahedral cage with additional internal density shown in the pentagonal faces. (Tang et al., 2001).

environmental sensing (Chhabra et al., 2010) and DNA computing (Sa-Ardyen et al., 2004).

7.1.1 Constructing DNA cages

To construct a DNA cage, a double stranded DNA molecule must run along each edge of the shape in question only once. The individual strands must therefore run in opposite 5' to 3' directions to ensure the correct hybridisation of the DNA double helix. Mathematically these cages are related to topological graph theory and such a DNA cage is termed an orientable thickened graph (Jonoska and Saito, 2002). Ideally the DNA cage would also be made out of a minimum number of individual DNA strands to increase the stability. A DNA cage using the minimal number of strands is also important in certain DNA computing applications (Sa-Ardyen et al., 2004).

7.2 Bead model

Previously, a systematic approach to designing DNA cages was proposed by Jonoska and Twarock (Jonoska and Twarock, 2008). This systematic approach is termed the bead model and was demonstrated for a dodecahedron of DNA. This dodecahedron of DNA was partly inspired by the dodecahedral cage of RNA found in pariacoto virus (Tang et al., 2001). The results of this chapter describe how, using the bead rule, a DNA cage may be designed in the shape of the icosidodecahedron, shown in figure 7.3. A DNA cage of the shape of an icosidodecahedron has a few advantages over previously constructed shapes. Firstly it has a greater volume to surface area ratio than the dodecahedron

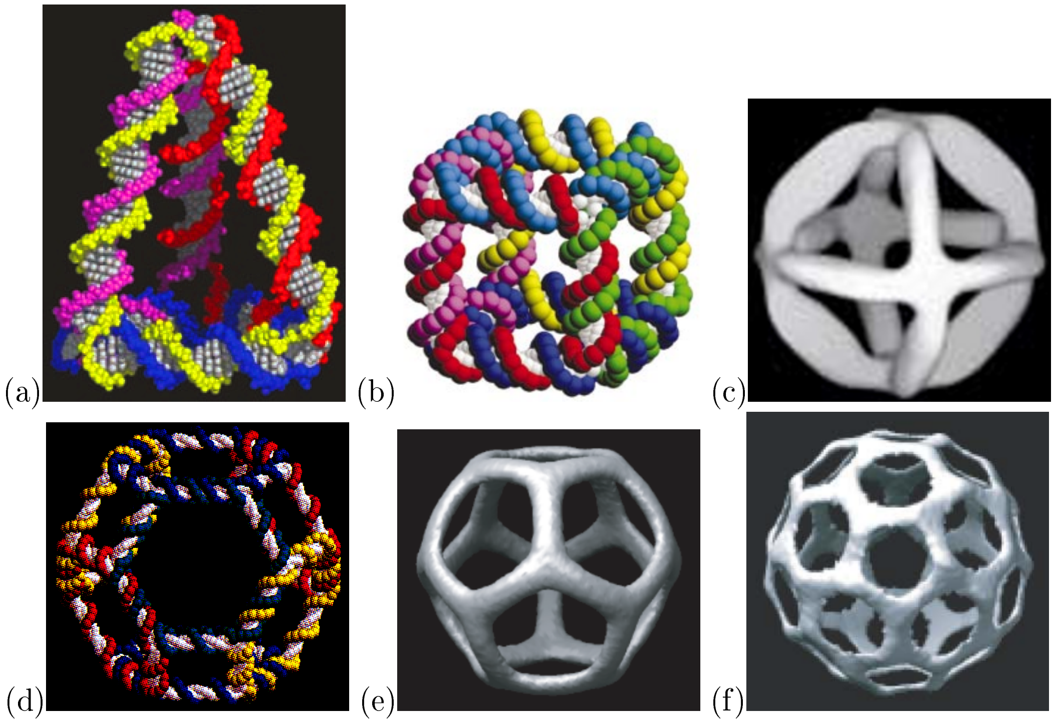

Figure 7.2: A DNA tetrahedron (a) (Goodman et al., 2005), cube (b) (Chen and Seeman, 1991), octahedron (c) (Shih et al., 2004), truncated octahedron (d) (Zhang and Seeman, 1994), (e) dodecahedron (He et al., 2008) and buckyball (f) (He et al., 2008). All images here have been reproduced from their respective papers.

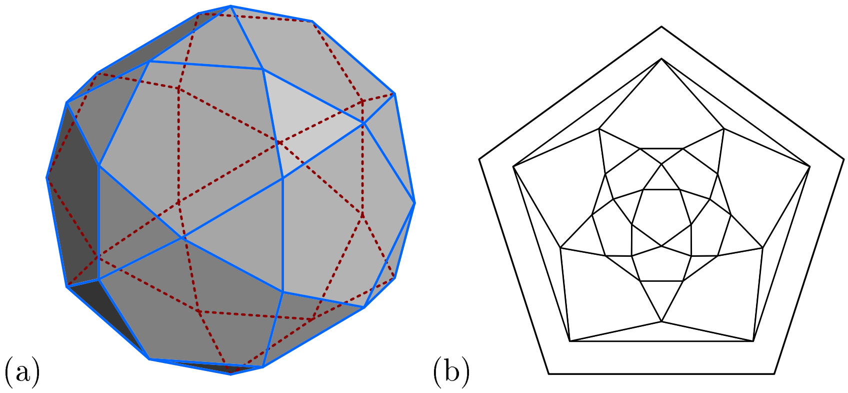

Figure 7.3: (a) The icosidodecahedron and (b) the corresponding planar Schlegel diagram with the pentagonal back face expanded.

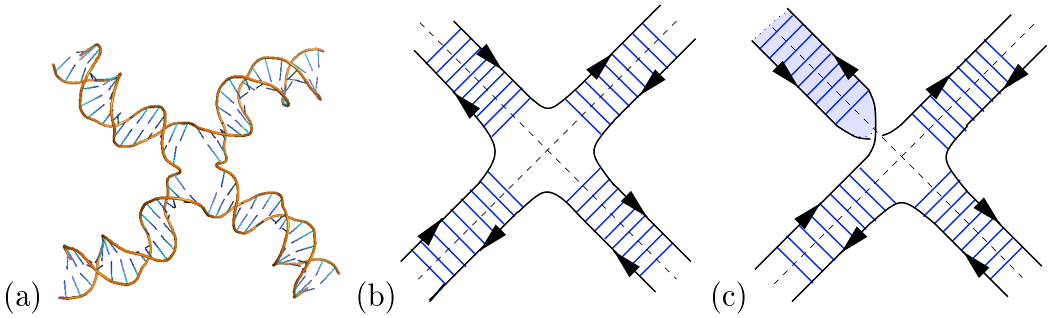

Figure 7.4: (a) DNA Holliday junction crystallised in (Gopaul et al., 1998), (b) the corresponding DNA four junction, (c) a DNA four junction with an extra half-turn. (b) and (c) have been reproduced from (Grayson et al., 2009).

and therefore may be more useful as a container for drug delivery purposes. Secondly, it is some-what simpler than the buckyball in figure 7.2(f). This simplicity is expected to lead to a greater yield upon manufacture when compared to the buckyball (He et al., 2008).

The icosidodecahedron has 20 triangular and 12 pentagonal faces along with 60 edges that are all of the same length. The edge length of the icosidodecahedron determines the size of the DNA cage and also the number of DNA helical turns along each edge. A full turn of the DNA helix has 10.4 base pairs and has a length of about 3.4nm. The edge length of the icosidodecahedral cage may be set such that an integer number of full DNA helix turns is possible on each edge. The junctions of this cage at the four coordinated vertices of the icosidodecahedron would then take on the structure of the well known Holliday junction (Gopaul et al., 1998). The Holliday junction is shown in figure 7.4(a) with its schematic layout shown in figure 7.4(b). This Holliday junction creates a correctly oriented DNA cage with 32 separate strands - one for each face of the icosidodecahedron. It is also possible to construct the icosidodecahedral DNA cage with an extra half-turn of the DNA helix on each edge. This allows for a greater choice of sizes for the cage. Introducing this extra half-turn on each edge creates a DNA cage of 12 separate strands. However, the individual strands are no longer correctly orientated such that base-pairing between them occurs. A cage that exhibits such an extra half-turn is shown in figure 7.5.

In order to correctly orientate the DNA strands when an extra half-turn is present on the edges, the bead rule has been introduced (Jonoska and Twarock, 2008). The bead rule describes where an extra half-turn must be introduced on the DNA cage to result in a correct orientation of the strands. A change in junction type resulting from the application of the bead rule is shown in figure

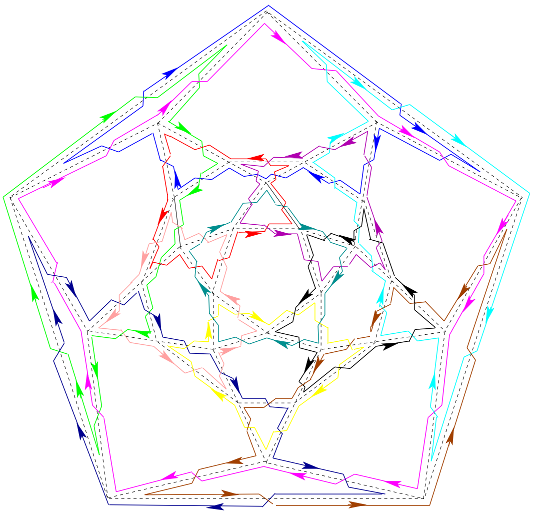

Figure 7.5: The planer icosidodecahedron with an extra half helix turn on each edge. This leads to 12 separate DNA strands that are not correctly orientated. Reproduced from (Grayson et al., 2009).

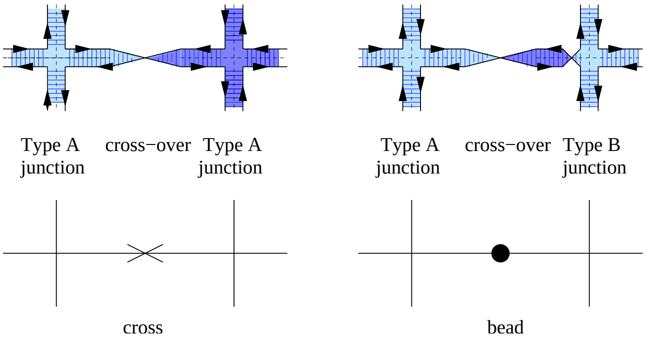

Figure 7.6: Demonstration of the bead rule: The change in DNA junction type resulting from placement of a cross (or a bead) on an edge of the icosidodecahedron. Reproduced from (Grayson et al., 2009).

7.6. Due to the extra mechanical stress on a junction where beads are present the number of bead placements is kept to a minimum.

Every face of the icosidodecahedron must have an even number of cross-overs to have the correct 5’ to 3’ orientations of the DNA. Therefore, each triangle must have a bead placed on one of its edges. Since there are 20 triangles the minimum number of beads is 20. Of course by placing a bead on the edge of a triangular face it is also placed on the edge of a pentagonal face that shares this edge. Following the rule that the pentagonal faces must have an even number of cross-overs, and therefore an odd number of beads, each pentagonal face must have either 1, 3 or 5 beads. From these constraints two equations have been produced, as shown in (7.1). In these equations, α

is the number of pentagonal faces that have one bead, β is the number with three beads and γ is the number with five beads. The number of beads must add up to 20 as shown, and the number of pentagons must add up to the 12 the icosidodecahedron contains. There are three solutions to these equations, termed cases I, II and III, which are shown in (7.2).

$$\begin{aligned} \alpha + 3\beta + 5\gamma &= 20 \\ \alpha + \beta + \gamma &= 12 \end{aligned} \tag{7.1}$$

$$\begin{array}{ll} \text{Case I} & \alpha = 8,\ \beta = 4,\ \gamma = 0 \\ \text{Case II} & \alpha = 9,\ \beta = 2,\ \gamma = 1 \\ \text{Case III} & \alpha = 10,\ \beta = 0,\ \gamma = 2. \end{array} \tag{7.2}$$

In Case I there are four pentagons that have 3 beads. It has been determined that there are nine different ways these four pentagons may be located on the icosidodecahedron. These nine layouts are shown in figure 7.7. The layouts A1, A2 and A3 in figure 7.7 correspond to all possible ways to arrange the four pentagon positions when three are adjacent to each other. Layouts B1, B2, B3 and B4 are all the possible positions when three of the four pentagons are arranged in a line. Finally layouts C1 and C2 are the possible positions when only two of the pentagons are adjacent. Also shown in figure 7.7 are the lines of symmetry and points of rotational symmetry that correctly rotate or reflect the four pentagons on to each other. These symmetries are later used to remove duplicate bead layouts.

For Case II there are seven possible layouts: D2, D4, D5, and D6 have the single pentagon with 5 beads adjacent to a pentagon with 3 beads. This is not the case for the further E1, E2, and E3 configurations. In the last possible case, Case III there are only two possible layouts, F2 and F6.

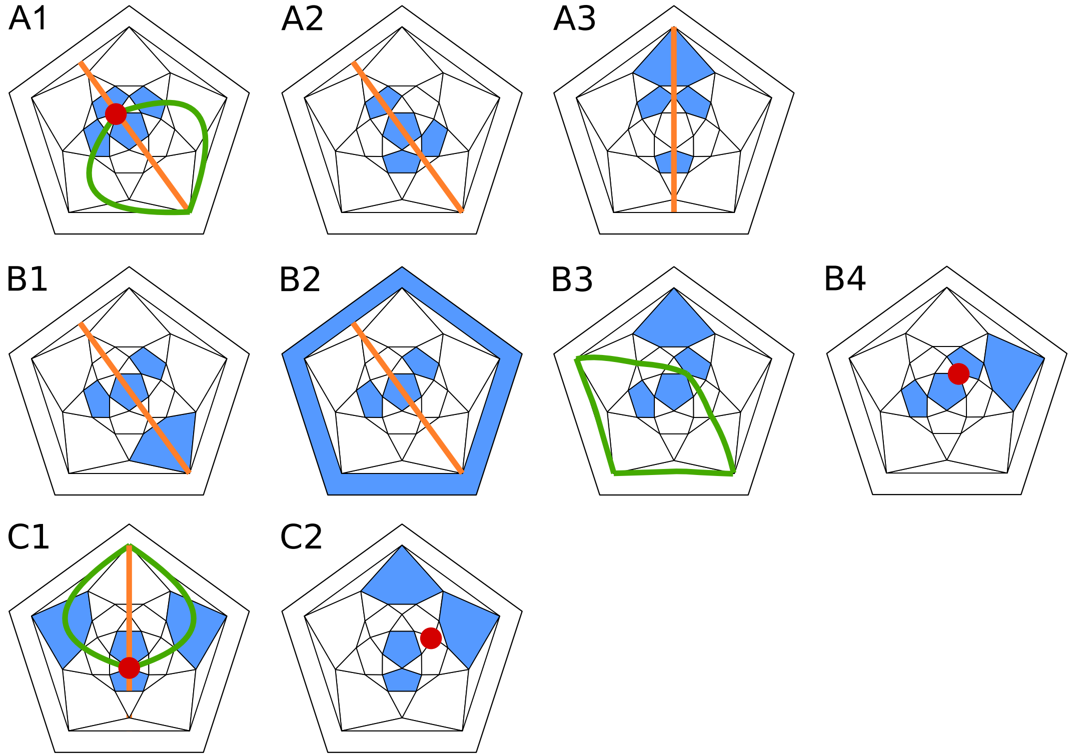

Figure 7.7: The layouts for Case I where four pentagons, shown in blue, have three beads. Mirror symmetry lines are shown in green and orange, while one end of each axis of rotational symmetry is shown by a red circle.

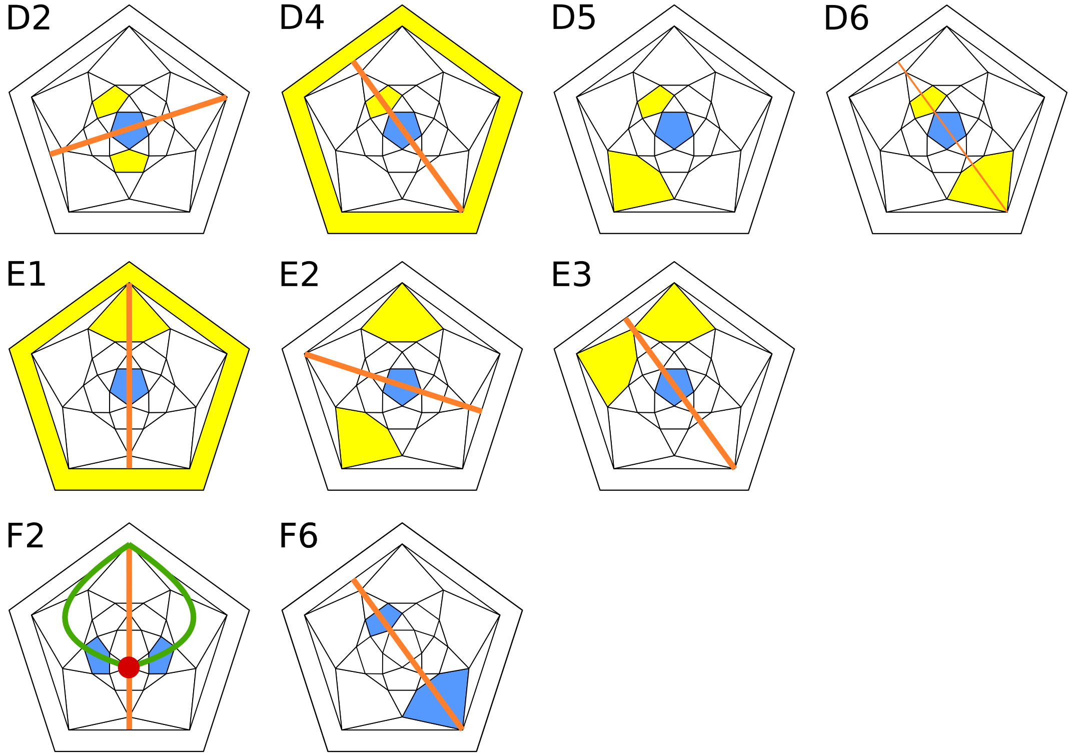

Figure 7.8: The layouts for Case II are labelled; D2, D4, D5, D6, E1, E2 and E3. Case III layouts are labelled F2 and F6. Mirror symmetry lines are shown in green and orange, while one end of each axis of rotational symmetry is shown by a red circle.

7.3 Computer Implementation

Now that the unique location of the pentagons has been determined, it is possible to start looking for bead configurations that fulfil the requirement that every face has an even number of crosses. This is not a simple constraint to satisfy, because each placement of a bead affects two faces. The only way to find all possible bead layouts is to check every combinatorial way of placing beads. For DNA cages in the shape of a dodecahedron this process was carried out manually (Jonoska and Twarock, 2008), however for the larger icosidodecahedron a computer program is required, because there are far more possible combinations. To store each configuration, the three edges of each of the 20 triangles are numbered 1, 2 or 3. Then a vector of length 20, corresponding to the 20 triangles, may be used to record which of the edges of each triangle has a bead placed upon it. Each possible position in this vector may contain the edge number 1, 2, or 3. With these three possibilities, for each position in the vector there are 3^{20} combinations which is, 3,486,784,401. To check whether this many bead layouts correspond to each face having an even number of beads would take far too long.

To increase the efficiency of the algorithm, admissible bead layouts are checked locally as the vector is filled. The first 5 triangles of the bead vector surround a single pentagon, this pentagon may then be checked at this early point to ensure it has the correct number of beads. Further pentagons are checked sequentially as their edges are assigned beads. This early checking, and likely rejection, avoids the need to construct every possible vector. With this algorithm, the computation is able to finish in only a few days. The associated programs may all be found on the accompanying CD, please refer to the read me file in the root directory.

Once all the possible bead locations have been determined, the mirror and rotational symmetries are removed in order to obtain a unique number of different layouts. The bead configurations are then converted into the number of DNA strands required to construct each layout. These DNA strands form loops. For a particular bead layout this process starts on an arbitrary edge of the icosidodecahedron and follows the beads or crosses until the starting edge is reached again. Then a new edge is picked and the walking along the edges repeated until all the edges have two DNA strands. Again a computer program has been written to convert the bead layouts to the DNA strands and produce a visualisation of the output. A re-drawing of one of these outputs is shown in figure 7.12.

Case	Configuration	Loop Number			
		10	12	14	16
Case I	A1	70	9		
	A2	325	13		
	A3	1025	65		
	B1	1743	39		
	B2	2248	47		
	B3	695	18		
	B4	1012	22		
	C1	2066	59	1	
	C2	2343	71	2	
	Case I Sum:	11527	343	3	
Case II	D2		13		
	D4		90		
	D5		100		
	D6		90		
	E1		274	6	
	E2		364	2	
	E3		20		
	Case II Sum:		951	8	
Case III	F2			9	
	F6			64	1
	Case III Sum:			73	1

Table 7.1: The loop numbers for each correctly orientable pentagon configuration.

7.4 Results

The numbers of correctly orientated DNA strand layouts are shown in table 7.1. The numbers of loops formed between the different configurations and bead layouts differ. The largest number of strands/loops was found to be 16, while the smallest is 10. There is also a large range in the numbers of loops, from 11527 with 10 loops to only 1 with 16 loops.

In the simple case where there are no beads or crosses on the edges each junction has four separate DNA loops. When there are beads present and the DNA cage forms 10 loops, notably every junction only has three or fewer different loops. When there are 12 loops present, there are either zero or between 3 and 17 junctions with four loops. For 14 loop cages there are also either zero, or between 10 and 18 four loop junctions. In the single 16 loop case there are 19 junctions with four strands and 11 junctions with three strands.

To create DNA cages with a lower number of total DNA strands it is possible to replace the junctions at the vertices of the icosidodecahedron with two energetically favourable alternatives (Condon et al., 2009), shown in figure 7.9. In the simpler case, where each edge of the icosidodecahedron does not have a cross-over or bead, it is possible to use either of the replacement junctions shown in figure 7.9. This reduces the number of loops to two. It is not possible to reduce the original 32 loops to a single loop, because either of the possible replacement junctions reduces the loop count by two on each application. Therefore, after 15 replacements, there still remain two loops that may not be combined. There are, however, many options of where to make these minimal 15 replacements which leaves a choice in the relative lengths of the remaining loops. If only one strand of DNA was required, a hairpin structure on one of the edges would create a single strand. The energetic considerations of placing a hairpin are discussed in (Jonoska and Twarock, 2008) and is demonstrated for a dodecahedral cage in figure 7.10. The stability of this hairpin would likely be low and require additional considerations to counter this. In viruses such a hairpin structure could easily be stabilised by the capsid proteins.

As an example, the 16 loop result is used (see figure 7.11). Due to there being 16 loops, 14 replacements are required to result in the final two. The first combinations create a single strand from 7 of the original loops, with the remaining 9 loops forming the second loop in the end configuration. Figure 7.11 shows the sequence of junction replacements that results in the final two loops. In figure 7.11(a) the numbered circles indicated the order of the junction replacements. Figure 7.11(b) shows the resulting red strand that comes from

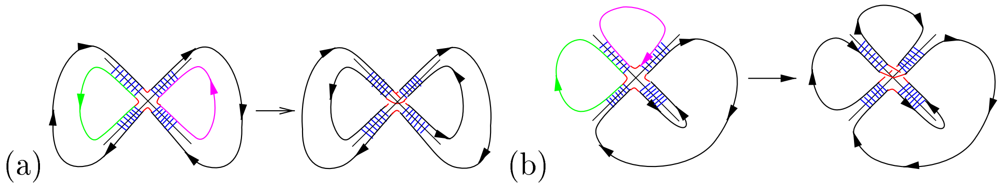

Figure 7.9: Two energetically favourable replacement junctions for the case that three different DNA strands meet at a junction. Different loops are colourd individually and blue lines represent base pairing between the loops. This figure has been reproduced from (Condon et al., 2009).

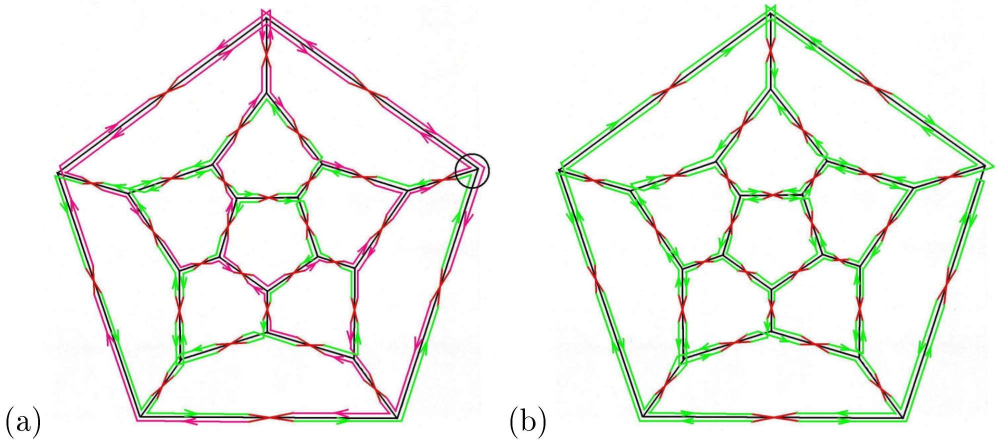

Figure 7.10: A demonstration of using a hairpin loop to combine two DNA strands in (a) to the single strand in (b). This figure has been reproduced from (Jonoska and Twarock, 2008).

the first four junction replacements. The remaining three replacements create the second (blue) strand shown in figure 7.11(c). This final structure is shown in 3D in figure 7.11(d). A further example is shown in figure 7.12 where, by using six junction replacements a 14 loop configuration may be reduced to two strands.

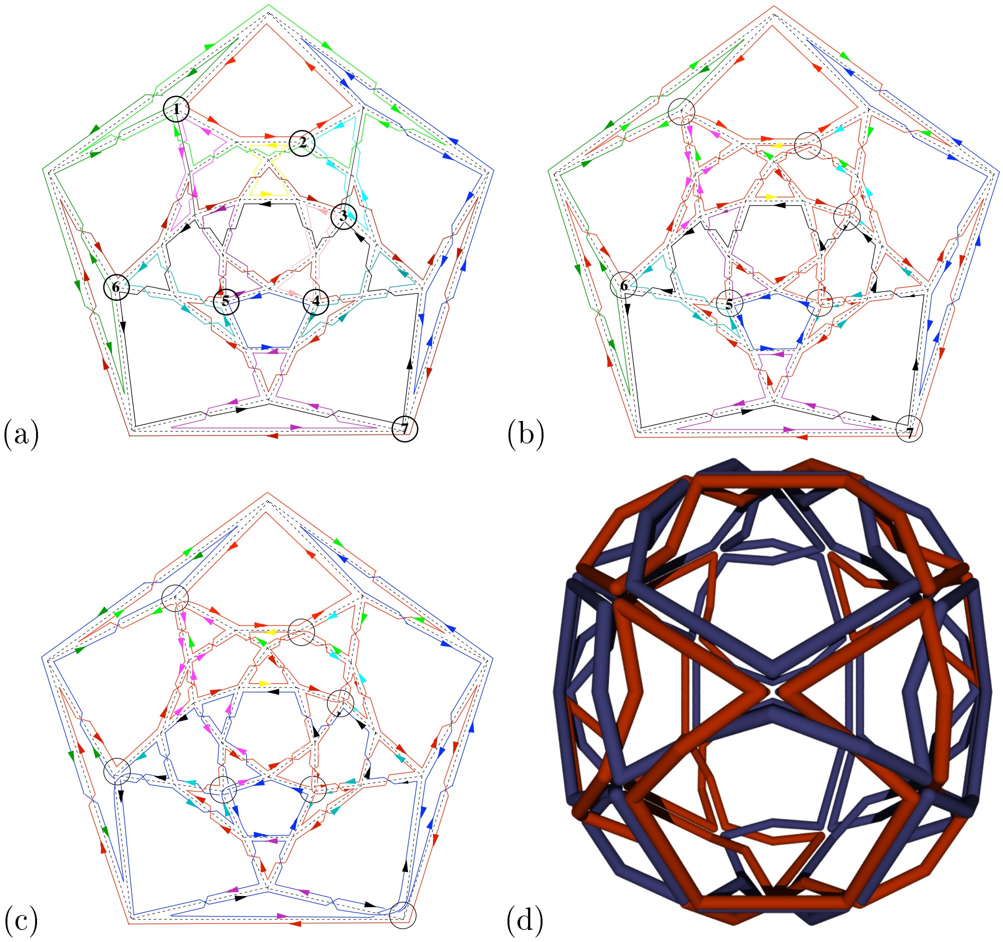

Figure 7.11: (a) The 7 possible locations of replacement junctions. (b) The DNA cage after the first 4 replacements, which create the DNA loop shown in red. (c) The final icosidodecahedral DNA cage constructed from two DNA strands, (d) is this same cage design in 3D, with the original seperate loops shown in their final colours. Parts (a), (b) and (c) have been reproduced from (Grayson et al., 2009).

By placing beads on some edges mechanical stress is introduced. To reduce this stress it may be possible to design the junctions such that there are extra base pairs in the unbound center, although this could make the junction less stable. Alternatively changing the length of the edge away from the icosidodecahedral ideal could also act to reduce the stress. The experimental set-up will determine which is the better option for a given application.

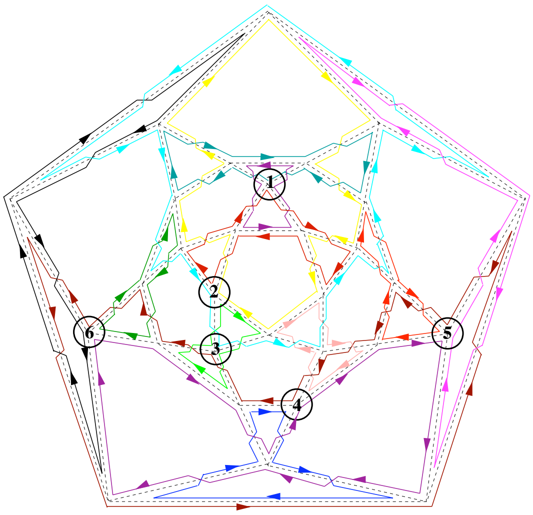

Figure 7.12: The example 14 loop layout reduced through six replacements to two DNA strands. Reproduced from (Grayson et al., 2009).

7.5 Conclusion

The systematic approach of using beads, developed by Jonoska and Twarock (Jonoska and Twarock, 2008) for the dodecahedron, has successfully been applied to the icosidodecahedron. This has lead to the design of many different possible DNA cages with the large volume to surface area of the icosidodecahedron. It has also been found that the minimal number of DNA strands required to construct an icosidodecahedral cage is two. In the design process attention has been paid to constructing DNA cages with minimum mechanical stress. This had lead to using the minimum number of beads and a minimal number of junction replacements. It is hoped these considerations will aid the manufacture of energetically favourable icosidodecahedral DNA cages similar to those in (Sa-Ardyen et al., 2004). DNA cages have a great deal of potential in various applications such as nanotechnology and medicine and it is hoped that the icosidodecahedron with its large volume and relative simplicity will be useful in these areas.

Chapter 8

Conclusion

In this thesis two models that are capable of molecular self-assembly have been described. Both models have a wide range potential in nanotechnology, such as containers for drug deliver (Ma et al., 2012) (Destito et al., 2007) and in nanofabrication (Chhabra et al., 2010) (Gerasopoulos et al., 2010).

The design processes developed for the construction of energetically favourable icosidodecahedral DNA cages will hopefully aid and inform their future construction. The designs modelled allow for many variations in the architecture, which should give more options in any future application. Energetic stability considerations used in the modelling of these DNA cages have also been taken into account, this should help ensure that their self-assembly is driven by the change in free energy.

The cage like structures of nucleic acid that form inside RNA viruses are an emergent property of the assembly process. It is the understanding of this process that is the subject of the majority of this thesis. In chapter 1 the well characterised viruses of the MS2 bacteriophage and STMV were described and their suitability for use as model systems shown. The consequences of the dimer switching model (Stockley et al., 2007) (Dykeman and Twarock, 2010) and cryo-EM density density (Van Den Worm et al., 2006) (Toropova et al., 2008) lead directly to a new way of modelling RNA virus assembly (Grayson et al., 2007). This new idea is that of using Hamiltonian paths to describe the assembly and final structure of the RNA.

Following a discussion of virus and polymer assembly models in the literature it was concluded that Zlotnick's model (Zlotnick, 1994) would be the most suitable to use as a basis to model the kinetics of Hamiltonian path based assembly. The subsequent extension of Zlotnick's kinetic model to include RNA effects proved to be very interesting and successful. This extended model has been shown to reproduce well known assembly behaviour, such as sigmoidal kinetics, hysteresis and also the kinetic effects of the RNA pathway constraints. Furthermore the results from this model show complex emergent behaviour, which is more than just a result of the assumptions used in creating the model.

In the previously simulated protein-only scenarios, kinetic trapping of material within smaller intermediates has been shown to significantly impact the assembly, to the detriment of capsid production (Endres et al., 2005). With the introduction of RNA this kinetic trap is resolved, certainly at RNA to protein amounts near the stoichiometric ratio. Although the introduction of assembly along an RNA Hamiltonian path allows for the possibility of forming dead-end intermediates, swapping one form of kinetic trap for another. However, this dead-end kinetic trap has been shown to be less detrimental to virus formation, over a range of bond values, in chapter 5.

The RNA assembly model introduced considers three main scenarios, that of assembly beginning at the end of the RNA and from the middle, with and without a constantly bound dimer on the TR position. There are also four sizes of virus particle modelled beginning with the cube and dodecahedron and leading up to the STMV 30-mer and 90-mer of MS2. The use and combinations of these different schemes has demonstrated the important features of the RNA assembly model and how the effects scale with virus size. The main consequence to assembly of using the Hamiltonian path model, especially in the larger model systems of STMV and MS2 is the number of dead-end pathways. For efficient virus assembly, using the Hamiltonian path model assumptions, it would be necessary for assembling intermediates to avoid these dead-end pathways. One such way of avoiding many of the dead-end pathways is simply to form well bonded, compact intermediates. Energetically favourable intermediate types have also been predicted, via an analysis of the Hamiltonian paths and further biological constraints in (Dykeman et al., 2011). The predominance of energetically favourable intermediates is one of the principles of a self-assembly process. With this also being the case for RNA viral assembly the result helps explain why it is such a robust process. It has also been shown that nucleating dimer addition in the centre of the RNA is more efficient than assembling from a single end, again as a result of dead-end pathways. This suggests that it may not be a coincidence that the TR stem-loop of MS2 is positioned within 2 % of the centre of the genome. This efficiency argument may even explain why the smaller STMV virus is able to assemble from one end of the RNA where as the larger MS2 virus must start in the middle.

Just by calculating the networks of intermediates required for this model, using the new algorithms developed, we have seen in chapters 3 and 6 that insights can be found into the putative assembly pathways. Coarse-graining of the kinetic model would be required for any future modelling of the larger STMV or MS2 capsids. These coarse grained models would have the advantage of being able to be tested on the smaller cube and dodecahedral systems. However these small systems in themselves have shown all the qualitative behaviour that might be expected of larger systems. The increase in scale from the cube to the dodecahedron shapes has also suggested a pattern of what would happen in larger kinetic networks. Rapaport suggests, "the robustness of self-assembly makes understanding the process in simplified environments a worthwhile endeavour" (Rapaport, 2012), which is probably why the small models work so well.

Virus growth and replication is, of course, a very important subject of research both medically and economically. One of the main reasons for studying virus assembly is to inform approaches to stop virus production. Indeed the results presented here suggest some possible therapeutic routes to preventing viral assembly. Firstly, it has been shown that in all the simulation scenarios the assembly is highly dependant on the availability of free building blocks. Any drug that acted to reduce the availability of the building blocks even slightly could have a significant impact on the number of viral particles produced. A promising area of current research focuses on using RNA aptamers to change the cell genetics and to act as alternatives to antibodies and small molecule drugs (Shigdar et al., 2011). Such aptamers have advantages in that they have low immunogenicity and can diffuse into cells. Aptamers already exist, such as the TR RNA sequence, that are capable of binding the MS2 coat protein. If the binding of such an aptamer was achieved *in vivo* the likely hood would be that the building block would not be incorporated into a capsid due to electrostatic repulsion and steric clashes.

It may also be possible to bind therapeutics such as antibodies to specific viral intermediates. This would block some of the pathways to full viral capsid. However as shown in chapter 6 there many be very many pathways to forming capsid. The kinetic analysis for the dodecahedron in chapter 5 also suggests that viral assembly would make use of these multiple pathways. However, for the MS2 virus it appears that certain individual pathways could be very much preferred when taking into account additional biological data. Therefore blocking this pathway could have a significant impact. For maximum effect in the more general case drugs could be designed to target intermediates early in assembly with relatively high stabilities for maximal effect. The quantitative analysis of removing intermediates and modelling possible therapeutic interventions is a question left for future work.

Finally, it is hoped that the productive nature of interdisciplinary research has been demonstrated throughout this thesis, and that the research presented here will be an inspiration for future self-assembly modelling and collaborations.

Abbreviations

BiRNA	Virus assembly initiating in the middle of the RNA that does not maintain a bound dimer to a specific location.
Cryo-EM	Cryo-Electron Microscopy
FM	Free Monomer
MD	Molecular Dynamics
STMV	Satellite Tobacco Mosaic Virus.
TR	Transcriptional Repressor, a stem loop in the MS2 RNA that binds a coat protein dimer and inhibits replicase production.
TrRNA	Virus assembly initiating in the middle of the RNA that maintains a bound dimer to a specific TR location.
UniRNA	Virus assembly initiating at the end of the RNA.

Bibliography

H Ackermann. *ICTVdB - The Universal Virus Database, version 4.* Management, Mailman School of Public Health, Columbia University, New York, NY, USA, 2006.

O M Becker and M Karplus. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *The Journal of Chemical Physics*, 106(4):1495–1517, 1997.

D L Casper and A Klug. Physical principles in the construction of regular viruses. *Cold Spring Harb Symp Quant Biol*, 27:1–24, 1962.

J H Chen and N C Seeman. Synthesis from DNA of a molecule with the connectivity of a cube. *Nature*, 350(6319):631–3, Apr 1991.

R Chhabra, J Sharma, Y Liu, S Rinker, and H Yan. DNA self-assembly for nanomedicine. *Advanced Drug Delivery Reviews*, 62(6):617 – 625, 2010. ISSN 0169-409X. From Biology to Materials: Engineering DNA and RNA for Drug Delivery and Nanomedicine.

F Chiti and C M Dobson. Amyloid formation by globular proteins under native conditions. *Nat Chem Biol*, 5(1):15–22, 2009.

A Condon, D Harel, J N. Kok, A Salomaa, and E Winfree, editors. *Algorithmic Bioprocesses.* Natural Computing. Springer, 2009.

M Cuillel, B Jacrot, and M Zulauf. A T = 1 capsid formed by protein of brome mosaic virus in the presence of trypsin. *Virology*, 110(1):63 – 72, 1981. ISSN 0042-6822.

M Cuillel, C Berthet-Colominas, B Krop, A Tardieu, P Vachette, and B Jacrot. Self-assembly of brome mosaic virus capsids: Kinetic study using neutron and X-ray solution scattering. *Journal of Molecular Biology*, 164(4):645 – 650, 1983. ISSN 0022-2836.

G Destito, P Singh, K J Koudelka, and M Manchester. Assembling viral nanoparticles for vascular imaging and tumor-specific targeting. In *Foundations of Nanoscience, Self-Assembled Architectures and Devices, Proceedings of FNANO07*, pages 2–4, 2007.

E C Dykeman and R Twarock. All-atom normal-mode analysis reveals an RNA-induced allostery in a bacteriophage coat protein. *Phys. Rev. E*, 81: 031908, Mar 2010.

E C Dykeman, N E Grayson, K Toropova, N A Ranson, P G Stockley, and R Twarock. Simple rules for efficient assembly predict the layout of a packaged viral RNA. *Journal of Molecular Biology*, 408(3):399 – 407, 2011. ISSN 0022-2836.

E C Dykeman, P G Stockley, and R Twarock. Identification of dispersed, cryptic packaging signals in two viral RNA genomes reveals a conserved assembly mechanism. (Awaiting publication). 2012.

O M Elrad and M F Hagan. Encapsulation of a polymer by an icosahedral virus. *CORD Conference Proceedings*, 7(4):45003–45034, 2010.

Oren M Elrad and Michael F Hagan. Mechanisms of size control and polymorphism in viral capsid assembly. *Nano Letters*, 8(11):3850–3857, 2008. PMID: 18950240.

K M ElSawy, L S D Caves, and R Twarock. The impact of viral RNA on the association rates of capsid protein assembly: Bacteriophage MS2 as a case study. *Journal of Molecular Biology*, 400(4):935 – 947, 2010. ISSN 0022-2836.

D Endres and A Zlotnick. Model-based analysis of assembly kinetics for virus capsids or other spherical polymers. *Biophysical Journal*, 83(2):1217 – 1230, 2002. ISSN 0006-3495.

D Endres, M Miyahara, P Moisant, and A Zlotnick. A reaction landscape identifies the intermediates critical for self-assembly of virus capsids and other polyhedral structures. *Protein Science*, 14(6):1518–1525, 2005. ISSN 1469-896X.

S N Fejer, T R James, J Hernandez-Rojas, and D J Wales. Energy landscapes for shells assembled from pentagonal and hexagonal pyramids. *Phys. Chem. Chem. Phys.*, 11:2098–2104, 2009.

W Fiers, R Contreras, F Duerinck, G Haegeman, D Iserentant, J Merregaert, W Min Jou, F Molemans, A Raeymaekers, A Van Den Berghe, G Volckaert, and M Ysebaert. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*, 260(5551): 500–507, 04 1976. 10.1038/260500a0.

E R Gansner, E Koutsofios, S C North, and K Vo. A technique for drawing directed graphs. *IEEE Transactions on software engineering*, 19(3):214–230, 1993.

GAP. *GAP – Groups, Algorithms, and Programming, Version 4.5.4.* The GAP Group, 2012.

K Gerasopoulos, M McCarthy, P Banerjee, X Fan, J N Culver, and R Ghodssi. Biofabrication methods for the patterned assembly and synthesis of viral nanotemplates. *Nanotechnology*, 21(5):055304, 2010.

D T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The-JournalofPhysicalChemistry*, 81(25):2340–2361, 1977.

R P Goodman, I A T Schaap, C F Tardin, C M Erben, R M Berry, C F Schmidt, and A J Turberfield. Rapid chiral assembly of rigid DNA building blocks for molecular nanofabrication. *Science*, 310(5754):1661–5, Dec 2005.

D N Gopaul, F Guo, and G D Van Duyne. Structure of the holliday junction intermediate in Cre-loxP site-specific recombination. *Biochemistry*, 17(14): 4175–4187, 1998.

N E Grayson, T Keef, S Severini, and R Twarock. Self-assembly of viral capsids via a hamiltonian paths approach: The case of bacteriophage MS2. In *Foundations of Nanoscience (FNANO)*, pages 183–186, 2007.

N E Grayson, A Taormina, and R Twarock. DNA duplex cage structures with icosahedral symmetry. *Theoretical Computer Science*, 410(15):1440 – 1447, 2009. ISSN 0304-3975. Aspects of Molecular Self-Assembly.

M F Hagan and D Chandler. Dynamic pathways for viral capsid assembly. *Biophys J*, 91(1):42–54, 2006.

Y He, T Ye, M Su, C Zhang, A E Ribbe, W Jiang, and C Mao. Hierarchical self-assembly of DNA into symmetric supramolecular polyhedra. *Nature*, 452(7184):198–201, 2008.

M Hemberg, S N Yaliraki, and M Barahona. Stochastic kinetics of viral capsid assembly based on detailed protein structures. *Biophysical Journal*, 90(9): 3029 – 3042, 2006. ISSN 0006-3495.

A C Hindmarsh. ODEPACK, a systematized collection of ODE solvers. *IMACS Transactions on Scientific Computation*, 1:55–64, 1983.

W T Horn, K Tars, E Grahn, C Helgstrand, A J Baron, H Lago, C J Adams, D S Peabody, S E V Phillips, N J Stonehouse, L Liljas, and P G Stockley. Structural basis of RNA binding discrimination between bacteriophages Qbeta and MS2. *Structure*, 14(3):487–95, Mar 2006.

K N Johnson, L Tang, J E Johnson, and L A Ball. Heterologous RNA encapsidated in pariacoto virus-like particles forms a dodecahedral cage similar to genomic RNA in wild-type virions. *J Virol*, 78(20):11371–8, Oct 2004.

I G Johnston, A A Louis, and J P K Doye. Modelling the self-assembly of virus capsids. *Journal of Physics: Condensed Matter*, 22(10):104101, 2010.

N Jonoska and M Saito. Boundary components of thickened graphs. In *DNA 7: Revised Papers from the 7th International Workshop on DNA-Based Computers*, pages 70–81, London, UK, 2002. Springer-Verlag. ISBN 3-540-43775-4.

N Jonoska and R Twarock. Blueprints for dodecahedral DNA cages. *Journal of Physics A: Mathematical and Theoretical*, 41(30):304043 (14pp), 2008.

S J Katen and A Zlotnick. The thermodynamics of virus capsid assembly. *Methods Enzymol.*, 455:395–417, 2009.

T Keef, A Taormina, and R Twarock. Assembly models for papovaviridae based on tiling theory. *Phys Biol*, 2(3):175–88, Sep 2005.

T Keef, C Micheletti, and R Twarock. Master equation approach to the assembly of viral capsids. *J Theor Biol*, 242(3):713–21, Oct 2006.

T Keef, R Twarock, and K M Elsawy. Blueprints for viral capsids in the family of polyomaviridae. *J Theor Biol*, 253(4):808–16, Aug 2008.

A Kivenson and M F Hagan. Mechanisms of capsid assembly around a polymer. *Biophys J*, 99(2):619–628, 2010.

T W Knapman, V L Morton, N J Stonehouse, P G Stockley, and A E Ashcroft. Determining the topology of virus assembly intermediates using ion mobility spectrometry–mass spectrometry. *Rapid Communications in Mass Spectrometry*, 24(20):3033–3042, 2010. ISSN 1097-0231.

R Koning, S Van Den Worm, J R Plaisier, J Van Duin, Jan Pieter A, and H Koerten. Visualization by cryo-electron microscopy of genomic RNA that binds to the protein capsid inside bacteriophage MS2. *J Mol Biol*, 332(2): 415–22, Sep 2003.

M S Kumar and R Schwartz. A parameter estimation technique for stochastic self-assembly systems and its application to human papillomavirus self-assembly. *Physical Biology*, 7(4):045005, 2010.

H Lago, A M Parrott, T Moss, N J Stonehouse, and P G Stockley. Probing the kinetics of formation of the bacteriophage MS2 translational operator complex: identification of a protein conformer unable to bind RNA. *Journal of Molecular Biology*, 305(5):1131–1144, 2 2001.

S B Larson and A McPherson. Satellite tobacco mosaic virus rna: structure and implications for assembly. *Current Opinion in Structural Biology*, 11 (1):59 – 65, 2001. ISSN 0959-440X.

S B Larson, J Day, A Greenwood, and A McPherson. Refined structure of satellite tobacco mosaic virus at 1.8 Å resolution. *Journal of Molecular Biology*, 277(1):37 – 59, 1998. ISSN 0022-2836.

C Levinthal. How to fold graciously. *In Mossbauer Spectros-copy in Biological Systems: Proceedings of a Meeting Held at Allerton House, Monticello, Illinois.*, page 22–24, 1969.

Yujie Ma, R J M Nolte, and J J L M Cornelissen. Virus-based nanocarriers for drug delivery. *Advanced Drug Delivery Reviews*, 64(9):811 – 825, 2012. ISSN 0169-409X. Approaches to drug delivery based on the principles of supramolecular chemistry.

A McPherson. Micelle formation and crystallization as paradigms for virus assembly. *BioEssays*, 27(4):447–458, 2005. ISSN 1521-1878.

P Moisant, H Neeman, and A Zlotnick. Exploring the paths of (virus) assembly. *Biophysical Journal*, 99(5):1350 – 1357, 2010. ISSN 0006-3495.

V L Morton, Dykeman E C, Stonehouse N J, Ashcroft A E, Twarock R, and Stockley P G. The impact of viral RNA on assembly pathway selection. *Journal of Molecular Biology*, 401(2):298 – 308, 2010. ISSN 0022-2836.

P Peterson. F2PY: a tool for connecting fortran and python programs. *Int. J. Comput. Sci. Eng.*, 4(4):296–305, November 2009. ISSN 1742-7185.

D C Rapaport. Self-assembly of polyhedral shells: a molecular dynamics study. *Phys Rev E Stat Nonlin Soft Matter Phys*, 70(5 Pt 1):51905–51905, 2004.

D C Rapaport. Molecular dynamics simulation of reversibly self-assembling shells in solution using trapezoidal particles. *CORD Conference Proceedings*, pages 1–13, 2012.

O Rolfsson, K Toropova, V Morton, S Francese, G Basnak, G S Thompson, S W Homans, A E Ashcroft, N J Stonehouse, N A Ranson, and P G Stockley. RNA packing specificity and folding during assembly of the bacteriophage MS2. *Computational and Mathematical Methods in Medicine*, 9(3):339 – 349, 2008.

J Rudnick and R Bruinsma. Icosahedral packing of RNA viral genomes. *Phys Rev Lett*, 94(3):038101, Jan 2005.

P Sa-Ardyen, N Jonoska, and N C Seeman. Self-assembly of irregular graphs whose edges are DNA helix axes. *J Am Chem Soc*, 126(21):6648–57, Jun 2004.

R Schwartz, P W Shor, P E Prevelige, and B Berger. Local rules simulation of the kinetics of virus capsid self-assembly. *Biophys J*, 75(6):2626–2636, 1998.

T Shiba and Y Suzuki. Localization of a protein in the RNA-A protein complex of RNA phage MS2. *Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis*, 654(2):249 – 255, 1981. ISSN 0005-2787.

S Shigdar, A C Ward, A De, C J Yang, M Wei, and W Duan. Clinical applications of aptamers and nucleic acid therapeutics in haematological malignancies. *British Journal of Haematology*, 155(1):3–13, 2011. ISSN 1365-2141. doi: 10.1111/j.1365-2141.2011.08807.x. URL http://dx.doi.org/10.1111/j.1365-2141.2011.08807.x.

W M Shih, J D Quispe, and G F Joyce. A 1.7-kilobase single-stranded DNA that folds into a nanoscale octahedron. *Nature*, 427(6975):618–21, Feb 2004.

T Shtatland, S C Gill, B E Javornik, H E Johansson, B S Singer, O C Uhlenbeck, D A Zichi, and L Gold. Interactions of escherichia coli RNA with bacteriophage MS2 coat protein: genomic SELEX. *Nucleic acids research*, 28:21:e93, 2000.

S Singh and A Zlotnick. Observed hysteresis of virus capsid disassembly is implicit in kinetic models of assembly. *J. Biol. Chem.*, 278(20):18249–18255, May 2003.

P G Stockley, O Rolfsson, G S Thompson, G Basnak, S Francese, N J Stonehouse, S W Homans, and A E Ashcroft. A simple, RNA-mediated allosteric switch controls the pathway to formation of a T=3 viral capsid. *J Mol Biol*, 369(2):541–52, Jun 2007.

L Tang, K N Johnson, L A Ball, T Lin, M Yeager, and J E Johnson. The structure of pariacoto virus reveals a dodecahedral cage of duplex RNA. *Nat Struct Biol*, 8(1):77–83, Jan 2001.

K Toropova, G Basnak, R Twarock, P G Stockley, and N A Ranson. The three-dimensional structure of genomic RNA in bacteriophage MS2: implications for assembly. *J Mol Biol*, 375(3):824–36, Jan 2008.

K Toropova, P G Stockley, and Ranson N A. Visualising a viral RNA genome poised for release from its receptor complex. *J Mol Biol*, 6(408):408–19, 2011.

K Valegård, J B Murray, N J Stonehouse, S Van Den Worm, P G Stockley, and L Liljas. The three-dimensional structures of two complexes between recombinant MS2 capsids and RNA operator fragments reveal sequence-specific protein-RNA interactions. *Journal of Molecular Biology*, 270(5):724 – 738, 1997. ISSN 0022-2836.

S H E Van Den Worm, R I Koning, H J Warmenhoven, H K Koerten, and J Van Duin. Cryo electron microscopy reconstructions of the leviviridae unveil the densest icosahedral RNA packing possible. *J Mol Biol*, 363(4): 858–65, Nov 2006.

D J Wales. Closed-shell structures and the building game. *Chemical Physics Letters*, 141(6):478 – 484, 1987. ISSN 0009-2614.

D J Wales. The energy landscape as a unifying theme in molecular science. *Phil. Trans. R. Soc. A*, 363:357–377, 2005.

T Zhang and R Schwartz. Simulation study of the contribution of oligomer/oligomer binding to capsid assembly kinetics. *Biophys J*, 90(1): 57–64, 2006.

Y Zhang and N C Seeman. Construction of a DNA-truncated octahedron. *Journal of the American Chemical Society*, 116(5):1661–1669, 1994.

A Zlotnick. To build a virus capsid. an equilibrium model of the self assembly of polyhedral protein complexes. *J Mol Biol*, 241(1):59–67, Aug 1994.

A Zlotnick. Are weak protein-protein interactions the general rule in capsid assembly? *Virology*, 315(2):269 – 274, 2003. ISSN 0042-6822.

A Zlotnick. Simple models and simple analyses of virus capsid assembly. *Journal of Theoretical Medicine*, 6(2):111–114, 2005a.

A Zlotnick. Theoretical aspects of virus capsid assembly. *Journal of Molecular Recognition*, 18(6):479–490, 2005b. ISSN 1099-1352.

A Zlotnick. Distinguishing reversible from irreversible virus capsid assembly. *Journal of Molecular Biology*, 366(1):14 – 18, 2007. ISSN 0022-2836.

A Zlotnick and S Mukhopadhyay. Virus assembly, allostery and antivirals. *Trends in Microbiology*, 19(1):14 – 23, 2011. ISSN 0966-842X.

A Zlotnick and S J Stray. How does your virus grow? Understanding and interfering with virus assembly. *Trends in Biotechnology*, 21(12):536 – 542, 2003. ISSN 0167-7799.

A Zlotnick, J M Johnson, P W Wingfield, S J Stahl, and D Endres. A theoretical model successfully identifies features of hepatitis B virus capsid assembly. *Biochemistry*, 38(44):14644–14652, 1999.