

**MODELING AND SIMULATION OF SINGLE STRANDED RNA
VIRUSES**

A Dissertation
Presented to
The Academic Faculty

by

Mustafa Burak Boz

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Chemistry and Biochemistry

Georgia Institute of Technology
August 2012

MODELING AND SIMULATION OF SINGLE STRANDED RNA VIRUSES

Approved by:

Dr. Stephen C. Harvey, Advisor
School of Biology
Georgia Institute of Technology

Dr. Roger Wartell
School of Biology
Georgia Institute of Technology

Dr. Rigoberto Hernandez
School of Chemistry & Biochemistry
Georgia Institute of Technology

Dr. Loren Willams
School of Chemistry & Biochemistry
Georgia Institute of Technology

Dr. Adegboyega Oyelere
School of Chemistry & Biochemistry
Georgia Institute of Technology

Date Approved: June 18, 2012

Dedicated to my parents.

ACKNOWLEDGEMENTS

I would like to thank my family for their incredible support and patience. I would not have been here without them. I especially would like to give my gratitude to my father who has been the most visionary person in my life leading me to towards my goals and dreams.

I would also like to thank to Dr. Harvey for being my wise and sophisticated advisor. I am also grateful to the all Harvey Lab members I have known during my Ph. D years, (Batsal Devkota, Anton Petrov, Robert K.Z. Tan, Geoff Rollins, Amanda McCook, Andrew Douglas Huang, Kanika Arora, Mimmin Pan, Thanawadee (Bee) Preeprem, John Jared Gossett, Kazi Shefaet Rahman) for their valuable discussions and supports.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	iv
LIST OF TABLES	vii
LIST OF FIGURES	viii
LIST OF SYMBOLS AND ABBREVIATIONS	x
SUMMARY	xi
<u>CHAPTER</u>	
1 Introduction	1
Icosahedral Symmetry	1
RNA Structure	4
Assembly of Icosahedral Viruses	10
Computational Studies	11
References	15
2 Computational Approaches to Modeling Viral Structure and Assembly	20
Introduction	20
Methods	21
Results	35
Discussion	45
Reference	60
3 Structural and Electrostatic Characterization of Pariacoto Virus: Implications for Assembly	65
Introduction	65
Methods	68

Results and Discussions	78
References	84
4 Capsid Assembly Simulations	88
Introduction	88
Methods	90
Results	93
Discussion	96
References	97
5 Assembly of T=1 Virus using Coarse-grained Models	98
Introduction	99
Methods	102
Results	107
Discussion	113
References	114
6 Conclusion and Future Work	118
YUP scripts	118
Modeling PaV	118
Capsid Simulations	119
Virus Simulations	120
APPENDIX A: Pre-equilibrated RNAs	123
APPENDIX B: YUP scripts	124
VITA	141

LIST OF TABLES

	Page
Table 2.1: Force constant	50
Table 2.2: Missing protein residues	52
Table 2.3: Energy terms	56
Table 3.1: Energy terms used for RNA	73
Table 5.1: Parameter sets for stability	107
Table 5.2: Parameter sets for assembly	110

LIST OF FIGURES

	Page
Figure 1.1: Icosahedron	1
Figure 1.2: Icosahedral capsids	2
Figure 1.3: h and k vectors on a hexameric plane	3
Figure 1.4: Icosahedral viruses	4
Figure 1.5: STMV crystal structure	5
Figure 1.6: Icosahedral ssRNA viruses	6
Figure 1.7: Secondary structure of STMV	8
Figure 1.8: CCMV RNA2 in different solutions	9
Figure 1.9: Capsid models	12
Figure 1.10: Assembly model	14
Figure 2.1: DNA models	22
Figure 2.2: The capsid model for epsilon 15	28
Figure 2.3: Ejection simulation of bacteriophage ϕ 29	32
Figure 2.4: Models of tRNA	39
Figure 2.5: RNA secondary structure conversion to 3D model	43
Figure 2.6: Minimization protocol of Pariacoto virus (PaV) RNA	46
Figure 2.7: Minimization protocol of PaV capsid protein	53
Figure 2.8: Final all-atom model of PaV	58
Figure 3.1: Secondary structure map for PaV	68
Figure 3.2: Stereo images of model junctions	70
Figure 3.3: Minimization protocol of PaV RNA	71
Figure 3.4: A 20 Å slice through of the center of PaV model_8	74

Figure 3.5: Comparison of model radial density distributions	79
Figure 3.6: Electrostatic potential mapped on the PaV surface	80
Figure 3.7: Model for assembly of icosahedral viruses	81
Figure 4.1: T=1 capsid models and capsid units	89
Figure 4.2: Capsid unit model	90
Figure 4.3: Potential variation of the capsid unit model	91
Figure 4.4: 20 super-imposed capsid units with different edge angles	91
Figure 4.5: Assembled T=1 capsid model	92
Figure 4.6: Results of edge angle and potential variations	94
Figure 4.7: Snapshots of edge angle and potential variation simulations	95
Figure 5.1: Proposed pathway for assembly	99
Figure 5.2: Capsid unit model	101
Figure 5.3: P model of STMV RNA	104
Figure 5.4: Energetic of the stability simulations	109
Figure 5.5: Snapshots of the stability simulations	109
Figure 5.6: Shapshots of the assembly simulations	111
Figure 5.7: Assembly of T=1 virus on two protocols	112
Figure A.1: Pre-equilibrated RNAs	122

LIST OF SYMBOLS AND ABBREVIATIONS

ATP	Adenosine Triphosphate
BPMV	Bean Pod Mottle Virus
CCMV	Cowpea Chlorotic Mottle Virus
CMCT	1-cyclohexyl-(2-morpholinoethyl)carbodiimide metho-p-toluene sulfonate
CU	Capsid Unit
DH	Debye-Hückle
DMS	Dimethyl Sulfate
DNA	Deoxyribonucleic Acid
LJ	Lennard-Jones
PaV	Pariacoto Virus
PT	Protein Tail
RNA	Ribonucleic Acid
ssRNA	Single-stranded Ribonucleic Acid
STMV	Satellite Tobacco Mosaic Virus
T-number	Triangulation Number
VLP	Virus Like Particle
YAMMP	Yet Another Molecular Mechanics Program
YUP	YAMMP Under Python

SUMMARY

The presented work is the application of recent methodologies on modeling and simulation of single stranded RNA viruses. We first present the methods of modeling RNA molecules using the coarse-grained modeling package, YUP. Coarse-grained models simplify complex structures such as viruses and let us study general behavior of the complex biological systems that otherwise cannot be studied with all-atom details.

Second, we modeled the first all-atom T=3, icosahedral, single stranded RNA virus, Pariaquito virus (PaV). The x-ray structure of PaV shows only 35% of the total RNA genome and 88% of the capsid. We modeled both missing portions of RNA and protein. The final model of the PaV demonstrated that the positively charged protein N-terminus was located deep inside the RNA. We propose that the positively charged N-terminal tails make contact with the RNA genome and neutralize the negative charges in RNA and subsequently collapse the RNA/protein complex into an icosahedral virus.

Third, we simulated T=1 empty capsids using a coarse-grained model of three capsid proteins as a wedge-shaped triangular capsid unit. We varied the edge angle and the potentials of the capsid units to perform empty capsid assembly simulations. The final model and the potential are further improved for the whole virus assembly simulations.

Finally, we performed stability and assembly simulations of the whole virus using coarse-grained models. We tested various strengths of RNA-protein tail and capsid protein-capsid protein attractions in our stability simulations and narrowed our search for optimal potentials for assembly. The assembly simulations were carried out with two different protocols: co-transcriptional and post-transcriptional. The co-transcriptional assembly protocol mimics the assembly occurring during the replication of the new RNA.

Proteins bind the partly transcribed RNA in this protocol. The post-transcriptional assembly protocol assumes that the RNA is completely transcribed in the absence of proteins. Proteins later bind to the fully transcribed RNA. We found that both protocols can assemble viruses, when the RNA structure is compact enough to yield a successful virus particle. The post-transcriptional protocol depends more on the compactness of the RNA structure compared to the co-transcriptional assembly protocol. Viruses can exploit both assembly protocols based on the location of RNA replication and the compactness of the final structure of the RNA.

CHAPTER 1

INTRODUCTION

Viruses are very diverse particles in structure; however, they can be categorized into four branches: helical, icosahedral, enveloped, and complex. We are interested in icosahedral single-stranded RNA (ssRNA) viruses. These viruses contain RNA as the genome and a protein capsid encapsulating the genome in an icosahedral symmetry.

Icosahedral Symmetry

An icosahedron (Figure 1.1) has 30 edges, 12 vertices and 20 faces. There are three symmetry axes: 2-fold axis on the edges, 3-fold axis on the faces, 5-fold axis on the vertices. An icosahedron is the dual partner of a dodecahedron. The consequence of that is each face and vertex in an icosahedron corresponds to a vertex and a face in a dodecahedron, respectively. This feature allows an icosahedron to be placed inside a dodecahedron, and visa versa. They are complementary platonic solids.

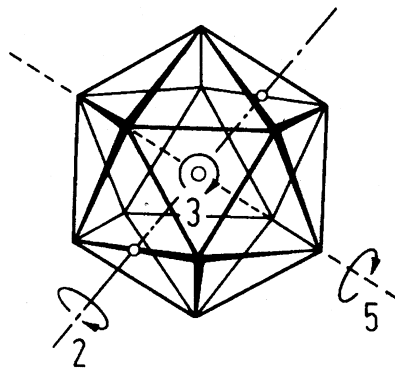


Figure 1.1: Icosahedron. Three symmetry axes; 2, 3, and 5 folds are shown. Reprinted from <http://viprdb.scripps.edu>.

Icosahedral virus structure has been classified by Caspar and Klug [1]. They proposed the quasi-equivalence theory to account for the protein arrangement on the capsid. They found out that 60 asymmetric units are enough to assemble an icosahedral

virus. These asymmetric units form only pentamers for the smallest virus structure and both pentamers and hexamers for bigger capsids. The pentamers form the vertices and the hexamers fill the rest of the capsid structure (faces and edges) (Figure 1.2).

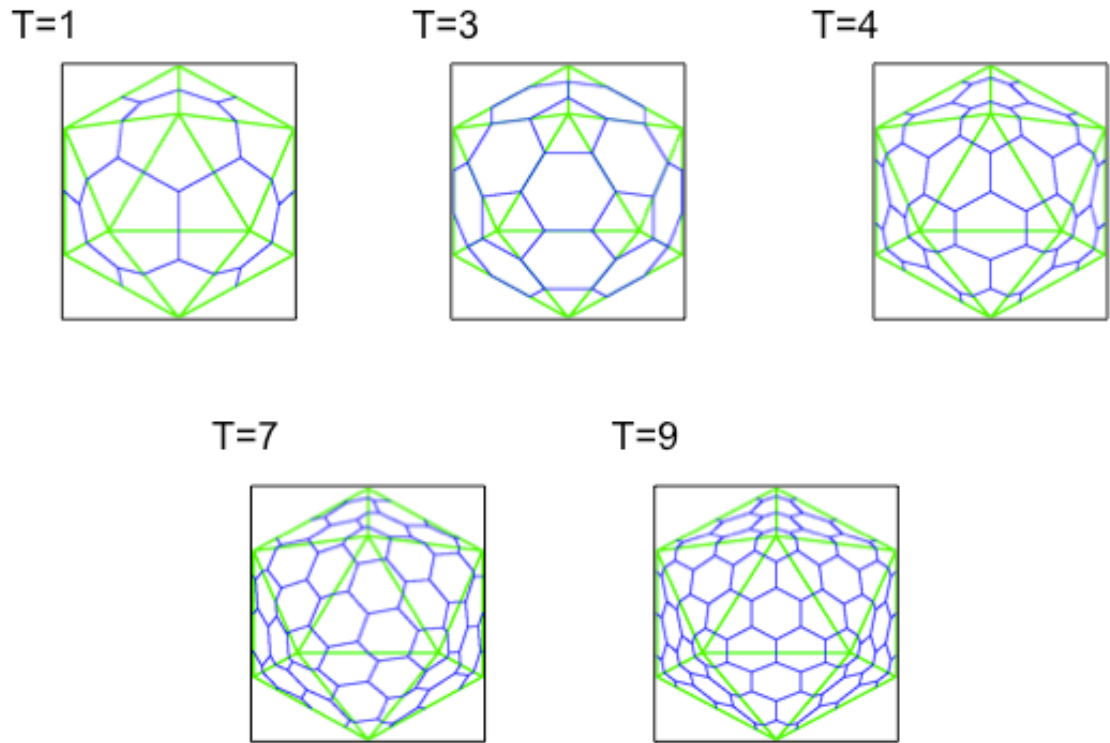


Figure 1.2: Icosahedral capsids with various T-numbers. Sizes are not proportional. Reprinted from www.viperdb.scripps.edu.

Icosahedral viruses are classified by the T-number corresponding to the number of proteins forming the asymmetric unit. T-number can also be defined as the following equation.

$$T=k^2+hk+h^2$$

Where h and k are two axes with a 60° separation on a hexameric plane (Figure 1.3). The values of h and k determine the length of a face of the icosahedral structure, therefore the size of the capsid. When neither h nor k is zero and h is not equal to k, right

or left handed capsids form i.e. pair of (2,1) and (1,2) yield $T=7d$ (right handed) and $T=7l$ (left handed) capsid structures where they differ by handedness.

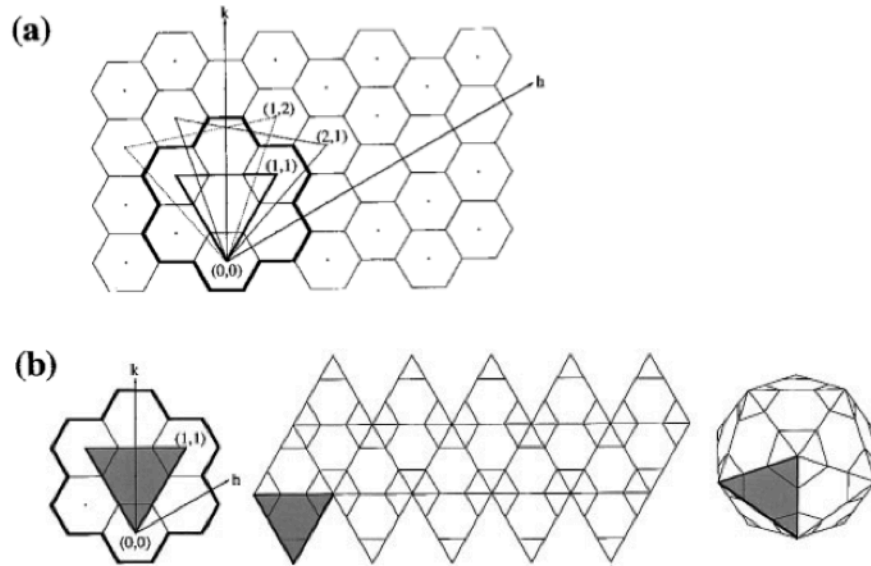


Figure 1.3: h and k axes on a hexameric plane (a). $T=3$ capsid generation using (1,1) h and k values (b). Reprinted from [2].

Structural features of an icosahedral virus can be calculated by knowing the T-number.

$$\text{Pentamers} = 12$$

$$\text{Hexamers} = 10 \times (T-1)$$

$$\text{Total number of proteins} = 60 \times T$$

The $T=1$ icosahedral virus capsid contains 60 copies of one protein with identical sequence and conformation. However, the $T=3$ icosahedral virus contains one protein with 3 slightly different conformations. As the T-number goes up, the number of proteins and the number of different conformations increases. This is due to curvature of the virus capsid. T-number also defines the number of protein conformations in a given

icosahedral virus. Figure 1.4 shows several icosahedral viruses with various T-numbers and the triangle of the icosahedral symmetry. For example, adenovirus is a T=25 icosahedral virus. It has 25 different conformational changes among proteins on the capsid.

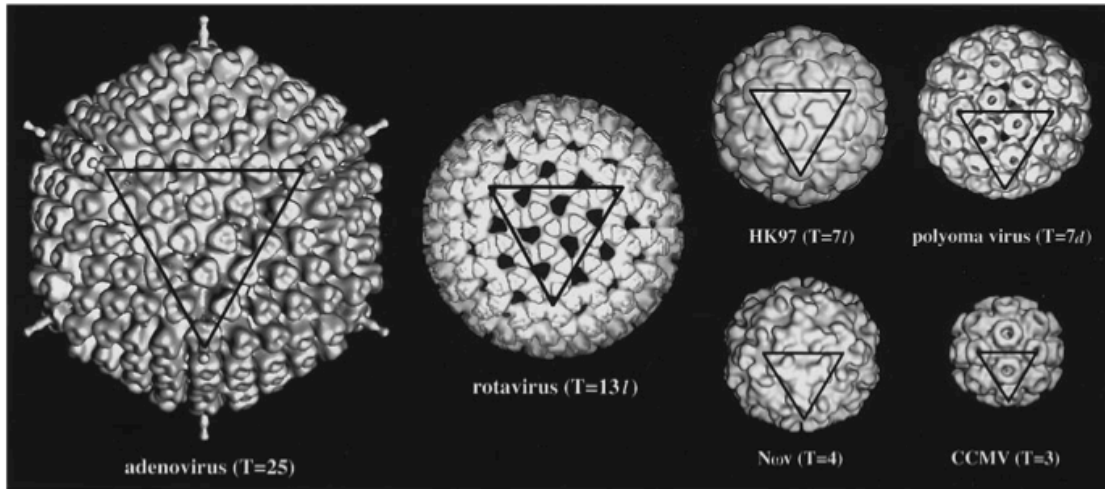


Figure 1.4: Icosahedral virus capsids with one face of the icosahedral symmetry shown. Reprinted from [2].

RNA Structure

Viruses contain either RNA or DNA as their genome. The presence of both nucleic acids as the genome has never been observed. We are interested in the ssRNA viruses. Being single-stranded rather than double-stranded changes the structure, the stability, and the assembly of both the genome and the virus. ssRNAs have many different secondary structures. The secondary structure is also dynamic based on the life cycle of the virus.

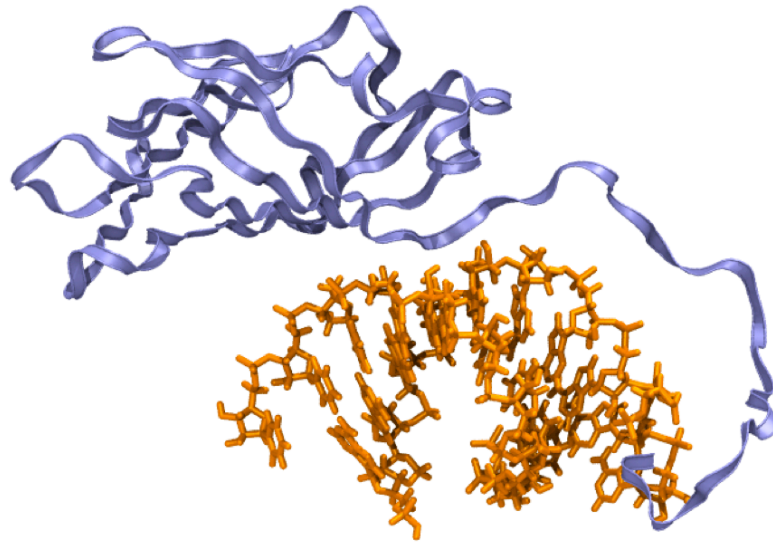


Figure 1.5: STMV crystal structure. Capsid protein is represented with purple ribbon and RNA duplex is shown in orange. [3]

Crystal structures of ssRNA viruses show a small portion of the genome and most of the capsid protein. There are several reasons for the structures of the ssRNA being not fully visible. One of the reasons is that the virus structures are icosahedrally averaged due to icosahedral symmetry of the capsid proteins. The RNA doesn't have icosahedral symmetry. Another reason is that there are flexible regions in both RNA and the protein that are completely invisible. In addition, there may not be a unique structure of the packaged genome inside the virus. It is also known that viruses mutate at a fast pace in order to evade degradation by host cell. These mutations might also change the structure of the genome.

Satellite Tobacco Mosaic Virus (STMV) ($T=1$) crystal structure presents 9 basepairs (Figure 1.5). It lies on the edge of the icosahedral capsid structure (Figure 1.6). All 60 copies of the RNA duplex represent 59% of the total genome and the rest is not visible. In Bean Pod Mottle Virus (BPMV) ($T=3$), the visible RNA lies on the face of the icosahedral capsid structure and over all visible RNA make up 20% of the whole

genome. Last, Pariacoto Virus (PaV) ($T=3$) structure has 25 basepairs on each edge of the icosahedral capsid structure and only 35% of the RNA genome is visible. The details of these viruses and more are reviewed by Anette Schneemann [4].

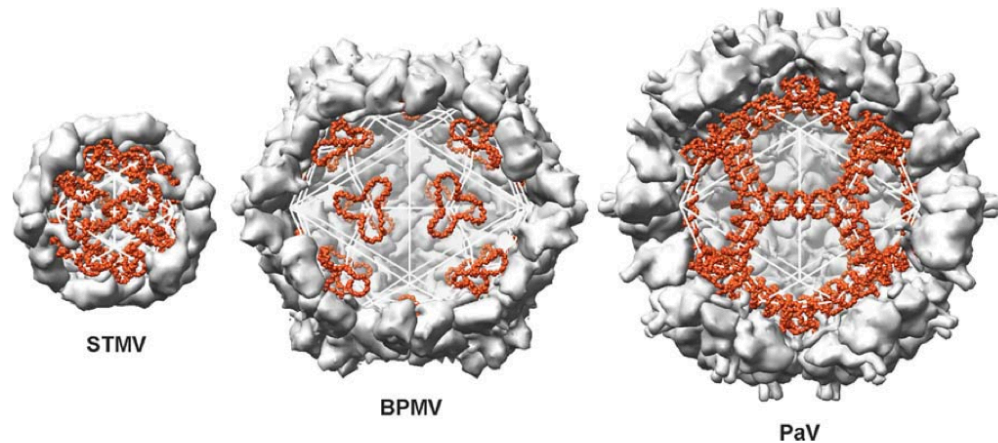


Figure 1.6: Icosahedral ssRNA viruses with crystallographically visible RNA structures. Reprinted from [4].

Secondary structure predictions suggest that some but not all viral RNAs tend to be more highly branched than shuffled sequences with the same composition, or with non-viral sequences of the same size, favoring compact three-dimensional structures compatible with viral assembly [6]. Yoffe and coworkers have pointed out that, since the RNA genomes are very densely packed in mature viruses, there could be a substantial advantage to sequences that favor secondary structures that are compact in three-dimensional space. To test this hypothesis, they predicted the secondary structures of viral RNAs, nonviral RNAs, and shuffled RNA sequences with the same composition as viral RNAs and calculated the maximum ladder distance for each secondary structure. The ladder distance between any two nucleotides in a secondary structure can be calculated by drawing the structure in the standard two-dimensional form and treating

each of the double-helical regions as a “ladder”, where the base pair lines are the rungs [7]. The maximum ladder distance is that of the longest direct path across the secondary structure and serves as a proxy for the extendedness of the molecule.

RNA secondary structure prediction programs are still struggling with predicting the correct secondary structure. Best-case scenarios on ribosomal RNAs are 70% correct. However, addition of experimental data improves the prediction. The most recent STMV secondary structure is proposed by Schroeder [8]. They chemically probed the RNA using CMCT, DMS and kethoxal in the intact virion. They chose a window size of 30 nucleotides to fold the RNA into local stem loops. They also allowed symmetric internal loops in the stems. They constructed an ensemble of secondary structures fitting the chemical data. They suggested the following best secondary structure among the ensemble of similar structures (Figure 1.7).

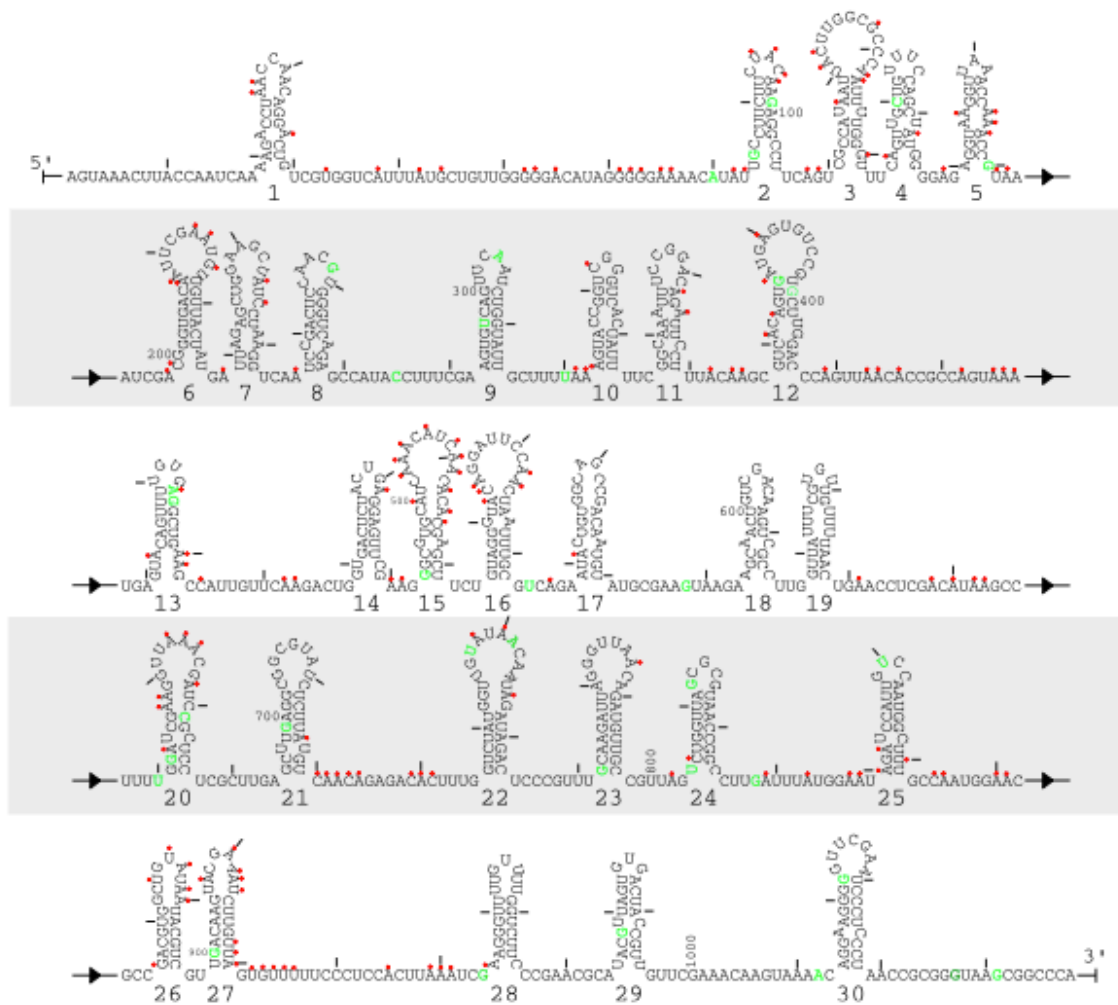


Figure 1.7: The proposed secondary structure of STMV. Red dots are the hit dots where the chemical data suggest these nucleotides are single stranded. The green nucleotides show the sequence variation in the STMV genome. Reprinted from [5].

Cations are also important in determination of the RNA structure. Gelbart and his collaborators [9] have recently shown that there is a very strong correlation between Mg^{2+} concentration and the radius of gyration of the RNA. They studied three long RNA molecules with 975, 1523 and 2777 nucleotides, respectively, from two non-coding sequences of the yeast and the CCMV RNA2. They observed flat prolate conformations with various branching due to electrostatic repulsion and coaxial stacking. In the presence of Mg^{2+} , the coplanar prolate conformations collapse into concave structures due to

tertiary interactions (salt bridges). The effect of Mg^{2+} on condensation is extensively studied [10-14].

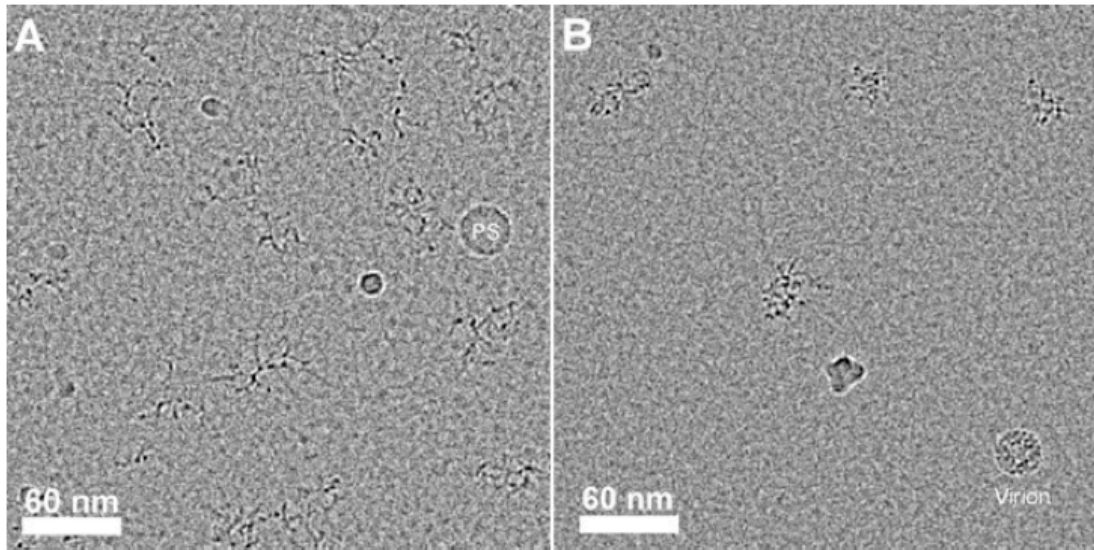


Figure 1.8: CCMV RNA2 genome in two different solutions. (A) The solution is Mg^{2+} free and the RNA is very branched. (B) The solution contains 5 mM Mg^{2+} (assembly buffer) and the RNA has more compact conformation. Polystyrene (PS) bead (30 nm in diameter) and the virion (28 nm in diameter) are placed for internal size comparison. Reprinted from [9].

Chaperones that can influence RNA structure have also been known for a number of years [15], and some viral proteins have proven RNA chaperone activities [16,17]. It appears likely that protein-binding plays a substantial role in the formation of specific genome structures required for the formation of some mature RNA viruses, at least in some cases.

Assembly of icosahedral viruses

The assembly of small, non-enveloped icosahedral RNA viruses does not require the hydrolysis of ATP, but isolated capsid proteins do not aggregate to form capsids under normal conditions; assembly is spontaneous, but it requires the simultaneous

presence of the genome and capsid proteins. The major challenges, then, are to explain how the relatively weak inter-protein forces become sufficient to promote assembly in the presence of the genome, to define the roles of RNA secondary structure and tertiary structure in assembly, and to determine how all these factors are integrated into the formation of the mature virion.

Cowpea Chlorotic Mottle Virus (CCMV) was the first ssRNA virus for which it was shown that the virus could be assembled *in vitro* and was still infectious [18]. CCMV can be assembled both as an empty capsid and with the RNA *in vitro*. The empty capsids are assembled at lower pH conditions. Virions are assembled at physiological conditions. The assembly was guided by the capsid protein interacting with the genome through the positively charged N terminus.

The biophysical studies done by McPherson and his colleagues [19] on the RNA cores of STMV has illustrated that even after the protein capsid is degraded with proteases, the positively charged tails stay with the RNA core and the RNA core remains in its compact form (10nm diameter) up to 12 to 24 hours. This suggests the importance of these positively charged residues on the stability and the assembly of STMV.

In addition, the tails' charge density seems to be conserved over the ssRNA and ssDNA viruses. Belyi and Muthukumar [20] reported that the ratio of total phosphate charge to the net charge on the protein tails is 1.61 ± 0.03 in 16 wild-type and three mutant ssRNA and ssDNA viruses whose genomes ranged from roughly 1 kb to 12 kb.

The crystal structure of STMV led to the proposal that the RNA genome is folded into a structure with many local stem-loops. Larson and McPherson [5] later suggested that co-transcriptional assembly could facilitate the formation of these structures, because

protein binding would inhibit the unfolding of hairpins and refolding into structures with long-range base pairs. This cannot be true for all small RNA viruses, however, because the formation of mature nodaviruses is delayed for about 30 minutes after replication is complete [21]. This suggests post-transcriptional assembly for nodaviruses.

Small RNA viruses can form virus-like particles (VLPs) around RNAs other than genomic RNAs [22]. These VLPs can have different sizes and morphologies than the native virus, and they can even be formed by the condensation of viral proteins around cargoes other than RNA [23-28]. In such cases, the size of the cargo influences the curvature of the capsid, thereby controlling the final size of the VLP [23,26,27] .

Computational Studies:

In addition to the experimental studies mentioned, computational studies are also important for understanding the structure and the assembly of icosahedral viruses. Computational studies focus on the stability and the assembly of both the empty capsid and the whole virus [29-34].

One of the simplest, yet important empty capsid simulations questions the source of icosahedral symmetry of the viruses [29]. Zandi *et. al.* performed Monte Carlo simulations of capsomers representing the capsid pentamers and hexamers on the fixed surface (2D) of a sphere demonstrating that the icosahedral symmetry comes from the minimum energy configurations of the hexamers and pentamers.

Later simulations are performed in three dimensions using more detailed coarse-grained models [30-34]. For instance, a capsid unit with three proteins was represented with 28 pseudo-atoms instead of five proteins represented with 1 pseudo-atom as in the

Zandi's study. Most of these simulations contained simple attractive potentials in the form of vectors or Lennard-Jones (LJ) (Figure 1.9).

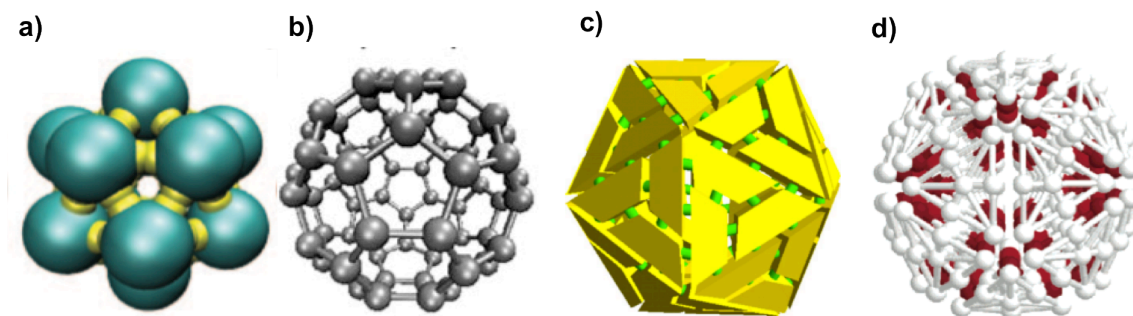


Figure 1.9: (a) Capsid is made of capsomers with five patchy particles [32]. (b) The capsomers forming the capsid have three vectors that are attracted to each other [30]. (c) Trapezoidal capsid units having five attractive LJ particles on the edges form the capsid [33]. (d) Wedge-shaped capsid unit is formed of 4 planes with 7 atoms on each plane. The capsid is formed using the two attractive LJ particles (red) on the corner of the capsid unit. Reprinted from [34].

These simulations indicate the importance of the protein concentration on the capsid assembly and demonstrate that there are many ways and models to assemble a T=1 capsid models. Brooks and his colleagues also assembled a T=3 empty capsid using specific attraction potentials by introducing three different capsid units [35].

There are also other sets of computational studies [36-41] focusing on the stability of the ssRNA viruses via the electrostatic nature of the virus. Most of these studies are coarse-grained simulations of RNA in a fixed capsid and finding the distribution of the RNA compared to capsid proteins. There is only one all-atom simulation of stability of T=1 virus (STMV) by Arkhipov *et.al* [39]. The model misses 110 RNA nucleotides from the structure and 720 capsid protein residues. The missing protein residues are positively charged and provide stability to the virus. The lack of the stability due to missing protein residues is compensated with addition of extra Mg^{2+} cations.

Hagan studied the assembly of virus-like particles in which capsid proteins from RNA viruses are used to encapsidate a charged cargo using coarse-grained models [42]. He modeled experiments in which bromine mosaic virus proteins encapsidate gold nanoparticles whose surfaces are covered by thiolalkylated tetraethylene glycol chains, some of which were terminated with carboxylate groups [28]. The charges on these groups were neutralized by positively charged model protein tails. Capsid proteins first condense on the gold nanoparticles, and later rearrange themselves to form icosahedral structures.

Hagan has recently published [43] coarse-grained assembly simulations of MS2 bacteriophage. As a genome, he used a single-strand polymer model attracted to the capsid units. He assembled a virus like particle with only 300 residues of the polymer. The repulsive nature of the electrostatics between polymer residues is completely ignored.

We have proposed a simple mechanism for the assembly of small icosahedral RNA viruses including the electrostatic nature of these interactions [38]. As seen in Figure 1.10a, we suggest that the positively charged protein tails bind to the RNA, leading to neutralization of a large fraction of the RNA charge and the collapse of the RNA-protein complex in a process reminiscent of DNA condensation. This squeezes the core domains of the capsid proteins into a shell on the outside of the condensed state (Figure 1.10b), leading to a sufficiently high local concentration that the capsid proteins can oligomerize to form the mature capsid (Figure 1.10c), in spite of the relatively weak protein–protein affinity.

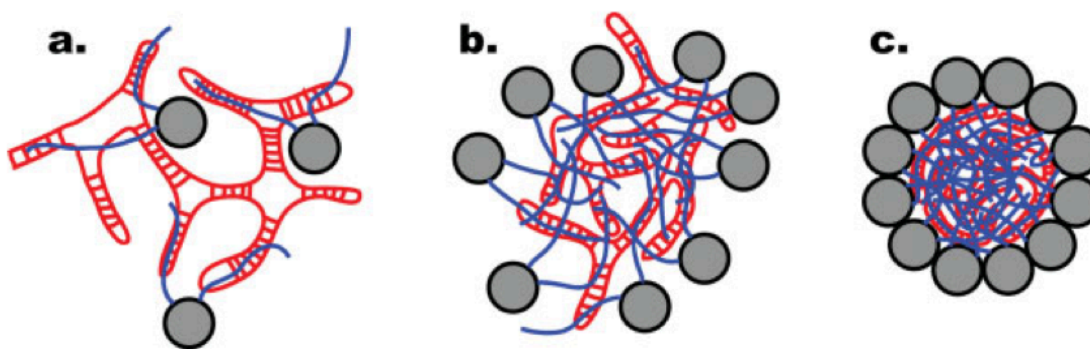


Figure 1.10: Proposed pathway for the assembly of small icosahedral RNA viruses. The positively charged protein tails (blue) bind non-specifically to RNA through electrostatic interactions. (b) When a sufficient fraction of the RNA charge has been neutralized by the polycationic protein tails, the complex of RNA plus protein tails collapses, following a pathway similar to that of the condensation of DNA by polyvalent cations. The protein/RNA condensate is dense enough to exclude the proteins' globular domains (grey), and these are concentrated in a shell around the condensate. When their concentration in the shell is sufficiently high, the weak inter-protein attractive forces are strong enough to lead to the formation of the mature capsid (c). Reprinted from [38].

This model exploits the fact that RNA replication, protein synthesis and RNA–protein binding occur close in time and space [44–46]. We hypothesize that most of the RNA–protein interactions are nonspecific. The initial collapse requires partial charge neutralization to overcome RNA–RNA repulsions. At the same time, neutralization should not be so extensive that the condensed state is locked into a rigid conformation, because final assembly of the mature capsid structure requires the globular domains of the proteins to retain some mobility.

References:

1. Caspar DLD, Klug A (1962) Physical Principles in the Construction of Regular Viruses. *Cold Spring Harbor Symposia on Quantitative Biology* 27:1-24.
2. Johnson JE, Speir J (1997) Quasi-equivalent viruses: a paradigm for protein assemblies. *Journal of molecular biology* 269:665-75.

3. Larson SB, Day J, Greenwood A, McPherson A (1998) Refined structure of satellite tobacco mosaic virus at 1.8 Å resolution. *Journal of molecular biology* 277:37-59.
4. Schneemann A (2006) The structural and functional role of RNA in icosahedral virus assembly. *Annual review of microbiology* 60:51-67.
5. Larson SB, McPherson A (2001) Satellite tobacco mosaic virus RNA: structure and implications for assembly. *Current opinion in structural biology* 11:59-65.
6. Yoffe AM, Prinsen P, Gopal A, Knobler CM, Gelbart WM and Ben-Shaul A (2008) Predicting the sizes of large RNA molecules, *Proc. Natl. Acad. Sci. U. S. A.*, 105(42):16153–8
7. Bundschuh R and Hwa T, (2002) Statistical mechanics of secondary structures formed by random RNA sequences, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 65(3), 031903
8. Schroeder SJ, Stone JW, Bleckley S, Gibbons T, Mathews DM (2011) Ensemble of secondary structures for encapsidated satellite tobacco mosaic virus RNA consistent with chemical probing and crystallography constraints. *Biophysical journal* 101:167-75.
9. Gopal A, Zhou ZH, Knobler CM, Gelbart WM (2012) Visualizing large RNA molecules in solution. *RNA (New York, N.Y.)* 18:284-99.
10. Buchmueller KL, Webb AE, Richardson DA, Weeks KM. (2000). A collapsed non-native RNA folding state. *Nat Struct Biol* 7: 362– 366.
11. Ribitsch G, Clercq RD, Folkhard W, Zipper P, Schurz J, Clauwaert J. (1985). Small-angle X-ray and light scattering studies on the influence of Mg²⁺ ions on the structure of the RNA from bacteriophage MS2. *Z Naturforsch C* 40: 234–241.
12. Caliskan G, Hyeon C, Perez-Salas U, Briber RM, Woodson SA, Thirumalai D. (2005). Persistence length changes dramatically as RNA folds. *Phys Rev Lett* 95: 268303–268307.
13. Chu VB, Bai Y, Lipfert J, Herschlag D, Doniach S. (2008). A repulsive field: Advances in the electrostatics of the ion atmosphere. *Curr Opin Chem Biol* 12: 619–625.
14. Draper DE. (2004). A guide to ions and RNA structure. *RNA* 10: 335– 343.
15. Herschlag D. (1995) RNA chaperones and the RNA folding problem, *J. Biol. Chem.*, 270(36), 20871–4.

16. Huang ZS and Wu HN. (1998) Identification and characterization of the RNA chaperone activity of hepatitis delta antigen peptides, *J. Biol. Chem.*, 273(41), 26455–61.
17. Rein A, Henderson LE and Levin GJ, (1998) Nucleic-acid-chaperone activity of retroviral nucleocapsid proteins: significance for viral replication, *Trends Biochem. Sci.*, 23(8), 297–301.
18. Bancroft JB, Hills GJ, Markham R (1967) A study of the self-assembly process in a small spherical virus formation of organized structures from protein subunits in vitro. *Virology* 31:354-379.
19. Day J, Kuznetsov YG, Larson SB, Greenwood a, McPherson a (2001) Biophysical studies on the RNA cores of satellite tobacco mosaic virus. *Biophysical journal* 80:2364-71.
20. Belyi VA and Muthukumar M, (2006) Electrostatic origin of the genome packing in viruses, *Proc. Natl. Acad. Sci. U. S. A.*, 103(46), 17174–8.
21. Gallagher TM and Rueckert RR, (1988) Assembly-dependent maturation cleavage in provirions of a small icosahedral insect ribovirus, *J. Virol.*, 62(9), 3399–406.
22. Schneemann A, Reddy V and Johnson JE, (1998) The structure and function of nodavirus particles: A paradigm for understanding chemical biology, *Adv. Virus Res.*, 50, 381–446.
23. Sun J, DuFort C, Daniel MC, Murali A, Chen C, Gopinath K, Stein B, De M, Rotello VM, Holzenburg A, Kao CC and Dragnea B, (2007) Core-controlled polymorphism in virus-like particles, *Proc. Natl. Acad. Sci. U. S. A.*, 104(4), 1354–9.
24. Dixit SK, Goicochea NL, Daniel MC, Murali A, Bronstein L, De M, Stein B, Rotello VM, Kao CC and Dragnea B, (2006) Quantum dot encapsulation in viral capsids, *Nano Lett.*, 6(9), 1993–9.
25. Chen C, Kwak ES, Stein B, Kao CC and Dragnea B, (2005) Packaging of gold particles in viral capsids, *J. Nanosci. Nanotechnol.*, 2005, 5(12), 2029–33.
26. Hu Y, Zandi R, Anavitarte A, Knobler CM and Gelbart WM, (2008) Packaging of a polymer by a viral capsid: the interplay between polymer length and capsid size, *Biophys. J.*, 94(4), 1428–36.
27. Chang CB, Knobler CM, Gelbart WM and Mason TG, (2008) Curvature dependence of viral protein structures on encapsidated nanoemulsion droplets, *ACS Nano*, 2(2), 281–6.

28. Chen C, Daniel MC, Quinkert ZT, De M, Stein B, Bowman VD, Chipman PR, Rotello VM, Kao CC and Dragnea B, (2006) Nanoparticle-templated assembly of viral protein cages, *Nano Lett.*, 6(4), 611–5.
29. Zandi R, Reguera D, Bruinsma RF, Gelbart WM, Rudnick J (2004) Origin of icosahedral symmetry in viruses. *Proceedings of the National Academy of Sciences of the United States of America* 101:15556-60.
30. Hagan MF, Chandler D (2006) Dynamic pathways for viral capsid assembly. *Biophysical journal* 91:42-54.
31. Zhang T, Schwartz R (2006) Simulation study of the contribution of oligomer/oligomer binding to capsid assembly kinetics. *Biophysical journal* 90:57-64.
32. Wilber AW et al. (2007) Reversible self-assembly of patchy particles into monodisperse icosahedral clusters. *The Journal of chemical physics* 127:085106.
33. Rapaport D (2004) Self-assembly of polyhedral shells: A molecular dynamics study. *Physical Review E* 70:1-13.
34. Nguyen HD, Reddy VS, Brooks CL (2007) Deciphering the kinetic mechanism of spontaneous self-assembly of icosahedral capsids. *Nano letters* 7:338-44.
35. Nguyen HD, Reddy VS, Brooks CL (2009) Invariant polymorphism in virus capsid assembly. *Journal of the American Chemical Society* 131:2606-14.
36. Zhang D, Konecny R, Baker N a, McCammon JA (2004) Electrostatic interaction between RNA and protein capsid in cowpea chlorotic mottle virus simulated by a coarse-grain RNA model and a Monte Carlo approach. *Biopolymers* 75:325-37.
37. Harvey SC, Petrov AS, Devkota B, Boz MB, (2009) Viral assembly: a molecular modeling perspective. *Physical Chemistry Chemical Physics* 11:10553-10564.
38. Devkota B, Petrov AS, Lemieux S, Boz MB, Tang L, Schneemann A, Johnson JE, Harvey SC, (2009) Structural and electrostatic characterization of pariacoto virus: implications for viral assembly. *Biopolymers* 91:530-8.
39. Arkhipov A, Freddolino PL, Schulten K (2006) Stability and dynamics of virus capsids described by coarse-grained modeling. *Structure (London, England : 1993)* 14:1767-77.
40. Angelescu DG, Bruinsma R, Linse P (2006) Monte Carlo simulations of polyelectrolytes inside viral capsids. *Physical Review E* 73:1-17.

41. Forrey C, Muthukumar M (2009) Electrostatics of capsid-induced viral RNA organization. *The Journal of Chemical Physics* 131:105101.
42. Hagan MF, (2009) A theory for viral capsid assembly around electrostatic cores, *J. Chem. Phys.*, 2009, 130(11), 114902.
43. Elrad OM, Hagan MF (2010) Encapsulation of a polymer by an icosahedral virus. *Physical biology* 7:045003.
44. Lanman J, Crum J, Deerinck TJ, Gaietta GM, Schneemann A, Sosinsky GE, Ellisman MH and Johnson JE, (2008) Visualizing Flock house virus infection in *Drosophila* cells with correlated fluorescence and electron microscopy, *J. Struct. Biol.*, 161(3), 439–46.
45. Venter, P.A. and Schneemann, A. (2007). Assembly of two independent populations of Flock House virus particles with distinct RNA packaging characteristics in the same cell. *J. Virol.* 81:613-9.
46. Venter PA, Krishna NK and Schneemann A, (2005) Capsid protein synthesis from replicating RNA directs specific packaging of the genome of a multipartite, positive-strand RNA virus, *J. Virol.*, 79(10), 6239–48.

CHAPTER 2

COMPUTATIONAL APPROACHES TO MODELING VIRAL STRUCTURE AND ASSEMBLY

Introduction

The simplest viruses have a nucleic acid genome that is surrounded by a protein capsid. Genomes can be single-stranded or double-stranded, and they may be either DNA or RNA. In some viruses, the capsid proteins will spontaneously assemble into a procapsid that is matured as the genome is inserted in an energy-consuming process. In others, capsid formation requires the proteins to bind to the genome, which has already been partially or completely synthesized. Assembly is a critical step in the life cycle of viruses, so a detailed understanding of assembly might offer new opportunities for the design antiviral agents. In addition, the design of novel nanoparticles might be based on principles of viral assembly.

A wide variety of experimental, theoretical and computational studies have been aimed at increasing our understanding of viral assembly (1-4). Wherever possible, atomic detail is desirable, but all-atom modeling is not always possible. Sometimes there are not sufficient data to provide an atomistic representation. Sometimes – even if the structure is known in atomistic detail – simulations on biologically relevant time scales are not possible, because of computational tractability. In these cases, investigators often resort to lower-resolution coarse-grained models. Here we review methods for studying the structure and assembly of small icosahedral DNA and RNA viruses, sometimes with coarse-grained approaches, and sometimes combining all-atom and coarse-grained methods.

Double-stranded DNA bacteriophage:

Bacteriophages are viruses that infect bacteria. They consist of a protein shell (capsid) surrounding a DNA or RNA genome. Bacteriophage capsids, vary in size (from several hundred to several thousand Ångstroms), shape (from isometric to highly elongated with axial ratios up to 5:1), and T number (from 1 to 7) (5, 6). The genome of most bacteriophages is in the form of double-stranded DNA (dsDNA) and ranges in size from about 20,000 to 150,000 base pairs. The genome generally occupies 30%-50% of the available volume inside the capsid (7).

Packaging of dsDNA into a highly compacted state requires energy, to overcome the electrostatic repulsions, hydration forces, and the loss of conformational entropy. DNA is forced into bacteriophage by an ATP-driven protein motor, located in one vertex of the icosahedral capsid (8). *In vivo*, packaging has a characteristic timescale on the order of minutes. Because of the large size of bacteriophage and the time scale of packaging, all-atom simulations of packaging using conventional molecular dynamics are not possible. Therefore, it is necessary to use coarse-grained models. This is not a serious limitation, however, as many of the structural, kinetic and thermodynamic aspects of DNA packaging can be well described by simplified low-resolution models.

Here we discuss coarse-grained models used to represent the constituents of bacteriophages (*i.e.*, dsDNA, capsid and the protein portal and core structure). We also summarize our studies on the packaging of DNA into bacteriophages, and our studies on ejection of DNA from the capsid and into the host bacterium.

DNA Models

In our simulations we have two distinct DNA models (9-11). The first model represents double-stranded DNA as a string of beads on a chain, with each spherical bead (pseudoatom) representing N consecutive base pairs. In our viral packaging studies, we most commonly use a model with $N=6$, which we designate 1DNA6 (Fig. 2.1a). The model accounts for the stiffness of stretching and bending, volume exclusion effects, and long-range interactions between DNA strands, but it excludes torsional stiffness from consideration. The elastic stretching and bending properties of DNA are reproduced by appropriately parameterized harmonic terms for bond stretching and bond angle bending:

$$E_{bond} = k_b (b - b_0)^2 \quad (2.1)$$

and

$$E_{angle} = k_\theta (\theta - \theta_0)^2 \quad (2.2)$$

k_b and k_θ are stretching and bending force constants, b_0 is the equilibrium value of the distance between two consecutive beads, and θ_0 is the equilibrium bending angle for consecutive triplets. The stretching modulus was parameterized from the variance in the distance between successive base pairs (rise) of B-DNA from Nucleic Acids Data Bank (www.pdb.org) (12), and the bending modulus was parameterized to reproduce the value of the DNA persistence length of 510 Å (13). The details of parameterization are given elsewhere (9, 14). In the 1DNA6 model, the numerical values of the parameters are $k_b = 3.5 \text{ kcal}/(\text{mol} \cdot \text{\AA}^2)$, $b_0 = 19.9 \text{ \AA}$, $k_\theta = 22.4 \text{ kcal}/(\text{mol} \cdot \text{rad}^2)$, and $\theta_0 = 0$.

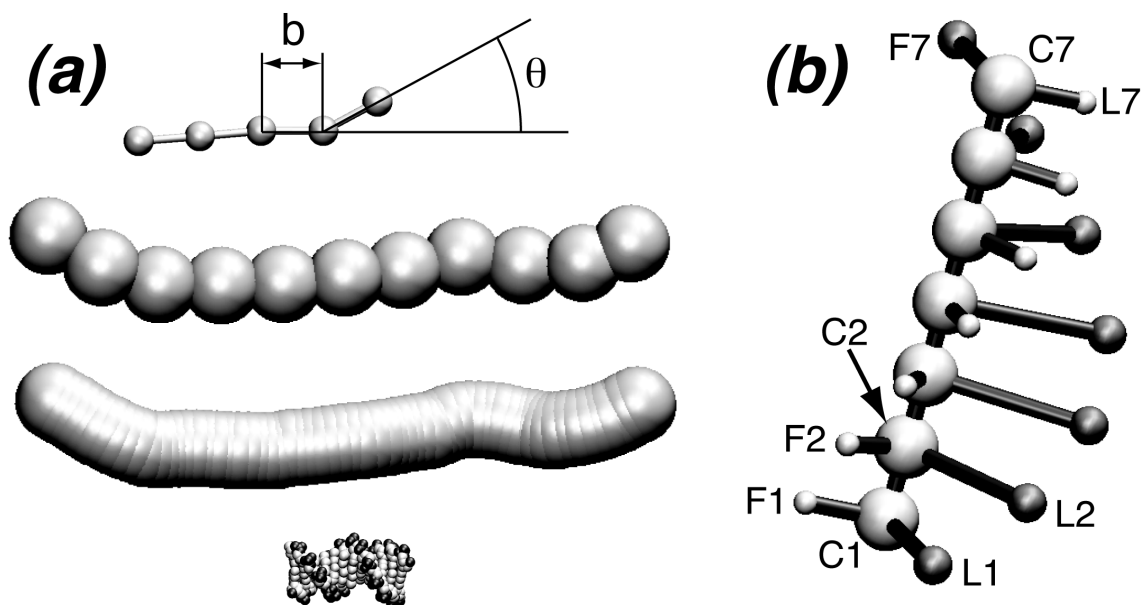


Figure 2.1. Coarse-grain models for DNA. (a) An all-atom representation of double-helical DNA (bottom) can be simplified to the 1DNA model with one spherical pseudoatom per base pair (lower middle). Further coarse-graining leads to the 1DNA6 model, with one bead for every six base pairs (upper middle). Both the 1DNA and 1DNA6 models have pseudoatoms with a diameter of 25Å. Chain stretching is opposed by elastic bonds, while bending is opposed by elastic bond angle terms. All four representations in this panel are shown to the same scale, but the radii of the beads have been scaled down for graphical purposes in the top representation of the 1DNA6 model, to permit visualization of one bond length (b) and one angle (θ). (b) The 3DNA model, in which the energetic cost of torsional deformations is included. Each base pair is represented by three pseudoatoms: the center atom (C), lying on the axis of the double-stranded DNA molecule; the “left” dummy atom (L), whose position approximates that of one phosphate group; and the “front” dummy atom (F), which lies somewhere in the major groove. The stretching elastic modulus determines the force constant for the harmonic bond between successive C atoms. Bending stiffness requires parameterization of several bond angles, *e.g.*, F1-C1-C2; L1-C1-C2; C1-C2-C3; C1-C2-F2; C1-C2-L2. Torsional stiffness requires parameterization of two improper torsions per base pair step, *e.g.*, F1-C1-C2-F2 and L1-C1-C2-L2. Volume exclusion is treated by the radius of the C atoms, since the dummy F and L atoms do not have volume; it is identical to the volume exclusion of the 1DNA model. We can generate a double helical graphical representation of any conformation by reversing each C-L vector to generate “right” dummy atoms located symmetrically opposite each L atom. This model can be further coarse-grained to the 3DNA6 model (not shown) by eliminating all pseudoatoms for base pairs 2-6 and making appropriate choices for parameters for bond stretching, angle bending, and improper torsions for the successive triads representing base pairs 1, 7, 13, and so on. The volume of the 3DNA6 model is essentially identical to that of the 1DNA6 model. Reprinted from [37]

To avoid interpenetration between DNA strands, each bead is spherical, with a radius of 12.5 Å. Non-bonded (volume exclusion) interactions are modeled by a semi-harmonic repulsive potential, often called a “soft sphere” potential:

$$E_{nb} = \begin{cases} k_{nb}(d_0 - d)^2, & \text{if } d < d_0 \\ 0, & \text{if } d \geq d_0 \end{cases} \quad (2.3)$$

where d is the distance between the two interacting pseudoatoms, $k_{DNA-DNA} = 11.0$ kcal/(mol·Å²), and $d_0 = 25.0$ Å. When modeling DNA as a simple elastic polymer (ignoring electrostatic effects), we used a cutoff of 50 Å for all volume exclusion calculations.

The second model allows the definition of a local DNA twist angle and allows the inclusion of torsional stiffness in the simulation (Fig. 2.1b). It contains two additional “left” and “front” dummy atoms attached to the central bead, and placed orthogonally to the DNA helical axis (10, 15, 16). In the original model (“3DNA1”), each triad of atoms defines a plane representing a single base pair; the “left” atom points toward the position of one backbone phosphate group, and the “front” atom defines the major groove of dsDNA. The torsional stiffness of DNA is represented by defining an improper torsion angle about the bond connecting successive backbone beads (Fig. 2.1b), with deformation energy,

$$E_{\text{improper}} = k_{\phi} (\phi - \phi_0)^2 \quad (2.4)$$

and by proper choice of the torsional force constant k_{ϕ} .

The 3DNA1 model is suitable for studying supercoiling in closed circular DNAs with lengths up to about 3000 base pairs (17), but its application to bacteriophage systems is impractical because of their sizes. We use a coarser version of this model for large DNA molecules, with N base pairs being represented by a single triad. The 3DNA6 model has $N=6$, and we used it in our investigations into the effects of torsional stiffness on viral packaging (10). The 1DNA and 3DNA models are easily parameterized for other values of N (11).

DNA is a charged polyelectrolyte, so it is essential to describe DNA-DNA interactions as accurately as possible. Experimental data on osmotic pressure show that this interaction is very complex (18, 19). In monovalent salts, DNA molecules are electrostatically repelled, though these repulsions are partially screened by counterions at long-range. At short distances (25-30 Å), hydration forces become important. These are due to the loss of conformational freedom of water molecules at the DNA surface. Trivalent or tetravalent cations in solution cause DNA condensation (20, 21). Because of the complexity of the problem and very large size of bacteriophage systems, we used a phenomenological approach to describe DNA-DNA interactions: instead of providing exact physical formulation for every component of this interaction, we derived a set of functions and parameters that accurately match the experimental potentials of mean force of DNA interactions *in vitro*. We treat two regimes: the repulsive regime is observed in presence of most monovalent and divalent cations, while the attractive regime appears upon addition of condensing agents (trivalent and tetravalent cations).

For the repulsive regime, we empirically derived the functional form of DNA-DNA interactions from the experimental data of Rau and Parsegian (22) and modeled them as a function of distance, r , by a modified Debye-Hückel function (23):

$$E_{DNA-DNA}^{rep}(r) = 0.59 L_b \frac{q_{eff}^2 \exp(-\kappa_{eff}(r-2a))}{r} \quad (2.5)$$

where $L_b = 7.135 \text{ \AA}$ is the Bjerrum length, and 0.59 is the conversion factor to kcal/mol. The other parameters (effective charge, $q_{eff} = -12.6 \text{ e}$ per pseudoatom, effective screening constant, $\kappa_{eff} = 0.31 \text{ \AA}^{-1}$, and DNA radius, $a = 10.0 \text{ \AA}$) correspond to the buffer containing 10 mM MgCl_2 , 100 mM NaCl and 10 mM TrisCl .

The interaction between DNA double helices in the attractive regime is described by the following empirical relationship, applied to pairs of DNA pseudoatoms in separate double helices, separated by a distance r :

$$E_{DNA-DNA}^{attr}(r) = A_1 \left[\exp\left(\frac{2(b_1 - r)}{c_1}\right) - 2 \exp\left(\frac{(b_1 - r)}{c_1}\right) \right] - A_2 \left[\exp\left(\frac{2(b_2 - r)}{c_2}\right) - 2 \exp\left(\frac{(b_2 - r)}{c_2}\right) \right] \quad (2.6)$$

with $A_1 = 0.011 \text{ kcal}/(\text{mol} \cdot \text{bp})$, $A_2 = 0.012 \text{ kcal}/(\text{mol} \cdot \text{bp})$, $b_1 = 30.5 \text{ \AA}$, $b_2 = 37.5 \text{ \AA}$, $c_1 = 2.6 \text{ \AA}$, and $c_2 = 2.2 \text{ \AA}$. The parameters were derived to match the data for the attractive interactions occurring in the range $r \sim 25\text{-}34 \text{ \AA}$, with a minimum of $\sim 130 \text{ cal}/(\text{mol} \cdot \text{bp})$ at $r \sim 27.2 \text{ \AA}$ (24), and the repulsive interactions in the range $35\text{-}50 \text{ \AA}$ as experimentally

observed by osmotic pressure data obtained in the presence of polycations (25). A cutoff of 70 Å was used to treat all long-range DNA-DNA interactions. We stress that parameterization was done to mimic properties of DNA free in solution, and there are no free parameters in our model that must be adjusted to match force-distance curves or other data from viral packaging experiments.

Capsid Models:

The protein-protein interactions in a bacteriophage capsid are relatively strong, and the capsid assembles spontaneously in the absence of genomic DNA. In contrast, the interactions between DNA and the walls of the capsid are relatively weak. The major role of capsid proteins is to keep DNA stored inside the capsid volume under high pressure after it is packaged. Thus, the capsids in our models play the role of a container to keep DNA confined within a volume of a defined geometry. We implemented two different approaches to model DNA capsids.

Many bacteriophage capsids have icosahedral isometric morphology. The simplest approximation for such a capsid is a sphere. We model spherical capsids by placing an additional dummy atom in the center of the spherical cavity of radius R and applying semiharmonic restraints between this pseudoatom and all DNA pseudoatoms. We call this energy function an “NOEN”, because of its resemblance to the semiharmonic restraint often used in refinement of NMR structures using contacts detected by the Nuclear Overhauser Effect. The energy is zero for any pseudoatom that lies within the sphere, and the energy penalty rises quadratically for pseudoatoms that violate the

spherical boundary. The dummy atom does not move in response to the NOEN forces, and the energy for a pseudoatom at a distance d from the center of the sphere is

$$E_{NOEN} = \begin{cases} k_{NOEN}(d - R)^2, & \text{if } d \geq R \\ 0, & \text{if } d < R \end{cases} \quad (2.7)$$

where $k_{NOEN} = 8.8 \text{ kcal}/(\text{mol} \cdot \text{\AA}^2)$.

Some spherical dsDNA viruses, *e.g.*, bacteriophage Lambda, undergo a significant capsid expansion process during maturation (26). Partially packed DNA pushes against the capsid walls and triggers the transition of capsid proteins to a new conformation. The expansion also affects the thermodynamics of the packaging process. In order to account for the expansion in a phenomenological fashion when modeling Lambda, we gradually increased the radius of confinement from 210 Å to 290 Å between 20% and 40% of Lambda genome packed, which is in the range where expansion occurs (27, 28). This simple model of capsid expansion is empirical and does not contain any regulatory feedback mechanism.

The second model describes the capsid as a polyhedron, either an icosahedron, an elongated icosahedron, or a more complex polyhedron (Fig. 2.2). We build such models from a set of triangular faces, edges and vertices, each of which is filled with a set of spherical pseudoatoms. The function of these spheres is to prevent the DNA chain from leaking out of the capsid, so the most important parameter of this model is the density of soft spheres: a low density runs the risk of DNA escape, while a high density increases simulation time. We cover the capsid surface with a hexagonal array of soft spheres, each of radius of 8 Å, and we have found that the minimum density required to keep DNA

inside the capsids corresponds to a separation between the spheres of 28 Å (23). The interactions between DNA and soft spheres is purely repulsive (Eq. 3); the parameters for the DNA-capsid interactions are $k_{nb} = 8.8 \text{ kcal}/(\text{mol} \cdot \text{Å}^2)$, and $d_0 = 20.5 \text{ Å}$.

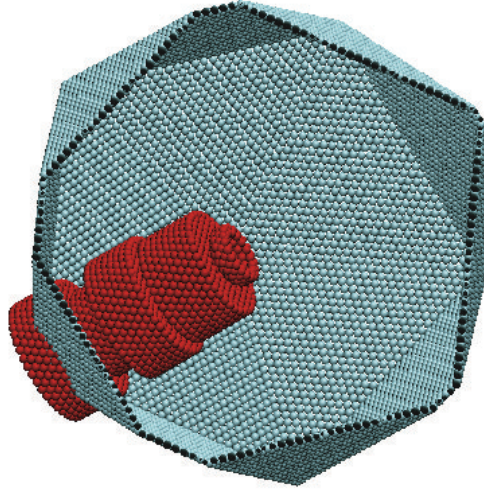


Figure 2.2. The model capsid for epsilon15. The triangular faces, edges and vertices of the icosahedral capsid are defined by collections of appropriately placed pseudoatoms, which are shown as opaque spheres. Reprinted from [30].

Both of the above capsid models may (optionally) have an additional feature. In bacteriophages such as T7 (29), epsilon15 (30), and P22 (31), there are other portal proteins at one of the capsid's vertices, in addition to the motor assembly. There is sometimes a well-developed structure (the core) that propagates into the viral interior, occupying as much as 15-20% of the inside volume of the capsid. The presence of a core structure can affect both DNA conformation inside the bacteriophage and the thermodynamics of DNA packaging, so we have included the cores in the models for those viruses where they are known to occur. The simplest model of the core structure is a hollow cylinder with an inner diameter of 30-40 Å, composed of soft spheres identical to those in the capsid walls. The outer radius and the length of the cylinder depend on the

particular bacteriophage. In a few bacteriophages, *e.g.*, epsilon15, the outer radius of the protein portal varies as it goes into the depth of capsid (30). We use a set of hollow, connected, coaxial cylinders to model such complex geometries, *e.g.*, Fig. 2.2 (32).

Packaging Protocols:

The packaging of the DNA genome into bacteriophages is not a spontaneous process, but is driven by a motor. The current level of the simulations cannot model the dynamics of the motor itself, but only the phenomenological result of its action. In the framework of our model, the packaging is driven by four auxiliary atoms (“stud atoms”) separated exactly by the equilibrium distance between DNA pseudoatoms, b_0 , and placed along the DNA axis, either outside of the capsid or inside the core structure, if present. Four successive DNA pseudoatoms (j through $j+3$) are attached via harmonic springs to the stud atoms (9, 23). The functional form of the stud energy function is identical to Eq. 1, with $b_0 = 0$ and a force constant of 0.01 pN/Å.

We ratchet the DNA forward into the capsid in a series of steps. The first half-step is achieved by moving the stud positions toward the center of the capsid a distance of $b_0/2$, followed by extensive equilibration using molecular dynamics (MD), to gradually move the DNA forward the same distance. The other half-step involves resetting the stud atoms back to their original positions and changing the harmonic restraints so that the studs are now attached to DNA pseudoatoms $j+1$ through $j+4$. Again, extensive MD equilibration moves the DNA forward by a distance of $b_0/2$.

All MD trajectories were generated using the YUP package (11), specifically designed for molecular modeling of coarse-grained systems. Simulations were performed with a time step of 1ps in the repulsive regime and 0.5 ps in the attractive regime.

Packaging was performed at 300K by coupling the systems to a Berendsen thermostat (33). The non-bonded lists were updated every ten steps.

Extensive equilibration is required during each step along the packaging trajectory to ensure that the structure and thermodynamic properties are not far from the equilibrium along the packaging trajectory. Each simulation begins with an equilibration time of 6 ns per half-step. As more DNA is crowded into the capsid, it takes longer to equilibrate the structure after each advance, so equilibration time is linearly increased by 4-8 ps per monomer as packaging progresses. Total trajectory time depends on the size of the model genome but typically ranges from $\sim 10 \mu\text{s}$ to $\sim 250 \mu\text{s}$ (23).

Data Analysis

The MD trajectories yield a range of structural and thermodynamic information. To determine the packaging forces, equilibrated intermediate conformations obtained at regular intervals along the packing trajectories (typically at intervals of 10% of the length of the DNA) are taken as starting points for a series of new MD runs, with DNA atoms at the entrance point held fixed. The time step during the force calculations is reduced to 0.1 ps. As the DNA tries to push its way out of the capsid, the springs connecting DNA beads with the stud atoms are stretched from their equilibrium lengths. To collect statistically uncorrelated data, 1000 of these displacements are collected at 500 ps intervals along the MD trajectory. The forces are calculated by multiplying the displacements by the force constants. Integrating the force-distance curve over the full genome length gives the work done during DNA packaging. Since the force is calculated in a series of simulations with a fixed amount of DNA held in the capsid, there is no net motion during force calculations, and the forces are equilibrium values. As a

consequence, the work that is done represents the free energy cost of packaging. The internal energies are extracted from the same MD trajectories, simply by summing the average component energies (Eqs. 1-6) and subtracting the corresponding values for free DNA at the same temperature and in the absence of capsid restraints. The entropic penalty associated with DNA confinement is then calculated as the difference between the free energy and the internal energy (23). Typically, ten to fifty independent packaging trajectories were carried out for each system that we investigated; by averaging over all of these, we obtained very accurate estimates of the forces and free energies.

Simulated low-resolution electron density maps are reconstructed by averaging over individual structures from ten to fifty independent packaging trajectories. In a single structure, each DNA segment between successive pseudoatoms along the chain is modeled as a cylinder with a radius of 10 Å. Each cylinder is uniformly filled with 2000 points (“atoms”), and the sets of these atoms are converted to corresponding values of single particle density maps with a voxel size of 3 Å using Spider (34). Superposing the individual densities generates average density maps that can be compared with experimental density maps from electron microscopy.

Ejection Protocols

The main difference between packaging and ejection is that the latter is a spontaneous process (at least at the initial stage) and does not require the help of an external motor. Ejection is driven by the high pressure of packaged DNA (35), which arises from hydration, electrostatic and entropic forces (23). The models and parameters for DNA and capsids applied to study ejection are essentially the same as those used to study packaging, except there are no stud atoms, so the DNA spontaneously escapes from

the capsid. In addition, we include a model of the bacterial cell by constraining the ejected portion of DNA in a sphere of appropriate volume (Fig. 2.3).

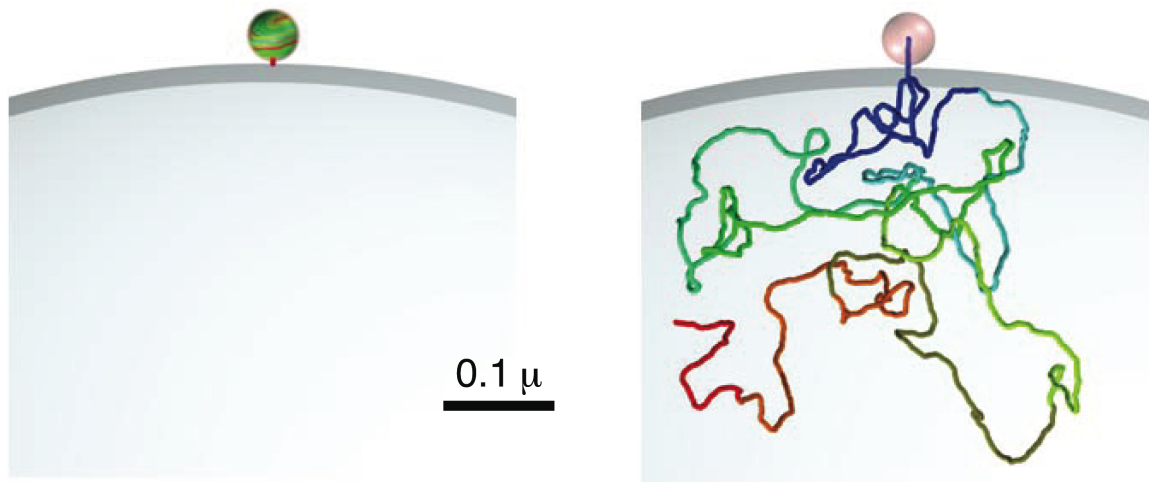


Figure 2.3. Simulation of the ejection of genomic dsDNA from bacteriophage f29. The genome was packaged into the spherical capsid as described in the text, and the full model is shown in the left panel, with the hollow core connecting the interior of the virus with the interior of a large sphere with the same radius (1 m) as a typical bacterium. Upon release of the restraint holding the DNA inside the virus, it is ejected into the bacterium, because of the combined electrostatic and entropic forces (right). Reprinted from [37]

The full ejection model includes DNA, the capsid, the connector channel, and a bacterial cell. For simplicity, we describe the capsid using the spherical approximation. The protein channel connecting the capsid and a bacterial cell is constructed as a hollow cylinder made of soft spheres, with inner diameter of 40 Å and length of 200 Å, similar to the protein cores used in the packaging simulations. The bacterial cell is modeled as a second NOE-like sphere with a radius of 1 m.

During the course of the simulation, we maintain and update a list pseudoatoms that have been ejected from the capsid; let us designate this list as containing beads 1 – $N_{ejected}$. We also maintain a list of twenty ejection candidates, atoms $N_{ejected}+1$ – $N_{ejected}+20$), which are still located inside the capsid. If a pseudoatom in this list is found

within 60Å of the capsid boundary, the spherical NOE-like capsid constraint for this atom is removed, so this bead is free to move down the connector channel and leave the capsid. After a pseudoatom comes out of the channel, enters the bacterial cell, and moves at least 100 Å into the cell, it is subjected to the spherical restraint of the bacterial cell. Thus, a pseudoatom cannot re-enter the capsid after entering the bacterial cell, because our model assumes that the probability of this event is very small. Addition and deletion of restraints are done on the fly during the course of the ejection simulations, which is possible due to the structure of YUP. The frequency of updating the ejection candidate list and modifying the spherical restraints varies between 0.5 ns and 10 ns, depending on the rate of ejection.

The viscosity of the medium inside the bacterial cell (or outside of the bacteriophage, if the bacterial cell is excluded from the model) strongly affects the kinetics of both packaging and ejection (36), so we carried out ejection simulations using the Langevin Dynamics (LD) protocol. The temperature was 298K, and the simulation time step was 0.5 ps. The frequency of applied stochastic forces (the collision frequency) varied over the range 0.001-0.02 ps⁻¹. Different viscosity regimes were studied to probe how the viscosity of the medium affects ejection kinetics.

Figure 2.3 shows the result of a typical ejection trajectory. We have analyzed these trajectories by plotting the amount of genome ejected vs. time. Numerical differentiation of this function gives the ejection rate along the trajectory. Additionally, the forces acting on the DNA were calculated according to a procedure similar to that described in the packaging protocol. Ejection was interrupted at every 10% of DNA genome ejected, and four successive DNA pseudoatoms inside the channel were

connected to four stud atoms placed inside the protein channel with harmonic restraints (Eq. 1, with $b_0 = 0$; recall that stud atoms are dummy atoms and do not move). We measured the average displacements of these DNA atoms with respect to the stud atoms along the packaging axis and converted these to forces by multiplying them by the stud force constant, in accordance with Hooke's law. No net motion of the DNA occurred during the force calculations, so these are equilibrium measurements. After the force measurements were complete, the stud atoms were detached and the ejection resumed.

The proposed model of DNA ejection could be further improved to account for the explicit presence of proteins, DNA, and organelles that occupy bacterial cells. It is known that the total volume fraction of DNA and proteins inside the bacterial cells is ~ 0.35 - 0.4 . The presence of these crowders is expected to affect both the thermodynamics and kinetics of ejection. A reduced void volume should result in the appearance of additional osmotic pressure that would act against the ejection force and eventually may stall ejection. A high concentration of crowders also changes the viscosity of the solvent, which is considered implicitly in the framework of our model. An increase of the collision frequency parameter in the Langevin Dynamics simulations would slow down the kinetics. All of these additional factors would increase the complexity of the model, resulting in a significant increase in required computational resources.

Results

We have recently summarized our understanding of DNA packaging inside bacteriophage systems elsewhere (3, 37); here we present only the highlights.

The high force developed by an ATP-driven motor is required to confine DNA inside the small volume of bacteriophage capsid. The free energy cost of packaging is primarily electrostatic and entropic in nature. These two components account for up to 90% of the total free energy cost, while the elastic bending energy accounts for most of the rest (3).

The confined DNA may fold into a number of conformations. All of these have significant disorder around certain idealized forms, including coaxial spools, concentric spools, twisted toroids, and folded toroidal structures (38). The specific DNA conformation inside a specific bacteriophage depends upon the size and shape of the capsid, the size and shape of the core at the portal (if any), and on the ionic composition of buffers in the surrounding media. Under fixed environmental conditions, the electrostatic and entropic costs of confinement are largely independent of the final conformation, so the optimum conformation minimizes the elastic bending energy (38). Simulations reproduce the multiple shell pattern of DNA density often seen in the experimental reconstructions. The latter reveal little about individual conformations, because the reconstructions are averages over thousands of individual viruses (23, 32), and the simulations provide these details. The current modeling method captures the essential physics of DNA packaging, but is not yet capable of describing complex features such as specific interactions between DNA and proteins in the capsid walls. Nor does it treat the interactions of DNA with the packaging motor in enough detail to understand the mechano-chemical transduction process behind the mechanism of DNA translocation.

Torsional stiffness does not significantly affect either the final DNA conformation or the thermodynamics of packaging, if one end of the DNA molecule is free (unattached) inside the bacteriophage, so it is free to rotate and relax torsional strain (10). When both ends are tethered, torsional stiffness has only a small effect on the thermodynamics of packaging, but the final conformations are different than for the untethered case (39).

Upon ejection of the first 50-60% of the ejected genome, the ejection forces drastically decrease, dropping to a few piconewtons. However, further ejection leads to a slight increase in the force that acts on DNA and pulls it outside of the capsid. This observation lends support to the dual “push-pull” mechanism of DNA ejection (35, 40). The initial decrease of the force during genome ejection is due to the drop in pressure inside the capsid. The subsequent increase of the force, which pulls the remaining DNA outside of the capsid, is due to the entropic force developed by the ejected portion of the genome. This force is on the order of a few piconewtons, and correlates well with the radius of gyration of the ejected DNA.

Single-Stranded RNA Viruses

A specific model system: pariacoto virus

Pariacoto virus (PaV) is an icosahedral T=3 RNA virus with a bipartite genome. The 4322 nucleotide genome consists of RNA1 (3011 nucleotides) and RNA2 (1311 nucleotides). The protein capsid is composed of 180 identical subunits, each containing 401 amino acids. There are 60 copies of the crystallographic asymmetric unit, each of which contains three copies of the capsid protein, in three different conformations, called A, B, and C (41). The asymmetric unit also contains an RNA segment of 25 nucleotides.

The RNA forms half of a double-stranded duplex that is perpendicular to the crystallographic two-fold axis and that lies just inside the protein capsid. The full structure of the virus can be generated from the asymmetric unit using the 60 matrices provided in REMARK 350 of the PDB file (1F8V.pdb), using the oligomer generator application from the VIPER website (42). The RNA forms a dodecahedral cage with a 25 base pair duplex lying on each of the 30 edges. Thus, the crystallographically resolved RNA accounts for about 35% ($25 \times 2 \times 30 = 1500$ nt) of the total genome. The remaining 65% of the RNA lies inside the dodecahedral cage and is not resolved in the crystal structure, because it lacks icosahedral symmetry. In addition, the RNA at the twenty vertices at which the duplexes are connected, are not crystallographically resolved, presumably because fragments at different vertices have different structures. Similarly, protein subunit A is missing 6 residues at the N-terminal end and 15 at the C-terminus in the crystal structure, while the B and C subunits are missing about 50 residues at the N-terminus and 19 residues at the C-terminus, due to the lack of clear electron density. Again, this almost certainly represents structural heterogeneity.

The challenge is to model the complete virus in as much detail as possible. The structure revealed by crystallography is very large, and there are only limited experimental data to guide modeling efforts on the rest of the structure. Because of the size of the system and the limited data on the protein tails and the RNA in the interior of the virus, coarse-grained modeling is appropriate for building and refining the model, although we converted the final coarse-grained model to an all-atom model at the end.

Conversion of RNA secondary structure into a 3D coarse-grained model

As will be seen presently, we based the model of the PaV RNA genome on a plausible secondary structure model (43). We built the three-dimensional model by connecting fragments from crystal structures with junctions that we built manually at the all-atom level, inter-converting all-atom and coarse-grained representations as appropriate. In some of our RNA modeling efforts, we use an entirely automated procedure for converting secondary structures into three-dimensional models. Although we did not use this procedure in our PaV model (44), we present this automated method here, for completeness.

RNA presents a more difficult modeling challenge than double-stranded DNA. Unlike dsDNA, ssRNA molecules contain various structural motifs, including double-stranded regions, single-stranded regions, stem-loops, and a variety of bulges and junctions. The simplest coarse-grained model of RNA is a linear beads-on-a-string model, but it cannot model the variety of structural motifs, and it does not describe RNA secondary structure, which plays a crucial role in defining RNA conformation in 3D space. Such a model necessarily has limited utility for investigating the structure and assembly of RNA viruses.

We previously developed a coarse-grained “PX” model of RNA that provides a good 3D description of RNA composed of different structural elements (16, 45). Figure 2.4b shows that model, which we have implemented in YUP as the *rrRNAv1* model (11). In the framework of this model each nucleotide is represented by one pseudoatom (P-atom). Single stranded regions are described by flexible strings composed of connected P-atoms, and helices are explicitly represented by semi-rigid fragments, in which hydrogen bonding between the strands are replaced by unbreakable bonds between P-

atoms on the two strands. There are terms in the energy function that describe the bond angle bending between successive triplets of P-atoms along the backbone, and other angular terms to define the ideal geometry of double-helical regions. An improper torsion $(j-1, j, k, k+1)$ is associated with the $j-k$ base pair, to enforce the right-handed chirality of double helices.

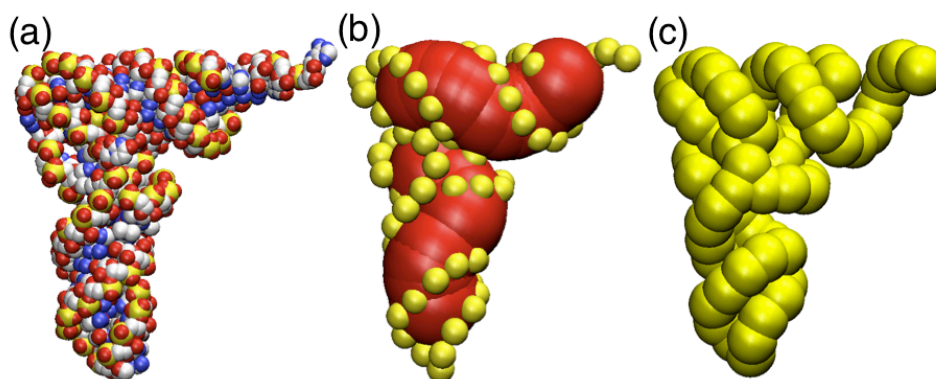


Figure 2.4. Models of tRNA. (a) All-atom model, with phosphorus atoms highlighted as small dark spheres. The larger grey spheres are the “2N” pseudoatoms, each representing two base pairs, and each placed at the midpoint of two successive glycosidic nitrogen atoms. **(b)** The PX model, also implemented as the *rrRNAv1* model. Each residue is represented by a single P-atom, centered at the position of the phosphate group (black). There is an additional pseudoatom (X-atom) for each base pair in the double-stranded regions. It is located at the geometric center of the base pair and has a sufficiently large radius to provide appropriate volume exclusion. **(c)** The 2N model, with one pseudoatom representing two successive nucleotides. Reprinted from [37]

A model containing only P-atoms would have hollow double helices, running the risk of artifactual inter-helical penetrations. Proper treatment of volume exclusion arises from the presence of a series of additional X-atoms along the axis of each double-helical fragment (Fig. 2.4b). Both the PX and *rrRNAv1* models have too many parameters to be given here; they are reported elsewhere (16, 45).

If the coordinates of all RNA atoms are known in 3D, then the positions of P-atoms can be easily extracted and the *rrRNAvI* model can be generated according to a previously described procedure (46). If the crystal structure is not known, small fragments can be built by manual modeling. For large systems of unknown structure, one of the common goals is to create a plausible 3D model that is compatible with a specified secondary structure. This is particularly important in studies on viral assembly and other properties of viral RNAs. We have developed an algorithm that generates the *rrRNAvI* model from a specified secondary structure. It can be used without providing any additional three-dimensional data, or, when such data are available, they can be incorporated into the model as restraints.

RNA secondary structure predictions from programs like Mfold (47) are often given in a CT file format. Columns 1 and 2 specify the index (residue number) and type (A,C,G,U) of each residue, while column 5 contains the index of the complementary base-pairing residue, if any (zero, otherwise). This information is extracted and converted to the BLUEPRINT format of the *rrRNAvI* model using the utility CT2BLUE.py located in the *rrRNAvI* folder of the YUP package (11).

The format of the BLUEPRINT file used by YUP to create the *rrRNAvI* model is described in the YUP documentation and will only be outlined here. Fragments of the secondary RNA structure must be given in hierarchal form. In the simplest case, all the elements of 2D RNA structure (loops, single- and double stranded regions) may be described at the same hierarchal level, but more complex organization containing multiple levels is also possible. The latter does not affect the properties of the *rrRNAvI* model but simply provides an additional amount of structural information for complex

RNA molecules containing multiple domains. The BLUEPRINT file (written in python) contains a dictionary “BLUE” with several keywords.

The first keyword “RNA_RNA” contains information about RNA secondary structure and given in the following format: (DOMAIN, 'all', (D_1, D_2,...)), is a tuple of tuples, where D_i is the label of the i^{th} region. For example, (DOMAIN, 'all', (S_1, H_1, S_2, H_2, S_3, H_3,...)) could specify a single-stranded region at the 5' end of the molecule, followed by a series of three double-helical regions connected by single-stranded regions, with other entries to identify the structure of the rest of the molecule. Here the entries S_1 and S_2 are labels for single stranded regions, and the entries H_1, H_2, and H_3 represent double-helical regions. (Other labels might be used for loops, bulges and strands that are part multi-branch junctions; these are all “single-stranded” in the sense that they do not have base-paired partners.) Each entry in the nested tuple is given in a format that defines the characteristics of the corresponding region, *e.g.*, S_1 = (TRACT, 'tract_1', (1,3)) and H_1 = (HELIX, 'helix_1', (4,7,45)) , where the first entry defines the type of the RNA fragment, the second entry labels it, and third entry provides the structural information. TRACT and HELIX define single-stranded and double-stranded domains, respectively. The third entry is a tuple that contains two or three residue indices for tracts and helices, respectively. For tracts, two indices define the beginning and the end of a single stranded fragment. (In this example, S_1 is single-stranded and contains nucleotides 1-3). For double helices, the first and third indices define 5' end positions of anti-parallel strands that form a double-helical region, and the second index defines the length of the double-stranded region. (Here, H_1 contains seven base pairs, between residues 4-10 and residues 51-45.)

The second keyword “RNA_BSQ” contains information about sequence in the format of a tuple: ('C','A','U','C','C',...). Finally, the last two keywords, RNA_XYZ and RNA_FIX, are by default empty tuples: (). They may contain information about the positions of the P-atoms and additional constraints (e.g. for loop regions), if such data are known from other sources.

The BLUEPRINT file is used as an input file to generate the *rrRNAv1* model. The model is generated in several steps using the YUP package. The first step:

`M=rrRNAFFA()` activates the model. The second and most important step reads the data from the BLUEPRINT file and creates the RNA: `R.addRNA(blueprint('BP_NAME'), modelname='M_NAME', randomize=1, dimensions=(5.6, 0.0, 180.0/n, 0.0, 0.0, 0.0))`.

The procedure *blueprint* reads python dictionary “BLUE” from the file ‘BP_NAME.py’, which contains the keywords describing the RNA secondary structure. The variable *modelname* is a string that defines the name of the molecule. If the variable *randomize* is set to 1, the coordinates of RNA are generated by an internal YUP routine. If it is set to 0, the coordinates will be read from the dictionary entry RNA_XYZ (if available).

The variable *dimensions* is a tuple that contains the average and standard deviation of the distances (Å) between two adjacent P-atoms, the average and standard deviation in the angles (degrees), and the average and standard deviation in the improper torsions (degrees). The *dimensions* argument is used to generate the initial coordinates of the RNA model in a form of a circular arc. It can generate a random chain using a random walk algorithm but we found that the random initial coordinates of RNA may result in topological traps once the constraints describing helical regions are applied,

whereas an initial conformation of RNA in the form of an arc avoids this problem. To generate an initial model where all P-atoms lie on a planar 180° circular arc, one sets the variable *dimensions* to (5.6, 0.0, $180.0/n$, 0.0, 0.0, 0.0). In this example, 5.6 Å is the equilibrium distance between adjacent P-atoms, n is the number of residues in the model, and the initial torsions and standard deviations are all set to zero. Figure 5a shows the result for a more open circular arc. The method `R.addRNA()` also activates all necessary force field terms. Note that the structure in Fig. 2.5a does not satisfy any of the restraints in the model, except for P-P bond lengths along the chain; optimization of the structure produces a model that does satisfy those restraints (Fig. 2.5b).

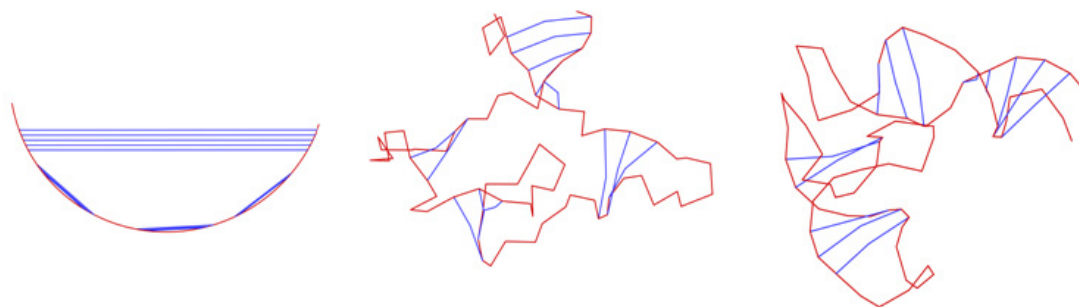


Figure 2.5. Conversion of the tRNA secondary structure model into a three-dimensional model. (a) 76 successive P-atoms are initially equally spaced along a circular arc in the xy plane, with pseudobonds corresponding to the secondary structure; although X-atoms are present, they are not shown, simply for graphical clarity. **(b)** Simulated annealing and minimization yields a three-dimensional structure that satisfies all the distance, angle and pseudotorision restraints of the secondary structure, as well as the volume exclusion requirements. **(c)** A plausible three-dimensional model of tRNA is produced by refinement after the addition of restraints representing the 18-55 and 19-56 base pairs between the D-loop and T-loop, along with restraints for correct stacking of the acceptor stem on the T-stem, and the anticodon stem on the D-stem. These restraints are not sufficient to completely define the three-dimensional structure of tRNA, because there are fewer restraints than degrees of freedom. The addition of a single distance restraint between the anticodon loop and the 3' tip of the acceptor stem does produce a model that resembles the crystal structure (not shown). Reprinted from [37]

Finally, the model is completed by the `M=R.finish()` method, which creates an object of the RNA model in YUP. The model object contains the detailed description of the model, including all force field terms and the initial coordinates. It exists virtually in the computer's memory, so its properties can be easily modified.

After creation, the model is optimized by extensive minimization (*e.g.*, 500,000 steps of steepest descent), followed by thermal equilibration using molecular dynamics (*e.g.*, simulated annealing; or, in the example of Fig. 2.5, 10ns at 300K with a time step of 10 fs.) After this procedure, RNA adopts a three-dimensional conformation that is folded in accordance with the secondary structure (Fig. 2.5b), plus any three-dimensional restraints (Fig. 2.5c), as enforced by the *rrRNAv1* force field.

At this point one may continue the simulations on RNA within the YUP package, or one can convert the *rrRNAv1* model (force field terms and XYZ coordinates) into the format for AMBER (48) or LAMMPS (49) for further simulations. AMBER is, of course, a very widely used package for biomolecular simulations; LAMMPS (<http://lammps.sandia.gov>) is a newer open source package, developed for simulating a wide range of condensed systems. We have previously published the AMBER conversion protocol (46) but have not yet done so for the LAMMPS conversion. Briefly, the conversions are done by executing the utility programs `AMBER.py` and `LAMMPS.py`, which are also contained in the *rrRNAv1* folder of the YUP package (11). Simulations in AMBER and LAMMPS significantly speed up the production stage of MD simulations, because these packages are available in parallel versions, while YUP is currently only available as single-processor code.

Pariacoto Virus: The RNA Model

To begin with, we converted the all-atom initial model to coarse-grain representation, with each nucleotide represented by a single pseudoatom at the phosphate position. A more complete description of this “all-P” model is available elsewhere (45).

We built the complete PaV model in two steps. First we modeled those parts of the viral genome that are not resolved in the crystal structure, attaching them to the 1500 crystallographically defined nucleotides in the RNA dodecahedral cage. Then we added the missing residues of the protein subunits.

Modeling the missing parts of the PaV RNA require us to visualize, manually manipulate, and refine the coarse-grained RNA model without tangling it. It is impossible to do this within the confines of the model capsid, because it is so small. Instead, we built the model in an expanded framework, then shrunk it down to the correct size in a series of scaling/optimization steps (Fig. 2.6). The initial, correctly scaled framework is defined by twenty pseudoatoms, each at the vertex of a virtual dodecahedron whose edges are coaxial with the RNA double helices that define the RNA dodecahedral cage in the crystal structure. Multiplying the coordinates of the twenty pseudoatoms by a factor of two provides a dodecahedral framework with eight times the volume of the virus, in which it is easy to build and manipulate the RNA model (Fig. 2.6a). Once that is done, we shrink the framework back down to its correct size by repeated scaling steps, each of which is followed by extensive minimization.

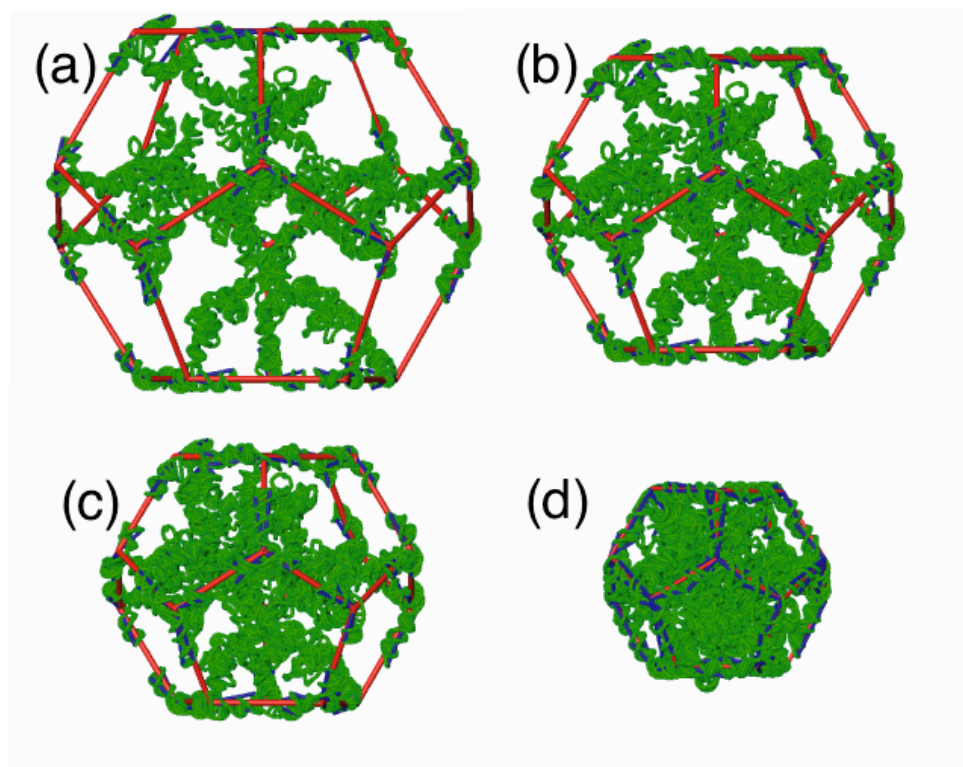


Figure 2.6. Optimization of the RNA model for pariacoto virus (PaV). It is not possible to manipulate the RNA model within the confines of the virus, so we define a dodecahedral framework that initially has twice the diameter and eight times the volume of the actual virus, build the RNA model in that framework, then refine by a series of shrinkage/minimization steps. (a) RNA is modeled in the expanded framework. Each RNA double helix on one edge of the original dodecahedral framework is cut into two fragments, with one attached to each vertex in the expanded framework. The “stalactites” of RNA that reach from twelve vertices into the interior of the virus are then attached, giving a complete model of the genome. (b) and (c) Two snapshots during the refinement, as the dodecahedral framework is shrunk stepwise to the correct size, followed by minimization of the RNA model at each step. (d) The final RNA model after complete contraction of the dodecahedral framework to the size it has in the crystal structure. Reprinted from [44].

To expand the RNA dodecahedral cage without deformation, we separated each RNA duplex between the twelfth and thirteenth nucleotides and moved each half duplex to the appropriate vertex of the expanded dodecahedral frame. This gave three pieces of RNA at each vertex. We had previously postulated a plausible secondary structure of the PaV RNA genome (43), based in part on the density of the cryo-electron microscopy map just below the vertices, which had suggested that there are approximately twelve

connections between the RNA dodecahedral cage and the remainder of the genome in the center of the virus. Our secondary structure model defines a set of two-, three- and four-way junctions at the twenty vertices of the dodecahedral cage, with twelve of these connecting to RNA in the center of the virus through short double-helical “stubs”.

We began 3D modeling by building all-atom models of the junctions and stubs. We then modeled the rest of the RNA inside the dodecahedral cage by attaching twelve identical copies of a globular RNA to the stubs. For this, we chose a 225-nucleotide fragment (residues 1764-1988) from domain IV of the large subunit of the *E. coli* ribosome (PDB id: 2WA4). We call these pieces of RNA “stalactites”. Within the expanded framework, it was relatively easy to add these stalactites without any steric clashes (Fig. 2.6a).

The model in the expanded framework has 4322 P-atoms (one per RNA residue) plus the twenty pseudoatoms at the vertices of the virtual dodecahedral framework. Some RNA fragments are based on crystal structures, while others are based on idealized double helices and junctions, so the RNA model is stereochemically correct, except that the double helices connecting adjacent vertices on the dodecahedral cage are split into two separate pieces on the expanded framework. To rejoin these, we scaled the framework downward in size (and moved the RNA radially inward) in a series of steps, each of which shortens the edges of the framework by 5 Å; the RNA model was re-minimized after each scaling. Figure 6 shows a series of snapshots from this process.

Minimization was done using *yammp* (50), which requires two input files. The archive file consists of the (x,y,z) coordinates of the structure. The descriptor file contains

the ideal values for different parameters (bonds, angles, etc.) and the force constants. These are given in Table 2.1.

Standard bond and angle energy functions are used for the connections between appropriate pairs and triplets of pseudoatoms. There are, for example, pseudobonds between successive P-atoms along the backbone of the molecule; there are also pseudoatoms connecting P-atoms representing the phosphate groups of a pair of nucleotides that interact through Watson-Crick base pairing. As in the case of the full PX and *rrRNAv1* models discussed above, the simplified all-P model also includes pseudotorsions to guarantee the proper chirality of the right-handed double helices (45).

There are two classes of bond, angle and pseudotorsion energy terms. The first class is designed to enforce idealized local geometry on the RNA model. In this model, “idealized” refers to values taken from the crystal structure of the RNA dodecahedral cage, from the crystal structure of the ribosomal RNA fragment used to model the stalactites, from model stem-loops, from the model three- and four-way junctions, and from the double helical stubs used to connect the dodecahedral cage to the stalactites. The second class consists of a set of restraints between the pseudoatoms of the expanded dodecahedral framework and pseudoatoms in the broken RNA double helices from the crystallographic dodecahedral cage; these keep the double helices correctly positioned as the framework is contracted, so that they are reconnected with the crystallographic geometry at the end of the contraction/refinement process.

As seen in Table 2.1, there are two different families of force constants (not to be confused with two different classes of bonds, angles and pseudotorsions). One family is applied to those distances and angles between atoms in the double-helical RNA cage,

while the other is applied to those in the stalactites. The former are ten times stronger than the latter, to prevent distortion of the cage away from the structure seen in the crystal; almost all deformations are thus forced onto the stalactites, since there are no data on the actual RNA structures in the viral interior.

As in the *rrRNAvI* model discussed above, a soft sphere semiharmonic repulsion is used for the nonbonded interaction between pairs of P-atoms that are not covalently connected through a bond or angle term, and that are not part of the same double helix. To reduce computational complexity, no X-atoms were included in the coarse-grained PaV model, so we used a rather large P-P contact distance (10\AA) to prevent interpenetration of double helices. This has the added advantage of keeping the RNA structure rather open, mimicking RNA-RNA electrostatic repulsions in the real world, and leaving room in the interior of the virus model for the penetration of positively charged protein tails to help neutralize the RNA and stabilize the structure (see below).

In early trials, we observed that the stalactite RNAs had a tendency to escape through the faces of the RNA dodecahedral cage during the contraction/minimization steps. To prevent this, we added an NOE-like restraint (NOEN in *yammp*) to confine all the RNA within a spherical boundary of radius R (Eq. 7). This parameter is decreased by $\sim 7\%$ during each step of scaling. This term also helps to keep the RNA helices attached at the vertex pseudoatoms properly oriented with respect to the dodecahedral framework during contraction. The NOEN is defined with respect to the center of the virus, which coincides with the origin of coordinates.

The twenty pseudoatoms defining the vertices of the dodecahedral framework are tethered to specified points in space with a harmonic “stud” energy function, as discussed

above. There are also thirty bonds between adjacent pairs of these pseudoatoms, to help rigidify the framework; they coincide with the edges of the dodecahedron (Fig. 2.6). The tethering positions of the vertex pseudoatoms were moved inward and the ideal bond lengths of the edges of the dodecahedral framework (b_0) were shortened in a series of 5 Å steps. The initial framework had an edge length 149.0 Å, and the final framework has $b_0 = 78.5$ Å. The model is minimized to convergence using the energy minimization protocol of *yammp* after each step. Since all the terms used in the potential energy function of all-P models are harmonic, full minimization of the model should lead to zero energy, if all restraints can be satisfied without steric overlaps.

Table 2.1: Force constants of each energy types.

Energy	Equation	Force constant
Bond ^a	Eq. (1)	RNA cage: 20 kcal/(mol Å ²) Stalactites: 2 kcal/(mol Å ²)
Angle ^a	Eq. (2)	RNA cage: 20 kcal/mol Stalactites: 2 kcal/mol
Improper torsion ^a	Eq. (4)	RNA cage: 20 kcal/mol Stalactites: 2 kcal/mol
Nonbond	Eq. (3), $d_0 = 10$ Å	2 kcal/(mol Å ²)
NOEN	Eq. (7)	2 kcal/(mol Å ²)
Stud	Eq. (3), $d_0 = 0$	40 kcal/(mol Å ²)

During minimization, the stalactite RNAs were free to move and adjust their conformations, to avoid steric overlap. They had softer force constants in the energy terms than did the RNA domains on the dodecahedral cage (Table 2.1). The crystallographic regions were restrained by using strong force constants in the energy terms, and by the addition of pseudobonds connecting each vertex pseudoatom to the ends of the RNA duplexes on each edge. These regions did not deviate significantly from

the crystal structure during the contraction/minimization cycled. The output file at the end of each step is a new archive file representing an intermediate model with the total energy converged to a minimum. This structure became the starting model for the next round of contraction/minimization, using a new descriptor file with ideal values for the edges and NOEN radius decreased appropriately.

Our collaborator Sébastien Lémieux (University of Montreal) converted the coarse-grained RNA model to an energy-refined all-atom model using a suite of programs that he had developed. This is quite straightforward for double-helical regions. In single stranded regions, conversion begins by generating candidate structures for fragments defined by four successive phosphate atoms along the backbone. Candidates are extracted from the same library that is used for modeling with MC-SYM (51), based on the requirement that the four phosphate groups in the library fragment must have a root-mean-square deviation of less than 1.5Å from the P-atom positions in the coarse-grained model. Once candidates are identified, the problem then becomes one of searching all combinations of candidates to identify which set will satisfy the RMSD restriction with the lowest non-bonded energy (van der Waals plus electrostatics). This optimizes base pairing and stacking, while minimizing steric clashes.

Pariacoto virus: Adding the capsid to the model

The final step in generating the model of PaV was to reconstruct the protein residues missing from the crystal structure. As mentioned before, the crystal structure of the asymmetric unit is missing residues from the N- and C-terminal tails of each protein because of the lack of clear electron density. The missing residues are shown in Table 2.2. The N-terminal tails contain an excess number of arginine and lysine residues

compared to rest of the protein, so the tails have a net positive charge. These basic residues interact with the RNA through electrostatic attractions, presumably stabilizing the structure of the virus. The C-terminal tails are composed of neutral residues.

Table 2.2: Protein residues that are not seen in the PaV crystal structure.

Protein	N-terminal	C-terminal	Total number
A	1–6	379–393	21
B	1–48	383–401	64
C	1–50	383–401	68

Our approach to modeling the capsid proteins was similar to the approach we used for modeling the RNA. We defined a framework with the same icosahedral symmetry as the virus, with 60 triangular faces, one for each copy of the asymmetric unit, then expanded it by a factor of three (a 27X expansion in volume). We radially translated a coarse-grained model of the crystallographically resolved parts of the capsid proteins to this expanded frame, keeping the RNA fixed at the center to generate enough space for us to place the missing protein residues (Fig. 2.7). We generated C and N-terminal tails, then compressed the capsid in radius in multiple steps, with minimization of the tails at each compression step, using YUP (11). This repeated compression/minimization protocol allows the protein tails to find their way into the fixed RNA.

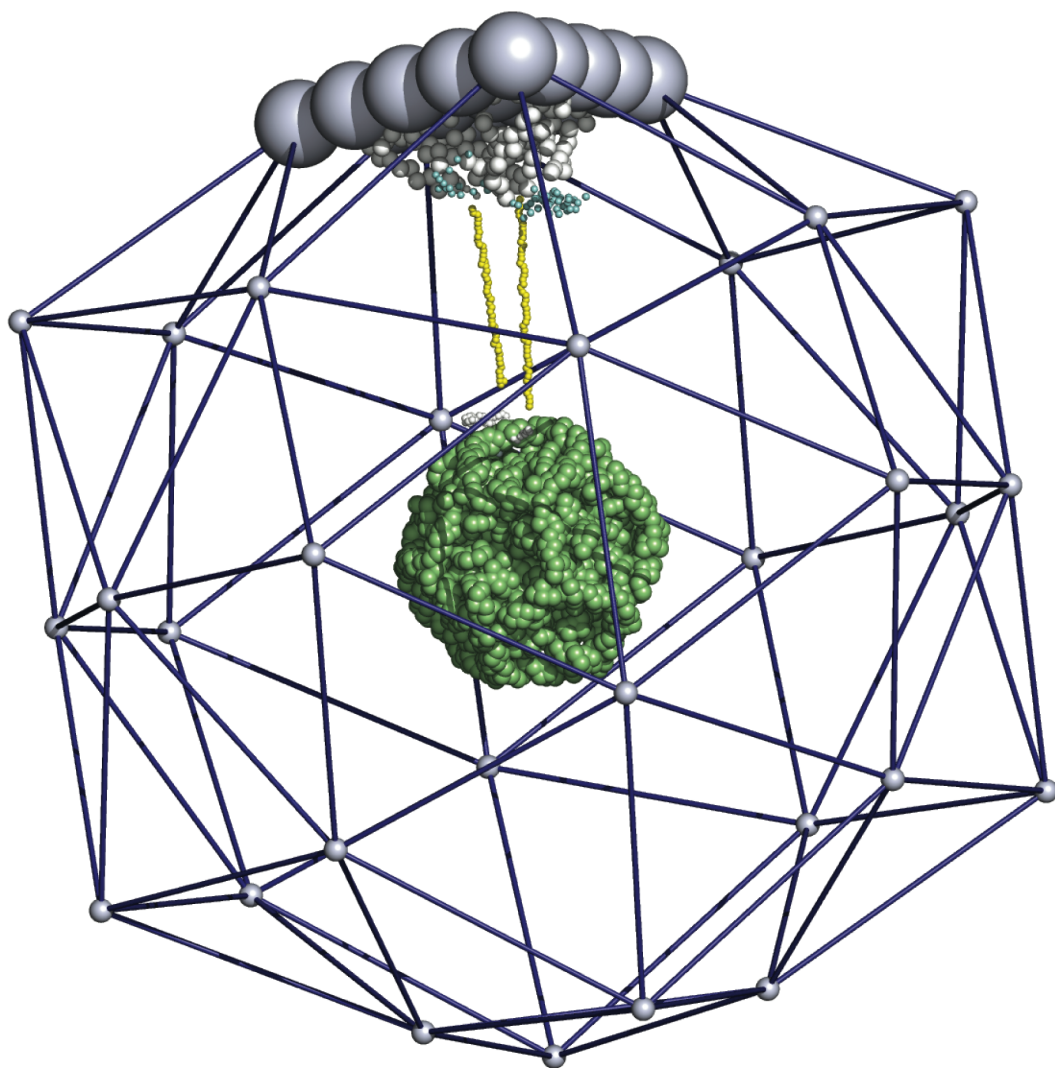


Figure 2.7. Addition of the capsid proteins to the RNA model for PaV, starting with a protein cage structure that is expanded to three times its final diameter. Coarse-grained models of different resolutions are used to model different regions of the proteins. Those parts of the crystallographically resolved regions that lie nearest the RNA are represented in a model with one pseudoatom representing two successive amino acids. The remaining residues are represented by a very coarse-grained model, with twelve pseudoatoms representing the face, side and vertices of the triangular asymmetric unit. The protein tails, whose conformations are not revealed in the crystal structure, are represented by one pseudoatom per amino acid and extend radially inward from the inside of the capsid toward the RNA genome. The expanded cage is shrunk to its crystallographic dimension in a series of steps, with energy minimization at each step. The protein tails are pulled toward the center of the virus during this process, and their ability to penetrate the porous RNA cage depends on the van der Waals radius assigned to these residues. Reprinted from [37].

The details of the modeling and simulation protocol are as follows. We first converted all of the RNA model and the crystallographically resolved protein residues into coarse-grain models to reduce the number of atoms. Our goal was to remove as many residues as possible from both the RNA and protein while maintaining their surface integrity. The inner side of the capsid proteins and the outer surface of the RNA are particularly important, because these surfaces are in contact with the missing amino acid residues.

We used a very coarse-grained model for regions of the protein on the outer surface of the capsid. To make this selection quantitative, we defined a triangle connecting the alpha carbons of residue 175 in the A, B and C proteins. Atoms outside this triangular plane were completely removed and replaced with twelve pseudoatoms (12C-model), each with a radius of 35 Å. These pseudoatoms covered the whole triangle, preventing any flexible chains from leaving the virus during the minimization protocol. The atoms below the triangular plane were converted into a 2C α -model by averaging the coordinates of successive pairs of C α atoms and replacing them with a single pseudoatom. Residues 7-50 of the protein A were exception to this conversion. These residues are in contact with the RNA in the crystal structure, so they were kept in their crystallographically defined positions; we modeled them with one pseudoatom per residue, placed at the position of the alpha carbon.

We converted the all-atom RNA model into a 2N model, with two consecutive nucleotides represented by one pseudoatom at the center of two consecutive glycosidic nitrogen atoms. This model conserves the minor and major grooves of the RNA double helices. It has less excluded volume than an actual RNA molecule, so a 2N model of a

viral genome is quite porous. In the case of PaV, this facilitates penetration of the polycationic tails of the capsid proteins into the RNA grooves in the viral interior.

After the RNA and the non-missing part of the asymmetric unit were converted into coarse-grain models, the asymmetric unit was moved out from the center of the RNA. This radial expansion was achieved by multiplying all coordinates by a factor of three, since the model is centered on the origin. This provided enough space for us to generate the missing tail residues, using one pseudoatom per amino acid (Fig. 2.7). Residues 7-50 of protein A were not moved, because these residues interact with the RNA. We generated the positively charged N-terminal tails of both proteins B and C as linear chains extending radially inward toward the center of the virus (Fig. 2.7). The C-terminal tails of proteins B and C were generated as random coils, because they are not charged.

The gap between residues 379 and 393 of protein A was closed by a random coil connected to those residues, using a Monte Carlo algorithm, as follows. Given the first pseudoatom in the chain, the algorithm first generates trial coordinates for the second pseudoatom at a fixed distance from the first, but in a random direction from it. If the new pseudoatom is within 3.0 Å of any other atom, the trial position is rejected, and a new one is generated. Repeating this process eleven times generates a twelve-residue chain of random configuration. This chain is rotated into a position where it lies in the gap between residues 379 and 393 of protein A; energy minimization yields a conformation that closes that gap.

After generating the missing residues, the complete coarse-grained capsid was generated by applying icosahedral transformation matrices to the asymmetric unit (Fig.

2.7). The coarse-grained capsid was compressed in a series of steps, with each step followed by steepest descent minimization of the protein tails, while keeping the rest of the capsid proteins and the RNA fixed. The protein tails are pulled toward the center of coordinates and penetrate into the genomic RNA. The force field terms and parameters used for the different components of the coarse-grain model are summarized in Table 2.3.

In the expanded framework, the interior of the capsid is a distance $D \sim 300 \text{ \AA}$ from the outside of the RNA (Fig. 2.7). We divided the process of compressing the capsid to its correct size into two stages. The first stage consisted of a series of nine scalings, each of which moved the capsid inward by a distance $0.1D$, and each of which was followed by extensive minimization. At this point, it becomes more difficult to resolve steric problems with large scaling steps, so the second stage consisted of a series of five scaling steps, moving the capsid inward $0.02D$ at each step, each followed by extensive minimization.

Table 2.3: Energy terms used for the protein component of the coarse-grain model for Pariacoto virus, and for the protein–RNA volume exclusion term.

Energy term	Atoms affected	Equation	Parameters
Bond	Flexible Tails ($C\alpha$ model)	Eq. (1)	$k_b = 3 \text{ kcal}/(\text{mol } \text{\AA}^2)$, $b_0 = 3.8 \text{ \AA}$
Angle	Flexible Tails ($C\alpha$ model)	Eq. (2)	$k_\theta = 3 \text{ kcal/mol}$, $\theta_0 = 1.94 (111.154^\circ)$
Volume exclusion	Outer Capsid (12C model)	Eq. (3)	$k_\theta = 3 \text{ kcal}/(\text{mol } \text{\AA}^2)$, $d_0 = 35.0 \text{ \AA}$
	Inner Capsid ($2C\alpha$ model)		$k_\theta = 3 \text{ kcal}/(\text{mol } \text{\AA}^2)$, $d_0 = 7.6 \text{ \AA}$
	Flexible Tails ($C\alpha$ model)		$k_\theta = 3 \text{ kcal}/(\text{mol } \text{\AA}^2)$, $d_0 = 3.8 \text{ \AA}$
	Tails/RNA ($C\alpha/2N$)		$k_\theta = 3 \text{ kcal}/(\text{mol } \text{\AA}^2)$, $d_0 = 12.5 \text{ \AA}$

After the final step of compression/minimization, we converted the protein tails into an all-atom model using PULCHRA (52) and connected these with the rest of the all-atom protein crystal structure. Since the RNA had not been allowed to move during the modeling of the protein tails, we simply replaced the coarse-grained RNA model with the all-atom model described above. The final all-atom model of PaV was further minimized with NAMD, using the CHARMM27 force field (53), with all protein and RNA atoms free to move. This eliminates any unacceptable steric conflicts and gives bond lengths and angles within standard ranges.

Pariacoto virus: Results

The final model of PaV is shown in Fig. 2.8. We generated two different models, to determine the energetic consequences of allowing the polycationic protein tails to penetrate deeply into the viral interior vs. having them associate predominantly with RNA in the outer regions. The first was achieved with the tail-RNA soft sphere contact distance $d_0 = 8\text{\AA}$, while a larger contact distance ($d_0 = 12\text{\AA}$) provides less penetration. We evaluated the electrostatic energies of these two models, finding that deep penetration does, as expected, provide substantial additional stabilization (44).

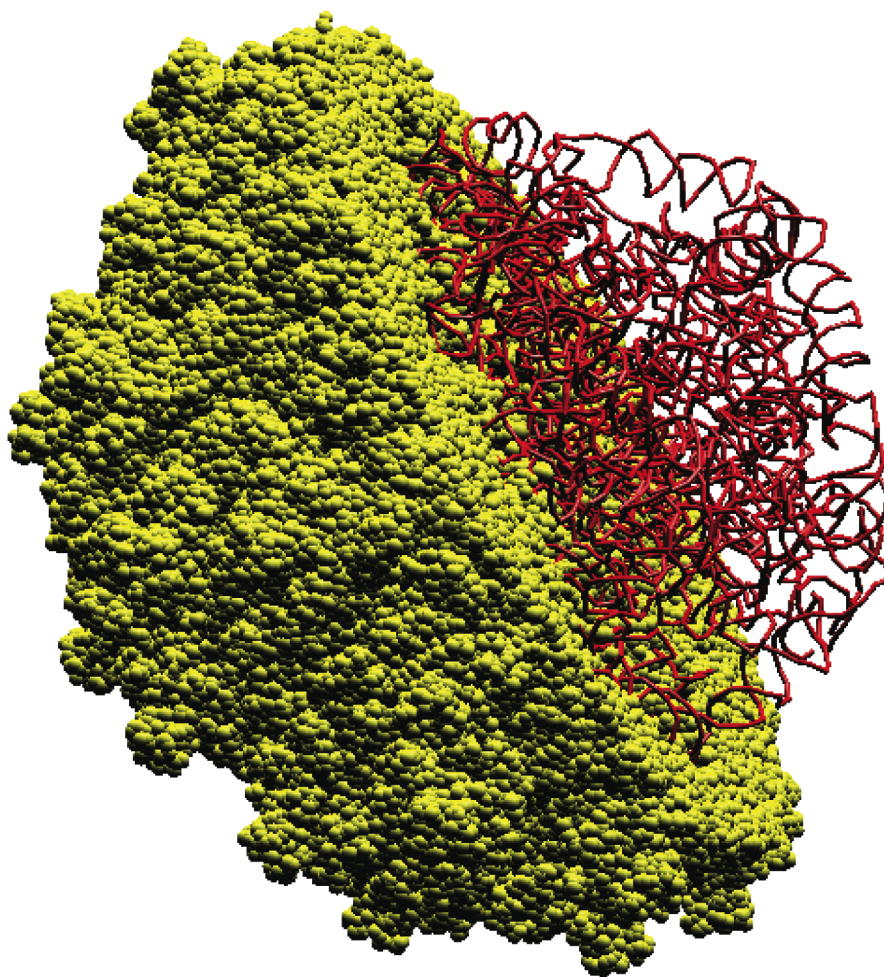


Figure 2.8. Final all-atom model of pariacoto virus. Half of the protein capsid is shown, with all non-hydrogen atoms represented as van der Waals spheres. The RNA model also specifies the coordinates of all non-hydrogen atoms, but only the backbone trace is shown here, for clarity. Some RNA double helices that are part of the dodecahedral cage are clearly seen around the periphery. Reprinted from [37]

This study also led to a new model for the assembly of icosahedral single-stranded RNA viruses like PaV, which are quite different from bacteriophage. Phage capsids are formed from proteins that interact strongly with one another, so that capsid formation is the first step in viral assembly, and the DNA must be loaded into the empty

capsid by an ATP-driven motor. In contrast, protein-protein interactions in PaV are weak, and capsid formation requires the presence of the viral genome. We have suggested that assembly begins with the condensation of the RNA by the polycationic protein tails, and that this compaction leaves the globular protein cores in a spherical shell surrounding the condensate, where their effective concentration is high enough to drive the cooperative association of those globular cores into the mature capsid (44).

References

1. Jardine PJ & Anderson DL (2006) DNA packaging in double-stranded DNA phages. *The Bacteriophages*, ed Calendar R (Oxford University Press, Oxford), 2nd Ed, pp 49-65.
2. Johnson JE & Chiu W (2007) DNA packaging and delivery machines in tailed bacteriophage. *Curr Opin Struct Biol* 17:237-243.
3. Petrov AS & Harvey SC (2008) Packaging double-helical DNA into viral capsids: structures, forces, and energetics. *Biophys J* 95:497-502.
4. Knobler CM & Gelbart WM (2009) Physical chemistry of DNA viruses. *Annu Rev Phys Chem* 60:367-383.
5. Ackermann H-W & DuBow MS (1987) *Viruses of prokaryotes* (CRC Press, Boca Raton, Fla.).
6. Granoff A & Webster RG eds (1999) *Encyclopedia of virology* (Academic Press, San Diego, Ca).
7. Purohit PK, *et al.* (2005) Forces during bacteriophage DNA packaging and ejection. *Biophys J* 88:851-866.
8. Smith DE, *et al.* (2001) The bacteriophage phi29 portal motor can package DNA against a large internal force. *Nature* 413:748-752.
9. Locker CR & Harvey SC (2006) A model for viral genome packing. *Multiscale Model Simul* 5:1264-1279.
10. Rollins GC, Petrov AS, & Harvey SC (2008) The role of DNA twist in the packaging of viral genomes. *Biophys J* 94:L38-40.

11. Tan RK, Petrov AS, & Harvey SC (2006) YUP: A molecular simulation program for coarse-grained and multi-scaled models. *J Chem Theory Comput* 2:529-540.
12. Berman HM, *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235-242.
13. Hagerman P (1988) Flexibility of DNA. *Annu Rev Biophys Biophys Chem* 17:265-286.
14. Locker CR, Fuller SD, & Harvey SC (2007) DNA organization and thermodynamics during viral packaging. *Biophys J* 93:2861-2869.
15. Tan RKZ & Harvey SC (1989) Molecular mechanics model of supercoiled DNA. *J Mol Biol* 205:573-591.
16. Tan RK-Z, Petrov AS, Devkota B, & Harvey SC (2009) Coarse-grained models for nucleic acids and large nucleoprotein assemblies. *Coarse-Graining of Condensed Phase and Biomolecular Systems*, ed Voth GA (CRC Press, Boca Raton, FL), pp 225-236.
17. Tan RK-Z, Sprous D, & Harvey SC (1996) Molecular dynamics simulations of small DNA plasmids: Effects of sequence and supercoiling on intramolecular motions. *Biopolymers* 39:259-278.
18. Parsegian VA, Rand RP, & Rau DC (1995) Macromolecules and water: Probing with osmotic stress. *Energetics Of Biological Macromolecules*, Methods In Enzymology), Vol 259, pp 43-94.
19. Parsegian VA, Rand RP, & Rau DC (2000) Osmotic stress, crowding, preferential hydration, and binding: A comparison of perspectives. *Proc Natl Acad Sci USA* 97:3987-3992.
20. Bloomfield VA (1991) Condensation of DNA by multivalent cations: Considerations on mechanism. *Biopolymers* 31:1471-1481.
21. Hud NV & Vilfan ID (2005) Toroidal DNA condensates: Unraveling the fine structure and the role of nucleation in determining size. *Annu Rev Biophys Biomol Struct* 34:295-318.
22. Rau DC, Lee B, & Parsegian VA (1984) Measurement of the repulsive force between polyelectrolyte molecules in ionic solution: Hydration forces between parallel DNA double helices. *Proc Natl Acad Sci USA* 81:2621-2625.
23. Petrov AS & Harvey SC (2007) Structural and thermodynamic principles of viral packaging. *Structure* 15:21-27.
24. Tzllil S, Kindt JT, Gelbart WM, & Ben-Shaul A (2003) Forces and pressures in DNA packaging and release from viral capsids. *Biophys J* 84:1616-1627.

25. Rau DC & Parsegian VA (1992) Direct measurement of the intermolecular forces between counterion-condensed DNA double helices. *Biophys J* 61:246-259.
26. Lander GC, *et al.* (2008) Bacteriophage lambda stabilization by auxiliary protein gpD: Timing, location, and mechanism of attachment determined by cryo-EM. *Structure* 16:1399-1406.
27. Fuller DN, *et al.* (2007) Measurements of single DNA molecule packaging dynamics in bacteriophage lambda reveal high forces, high motor processivity, and capsid transformations. *J Mol Biol* 373:1113-1122.
28. Dokland T & Murialdo H (1993) Structural transitions during maturation of bacteriophage-lambda capsids. *J Mol Biol* 233:682-694.
29. Agirrezabala X, *et al.* (2005) Structure of the connector of bacteriophage T7 at 8 angstrom resolution: Structural homologies of a basic component of a DNA translocating machinery. *J Mol Biol* 347:895-902.
30. Jiang W, *et al.* (2006) Structure of epsilon15 bacteriophage reveals genome organization and DNA packaging/injection apparatus. *Nature* 439:612-616.
31. Lander GC, *et al.* (2006) The structure of an infectious P22 virion shows the signal for headful DNA packaging. *Science* 312:1791-1795.
32. Petrov AS, Lim-Hing K, & Harvey SC (2007) Packaging of DNA by bacteriophage epsilon15: structure, forces, and thermodynamics. *Structure* 15:807-812.
33. Berendsen HJC, Postma JPM, Vangunsteren WF, Dinola A, & Haak JR (1984) Molecular-Dynamics With Coupling To An External Bath. *J Chem Phys* 81:3684-3690.
34. Frank J (2002) Single-particle imaging of macromolecules by cryo-electron microscopy. *Annu Rev Biophys Biomol Struct* 31:303-319.
35. Jeembaeva M, Castelnovo M, Larsson F, & Evilevitch A (2008) Osmotic pressure: Resisting or promoting DNA ejection from phage? *J Mol Biol* 381:310-323.
36. Evilevitch A, Lavelle L, Knobler CM, Raspaud E, & Gelbart WM (2003) Osmotic pressure inhibition of DNA ejection from phage. *Proc Natl Acad Sci U S A* 100:9292-9295.
37. Harvey SC, Petrov AS, Devkota B, & Boz MB (2009) Viral assembly: a molecular modeling perspective. *Phys Chem Chem Phys* 11:10553-10564.

38. Petrov AS, Boz MB, & Harvey SC (2007) The conformation of double-stranded DNA inside bacteriophages depends on capsid size and shape. *J Struct Biol* 160:241-248.
39. Spakowitz AJ & Wang ZG (2005) DNA packaging in bacteriophage: is twist important? *Biophys J* 88:3912-3923.
40. Grayson P & Molineux IJ (2007) Is phage DNA 'injected' into cells-biologists and physicists can agree. *Curr Opin Microbiol* 10:401-409.
41. Tang L, *et al.* (2001) The structure of pariacoto virus reveals a dodecahedral cage of duplex RNA. *Nature Struct Biol* 8:77-83.
42. Shepherd CM, *et al.* (2006) VIPERdb: a relational database for structural virology. *Nucleic Acids Res* 34:D386-389.
43. Tihova M, *et al.* (2004) Nodavirus coat protein imposes dodecahedral RNA structure independent of nucleotide sequence and length. *J Virol* 78:2897-2905.
44. Devkota B, *et al.* (2009) Structural and electrostatic characterization of Pariacoto virus: Implications for viral assembly. *Biopolymers* 91:530-538.
45. Malhotra A, Tan RK, & Harvey SC (1994) Modeling large RNAs and ribonucleoprotein particles using molecular mechanics techniques. *Biophys J* 66:1777-1795.
46. Cui Q, Tan RK, Harvey SC, & Case DA (2006) Low-resolution molecular dynamics simulations of the 30S ribosomal subunit. *Multiscale Modeling Simul* 5:1248-1263.
47. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 31:3406-3415.
48. Pearlman DA, *et al.* (1995) AMBER: A computer program for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to elucidate the structures and energies of molecules. *Comput Phys Commun* 91:1-41.
49. Plimpton S (1995) Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J Comp Phys* 117:1-19.
50. Tan RK-Z & Harvey SC (1993) Yammp: Development of a molecular mechanics program using the modular programming method. *J Comput Chem* 14:455-470.
51. Parisien M & Major F (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* 452:51-55.

52. Rotkiewicz P & Skolnick J (2008) Fast procedure for reconstruction of full-atom protein models from reduced representations. *J Comput Chem* 29:1460-1465.
53. MacKerell AD, Jr., Banavali N, & Foloppe N (2000) Development and current status of the CHARMM force field for nucleic acids. *Biopolymers* 56:257-265.

CHAPTER 3

STRUCTURAL AND ELECTROSTATIC CHARACTERIZATION OF PARIACOTO VIRUS: IMPLICATIONS FOR VIRAL ASSEMBLY

Abstract:

We present the first all-atom model for the structure of a T=3 virus, Pariacoto virus (PaV), which is a non-enveloped, icosahedral RNA virus and a member of the *Nodaviridae* family. The model is an extension of the crystal structure, which reveals about 88% of the protein structure but only about 35% of the RNA structure. Evaluation of alternative models confirms our earlier observation that the polycationic protein tails must penetrate deeply into the core of the virus, where they stabilize the structure by neutralizing a substantial fraction of the RNA charge. This leads us to propose a model for the assembly of small icosahedral RNA viruses: the nonspecific binding of the protein tails to the RNA leads to a collapse of the complex, in a fashion reminiscent of DNA condensation. The globular protein domains are excluded from the condensed phase but are tethered to it, so they accumulate in a shell around the condensed phase, where their concentration is high enough to trigger oligomerization and formation of the mature virus.

Introduction:

Pariacoto virus (PaV), a T=3, non-enveloped, icosahedral virus is a member of the *Nodaviridae* family. It was originally isolated in Peru from the Southern armyworm, *Spodoptera eridania* [1]. Its genome consists of two positive-sense ssRNAs [2]. RNA1 (3011 nucleotides) codes for protein A, the catalytic subunit for the host RNA replicase,

which enables the RNA-dependent RNA replicase to start replicating the viral RNA. RNA2 (1311 nucleotides) codes for capsid precursor protein α . 180 of these α proteins and the genome assemble together to make up the virus. Ever since it was isolated, PaV has been extensively studied using various techniques [3-6]. The relatively small size (20nm in diameter) compared to other RNA viruses, and the ease by which it can be produced in various cell lines [7], make PaV and other members of the Nodaviridae family easy to characterize at the molecular level [8-10].

Structural studies of viruses are very important to understand the protein-protein and protein-RNA interactions as well as to understand assembly pathways in RNA viruses [11-14]. In the last few years, many studies have been done on RNA viruses using molecular modeling as a supplementary method when other methods such as x-ray crystallography and cryo-electron microscopy (cryo-EM) do not give sufficient structural information. An all-atom model was derived for a Satellite Tobacco Mosaic Virus (STMV), a T=1 virus, using molecular modeling [15]. Subsequently, molecular dynamics was done on the model to study the stability of the protein capsid and the RNA genome [15]. Electrostatic interactions between RNA and the protein capsid were studied in Cowpea Chlorotic Mottle Virus (CCMV) by modeling the virus using coarse-grained modeling and representing RNA nucleotides by spheres that were distributed using the Monte Carlo method [16]. In addition, electrostatic properties of virus capsids and RNA have also been studied to understand the structural properties and the molecular interactions within the virus [17, 18].

The 3.0Å x-ray crystal structure of PaV reveals an asymmetric unit with three quasi-equivalent protein subunits (A, B and C) and one strand of a 25 base pair RNA

duplex [6]. Sixty of these units combine to form the icosahedral capsid, with 30 RNA duplexes lying along subunit contacts across the icosahedral 2-fold axes, forming a dodecahedral cage inside the capsid. The A, B, and C subunits (residues 83-321) are folded into an eight-stranded antiparallel β -sandwich, similar to proteins in other nodaviruses. Complementing the x-ray studies, cryo-electron microscopy showed the general overall structure of PaV at 23Å resolution, which matched well with the low-resolution model calculated from the atomic coordinates [6]. Cryo-EM also confirmed that the part of the RNA genome that was resolved in the x-ray structure forms the edges of the dodecahedral cage inside the protein capsid.

Although x-ray crystallography and cryo-EM provided a lot of information regarding the PaV structure, they were not able to determine the atomic structure of the complete virus. RNA at the dodecahedral edges accounts for only 35% of the total genome. The remaining 65% of the RNA lies inside the dodecahedral cage and is not resolved in the crystal structure because it lacks icosahedral symmetry. In addition, the 20 vertices at which the RNA duplexes are connected could not be resolved, presumably because different vertices have different structures. Similarly, protein subunit A is missing 6 residues at the N terminal end and 15 at the C-terminal in the crystal structure, while the B and C subunits are missing about 50 residues at the N-terminus and 19 residues at the C-terminus in the crystal structure [6].

In this paper, we report a model for the complete virus and examine the interactions of the basic N-terminus tails with the RNA genome, and their role in the stability of PaV. We used molecular modeling to model the missing 65% of the genome and the unresolved protein residues. We built our models using coarse-grained modeling,

representing unresolved nucleotides and amino acids by pseudoatoms and interpolating the pseudoatomic models to all-atom using special algorithms. We generated two all-atom models for the virus that differed in the conformations of the N-terminus protein tails and the extent to which they penetrate into the RNA genome. We tested these against the experimental radial density distributions from cryo-EM, and we evaluated the relative stabilities of the two models by comparing their energies. The result is the first all-atom model for a complete T=3 virus. Further, this effort has led to a new model for the assembly of small, non-enveloped icosahedral RNA viruses.

Methods:

RNA modeling:

The modeling of the Pariacoto virus genome posed several challenges because of the limited amount of available structural data. To begin with, the secondary structure for the PaV genome is not known. We used a hypothetical secondary structure mapped onto the dodecahedral cage (Figure 3.1). This is the same secondary structure that we proposed earlier [19]. Those parts of the RNA genome that do not form the edges of the dodecahedral cage drop inwards towards the center of the capsid as “stalactites”. The exact number of these connections is not known, but we used a combination of 3-way junctions and 4-way junctions as structural motifs connecting the RNA on the dodecahedral cage with the RNA in the interior (Figure 3.2). Nothing at all is known about the RNA structure in the interior, so we have to postulate a collection of plausible structures for the stalactites. We used twelve copies of a structure derived from the *E. coli* ribosome domain IV (residues 1764-1988) to represent these. Although the twelve

the secondary structure of RNA1 and RNA2, nor the structure of the interactions between them, if any. Pink and green dots represent the 5' and 3' ends, respectively. Red circles with blue borders are the junctions where the stalactites were added to connect with RNA deeper in the interior of the capsid (see text). Reprinted from [32].

The crystal structure of PaV (1F8V.pdb) is available from the RCSB Protein Data Bank [20]. The dodecahedral RNA cage was generated by applying the BIOMT TRANSFORMATION matrix given in the file, using the oligomer generator tool in the Viper database [21]. The vertex structures were defined by the secondary structure (Figure 3.1). Each vertex had either three or four extensions of RNA coming out of it (Figure 3.2). Small hairpin loops were added at twelve vertices, as stubs to which the stalactites were subsequently added. We cut the RNA duplex on each edge in half, fixing each half to the appropriate vertex. This initial model was generated on a Silicon Graphics workstation using the Builder module of INSIGHT II graphics software. This initial model (Figure 3.3a) contained all 4322 RNA residues.

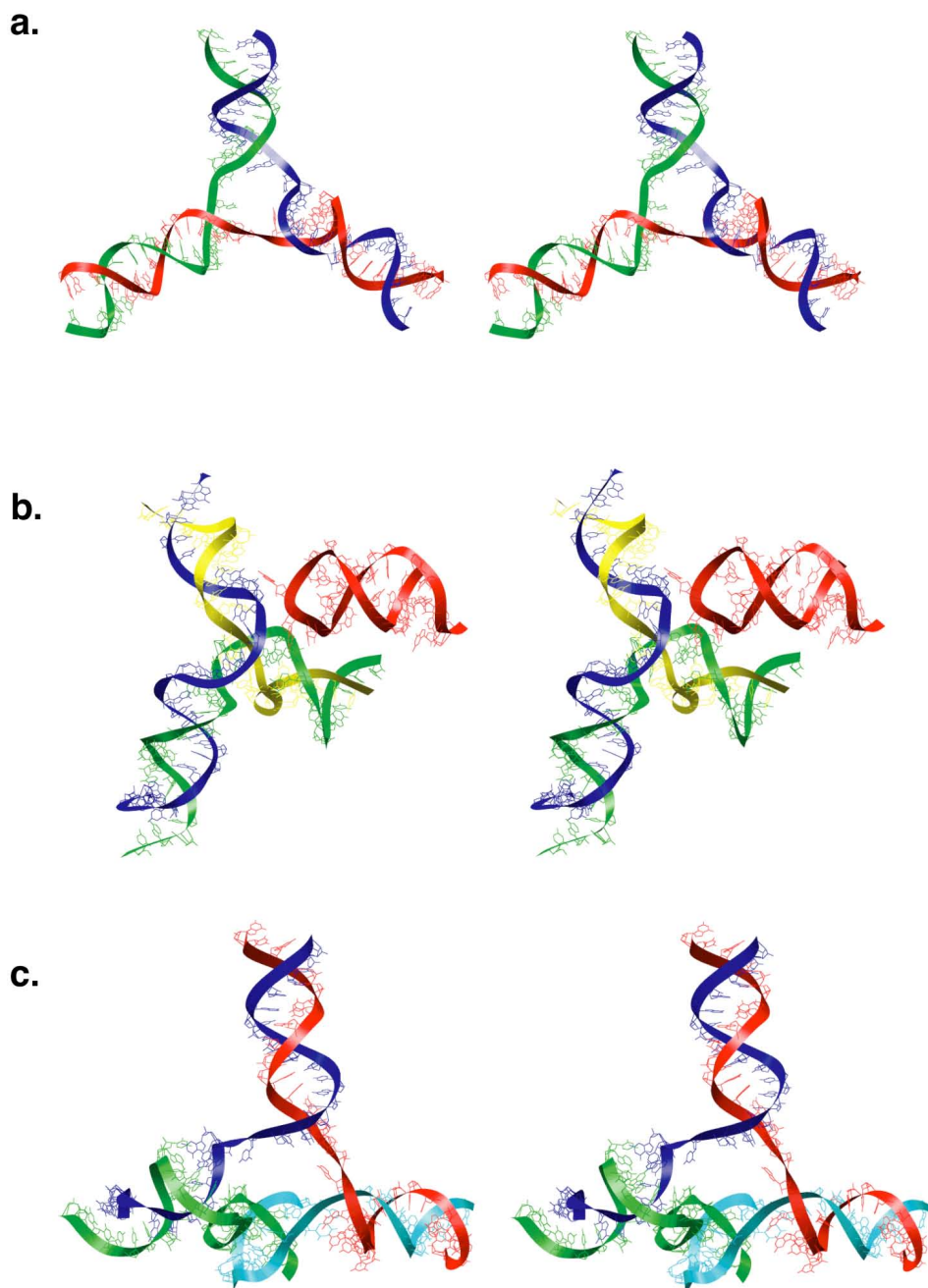


Figure 3.2: Stereo images of model junctions. a. A typical three-way junction. RNA duplexes line on three adjacent edges of dodecahedral cage, and there is no stalactite at the vertex. b. Another type of three-way junction, connecting duplexes on two edges with a stalactite. The stalactite is attached to the green and yellow helix. There is a stem-loop on the third edge, coming from a neighboring vertex (red). c. A four-way junction, connecting duplexes on three edges with a helix (blue and green) that is the attachment point for a stalactite. Reprinted from [32].

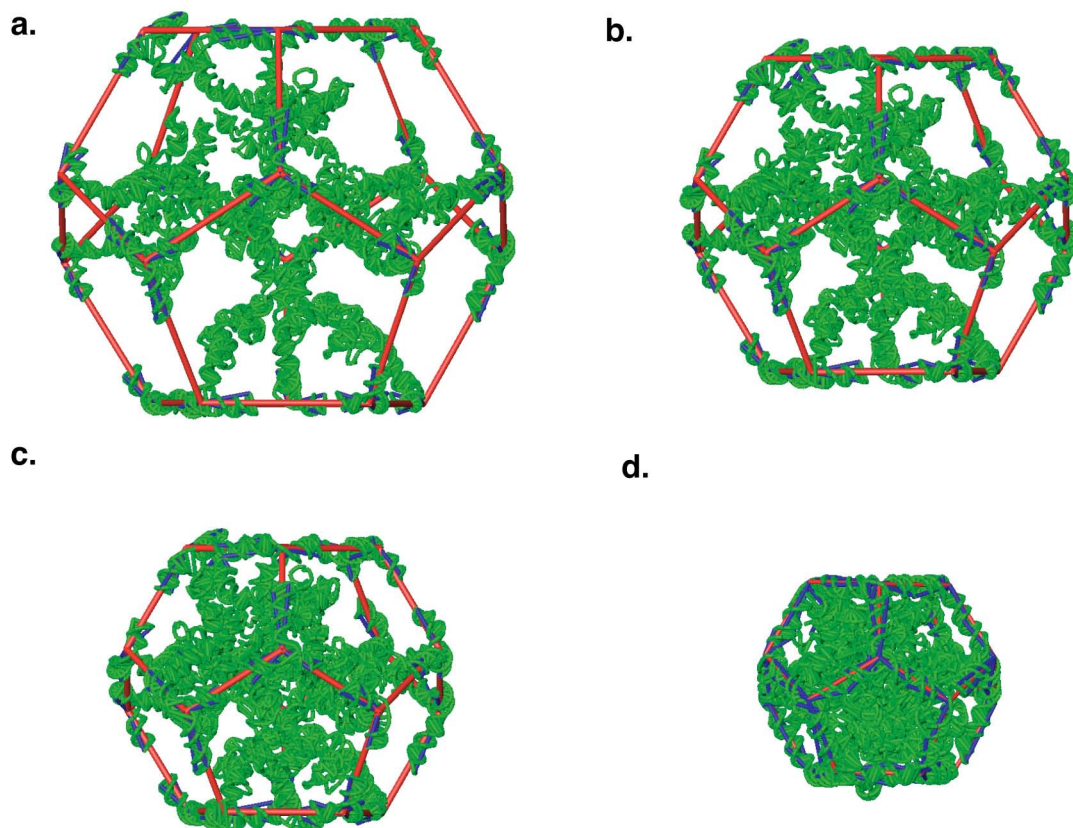


Figure 3.3: Minimization protocol for the viral RNA. a. The initial model with the diameter of the dodecahedral cage doubled (red lines). RNA duplexes are cut at the middle and rigidly attached to their corresponding vertex atoms. Pseudo-bonds from each vertex atom to the edge of the RNA duplex are represented as blue lines. These bonds restrain the crystallographic regions during minimization. The stalactites can be seen inside the dodecahedral cage. The volume of the cage is eight times the volume of the actual cage. b. The model after four rounds of minimization, at about six times the actual volume. c. The model after eight rounds of minimization, at about three times the actual volume. d. The final model, after twelve rounds of minimization. Reprinted from [32].

The initial model is quite large and the experimental data available for modeling are quite limited, so coarse-grained modeling is appropriate for refining the model. We converted the all-atom initial model to coarse-grain representation, with each nucleotide represented by a pseudo-atom at the phosphate position. A more complete description of this “all-P” model is available elsewhere [22], along with a full description of the corresponding force field. Twenty pseudo-atoms were also added at the vertices of the

dodecahedral cage, to form a framework that could be easily expanded and contracted; we call these “vertex pseudo-atoms”.

The edges of the dodecahedral cage were decreased to the original length in multiple steps, decreasing the ideal bond length (b_0) of the expanded framework in 5 Å steps and minimizing until convergence after each step (Figure 3.3). The minimization was done using our in-house molecular mechanics package, YAMMP [22]. The harmonic energy terms used in the minimization are tabulated in Table 3.1. Since all the terms used in the potential energy function of all-P models are harmonic, full minimization of the model should lead to zero energy, if all restraints can be satisfied.

During minimization, the stalactite RNAs were free to move and adjust their conformations, to avoid steric overlap. They had softer force constants in the energy terms than did the RNA domains on the dodecahedral cage (Table 3.1). The crystallographic regions were restrained by using strong force constants in the energy terms, and by the addition of pseudo-bonds connecting each vertex pseudo-atom to the ends of the RNA duplexes on each edge (Figure 3.2). These regions did not deviate significantly from the crystal structure during the contraction/minimization cycles.

Table 3.1: Energy terms used in the RNA modeling.

Energy terms	Types	Equation	Force constant
Bond	Crystallographic	$E_b = k_b (b - b_0)^2$	$k_b = 20 \text{ kcal/mol}$
	Stalactites	where b_0 is the distance in the initial model derived from crystal structure.	$k_b = 2 \text{ kcal/mol}$
Angle	Crystallographic	$E_\theta = k_\theta (\theta - \theta_0)^2$	$k_\theta = 20 \text{ kcal/mol}$
	Stalactites	where θ_0 is the angle in the initial model	$k_\theta = 2 \text{ kcal/mol}$
Improper torsion	Crystallographic	$E_i = k_i (i - i_0)^2$	$k_i = 20 \text{ kcal/mol}$
	Stalactites	where i_0 is the improper torsion between four atoms in the initial model	$k_i = 2 \text{ kcal/mol}$
Non-bond exclusion		$E_{nbn} = k_{ij} (d - d_{ij})^2$, if $d < d_0$, $d_0 = 10 \text{ \AA}$	$k_{ij} = 2 \text{ kcal/mol}$
NOE term		$E_{noe} = \begin{cases} k_{hi} (r - r_{hi})^2, & \text{if } r > r_{hi} \\ 0, & \text{if } r \leq r_{hi} \end{cases}$ where r is the distance from the center to atom i . r_{hi} was changed during each step of minimization.	$k_{hi} = 2 \text{ kcal/mol}$
Stud	Stud atom was kept at the center to keep the RNA within a certain radius.	$E_{st} = k_{st} [(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2]$ where (x, y, z) is the current position of atom i and (x_0, y_0, z_0) is the desired position.	$k_{st} = 40 \text{ kcal/mol}$

Generating an all-atom model from phosphate positions is a challenging problem. The bond and angle restraints in the all-P models are based on observed distributions of P-P distances and P-P-P angles in the Nucleic Acid Database [22]. With only these restraints, there is no way to guarantee that groups of four or more successive P atoms in any all-P model will have a conformation that corresponds to any real RNA structure. As a consequence, all-atom models can be generated fairly easily in double-helical regions, but all-atom models for other regions (loops, bulges, single-strands) are necessarily more speculative. This is not inappropriate, considering the modesty of our overall goal: generate a plausible RNA model, in terms of connectability along the backbone and the

absence of serious steric problems. A more rigorous structural effort would not be justified, because we don't know the actual secondary structure of the PaV RNAs, and there are no high-or intermediate-resolution data on the RNA structure, except within the dodecahedral cage.

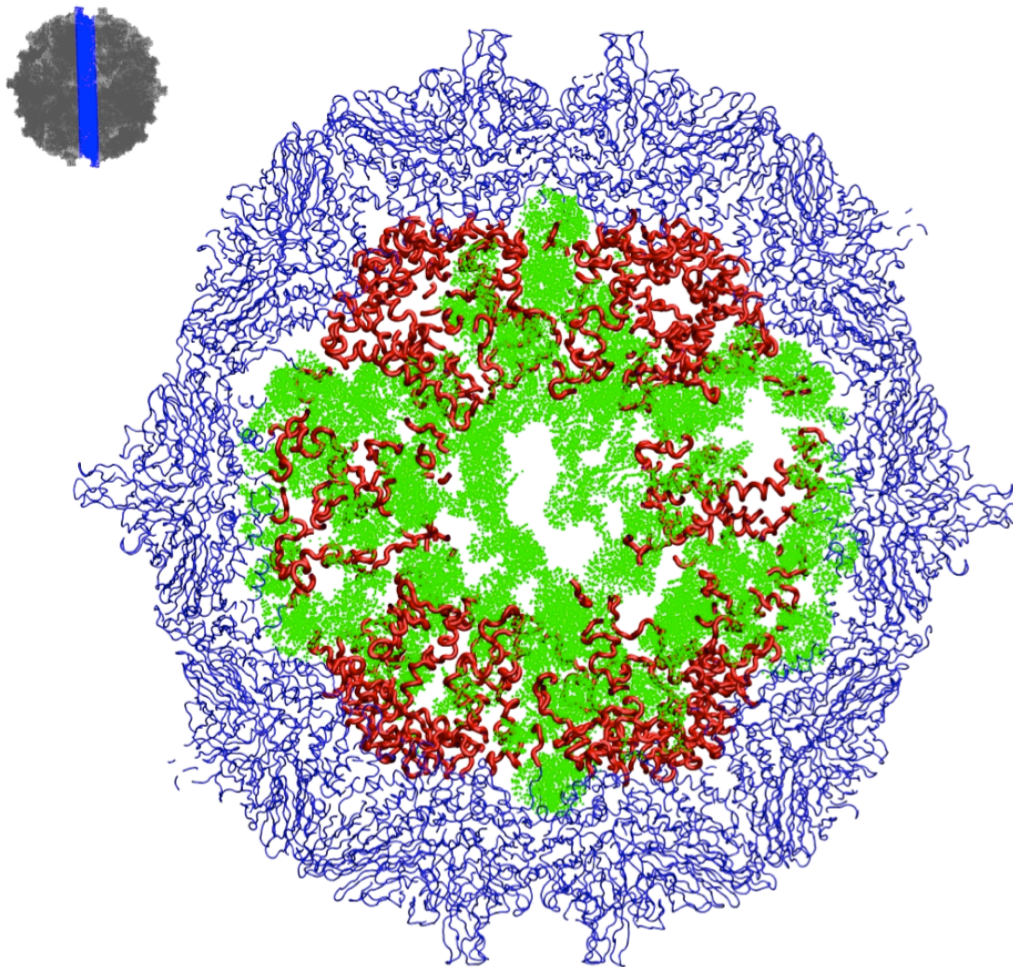


Figure 3.4: A 20Å slice through the center of Model_8. Protein residues seen in the crystal structure are colored blue, while noncrystallographic residues are red. The RNA is green. The protein tails reach very close to the center of the structure. Reprinted from [32].

Briefly, the procedure used here builds all-atom models using a database of nucleotide conformations derived from all RNA-containing structures in the PDB as of April, 2006. In base-paired regions, four phosphate positions (0 and +1 on each strand)

serve as anchor points, and a pair of nucleotides from the database must be fit to the structure, one on each strand. In non-base-paired regions, the four anchor phosphates are those -1, 0, +1 and +2 relative to the nucleotide being placed. The compatibility of all examples in the database with a particular position is assessed by requiring that the base be identical to the one being modeled, and that the root mean square deviations of the four phosphate positions in the example be within 1.5 Å of the anchor phosphates in the all-P model. Only examples that pass this compatibility test are kept within the search space of each nucleotide.

The modeling problem then becomes one of exploring the search space of the whole molecule to determine which combination of examples gives the most plausible structure, where plausibility is defined as the lowest energy (van der Waals plus electrostatics, using the AMBER 8 force field). This optimizes base pairing and stacking, while minimizing steric clashes. Searching is done in a piecewise fashion, focusing on individual regions, to optimize performance. The most plausible structure is then refined by optimization of the ribose conformations, followed by energy minimization and a short annealing of the entire model, using molecular dynamics.

Protein modeling:

For modeling the missing protein residues, we followed a similar methodology as in the case of RNA modeling, expanding the capsid, adding missing amino acids, and then shrinking the capsid back to its original size in multiple steps, with minimization at each step. Coarse-grain modeling was the initial step in modeling the missing residues of the capsid proteins. After refinement of the coarse-grain model was complete, it was converted to an all-atom model, followed by final refinement.

First, the capsid was expanded three times in length by simply multiplying the coordinates of the capsid atoms by 3. The crystallographic residues facing towards the RNA were converted into a model where two consecutive residues are represented by a pseudo-atom (2C-model). The rest of the crystallographic residues were represented by twelve pseudo-atoms each, defining the face, edge and the vertices of the equilateral triangle of each asymmetric unit. The missing N-terminal residues were generated in extended linear form pointing towards the RNA genome at the center. C-terminal residues were generated as a random coil. Residues for both the N-and C-terminal tails were represented by one pseudo-atom per residue (Figure C.2).

The starting capsid model was scaled back down to the original size in a series of steps, testing different scaling factors and Van der Waals (vdw) diameters for the pseudoatoms of the protein tails. We examined scaling ratios between 0.95 to 0.99, finding that different scaling ratios did not significantly affect the configurations of the protein tails (data not shown). However, changing the vdw diameters from 8 to 12 Å significantly affected the penetration of the protein tails into the RNA genome (Figure 3.5b). The resulting structures, designated model_8 and model_12, have dramatically different conformations for the protein tails. In model_8, the tails penetrate deeply into the RNA core, while they lie on the outside of the RNA core in model_12.

Model_8 and model_12 were converted into all-atom representation using PULCHRA (22). This program converts $C\alpha$ models to all-atom models using a rotamer library prepared from the statistics of $C\alpha$ distances in the PDB. The complete all-atom models, including all residues of the RNA genome and the capsid proteins, were energy minimized with NAMD, using the CHARMM forcefield.

Calculations of the electrostatic potential were performed using the Adaptive Poisson-Boltzmann Solver (APBS) (23). CHARMM27 forcefield radii and charges were assigned to the minimized all-atom structures of Model_8 and Model_12 using the PDB2PQR (24) routine, yielding a charge of $+46e$ for each of the 60 capsomers and $-4320e$ for the RNA genome, where e is the charge on the proton. This resulted in a net charge of $-1560e$ for the complete virus. The nonlinear version of the Poisson-Boltzmann equation was solved numerically on a $225 \times 225 \times 225$ grid with an initial grid spacing of 2.0 \AA , followed by focusing with the grid spacing reduced to 1.5 \AA . The dielectric constants of the interior and exterior of the macromolecules were set to 10 and 78.5, respectively. The ionic strength was set to 100mM, using only monovalent ions. The resulting potentials were mapped onto the solvent accessible surface area of the models generated at the coarse-grained level and visualized using Chimera (25).

The coarse-grained pseudoatomic model of the genome was checked for the presence of possible knots using the “knot” program (26). Our RNA model does not contain any knots. The all-atom genome model reconstructed from the pseudoatomic model was also checked for interpenetration of rings and correct stereochemistry using PROCHECK, provided in the RCSB PDB website (<http://www.pdb.org>). There are no ring penetrations or other stereochemical problems. The RNA and protein distributions inside the complete all-atom models of the virus were compared with the native virus by generating density maps and corresponding radial density distribution functions (Figure 3.5) from the final all-atom models, using SPIDER (27).

Results and Discussions

The 65% of the genome that was not resolved in the crystal structure was generated and packaged within the dodecahedral cage. Even though all twelve stalactites had the same starting structures, they have significantly different conformations in the final model (Figure C.1). The protein tails missing in the crystal structure were also generated, and their final conformations also vary significantly from one another in the final model.

The generation of two models for PaV that differ in the distribution of the N-terminus protein tails offers an opportunity to study their role in stabilizing the virus. The different positions of the tails in the two models are reflected in different density distributions (Figure 3.5). In model_12 most of the tails are packed in a shell around 100Å from the center, which is between the genome and capsid. For model_8, many protein tails were able to penetrate deep inside the genome, and they contribute significantly to the density peak at a radius of about 50Å (Figure 3.5a). Peaks around this radius have been found in PaV (Figure 3.5b) and in other nodaviruses (19). Thus, structurally model_8 is structurally more consistent with native viruses than model_12. This is also consistent with density maps in Flock House virus (FHV), which is closely related to PaV. The radial density distribution for wild type FHV has a peak at $R \sim 32\text{\AA}$, but that peak is missing in mutant FHV in which 30 amino acids have been deleted from the amino terminus (19).

Single point energy calculation of the two models showed that model_8 is also energetically more favorable than model_12. The electrostatic interaction energy between the RNA and the capsid of PaV is much lower for model_8 (-3910 kcal/mol) than for model_12 (-523 kcal/mol). This agrees with the observations drawn from the structural

data (Figure 3.5): the protein tails that penetrate deep into the core of the virus stabilize PaV by neutralizing a large fraction of the charge of the RNA genome.

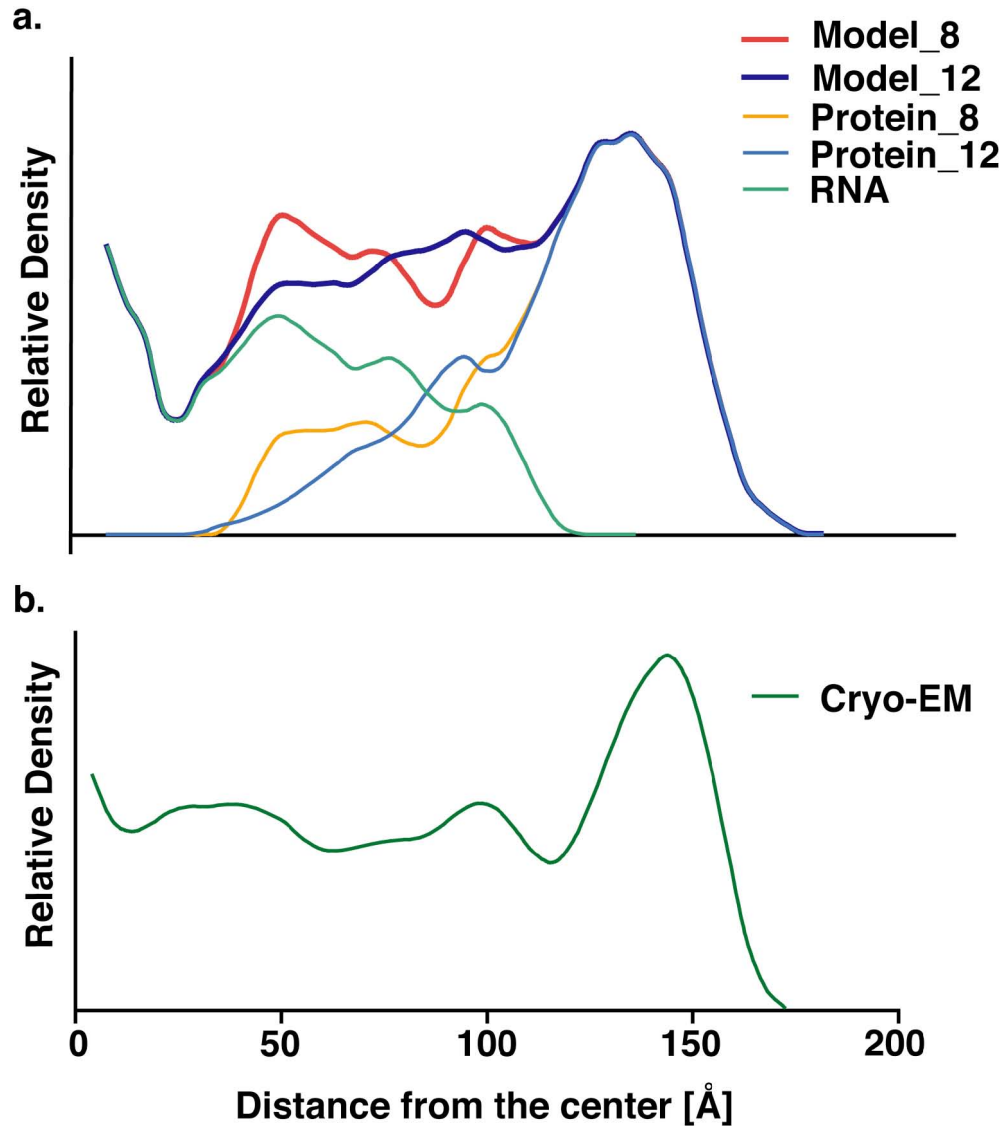


Figure 3.5: Comparison of model radial density distributions with the experimental distribution. a. Density distributions have been separated into RNA and protein components for model_8 and model_12. The peak at around 50Å for model_8 is due to the major contribution of the protein tails that penetrate deeply into the RNA core. For model_12, most of the protein tails are packed in a shell at a radius of ~100Å. **b.** Experimental cryo-EM density distribution.

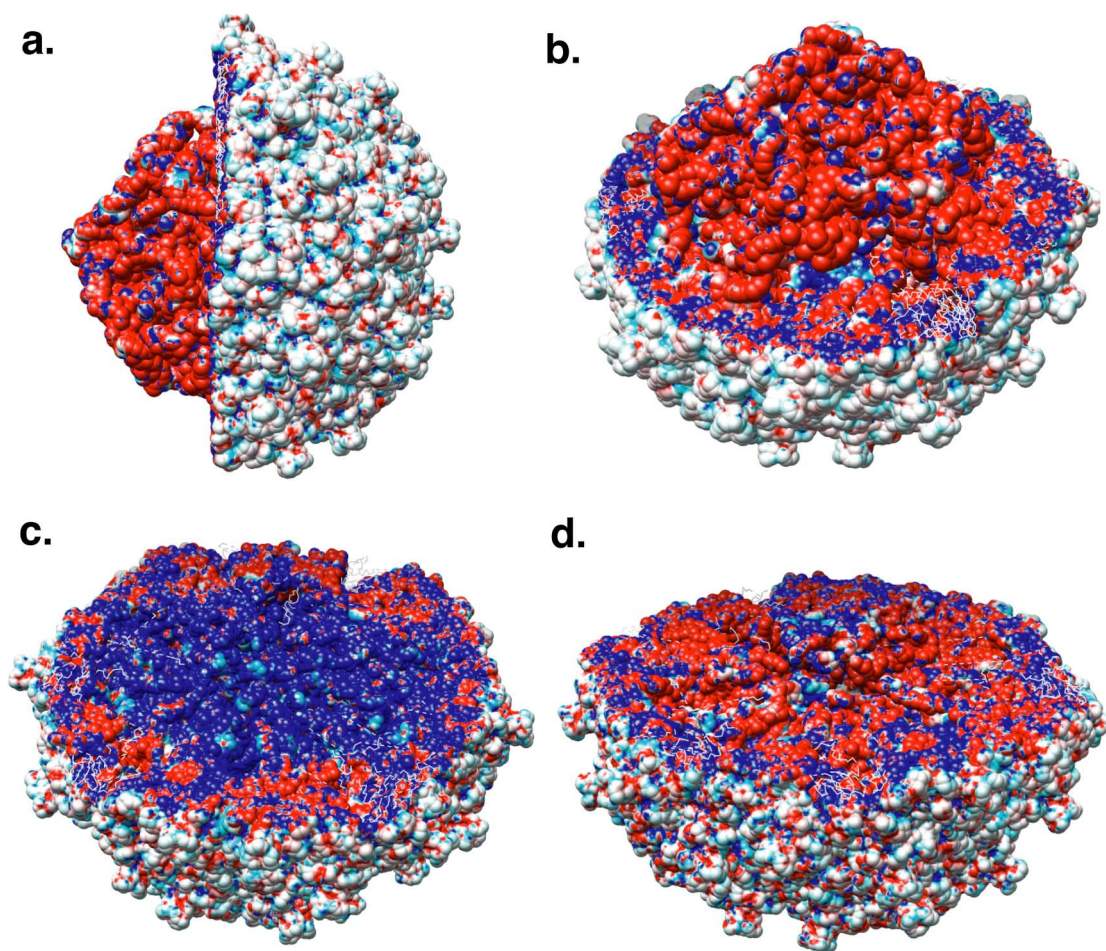


Figure 3.6: Electrostatic potential mapped onto the solvent-accessible surface area of PaV. The potential of the entire virus is mapped onto the surface of the RNA and one hemisphere of the capsid shell: a. side view; b. top view. c. Potential of the empty capsid mapped onto the surface of one hemisphere of the capsid. d. Potential of the entire virus mapped onto the surface of an empty hemisphere of capsid proteins. The color code of the electrostatic potential ranges from -5 kT/e (red) to 5 kT/e (blue).

Figure 3.6 depicts the electrostatic potential mapped onto the solvent accessible surface area of PaV. The external surface of PaV is almost neutral (Figure 3.6a), whereas the interior of the virus bears both positive charges (the protein tails) and negative charges (RNA). The lower panels of Figure 3.6 show the potential calculated for the virus without (Figure 3.6c) and with (Figure 3.6d) RNA, mapped onto the surface of the empty capsid. The positively charged tails (blue in Figure 3.6c) are fully neutralized and even reveal some negative potential on their surface due to the close proximity of RNA. The

latter observation is probably due to the fact that the total charge of RNA is almost factor of two greater than that of the capsid.

Conclusions:

There are three pieces of evidence that the polycationic protein tails penetrate deeply into the interior of nodavirus capsids. First, mutant FHV that lack 30 N-terminal amino acids lack the 32Å peak seen in cryo-EM radial density distribution profiles for wild-type FHV [19]. Second, our model 8 reproduces the experimental radial density distribution much better than model 12, and tails in the former penetrate much deeper into the capsid than those in the latter model. Finally, electrostatic calculations show that deep penetration of the tails has a stabilizing effect, because of more efficient neutralization of the RNA charge. This observation has important implications for viral assembly.

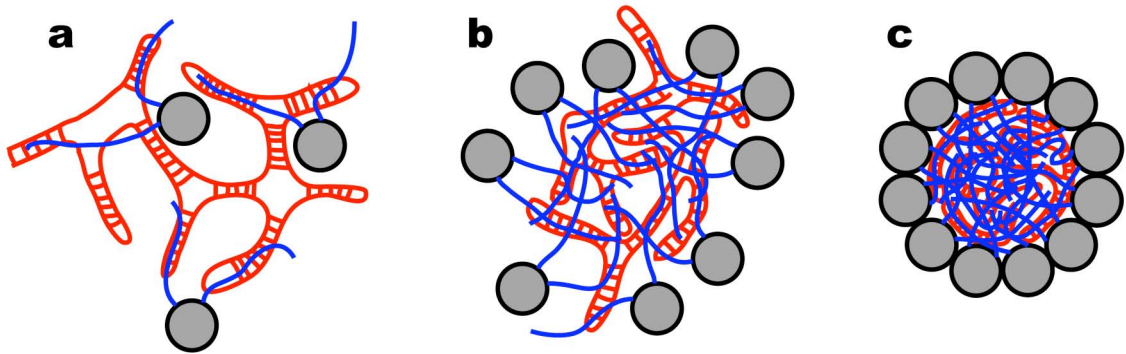


Figure 3.7: Model for assembly of small icosahedral RNA viruses. a. The polycationic N- and C-terminal protein tails bind nonspecifically to the RNA genome. **b.** When enough proteins are bound and the RNA charge is sufficiently neutralized, the complex collapses, in a process much like the condensation of DNA by polyvalent cations. The globular domains of the capsid proteins are tethered to the condensed RNA but are squeezed out and form a shell around it. **c.** The local concentration of the globular domains is high enough to promote oligomerization, leading to the formation of the mature capsid. Reprinted from [32].

The assembly of small icosahedral RNA viruses like PaV and FHV is quite different from bacteriophage. Interactions between phage capsid proteins are strong enough that capsids assemble spontaneously. The DNA genome is then forced into the pre-formed capsid by an ATP-dependent motor; there is little or no attraction between the DNA and the capsid proteins, in order to promote ejection of the genome upon infection of the host bacterium. In contrast, protein-protein interactions are weak in nodaviruses (capsids do not assemble spontaneously), and RNA-protein interactions are strongly attractive.

We propose a simple mechanism for the assembly of nodaviruses. Positively charged protein tails bind to the RNA (Figure 3.7a), with RNA replication, protein synthesis and RNA-protein binding occurring very closely in time and space (28, 29). When a sufficient quantity of the RNA charge is neutralized, the resulting complex collapses in a process reminiscent of DNA condensation (Figure 3.7b). We believe that most of these interactions are nonspecific, although in the mature virus there is evidence of a specific interaction between RNA2 and the N-terminal tail (30). In addition, the crystal structure (6) shows ordered interactions between the RNA and 36 N-terminal residues of subunit A, and between the RNA and eight residues of the C-terminus of subunit A, although the identity of the RNA in those interactions cannot be determined. We hypothesize that the globular domains of the capsid proteins are squeezed to the outside of the collapsed state, as shown in figure 3.7b. This provides a sufficiently high local concentration that the relatively weak protein-protein affinity is overcome, leading to oligomerization and the formation of the mature capsid (Figure 3.7c).

One remarkable observation suggests that this mechanism might apply to many single-stranded viruses. Belyi and Muthukumar examined 16 wild-type and 3 mutant viruses (both DNA and RNA viruses) with genomes ranging from about 1 kb to 12 kb (31). They found that the ratio of the genome size to the net charge on the terminal protein tails is 1.61 ± 0.03 , an unexpectedly uniform ratio. Such a narrow range might be

explained by our model, because the initial collapse would require sufficient charge neutralization to overcome RNA-RNA repulsions, but not so much as to lock the condensed state into a fixed configuration that could preclude the structural flexibility necessary for fitting the condensed mass into the final capsid structure.

This model provides a simple mechanistic basis for explaining how the relatively weakly associating proteins can force RNA into a small compact volume: the very strong electrostatic interactions between the polyanionic RNA and the polycationic protein tails provide a sufficiently favorable change in enthalpy to overcome the unfavorable entropic penalty associated with the dramatic reduction in RNA conformational space. It seems highly unlikely that a compact RNA structure would form first, followed by the formation of the protein capsid around it, as suggested earlier (15). The former is opposed by very strong forces, while the latter is driven by only weak ones.

In summary, we present the first all-atom model of a complete T=3 virus. Although there are insufficient experimental data to allow the development of a completely rigorous model, our model is consistent with all the available data, and it is sterically plausible. Most important, it leads to a simple mechanistic explanation of the assembly of small icosahedral RNA viruses. It will be exciting to test this model both experimentally and computationally.

Acknowledgments

We thank Robert K.-Z. Tan and Thomas R. Caulfield for their valuable insights and discussions. Supported by a grant from the NIH (GM70785) to SCH.

References

1. Zeddam, J. L., J. L. Rodriguez, M. Ravallec, and A. Lagnaoui. 1999. A noda-like virus isolated from the sweetpotato pest *spodoptera eridania* (Cramer) (Lep.; noctuidae). *J Invertebr Pathol* 74:267-274.

2. Krishna, N. K., and A. Schneemann. 1999. Formation of an RNA heterodimer upon heating of nodavirus particles. *Journal of virology* 73:1699-1703.
3. Johnson, K. N., J. L. Zeddam, and L. A. Ball. 2000. Characterization and construction of functional cDNA clones of Pariacoto virus, the first Alphanodavirus isolated outside Australasia. *Journal of virology* 74:5123-5132.
4. Johnson, K. N., L. Tang, J. E. Johnson, and L. A. Ball. 2004. Heterologous RNA encapsidated in Pariacoto virus-like particles forms a dodecahedral cage similar to genomic RNA in wild-type virions. *Journal of virology* 78:11371-11378.
5. Johnson, K. N., and L. A. Ball. 2003. Virions of Pariacoto virus contain a minor protein translated from the second AUG codon of the capsid protein open reading frame. *The Journal of general virology* 84:2847-2852.
6. Tang, L., K. N. Johnson, L. A. Ball, T. Lin, M. Yeager, and J. E. Johnson. 2001. The structure of pariacoto virus reveals a dodecahedral cage of duplex RNA. *Nature structural biology* 8:77-83.
7. Schneemann, A., W. Zhong, T. M. Gallagher, and R. R. Rueckert. 1992. Maturation cleavage required for infectivity of a nodavirus. *Journal of virology* 66:6728-6734.
8. Fisher, A. J., B. R. McKinney, J. P. Wery, and J. E. Johnson. 1992. Crystallization and preliminary data analysis of Flock House virus. *Acta crystallographica* 48 (Pt 4):515-520.
9. Mori, K., T. Nakai, K. Muroga, M. Arimoto, K. Mushiake, and I. Furusawa. 1992. Properties of a new virus belonging to nodaviridae found in larval striped jack (*Pseudocaranx dentex*) with nervous necrosis. *Virology* 187:368-371.
10. Reinganum, C., J. B. Bashiruddin, and G. F. Cross. 1985. Boolarra virus: a member of the Nodaviridae isolated from *Oncopera intricoides* (Lepidoptera: Hepialidae). *Intervirology* 24:10-17.
11. Klug, A. 1999. The tobacco mosaic virus particle: structure and assembly. *Philosophical transactions of the Royal Society of London* 354:531-535.
12. Namba, K., R. Pattanayek, and G. Stubbs. 1989. Visualization of protein-nucleic acid interactions in a virus. Refined structure of intact tobacco mosaic virus at 2.9 Å resolution by X-ray fiber diffraction. *Journal of molecular biology* 208:307325.
13. Namba, K., and G. Stubbs. 1986. Structure of tobacco mosaic virus at 3.6 Å resolution: implications for assembly. *Science (New York, N.Y.)* 231:1401-1406.

14. Reddy, V. S., H. A. Giesing, R. T. Morton, A. Kumar, C. B. Post, C. L. Brooks, 3rd, and J. E. Johnson. 1998. Energetics of quasiequivalence: computational analysis of protein-protein interactions in icosahedral viruses. *Biophysical journal* 74:546-558.
15. Freddolino, P. L., A. S. Arkhipov, S. B. Larson, A. McPherson, and K. Schulten. 2006. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure* 14:437-449.
16. Zhang, D., R. Konecny, N. A. Baker, and J. A. McCammon. 2004. Electrostatic interaction between RNA and protein capsid in cowpea chlorotic mottle virus simulated by a coarse-grain RNA model and a Monte Carlo approach. *Biopolymers* 75:325-337.
17. Konecny, R., J. Trylska, F. Tama, D. Zhang, N. A. Baker, C. L. Brooks, 3rd, and J. A. McCammon. 2006. Electrostatic properties of cowpea chlorotic mottle virus and cucumber mosaic virus capsids. *Biopolymers* 82:106-120.
18. Chin, K., K. A. Sharp, B. Honig, and A. M. Pyle. 1999. Calculating the electrostatic properties of RNA provides new insights into molecular interactions and function. *Nature structural biology* 6:1055-1061.
19. Tihova, M., K. A. Dryden, T. V. Le, S. C. Harvey, J. E. Johnson, M. Yeager, and A. Schneemann. 2004. Nodavirus coat protein imposes dodecahedral RNA structure independent of nucleotide sequence and length. *Journal of virology* 78:2897-2905.
20. Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. *Nucleic acids research* 28:235-242.
21. Shepherd, C. M., I. A. Borelli, G. Lander, P. Natarajan, V. Siddavanahalli, C. Bajaj, J. E. Johnson, C. L. Brooks, 3rd, and V. S. Reddy. 2006. VIPERdb: a relational database for structural virology. *Nucleic acids research* 34:D386-389.
22. Malhotra, A., R. K. Tan, and S. C. Harvey. 1994. Modeling large RNAs and ribonucleoprotein particles using molecular mechanics techniques. *Biophysical journal* 66:1777-1795.
23. Baker, N. A., D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. 2001. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences of the United States of America* 98:10037-10041.

24. Dolinsky, T. J., J. E. Nielsen, J. A. McCammon, and N. A. Baker. 2004. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic acids research* 32:W665-667.
25. Pettersen, E. F., T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. 2004. UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of computational chemistry* 25:1605-1612.
26. VanLoock, M. S., B. A. Harris, and S. C. Harvey. 1998. To knot or not to knot? Examination of 16S ribosomal RNA models. *J Biomol Struct Dyn* 16:709-713.
27. Frank, J., M. Radermacher, P. Penczek, J. Zhu, Y. Li, M. Ladjadj, and A. Leith. 1996. SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *Journal of structural biology* 116:190-199.
28. Venter, P. A., N. K. Krishna, and A. Schneemann. 2005. Capsid protein synthesis from replicating RNA directs specific packaging of the genome of a multipartite, positive-strand RNA virus. *Journal of virology* 79:6239-6248.
29. Venter, P. A., and A. Schneemann. 2007. Assembly of two independent populations of flock house virus particles with distinct RNA packaging characteristics in the same cell. *Journal of virology* 81:613-619.
30. Marshall, D., and A. Schneemann. 2001. Specific packaging of nodaviral RNA2 requires the N-terminus of the capsid protein. *Virology* 285:165-175.
31. Belyi, V. A., and M. Muthukumar. 2006. Electrostatic origin of the genome packing in viruses. *Proceedings of the National Academy of Sciences of the United States of America* 103:17174-17178.
32. Devkota B, *et al.* (2009) Structural and electrostatic characterization of Pariacoto virus: Implications for viral assembly. *Biopolymers* 91:530-538.

CHAPTER 4

CAPSID ASSEMBLY SIMULATIONS

Abstract:

Assembly of T=1 virus using coarse-grained models is an ambitious and demanding challenge. Before moving into assembly of the whole virus, we tested one of the earlier capsid simulation studies [5]. Brooks' and his colleagues performed discontinuous molecular dynamics using a coarse-grained capsid unit representing three proteins. We performed classical molecular dynamics with a similar capsid unit. We also further investigated the effects of edge angle variations and two different potentials: non-specific and specific. There were two conformations of dimers formed: a flat and curved one. The curved dimers have lower energy compared to the flat dimers using specific potential. The non-specific potential cannot distinguish the two conformations of the dimers energetically and the simulations using non-specific potential result in kinetic traps. The capsid unit model was studied for stability and chosen for further improvements for whole virus assembly simulation.

Introduction:

Computational studies of empty capsids have been studied with many coarse-grained models [1-5]. These studies vary in the level of details of their coarse-grained models. Some of these models use lower resolution coarse-grained models where a pseudo-atom represents a hexamer or pentamer conformation of the proteins. Zandi *et. al.* used this type of model on a restricted 2D spherical surface where the radius of the sphere is fixed. They simulated different numbers of pseudo-atoms on a fixed surface and

they found out that specific numbers of pseudo-atoms (12, 32, 42, 72) form icosahedral symmetry on the spherical surface and these structures are lower in energy compared to other numbers. These numbers are equal to the number of hexamers and pentamers of T=1, T=3, T=4 and T=7 capsids. They conclude that the icosahedral symmetry of the virus capsid comes from the minimum energy arrangement of capsid proteins.

Other computational studies of capsid assembly use higher-level (shape-based) coarse-grained models. Shape based coarse-grained models mimic the main structure however, there is no direct one-to-one or one-to-many correspondence of the pseudo-atom and the real atoms of the protein. Brooks and his colleagues assembled a T=1 virus using two different shape based models [5] (Figure 4.1). Capsid unit of the first model represents three proteins and it has 28 pseudo-atoms. The red pseudo-atoms are attracted with a square-well potential and the white ones are hard spheres with volume exclusion. They performed discontinuous molecular dynamics to accelerate the simulation using both models. Both models resulted in successful T=1 capsid assembly. The first model yields the same results with less computational demand.

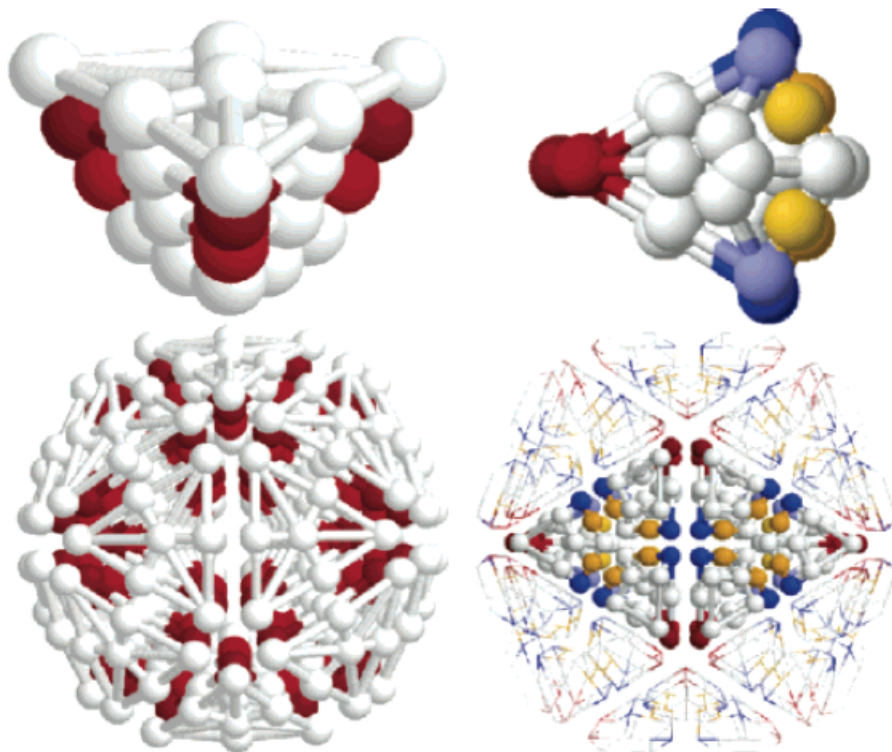


Figure 4.1: Capsid units and T=1 capsid models. First capsid unit represents three proteins using 28 pseudo-atoms. Red pseudo-atoms are attracted to each other with a square-well potential. The second model represents one protein, and each color attracts one another. Reprinted from [5].

We proposed a mechanism for single-stranded RNA virus assembly using coarse-grained models [6]. We chose to replicate Brooks's first shape based coarse-grained model of the capsid unit for our virus simulations.

Methods:

We built a similar capsid unit (Figure 4.2) having 4 layers of 7 pseudo-atoms in each layer. The attractive pseudo-atoms are on the corner of the 2nd and 3rd layers. The pseudo-atoms are connected to each other via bond, angle, and torsion potentials. We performed traditional molecular dynamics simulations using LAMMPS at 300 K.

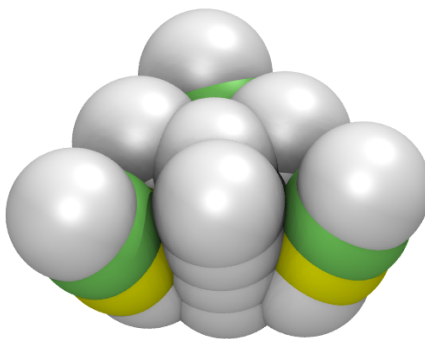


Figure 4.2: The capsid unit containing 28 pseudo-atoms. The colored pseudo-atoms are attracted to each other via LJ potential.

In addition, we tested two different configurations of our capsid unit by changing the way that the attractive pseudo-atoms interact. We called these potentials specific and non-specific based on the conformations they form. The specific potential has 2 different types of LJ particles on the corner of the 2nd and 3rd layers of the capsid unit. Figure 4.3 shows two pseudo-atoms colored red and blue, respectively, for the specific potential. The red atoms attract other red atoms and the blue atoms interact other blue atoms. The attraction of blue and the red is not allowed. This method ensures that a dimer with an inward curvature has a lower energy conformation than the flat formation. Thus the inward conformation of the dimer dominates in the simulation. In the non-specific potential, the attractive pseudo-atoms have only one type, and this allows attraction between the particle on the 2nd layer of the capsid and the particle on the 3rd layer of the capsid. This extra attraction energetically balances the inward dimer conformation and the flat dimer conformation.

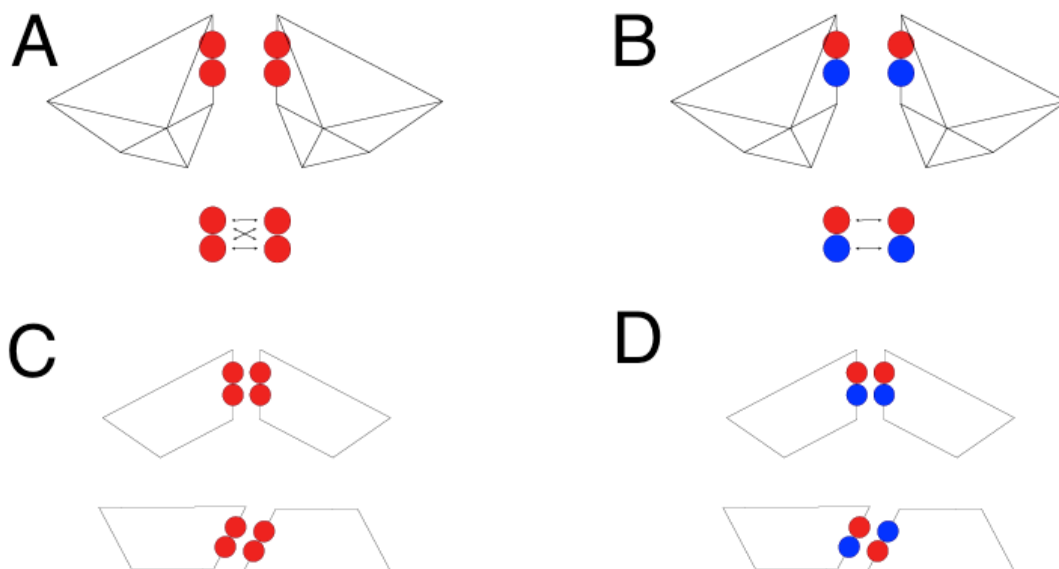


Figure 4.3: The Non-specific potential (A) has the same type of atom attracting each other via 4 directions. The Specific potential (B) has two types of atoms and each type attract each other. Attraction between different types is not allowed. C and D show the two distinct conformations of dimers. The Specific potential energetically favors the curved conformation over the flat conformation.

We also studied conformational changes over variation of the edge angle of the capsid unit. The edge angle of the perfect icosahedron is 20.9° . We varied this angle from 2.9° to 20.9° that changes the wedged triangular prism to almost perfect triangular prism (Figure 4.4).

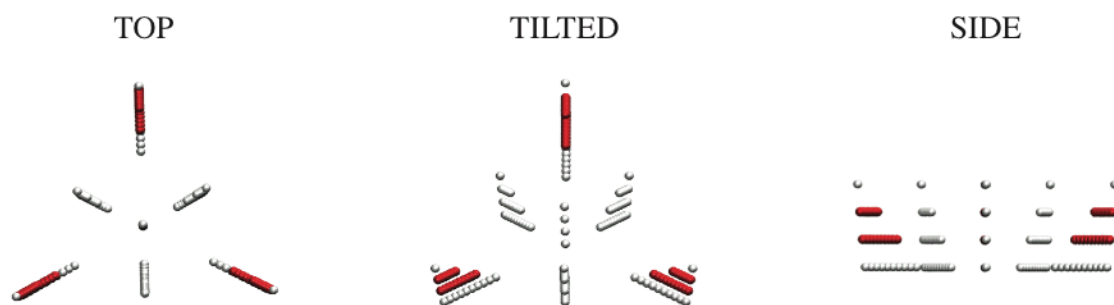


Figure 4.4: 20 capsid units with varying edge angles are superimposed and shown in three views; top, tilted and side.

Results:

We observed T=1 capsid formation using our capsid model (Figure 4.5). The final minimized conformation is slightly distorted. The pseudo atoms at the five fold axes form trapezoidal conformations rather than the pentagonal conformations. This is a direct consequence of the LJ potential. The trapezoidal conformation is lower in energy when compared to the pentagonal conformation.

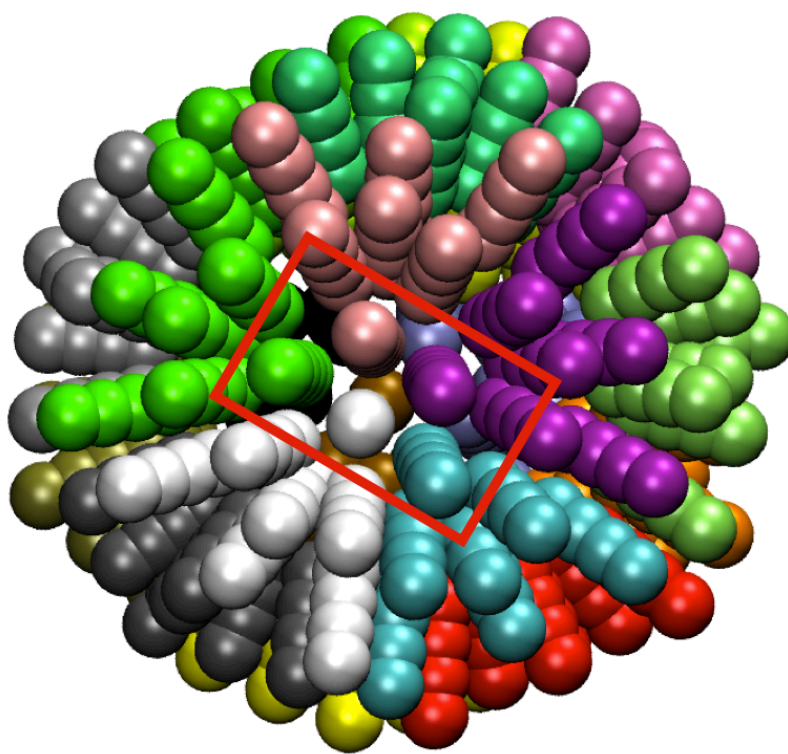


Figure 4.5: T=1 capsid model. Each capsid unit is colored differently. The capsid is made of 20 capsid units each having 28 pseudo-atoms. The trapezoidal 5-fold axis is shown in the red rectangle.

We have also varied the edge angle of the capsid unit and studied its effect on capsid geometry using two different pairwise potentials, non-specific and specific potentials. The specific potential helps to form the icosahedron and the non-specific

potential causes many kinetic traps during formation of the icosahedron. The specific interaction is introduced to overcome the entropy problem and giving curved conformation a lower energy thus increasing the probability of having curved conformation.

We performed simulations with capsid units with various edge angles using both specific and non-specific potentials (Figure 4.6). The spherical aggregates having 20 capsid units form T=1 empty capsids between edge angle 16.9° and 20.9° . The bigger spherical aggregates having more than 20 capsid units are not equal to any higher T-number icosahedral capsids. The range of observing bigger spherical aggregates differs in specific and non-specific potentials. This range in the non-specific potential is between 10° and 15° and the increase of the size is very sharp. The reason is that the probability of having a flat dimer and curved dimer is equal to each other. However the probability of having curved dimer is higher in specific potential. It results a broader range of having bigger spherical aggregates. It is between 6° and 15° .

Figure 4.7 shows the snapshot of the aggregates at 20.9° , 12.9° , and 4.9° edge angles, respectively, with specific and non-specific potentials. At 20.9° edge angle both of them yield T=1 capsids. However, at the 12.9° edge angle the spherical aggregate sizes differ. The specific potential resulted in one big spherical aggregate. In the non-specific potential, there are two smaller spherical aggregates. Finally, at the 4.9° edge angle, the specific potential yielded a very large curved sheet and the non-specific potential resulted in flatter sheet with a mixture of up and down curved aggregates.

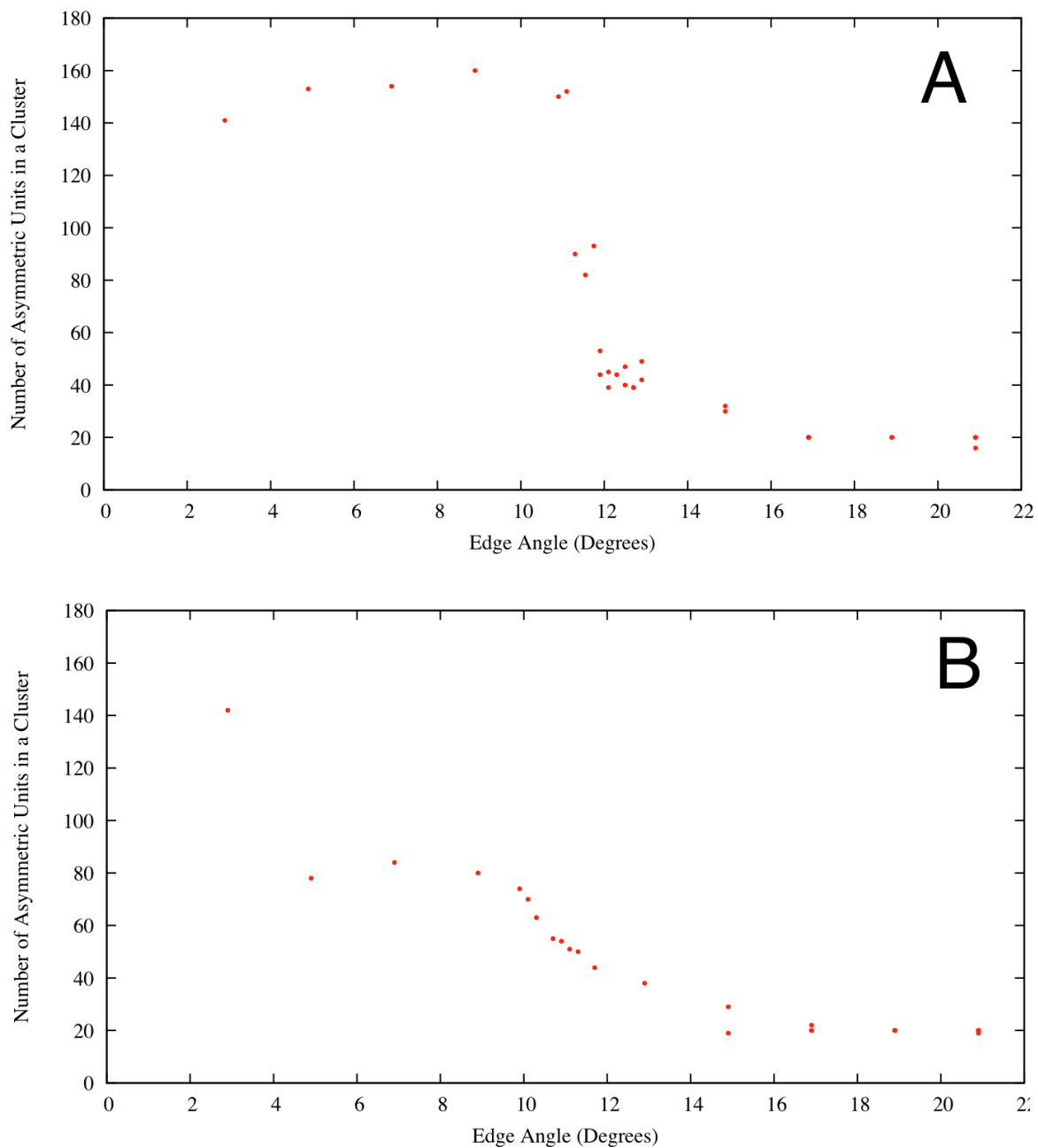


Figure 4.6: Effects of the specific (B) and non-specific (A) potentials with the angle variation is shown. The non-specific potential demonstrates sharp phase change from spherical to flat conformation where this phase change is broader in the specific potential.

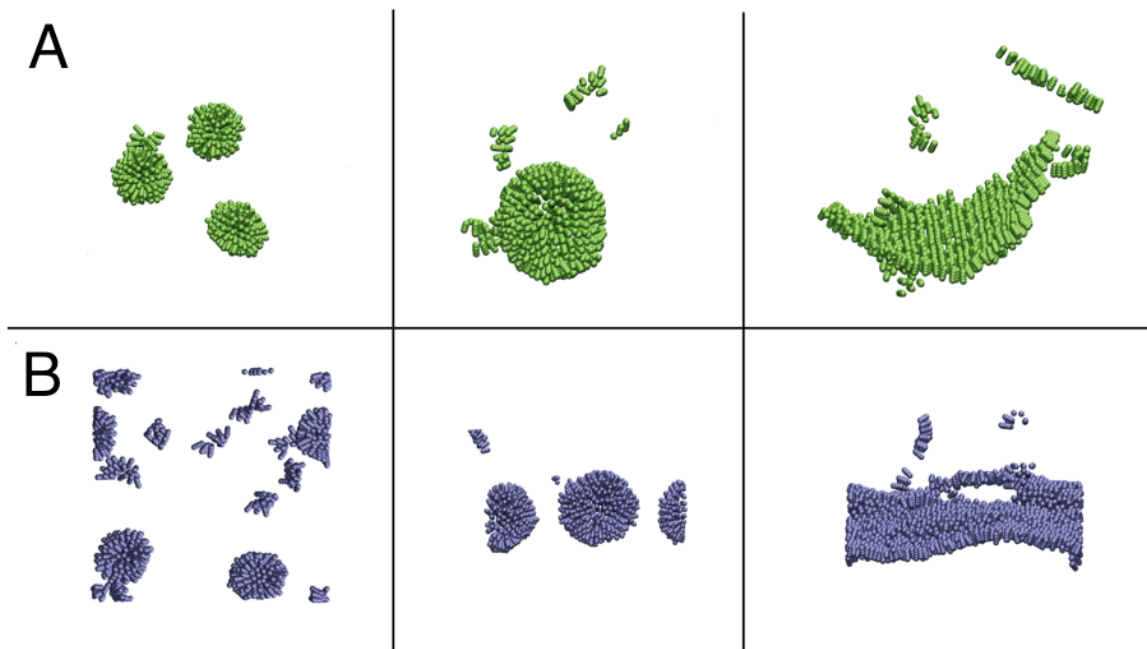


Figure 4.7: Three snapshots of assembly simulations at 20.9° , 12.9° , and 4.9° edge angle with the specific (A) and non-specific (B) potentials respectively. At the 20.9° edge angle, T=1 capsids forms. At the 12.9° edge angle, bigger, irregular capsids form. At the 4.9° edge angle, the specific potential forms a curved sheet, however the non-specific potential forms flat sheet with mixed curvature.

Discussion:

We performed empty capsid simulations using a simple wedge-shaped triangular prism and achieved the assembly of T=1 capsid. We varied both the potential and the edge angle of the capsid unit. Edge angle variation demonstrated that T=1 capsids are very stable up to 4.0° variation from 20.9° to 16.9° . The potential variation from non-specific to specific potential changes the size and the type of the aggregates. Using specific potential ensures the curved dimers are dominant over flat dimers and ensures the formation of the spherical aggregates by lowering the entropy. The specific potential is chosen to be improved for our proposed [6] T=1 virus coarse-grained model assembly.

References:

1. Hagan MF, Chandler D (2006) Dynamic pathways for viral capsid assembly. *Biophysical journal* 91:42-54.
2. Zhang T, Schwartz R (2006) Simulation study of the contribution of oligomer/oligomer binding to capsid assembly kinetics. *Biophysical journal* 90:57-64.
3. Wilber AW et al. (2007) Reversible self-assembly of patchy particles into monodisperse icosahedral clusters. *The Journal of chemical physics* 127:085106.
4. Rapaport D (2004) Self-assembly of polyhedral shells: A molecular dynamics study. *Physical Review E* 70:1-13.
5. Nguyen HD, Reddy VS, Brooks CL (2007) Deciphering the kinetic mechanism of spontaneous self-assembly of icosahedral capsids. *Nano letters* 7:338-44.
6. Petrov AS, Boz MB, & Harvey SC (2007) The conformation of double-stranded DNA inside bacteriophages depends on capsid size and shape. *J Struct Biol* 160:241-248.

CHAPTER 5

ASSEMBLY OF T=1 VIRUS USING COARSE-GRAINED MODELS.

ABSTRACT

The spontaneous assembly of small, icosahedral RNA viruses involves a delicate balance between the attractive forces between the capsid proteins, the attractive forces between those proteins and the genomic RNA, and the repulsive RNA-RNA forces. We investigate the roles of RNA- protein and capsid protein-capsid protein attractions on the stability and the assembly of Satellite Tobacco Mosaic Virus (T=1), using a coarse-grained model. The RNA is assumed to have a fixed secondary structure containing a series of stem-loops, connected by flexible single-stranded regions. The protein model consists of a rigid wedge-shaped region representing the globular domain of the protein trimer defining one face of the icosahedral structure, along with three positively charged flexible N-terminal tails. We carried out a collection of stability simulations to define the possible ranges of the parameters describing these interactions. We then examined two different approaches to assembly: one set of simulations ("co-transcriptional assembly") mimics viral assembly assuming that the capsid proteins are available and interact with the RNA during transcription, while the other ("post-transcriptional assembly") mimics assembly under the assumption that RNA replication is completed in the absence of the capsid proteins. We find successful assembly of a model T=1 virus model for a narrow range of parameters with both protocols. The results of the post-transcriptional assembly simulations also depend on the three-dimensional structure of the RNA, with successful assembly only being obtained when the initial RNA conformation is quite compact.

INTRODUCTION

Small icosahedral single-stranded RNA (ssRNA) viruses are of interest, both because they are among the simplest of all viruses, and because they are important model systems for spontaneous assembly. The structures of ssRNA viruses have been reviewed in great detail elsewhere^{1 2}. All of the crystal structures of ssRNA viruses are based on icosahedral averaging, which clearly reveals the structures of the globular domains of the capsid proteins (and sometimes some of the viral RNA), but it obscures the structures of those regions that do not have icosahedral symmetry; this generally includes most of the RNA, and all or part of the N- and C-terminal protein tails. The protein tails generally contain a substantial number of positively charged residues, which are known to be critical for viral assembly and stability.

In contrast to the packaging of double-stranded DNA (dsDNA) in bacteriophage, which is driven by ATP hydrolysis³, the assembly of small icosahedral ssRNA viruses requires no energy. Assembly is a slow condensation process of capsid proteins with the RNA. It is difficult to track the assembly process and the intermediate structures in vivo, and in vitro assembly is difficult to achieve. Cowpea Chlorotic Mottle Virus (CCMV) was the first ssRNA virus that was assembled in vitro⁴. The experiments on CCMV^{5 6 7} emphasized the importance of the positively charged proteins and the solvent conditions (pH, ionic strength and type of cation) in the in vitro assembly process. The effects of solvent conditions on RNA structure and dynamics have also been extensively studied^{8 9 10 11 12 13 14}.

Beyli and Muthukumar¹⁴ made the interesting observation that the ratio of the genome's negative charge to the sum of the positive charges on the terminal protein tails

is 1.61 ± 0.03 in 16 different ssRNA and ssDNA viruses. (STMV is an outlier with the value of 2.2.) We have previously proposed a specific assembly mechanism (Figure 5.1). In this model, the polycationic protein tails first interact nonspecifically with the genomic RNA, leading to charge neutralization and a structural collapse similar to that of DNA condensation by cations of charge +3 and higher; this concentrates the proteins' globular domains in a spherical shell surrounding the genome, where the weak inter-protein attractions are sufficient to lead to formation of the mature particle ¹⁵.

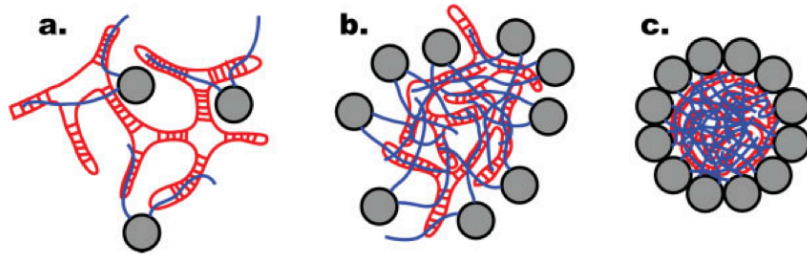


Figure 5.1: Proposed pathway for the assembly of small icosahedral RNA viruses. The positively charged protein tails (blue) bind non-specifically to RNA through electrostatic interactions. (b) When a sufficient fraction of the RNA charge has been neutralized by the polycationic protein tails, the complex of RNA plus protein tails collapses, following a pathway similar to that of the condensation of DNA by polyvalent cations. The protein/RNA condensate is dense enough to exclude the proteins' globular domains (grey), and these are concentrated in a shell around the condensate. When their concentration in the shell is sufficiently high, the weak inter-protein attractive forces are strong enough to lead to the formation of the mature capsid (c). Reprinted from [15].

There have been several theoretical and computational studies ^{16 17 18 19, 20 21 22 23} focused on the stability of ssRNA viruses in all-atom and coarse-grained models. It is computationally challenging to study the assembly of these viruses because the size of the system and the duration of the simulation. A number of simulations have examined capsid assembly, but only one has successfully packaged a model genome into a capsid ²³; in this coarse-grained simulation from Michael Hagan's laboratory, the capsid-capsid and capsid-genome interactions were all modeled with a Lennard-Jones potential (a 4-8 potential, rather than that 6-12 potential commonly used for van der Waals interactions).

While this was a significant achievement, the capsid-RNA model is a poor mimic of the electrostatic interactions that drive assembly in the real system.

In the present work, we introduce a coarse-grained model for examining electrostatic contributions to the stability and assembly of a model for a T=1 virus, Satellite Tobacco Mosaic Virus (STMV). The STMV crystal structure revealed 30 RNA duplexes, each 9 base pairs long, centered on the two-fold axis; there is an additional non-paired nucleotide at both 3' ends of the duplex ²⁴. This represents over 55% of the 1058 nucleotide viral RNA genome. Schroeder et al. ²⁵ proposed a secondary structure for STMV RNA, based on a combination of chemical probing and the requirement that the secondary structure have 30 short symmetric double helices connected by single-stranded regions, as proposed by Larson and McPherson ²⁶. Zeng et al. used the crystal structure and Schroeder's model for the RNA secondary structure to develop an all-atom model of STMV, including every amino acid and every single nucleotide ²⁷.

We based our RNA model on Schroeder's proposed secondary structure, and we tested the stability of a coarse-grained STMV model whose structure is based on the RNA conformation from Zeng's all-atom model, coupled to idealized capsid units. We examined the stability of the complete virus by varying the strength of the two important pair-wise interactions: RNA-protein and protein-protein attractions. This involved the characterization of a three-dimensional parameter space, leading to the identification of a "stability island", defined by the ranges over which the three parameters can be varied without disrupting the viral structure. We then explored the assembly of the STMV model using two different protocols: post-transcriptional and co-transcriptional assembly. The post-transcriptional protocol assumes that the condensation of RNA and protein tails

happens after all the RNA has been fully transcribed. The co-transcriptional protocol mimics assembly during RNA synthesis, by using a scenario in which the 30 stem-loops are generated sequentially in three-dimensional space, with the complete pool of capsid proteins being allowed to interact with each RNA stem-loop as it appears.

METHODS

Capsid Model

Figure 5.2 shows the coarse-grained capsid unit (CU) a wedge-shaped triangular prism, with three flexible tails (each 16 pseudo-atoms long) attached to the inner side of the CU representing 3 proteins. There are 32 pseudo-atoms (16 in each of two planes) on the outer shell representing roughly 15,000 atoms. Corner pseudo-atoms at the first and the second layers of two different CUs are attracted to each other via Lennard-Jones (LJ) potential (Eq. 2); the strength of this interaction is one of the parameters to be examined. The remainders of the inter-subunit interactions are treated as hard spheres, using the Yukawa potential (Eq. 3). The protein tails (PT) have a higher-level coarse-grained model (C α -model) with each pseudo-atom representing one residue. Each tail has 8 positively charged pseudo-atoms alternating with neutral pseudo-atoms. (This preserves the net RNA:protein charge ratio of 2.2 described above.) Charged pseudo-atoms are treated via Debye-Huckel (DH) electrostatics (Eq. 1) and LJ potential (Eq. 2). The Debye length is 8.0 Å, corresponding to an ionic strength of 150 mM (roughly physiological ionic strength). Interactions between pairs of neutral pseudo-atoms on the tails are treated with a Yukawa potential (Eq. 3).

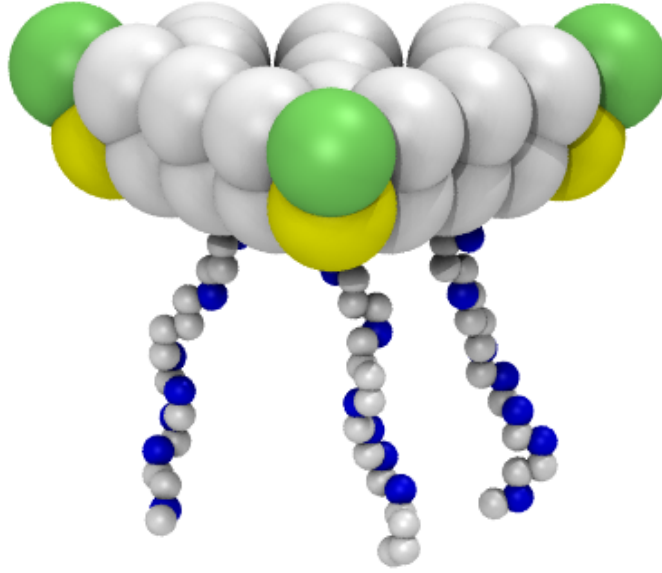


Figure 5.2: Capsid model is a multilevel model having outer shell and protein tails. Outer shell is composed of 32 pseudo atoms. Green attracts green and yellow attracts yellow atoms. Blue protein tail atoms are (+) charged and all white atoms are neutral and represented as hard spheres.

The Debye-Hückel, Lennard-Jones and Yukawa terms in the energy function are, respectively:

$$E_{DH} = C \frac{q_i q_j}{Dr} e^{(-\kappa r)} \quad r < r_c \quad (1)$$

$$E_{LJ} = 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right] \quad r < r_c \quad (2)$$

$$E_Y = A \frac{e^{-\kappa r}}{r} \quad r < r_c \quad (3)$$

q_i, q_j are the charges of RNA pseudo-atom i and PT pseudo-atom j , respectively.

D is the dielectric constant, and C is a conversion factor to express the energy units to kcal/mol (with charges in units of proton charge, distances in Ångstroms, and a dimensionless dielectric constant, $C=332$). κ is the inverse of the Debye length; the

Debye length was 8.0 Å corresponding to roughly 100-150 mM ionic strength. In the LJ

potential (Eq. 2), ϵ is the well-depth of the LJ attraction, and σ is the distance where two pseudo-atoms touch each other and the energy is zero. A is a constant to change the energy units to kcal/mol. κ for the Yukawa potential is used to determine the radius of the pseudo-atoms.

RNA Model

The coarse-grained RNA model used in this study is based on a model previously developed in the Harvey laboratory²⁸ and used for examining problems ranging from the ribosome²⁹ to viruses¹⁵. The full set of parameters are described in a recent review³⁰.

In this model (Figure 5.3), each residue of the RNA is represented with a pseudo-atom (P-atom) located at the phosphorus atom. For double helical regions, another pseudo-atom (X-atom) is introduced between pairs of P-atoms that form Watson-Crick base pairs. X-atom provides the volume exclusion. This model describes the secondary structure by distinguishing double helices from single-stranded regions. Interactions between pairs of P-atoms are treated with DH electrostatics (Eq. 1), along with an LJ potential (Eq. 2) to guarantee volume exclusions. Interactions between pairs of X-atoms are treated with the Yukawa potential (Eq. 3).

In exploratory studies, we used an RNA secondary structure identical to that of the Schroeder model²⁵. The long, floppy single-stranded connectors hindered assembly, however, often sticking out of partially assembled capsids. In real viruses, the RNA undoubtedly attracts polyvalent cations whose effects that facilitate the compact conformations needed for viral assembly. Polyvalent cations are not well represented by the DH potential. In addition, the Schroeder secondary structure model almost certainly understates the actual secondary structure content. In the absence of proper treatment of

the effects of polyvalent cations, and in the absence of some probable double-helical regions, this initial RNA model is almost certainly more extended in three-dimensional space than is the actual viral RNA. To generate a more compact RNA model, we deleted some residues of the connecting single-stranded regions between the stem-loops of the Schroeder model. Overall, the genomic RNA was shortened by a total of 212 nucleotides.

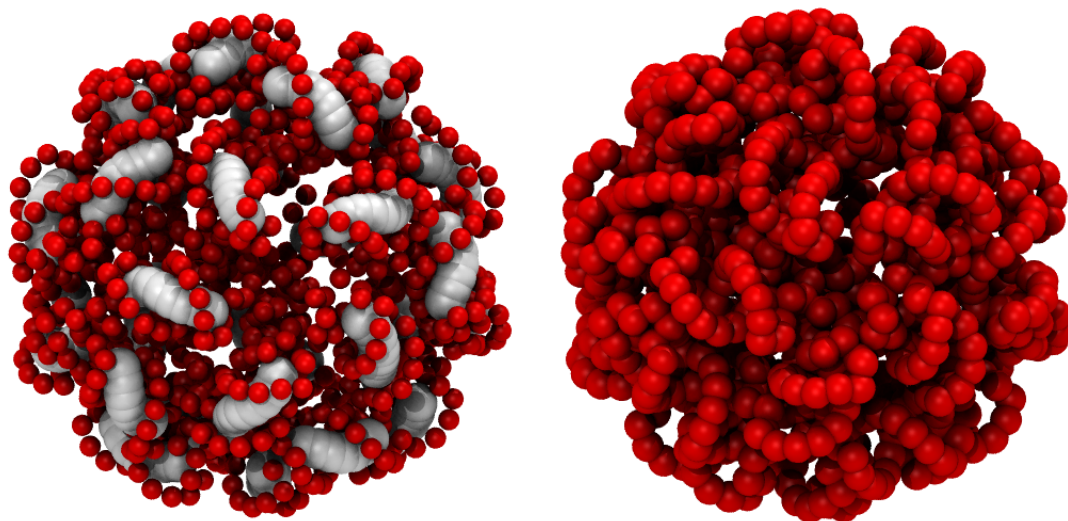


Figure 5.3: The P-model of proposed all-atom model of the STMV RNA [yingyings ref]. Red pseudo-atoms (P-atoms) are (-) charged and white pseudo-atoms (X-atoms) are neutral. The X-atoms are removed and the radius of the P-atoms are increased for visibility of the other Figures.

Simulations and Protocols

We carried out three kinds of simulations, one examining the dependence of the model's stability on the parameters of the energy function. The second and the third simulations examined assembly with two different protocols: post-transcriptional and co-transcriptional assembly.

Stability Simulation

The capsid model and the RNA model are assembled by first, capsid model expansion and then series of compression steps following minimization of tails at each

compression steps similar to the protocol described in our earlier work on Pariacoto virus¹⁵. This protocol allows the protein tails to find the gaps and minimal energy conformations inside the RNA model. The stability of the final STMV coarse-grained model was tested with the variation of the attractive potentials between RNA-PT and CU-CU. We varied the RNA-PT attraction by changing the dielectric constant and the charges of the P-atom. The CU-CU attraction was varied by changing the well-depth (ϵ) of the LJ potential (Eq. 2). Stability simulations were carried out using Langevin molecular dynamics simulations, each of length 100 ns.

Assembly Simulations

Post-transcriptional Assembly Protocol

This protocol assumes the assembly occurs after the transcription is completely finished and capsid proteins bind to the RNA after it is fully transcribed. To mimic these conditions, the RNA model is equilibrated in the absence of CUs. Since the RNA is repulsive itself, it is equilibrated in three different sized boxes with fixed boundaries resulting three different radii of gyration (97.4 Å, 69.7 Å, 51.1 Å) of the RNA (Figure Supp. 1). Later the RNA model is centered at the simulation box and 100 CUs are randomly generated around the RNA with different orientations. Langevin molecular dynamics is performed for 200 ns at 300 K.

Co-transcriptional Assembly Protocol

This protocol assumes the assembly starts during the RNA transcription by each transcribed RNA stem condensing with capsid proteins. To mimic these conditions, the RNA model is gently squeezed into cubic box with a length of 30 Å using moving

harmonic repulsive boundaries with all the repulsive pair-wise potentials completely turned off. Then, 100 CUs are randomly generated around the RNA model. Three types of RNA stems are defined; visible, semi-visible and invisible. When the simulation starts all RNA stems are invisible to the capsid proteins and to one another except the 1st stem and the 2nd stem. First stem is in visible stem group that has both DH electrostatic and LJ potential turned on. The second stem is in the semi-visible group that has only LJ potential turned on and diameter of the P-atoms (σ) is increased linearly starting from 0 to the its original value within 1 ns. The rest of the stems are in the invisible stem group that both DH electrostatic and LJ potential turned off. At every 1 ns, P-atoms of the semi-visible stem moves to visible stem group by turning on the DH electrostatics and the next stem is moved to the semi-visible group from the invisible stem group by turning on the LJ potential and increasing diameter of the P-atoms (σ). This method mimics the transcription of each stem one by one. Therefore up to 30 ns of the simulations, there are stems transcribed (the visible stem group) interacting everything, stems being transcribing (the semi-visible stem group) interaction everything via LJ potential and slowly growing and stems will be transcribed later (the invisible stem group) interacting with anything. All stems are moved to visible stem group after 30 ns and total simulation takes 200 ns.

RESULTS

Stability Results

We surveyed parameter space with three different values for the dielectric constant (D), four different values for the net charge of the P-atom (q_1), and four different

values for the well-depth (ϵ) of LJ potential between CUs. This yielded a total of 48 parameter sets, each of which was tested in a single stability simulation (Table 5.1).

Table 5.1: Parameters of dielectric constant (D) charge of the P-atom (q_l) and the well-depth (ϵ). Every possible combinations of three have been tested.

D	q_l	ϵ_{ij} (kcal/mol)
4	-0.10	1.0
7	-0.15	1.5
15	-0.20	2.0
	-0.30	2.5

Figure 5.4 reports the results of the stability simulations. Among the 48 sets of parameters, we have found boundaries of the strengths of RNA-protein and CU-CU attractions. Lower dielectric and high charge of P-atom increases the electrostatics. We varied RNA-TP attraction between -0.1 kcal/mol to -1.4 kcal/mol with the variation of the parameter sets. The STMV coarse-grained model is stable over -0.4 kcal/mol of RNA-PT interaction regardless of the strength CU-CU interaction. When the RNA-PT interaction is weaker than -0.4 kcal/mol, a stronger CU-CU interaction is required. On the other hand, -5.0 kcal/mol of CU-CU attraction makes the capsid so stable that even at high value of P-atom charges doesn't create enough repulsion to break CUs apart.

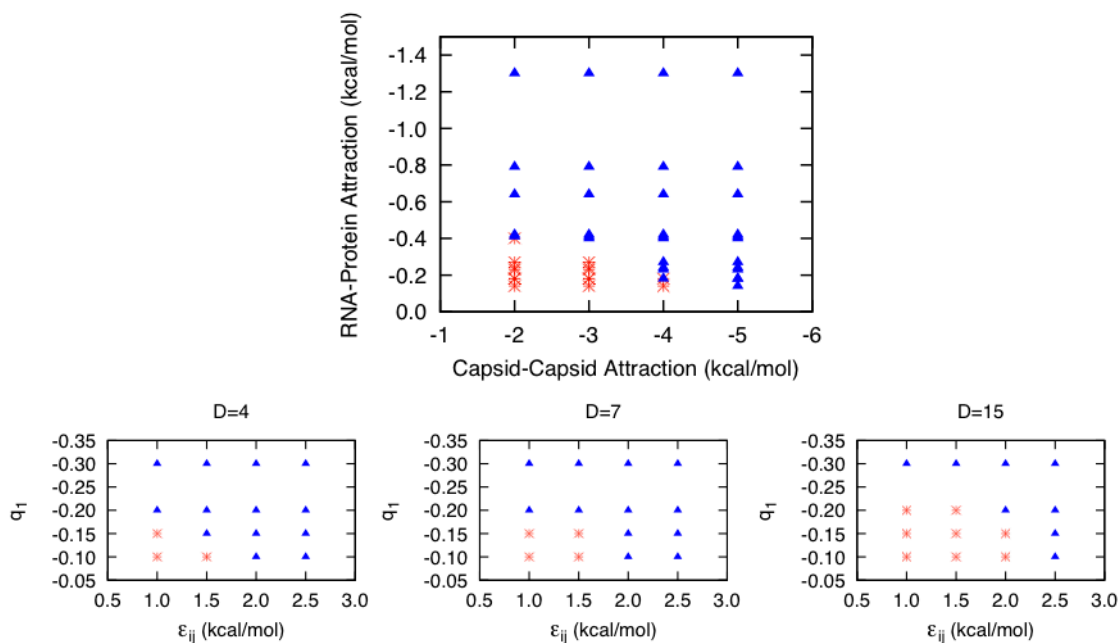


Figure 5.4: Unstable (*) and (▲) stable simulations in the 2-dimensional parameter space defined by the RNA-PT and the CU-CU attractions. Decomposition of the potentials to variables; D (the dielectric constant), q_1 (the charge of the P-atom) and ϵ_{ij} (well-depth) are also shown.

We found five distinct unstable structures when using suboptimal values for the RNA-PT and CU-CU attractions (Figure 5.5). When both of these attractions are weak, the model disassembles completely. If CU-CU attraction grows in the presence of weak RNA-PT attraction, an empty capsid is observed where the RNA is fully ejected. Visa versa, CU collapses on the RNA. When both attractions are close to optimal strengths, only one of the CU pops up with several stems condensed with the PT. Once the optimal conditions (Table 5.2) are met, we observe a stable model. However, the final minimized stable model loses the 5-fold symmetry axis, with the pentagonal conformation shifting to a trapezoidal conformation (green spheres in Figure 5.5e). This is a consequence of using a Lennard-Jones potential to stabilize the capsid, rather than a potential that would enforce five-fold symmetry: the trapezoidal conformation is lower in energy than a pentagonal conformation, because it maximizes the number of contacts between pairs of

pseudoatoms. This behavior also occurred in the empty capsid simulations performed by Brooks and his colleagues³¹, although those authors did not comment on it. (See the right-hand member of the aggregate in Figure 3b of that paper.)

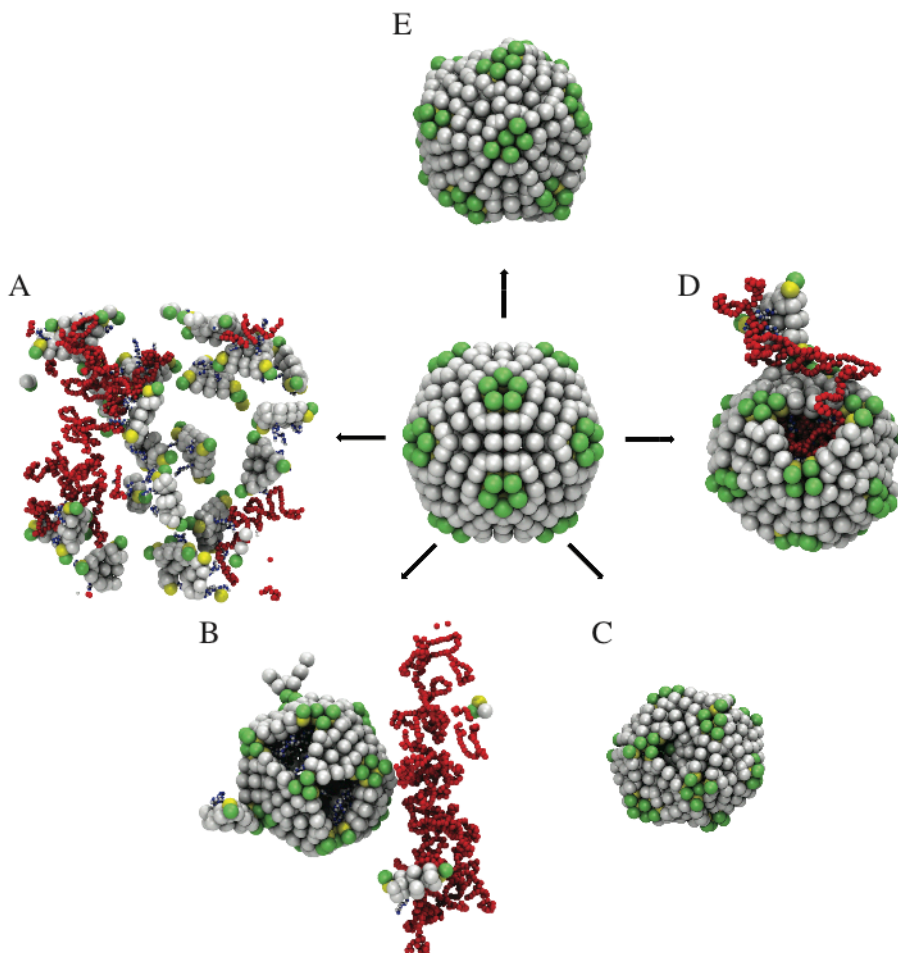


Figure 5.5: The coarse-grained model of STMV before the stability simulation is in the center. (A) Weak interactions of both RNA-protein tail and CU-CU ($D=4, q_1=-0.10, \epsilon_{ij}=1.0$). (B) Weaker RNA-protein tail and stronger CU-CU attractions ($D=15, q_1=-0.10, \epsilon_{ij}=2.0$). (C) Stronger RNA-protein tail and weaker CU-CU attractions. (D) Relatively stronger RNA-protein tail and CU-CU attractions ($D=7, q_1=-0.15, \epsilon_{ij}=1.5$). (E) Both strong RNA-protein tail and CU-CU attractions ($D=4, q_1=-0.20, \epsilon_{ij}=2.5$).

For the assembly simulations, the parameter sets of interest are the sets near the edges of the stability islands in Figure 5.4. Those parameters in the center of the stability islands give very stable models, but they pose the risk of rapid condensation into large

aggregates and kinetically trapped structures. Parameter sets near the edges of the stability island are more likely to create metastable intermediates that can be kinetically reorganized into lower energy structures, giving greater prospects of successful assembly of the complete particle.

Table 5.2: Parameter sets chosen for assembly simulation. Only the sets in bold resulted in the formation of T=1 virus.

$q_1, \epsilon, D=4$	$q_1, \epsilon, D=7$	$q_1, \epsilon, D=15$
-0.20, 1.0	-0.20, 1.0	-0.30, 1.0
-0.15, 1.5	-0.20, 1.5	-0.30, 1.5
-0.10, 2.0	-0.10, 2.0	-0.20, 2.0

Assembly Results

Having eliminated 39 sets of parameters in the stability simulations due either weak or strong interactions, we carried out assembly simulations on the remaining 9 sets of parameters. Only 2 of them yielded successful T=1 viruses using the co-transcriptional protocol (Table 5.2). We classified four different kinetic traps based on the strength of CU-CU and RNA-TP attractions (Figure 5.6). When the well depth (ϵ) is low (~ 1.0 kcal/mol), the aggregate becomes more like a rod rather than being spherical regardless of the strength of the RNA-TP interaction (Figure 5.6B). If the well depth (ϵ) is strong (~ 2.0 kcal/mol) with strong RNA-TP interaction, all CUs and the RNA condense in a big aggregate (Figure 5.6D). If the RNA-TP attraction is low, RNA is independent from this big aggregate (Figure 5.6A). The well-depth (ϵ) value for CU-CU interaction is found to be optimal at 1.5 kcal/mol. The structure of the aggregate depends on the RNA-TP attraction when CU-CU attraction is optimal at 1.5 kcal/mol. If The RNA-TP attraction is

lower, we observed conjoint virus capsids. The optimal RNA-TP attractions correspond to two sets of parameters ($D=4$, $q_1=-0.15$, and $D=7$, $q_1=-0.20$). These give energies for optimized RNA-TP complexes of -0.40 and -0.41 kcal/mol, respectively, with -0.1 kcal/mol coming from the Lennard-Jones potential.

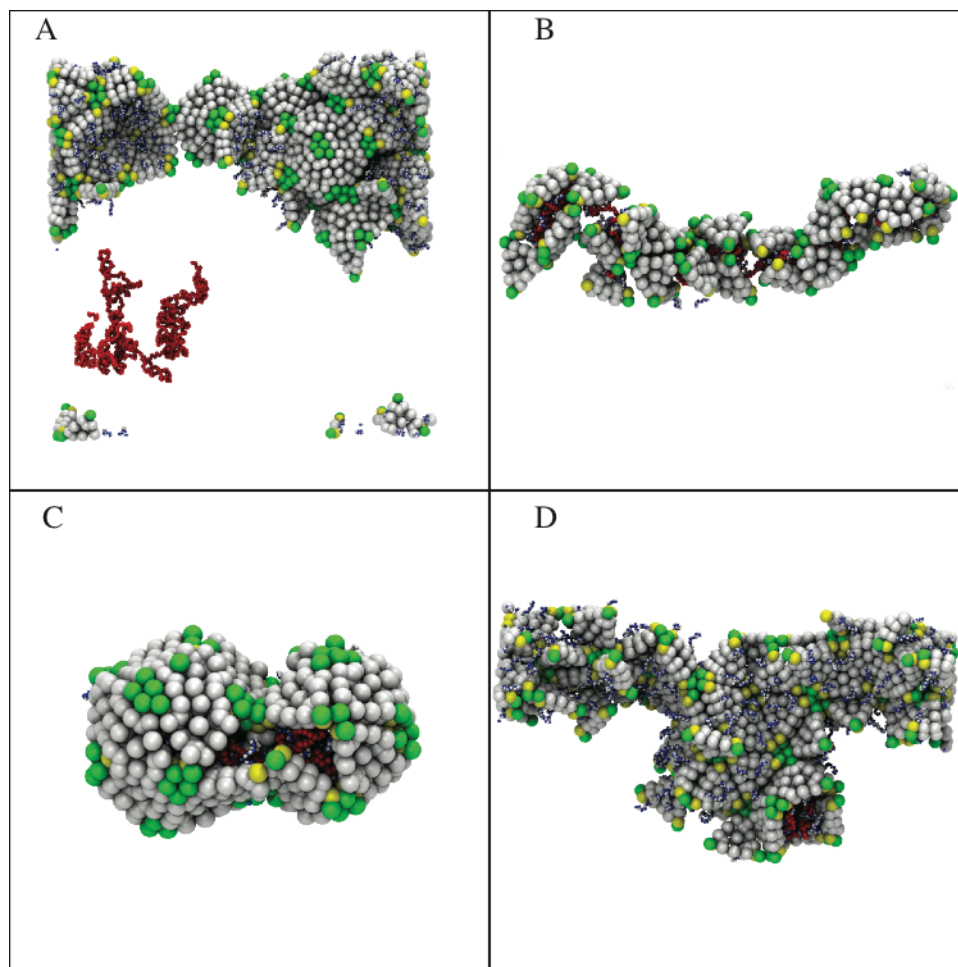


Figure 5.6: (a) Aggregation of all CUs due to strong CU-CU and weak RNA-PT attraction ($D=15, q_1=-0.10$, $\epsilon_{ij}=2.5$). (b) Linear condensation of CUs with RNA due to weak CU-CU and strong RNA-PT attraction ($D=15, q_1=-0.30$, $\epsilon_{ij}=1.0$). (c) Conjoint capsids aggregation at the moderate level of both CU-CU and RNA-PT attractions. ($D=15, q_1=-0.30$, $\epsilon_{ij}=1.5$) (d) Aggregation of all CUs and RNA into one giant condensate ($D=4, q_1=-0.10$, $\epsilon_{ij}=2.0$).

For the post-transcriptional protocol, we could not initially assemble a successful T=1 virus model even using the RNA model with 212 nucleotides. Following the lead of Yoffe *et al.*, who suggested that compact RNA conformations facilitate viral assembly³²,

we examined three different conformations for the RNA at the start of the simulation. The initial compactness (smaller radius of gyration) of the RNA is not relevant, when the simulation starts the RNA expands and binds too many CUs. This RNA structure with short local helices yields conjoint virus capsids instead of T=1 virus at the optimal value we have found in the co-transcriptional protocol. However, we can assembly T=1 virus model using smaller RNA stems (25 stems).

Figure 5.7 shows snapshots from the assembly simulations from both protocols. Post-transcriptional assembly occurs more rapidly than co-transcriptional assembly. CUs bind to the RNA to form an initial aggregate, and this aggregate forms an icosahedral structure slowly. The last (twentieth) CU binds slowly to the aggregate in both protocols.

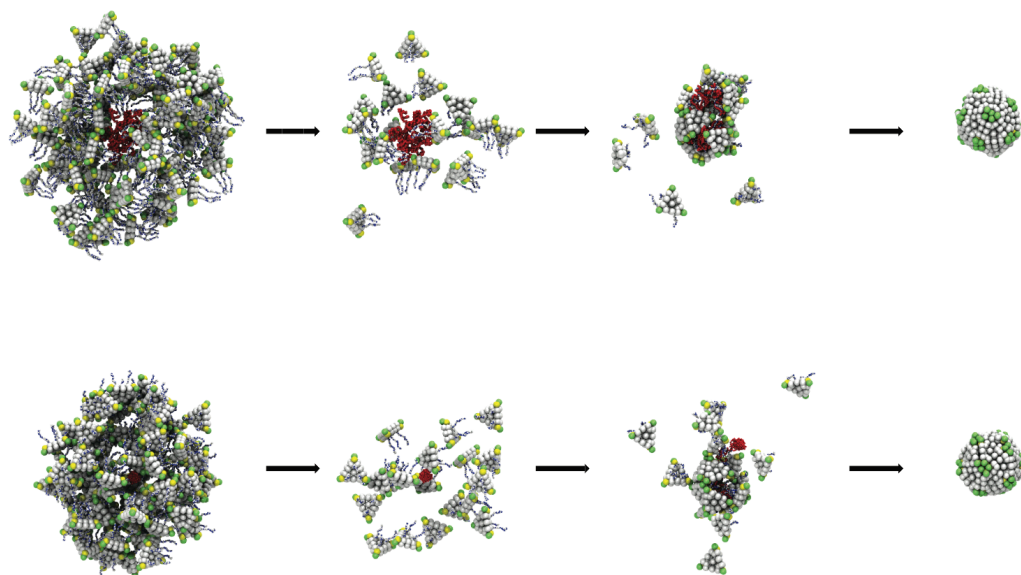


Figure 5.7: Assembly process with the post-transcriptional protocol (upper panel) and with the co-transcriptional protocol (lower panel). First image is the snapshot of starting of the simulation with 100 CUs and the RNA. Second image is the starting of the simulations with 20 CUs that will form the capsid and the RNA. The third image is the snapshot from the middle of the simulation. The last image is the end of the simulation.

Discussion

We have successfully modeled electrostatic effects in the assembly of a model T=1 virus using multi-level coarse-grained models of RNA and capsid proteins. This is a significant advance over earlier simulations. The first simulations on capsid assembly¹⁷ required the use of directional potentials to guarantee the production of the five-fold and six-fold symmetries characteristic of icosahedral viruses. Subsequent simulations of capsid assembly were based on simple pair-wise attractive potentials³¹, and similar potentials were used in the first simulation in which model proteins encapsidated a model polymer²³. But the work reported here is the first to incorporate electrostatic effects into the energy function.

We first examined the range of parameters consistent with stability of the fully assembled model (Figures 5.4 and 5.5), finding that the range of parameters in the optimal attraction region of CU-CU and RNA-PT interactions is fairly narrow. Next, we examined the feasibility of assembly with sets of parameters that provided marginal stability. This tested our hypothesis that there the parameters must lie somewhere between those that describe strong protein-protein and protein-RNA interactions (guaranteeing the stability of the final model), and those that describe weak intermolecular interactions (allowing transient structures to be rearranged, and preventing kinetic capture in local energy minima). We examined both co-transcriptional and post-transcriptional protocols.

One very important feature of this work is that the model is based on the structure of a real virus, STMV. In particular, we implemented the accepted secondary structure

model for the genomic RNA of STMV in the mature virus ²⁵, and the model of the mature virus that was examined in the stability simulations is based on our recent all-atom STMV model ²⁷; this model incorporates the 30 stem-loops of the Schroeder RNA secondary structure model. Most significantly, we find that the law of mass action can drive the assembly process, even in the absence of terms in the energy function that enforce specific tertiary interactions from the model of the mature virus, and even without terms that enforce five-fold symmetry at the vertices of the capsid.

RNA structure is an important factor for choosing the assembly protocol. Co-transcriptional assembly mostly likely occurs with RNA secondary structure having short-range local helices, because initial CU binding to short-range stem prevents the RNA from expanding. A smaller RNA model (25 stems) is required for the post-transcriptional assembly protocol. Post-transcriptional assembly probably requires an RNA secondary structure having a mixture of short and long-range helices, but there are no long-range helices in Schroeder's secondary structure. RNA secondary structures with a mixture of long and short-range helices are more compact relative to structures only dominant with short-range local helices. As Yoffe *et al.* ³² demonstrated, the secondary structures of genomic RNAs of small icosahedral RNA viruses are more highly branched than secondary structures of RNA with random sequences, which almost certainly guarantees that RNAs from these viruses are more compact in three dimensions than other RNAs. This structural constraint might be an advantage for the post-transcriptional assembly protocol. In addition to having long-range helices, the other physiological factors *i.e.* pH, ionic strength and type of cations affect the structure of the RNA.

Localization is another factor on the assembly process. The localization of two processes, the protein translation and the RNA replication may play an important role on the preference of the assembly protocols. If these two processes occur in separate parts of the cell, post-transcriptional assembly might be dominant, if not co-transcriptional assembly protocol might be the dominant.

References

1. Schneemann, A. *Annu Rev Microbiol* **2006**, 60, 51-67.
2. Johnson, J. E.; Speir, J. A. *Journal of molecular biology* **1997**, 269, (5), 665-75.
3. Guo, P.; Peterson, C.; Anderson, D. *Journal of molecular biology* **1987**, 197, (2), 229-36.
4. Bancroft, J. B.; Hills, G. J.; Markham, R. *Virology* **1967**, 31, (2), 354-79.
5. Bancroft, J. B. *Advances in virus research* **1970**, 16, 99-134.
6. Speir, J. A.; Munshi, S.; Wang, G.; Baker, T. S.; Johnson, J. E. *Structure* **1995**, 3, (1), 63-78.
7. Zhao, X.; Fox, J. M.; Olson, N. H.; Baker, T. S.; Young, M. J. *Virology* **1995**, 207, (2), 486-94.
8. Gopal, A.; Zhou, Z. H.; Knobler, C. M.; Gelbart, W. M. *RNA* **2012**, 18, (2), 284-99.
9. Buchmueller, K. L.; Webb, A. E.; Richardson, D. A.; Weeks, K. M. *Nature structural biology* **2000**, 7, (5), 362-6.
10. Ribitsch, G.; De Clercq, R.; Folkhard, W.; Zipper, P.; Schurz, J.; Clauwaert, J. *Zeitschrift fur Naturforschung. Section C: Biosciences* **1985**, 40, (3-4), 234-41.
11. Caliskan, G.; Hyeon, C.; Perez-Salas, U.; Briber, R. M.; Woodson, S. A.; Thirumalai, D. *Physical review letters* **2005**, 95, (26), 268303.
12. Chu, V. B.; Bai, Y.; Lipfert, J.; Herschlag, D.; Doniach, S. *Current opinion in chemical biology* **2008**, 12, (6), 619-25.
13. Draper, D. E. *RNA* **2004**, 10, (3), 335-43.

14. Belyi, V. A.; Muthukumar, M. *Proceedings of the National Academy of Sciences of the United States of America* **2006**, 103, (46), 17174-8.
15. Devkota, B.; Petrov, A. S.; Lemieux, S.; Boz, M. B.; Tang, L.; Schneemann, A.; Johnson, J. E.; Harvey, S. C. *Biopolymers* **2009**, 91, 530-8.
16. Zandi, R.; Reguera, D.; Bruinsma, R. F.; Gelbart, W. M.; Rudnick, J. *Proceedings of the National Academy of Sciences of the United States of America* **2004**, 101, (44), 15556-60.
17. Hagan, M. F.; Chandler, D. *Biophys J* **2006**, 91, (1), 42-54.
18. Zhang, D.; Konecny, R.; Baker, N. A.; McCammon, J. A. *Biopolymers* **2004**, 75, (4), 325-37.
19. Arkhipov, A.; Freddolino, P. L.; Schulten, K. *Structure* **2006**, 14, (12), 1767-77.
20. Angelescu, D. G.; Bruinsma, R.; Linse, P. *Phys Rev E Stat Nonlin Soft Matter Phys* **2006**, 73, (4 Pt 1), 041921.
21. Forrey, C.; Muthukumar, M. *Biophys J* **2006**, 91, 25-41.
22. Hagan, M. F. *J Chem Phys* **2009**, 130, (11), 114902.
23. Elrad, O. M.; Hagan, M. F. *Physical biology* **2010**, 7, (4), 045003.
24. Larson, S. B.; Day, J.; Greenwood, A.; McPherson, A. *Journal of molecular biology* **1998**, 277, (1), 37-59.
25. Schroeder, S. J.; Stone, J. W.; Bleckley, S.; Gibbons, T.; Mathews, D. M. *Biophys J* **2011**, 101, (1), 167-75.
26. Larson, S. B.; McPherson, A. *Curr Opin Struct Biol* **2001**, 11, 59-65.
27. Zeng, Y.-Y.; Larson, S. B.; Heitsch, C. E.; McPherson, A.; Harvey, S. C. *J Struct Biol* **in press**.
28. Malhotra, A.; Tan, R. K.; Harvey, S. C. *Biophys J* **1994**, 66, 1777-95.
29. Malhotra, A.; Harvey, S. C. *Journal of molecular biology* **1994**, 240, (4), 308-40.
30. Harvey, S. C.; Petrov, A. S.; Devkota, B.; Boz, M. B. *Meths Enzymol* **2011**, 487, 513-543.
31. Nguyen, H. D.; Reddy, V. S.; Brooks, C. L., 3rd. *Nano letters* **2007**, 7, (2), 338-44.

32. Yoffe, A. M.; Prinsen, P.; Gopal, A.; Knobler, C. M.; Gelbart, W. M.; Ben-Shaul, A. *Proceedings of the National Academy of Sciences of the United States of America* **2008**, 105, (42), 16153-8.

CHAPTER 6

Conclusion and Future work

YUP scripts

YUP is one of the most efficient coarse-grained modeling programs capable of creating DNA, RNA and protein models using simple python libraries. I wrote two scripts: *ct_to_blue.py* and *LAMMPS.py* to further improve the RNA modeling package of YUP: *rrRNA.py*. The first script, *ct_to_blue.py*, converts the secondary structure information of an RNA molecule from CT file to BLUEPRINT file. The second script, *LAMMPS.py*, converts the YUP model of the RNA to LAMMPS model. With the addition of these scripts, we can generate a coarse-grained model of any RNA with a given CT file easily. The codes of these scripts are attached in Appendix B. YUP and LAMMPS have been extensively used in our virus modeling and simulations.

Modeling PaV

We present the first all-atom model of T=3 virus, PaV. This model provides insights about the importance of the positively charged protein tails on the stability of the virus. We generated two models with the location variation of the protein tails. In the first model protein tails penetrate deeply into the viral interior with the protein tail-RNA soft sphere contact distance $d_0 = 8\text{\AA}$. The second model has the protein tails predominantly associated with RNA near the outer regions with a larger contact distance of RNA-protein tail ($d_0 = 12\text{\AA}$) providing less penetration. Calculating the final energy of the two models clearly demonstrated that the first model with lower energy is more stable than the second model.

The interesting observation of Belyi and Muthukumar [1] is that the ratio of the genome size to the net charge on the terminal protein tails is 1.61 ± 0.03 among 16 single-stranded RNA and DNA viruses. This suggests that the mechanism described above may apply to many single-stranded viruses since the ratio seems to be consistent and narrow. This observation may also imply that the attraction between the RNA and the positively charged protein tails is very sensitive to change. The attraction of these should be strong enough that the initial collapse neutralizes the charges of RNA. Otherwise RNA-RNA repulsion will disturb the aggregation. These attractions cannot be so strong either, because the aggregate will be stuck with fixed configurations that never lead to a successful assembly.

Our proposed model [2] provides a simple mechanistic basis for explaining how the relatively weakly associating proteins can force RNA into a small compact volume: the very strong electrostatic interactions between the negatively charged RNA and the positively charged protein tails provide a sufficiently favorable change in enthalpy to overcome the unfavorable entropic penalty associated with the dramatic reduction in RNA conformational space.

Capsid Simulations:

The empty capsid simulation is the first step towards to the whole virus simulations. These simulations helped us to understand the kinetics of the empty capsid assembly. We have experimented with the capsid model and potentials. We generated a wedge-shaped capsid unit similar to Brooks' model [3] and achieved the assembly of a T=1 capsid. We varied both the potential and the edge angle of the capsid unit. The simulations with the edge angle variation demonstrated that T=1 capsid assembly is

achieved in the range from 16.9° to 20.9°. We observed the formation of two dimers: curved and flat in our simulations. Curved dimers were desired to form T=1 capsids, however the flat dimers were not needed for the assembly of the capsid. Applying the specific potential lowers the energy of curved dimers and increases the probability of the having curved dimers. We preferred to apply the specific potential in our whole virus simulations.

Virus Simulations

We have demonstrated that the optimal attraction region of CU-CU and RNA-PT is very narrow, and we have achieved assembly of the model T=1 virus using the co-transcriptional assembly protocol. We used an RNA model having 30 stem-loops and short single-stranded regions between stems. This model is 212 nucleotides shorter than the RNA model we used in the stability simulations. This is due to the lack of tertiary information for these long single-stranded (~20-25 nucleotides) regions.

The post-transcriptional protocols did not yield a T=1 virus using the 30 stem-loops. We have tried to bias and confine the RNA structure. We equilibrated the RNA in smaller boxes with fixed boundaries yielding three different radii of gyration: 97.4, 69.7, 51.1. As soon as the assembly simulation starts with the presence of the capsid units, the RNA-RNA repulsion dominates the kinetics and the RNA expands before the capsid units bind and stabilize the compact RNA. The simulations of post-transcriptional assembly yielded conjoint virus formation even in the optimal strength of RNA-TP and CU-CU interactions.

We shortened the number of stem from 30 to 25 and then we were able to achieve the assembly of a T=1 virus. This result demonstrates that our secondary structure is not suitable for the post-transcriptional assembly protocol.

RNA structure is an important factor for choosing the proper assembly protocol. Co-transcriptional assembly works better for an RNA secondary structure with short-range local helices, because initial CU binding to short-range stems prevents RNA expansion. A smaller RNA model (25 stems) is required for the post-transcriptional assembly protocol. Post-transcriptional assembly may work better with an RNA secondary structure having mixture of long-range and short-range helices. These types of RNA secondary structures are more compact relative to structures with only short-range local helices. As Yoffe *et.al* [4] demonstrated, most of the virus RNAs are more compact than the random RNA secondary structures. This structural constraint might be an advantage for the post-transcriptional assembly process. In addition to having long-range helices, the other physiological factors *i.e.* pH, ionic strength and type of cations affect the compactness of the RNA structure.

Localization is another factor for the assembly and the structure of the RNA. The localization of two processes, protein translation and RNA replication, may play an important role on the preference of the assembly protocols. If these two processes occur separate in parts of the cell, the RNA may fold slowly and end up with a more compact structure. Thus, post-transcriptional assembly might be dominant. If RNA replication occurs at the presence of a high concentration of proteins, the RNA may fold into short-ranged local helices with the help of the proteins. Therefore assembly continues via the co-transcriptional protocol.

Future work:

Deciphering the assembly of viruses is a very challenging task and it involves both experimental and computational efforts. My work on virus assembly using coarse-grained models can be further improved in the presence of new experimental information about tertiary structures of both the capsid unit and the RNA. The further information can even lead to all-atom simulations of viruses with improved computational power and can yield more accurate thermodynamic and kinetic understanding of the virus assembly.

References:

1. Belyi VA and Muthukumar M, (2006) Electrostatic origin of the genome packing in viruses, *Proc. Natl. Acad. Sci. U. S. A.*, 103(46), 17174–8.
2. Devkota B, Petrov AS, Lemieux S, Boz MB, Tang L, Schneemann A, Johnson JE, Harvey SC, (2009) Structural and electrostatic characterization of pariacoto virus: implications for viral assembly. *Biopolymers* 91:530-8.
3. Nguyen HD, Reddy VS, Brooks CL (2007) Deciphering the kinetic mechanism of spontaneous self-assembly of icosahedral capsids. *Nano letters* 7:338-44.
4. Yoffe AM, Prinsen P, Gopal A, Knobler CM, Gelbart WM and Ben-Shaul A (2008) Predicting the sizes of large RNA molecules, *Proc. Natl. Acad. Sci. U. S. A.*, 105(42):16153–8

APPENDIX A

PRE-EQUILIBRATED RNAS

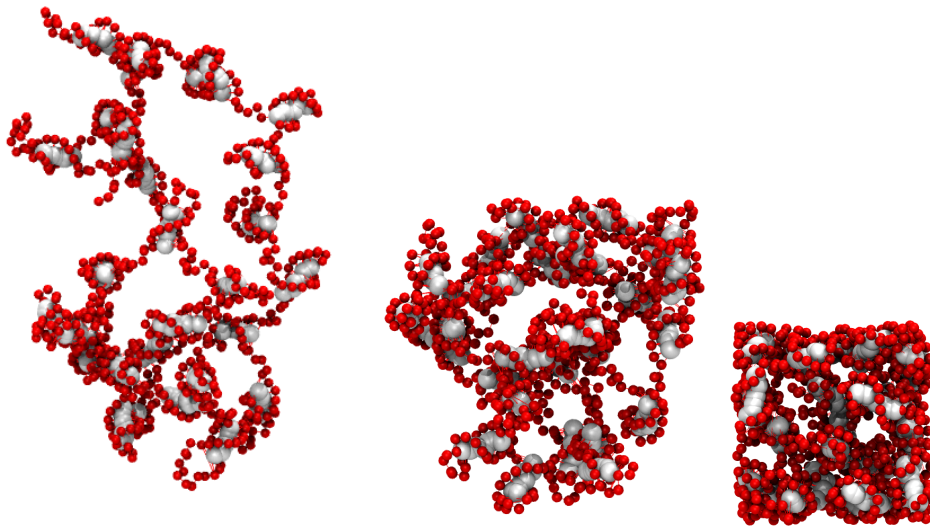


Figure A.1: 3 pre-equilibrated RNA models.

APPENDIX B

YUP SCRIPTS

Three scripts: *ct_to_blueprint.py*, *LAMMPS.py*, *run_rrRNA.py* are listed.

ct_to_blueprint.py converts ct files to blueprint files required for YUP *rrRNA.py* module.

LAMMPS.py converts RNA model generated in YUP to LAMMPS ready model. Last script is an example of how YUP model is created and converted to LAMMPS model.

ct_to_blueprint.py

```
#!/usr/bin/env python
```

```
# This script is written by Mustafa Burak Boz (04/05/09). It is modified at 05/24/09.
```

```
# It reads MFOLD/UNAFOLD CT file and creates a very simple blueprint file required  
for YUP rrRNA module
```

```
import sys
```

```
def read_pdb(pdb):
```

```
    input=open(pdb,"r")
```

```
    p_coord=[]
```

```
    check=0
```

```
    for line in input:
```

```
        if line[0:4] == "ATOM" and line[13:14] == "P":
```

```
            x,y,z=line[30:38],line[38:46],line[46:55]
```

```
            p_coord.append([float(x),float(y),float(z)])
```

```
            check=1
```

```
    if check==0:
```

```
        print "Warning: P atoms couldn't be found in the file"
```

```
    input.close()
```

```
    return p_coord
```

```
def read_ct_file(ifn):
```

```
    input=open(ifn,"r")
```

```

seq=[]
base_pairs={}
first_line=input.readline().split()
helix_index={}
i=1
for line in input:
    a=line.split()
    seq.append(a[1])
base_pairs[i]=int(a[4])
    if base_pairs[i] != 0:
        helix_index[i]=int(a[4])
    i=i+1
input.close()
return seq, base_pairs, helix_index

def check_tract(i):
    check =0
    if i == 0:
        check=1
    return check

def look_for_stem(strand,helix_index):
    stems=[]
    temp=[]
    i=0
    while i < len(strand)-1:
        if helix_index[strand[i]]-1==helix_index[strand[i+1]]:
            temp.append([strand[i],helix_index[strand[i]]])
            i=i+1
        else:
            temp.append([strand[i],helix_index[strand[i]]])
            stems.append(temp)
            temp=[]

```

```

        i=i+1

    temp.append([strand[i],helix_index[strand[i]]])
    stems.append(temp)

return stems

def look_for_tracts(tracts_list):
    tracts=[]
    temp=[]
    i=0
    while i < len(tracts_list)-1:
        if tracts_list[i]+1==tracts_list[i+1]:
            temp.append(tracts_list[i])
            i=i+1
        else:
            temp.append(tracts_list[i])
            tracts.append(temp)
            temp=[]
            i=i+1

    temp.append(tracts_list[i])
    tracts.append(temp)
    return tracts

def remove_duplication(stems):
    stems_copy=stems[:]
    for i in range(len(stems)):
        for j in range(i+1,len(stems_copy)):
            if stems_copy[i][-1][1] == stems_copy[j][0][0]:
                stems.remove(stems_copy[j])

def tracts_and_helices(base_pairs,helix_index,helices):

```

```

tracts=[]
stems=[]
for strand in helices:
    temp=look_for_stem(strand,helix_index)
    stems.extend(temp)

remove_duplication(stems)
helix_list=[]
stems_copy=stems[:]
for stem in stems_copy:
    if len(stem) < 3:
        stems.remove(stem)
for stem in stems:
    for each_pair in stem:
        for base in each_pair:
            helix_list.append(base)

tracts_list=[]
for i in range(1,len(base_pairs)+1):
    if not i in helix_list:
        tracts_list.append(i)
tracts=look_for_tracts(tracts_list)
return tracts,stems
def pick_helices(base_pairs):
    tracts=[]
    helices=[]
    temp=[]
    temp_2=[]
    change=0
    if check_tract(base_pairs[1]) == 1:
        temp.append(1)
    else:
        temp_2.append(1)
        change=1

```

```

for i in range(2,len(base_pairs)+1):
    change_old=change
    if check_tract(base_pairs[i]) == 1:
        temp.append(i)
        change=0
    else:
        temp_2.append(i)
        change=1
    if change_old != change:
        if change_old ==0:
            tracts.append(temp)
            temp=[]
        else:
            helices.append(temp_2)
            temp_2=[]

    if change ==0:
        tracts.append(temp)
    else:
        helices.append(temp_2)
return helices
def pick_smaller(a,b,helix):
    if a < b:
        smaller=a
    elif b < a:
        smaller=b
    else:
        print "there is a problem with helix selection!!!"
        sys.exit()
    return smaller

def helix_yup_format(helices):

```

```

yup_helix_format=[]
for helix in helices:
    if helix[0][0] < helix[-1][1]:
        yup_helix_format.append([pick_smaller(helix[0][0],helix[-1][0],helix),len(helix),pick_smaller(helix[0][1],helix[-1][1],helix)])

    else:
        yup_helix_format.append([pick_smaller(helix[0][1],helix[-1][1],helix),len(helix),pick_smaller(helix[0][0],helix[-1][0],helix)])

return yup_helix_format

def reconstruct_tracts(tracts):
    yup_format_tracts=[]
    for tract in tracts:
        yup_format_tracts.append([tract[0],tract[-1]])
    return yup_format_tracts

def write_blue_print(yup_helix_format,yup_tracts_format,seq,ofn,coordinates,pdb):
    output=open(ofn,"w")
    temp="from Yup.Models.rrRNAv1.const import DOMAIN, TRACT, HELIX, RNA_RNA, RNA_BSQ, RNA_FIX, RNA_XYZ\n\n"
    output.write(temp)
    temp="BLUE = {}\n"
    output.write(temp)

    all_helices_tracts=[]
    all_helices_tracts.extend(yup_helix_format)
    all_helices_tracts.extend(yup_tracts_format)

    all_helices_tracts.sort( lambda x, y: cmp( x[0], y[0] ) )

```

```

helix_i=1
tract_i=1
domain=""

for element in all_helices_tracts:
    if len(element) == 3: # It is HELIX
        h_name="H_"+str(helix_i)
        temp=h_name+" = ( HELIX, 'helix_'+str(helix_i)+",
("+str(element[0])+", "+str(element[1])+", "+str(element[2])+"))\n"
        output.write(temp)
        domain=domain+h_name+', '
        helix_i=helix_i+1

    else: # It is TRACT
        t_name="T_"+str(tract_i)
        temp=t_name+" = ( TRACT, 'tract_'+str(tract_i)+",
("+str(element[0])+", "+str(element[1])+"))\n"
        output.write(temp)
        domain=domain+t_name+', '
        tract_i=tract_i+1

output.write("\n")
temp="BLUE[RNA_RNA]= (DOMAIN, 'all',"
temp=temp+"("+domain[:-1]+"))\n"
output.write(temp)


temp="BLUE[RNA_BSQ] = ("
temp_2=""
for base in seq:
    temp_2=temp_2+"""+base.upper()+""", "
temp=temp+temp_2[:-1]+"))\n"

```

```

        output.write(temp)

        if coordinates=="on":
            Coords=read_pdb(pdb)
            temp="BLUE[RNA_XYZ] = ("+" ,\n".join(map(str,(tuple(i) for i in
Coords))))+" )\n"
        else:
            temp="BLUE[RNA_XYZ] = ( )\n"
        output.write(temp)
        temp="BLUE[RNA_FIX] = ()"
        output.write(temp)
        output.close()

```

```

def run():

```

```

    # Arguments: [1] input .ct file, [2] output blueprint file,

```

```

    L = len( sys.argv )

```

```

    coordinates="off"

```

```

    pdb = "None"

```

```

    if L == 3 :

```

```

        ifn = sys.argv[1]

```

```

        ofn = sys.argv[2]

```

```

    elif L==4:

```

```

        ifn=sys.argv[1]

```

```

        ofn=sys.argv[2]

```

```

        pdb=sys.argv[3]

```

```

        coordinates="on"

```

```

    else:

```

```

        print '\n\tUsage: name_of_the_script .ctfile outputfile pdbfile\n'

```

```

print '\tctfile\t\t: name of .ct file including ct extension'
print '\toutputfile\t: name of output file with a py extension '
print '\tpdbfile\t\t: name of the pdbfile containing P atom coordinates \n'
print '\tNote\t\t: pdbfile is optional. If it is not provided, the coordinates will be
generated later by rrRNAv1 module of YUP'
sys.exit()

```

```

seq,base_pairs,helix_index    = read_ct_file(ifn)
helix_strand_list             = pick_helices(base_pairs)
tracts,helices                 =
tracts_and_helices(base_pairs,helix_index,helix_strand_list)
yup_helix_format              = helix_yup_format(helices)
yup_tracts_format              = reconstruct_tracts(tracts)
write_blue_print(yup_helix_format, yup_tracts_format, seq, ofn,coordinates,pdb)

run()

```

LAMMPS.py

""""LAMMPS.py: LAMMPS class to convert a YUP Model object into input data and config files for LAMMPS. This file is created by Mustafa Burak Boz from the original file ParmTop.py wrtitten by Robert Tan. This script is optimized for rrRNAv1 model""""

```

from Yup.Tools.Atoms import EveryAtom
from Yup.Taro.Model import Model

```

```

def _inter_const( terms ):
# <terms> is assumed to be a list of two tuples, the first tuple contains the interacting
atoms and the second tuple
# the parameters of the energy term. The returned values are two lists. The first contains
the tuple of interacting
# atoms with the addition of the index to the parameter in the second list. The second is a
slimmed down list of
# unique force parameters.
# --- split <terms> into <inter> and <parms> both with added placeholders

```

```

parms = []
inter = []
L = len( terms )
i = 0
while i < L:
    T = terms[i]
    inter.append( [ T[0], 0 ] )
    parms.append( [ T[1], i, 0 ] )
    i += 1

# --- sort <parms> according to the actual parameters - the first item
parms.sort( lambda x, y: cmp( x[0], y[0] ) )
# --- for <parms>: index the unique items, duplicates get index of original item ...
p = 1
parms[0][2] = p
const = [ parms[0][0] ]
# ... and collect the unique parameters into another list <const>
i = 1
while i < L:
    P = parms[i]
    if P[0] != parms[i-1][0]:
        p += 1
        const.append( P[0] )
    P[2] = p
    i += 1

# --- sort <parms> by its index, i.e. return to its original order ...
parms.sort( lambda x, y: cmp( x[1], y[1] ) )
# ... which allows us to assign the index item for <inter>
i = 0
while i < L:
    inter[i][1] = parms[i][2]
    i += 1

# --- return the lists
return inter, const

```

class LAMMPS:

```
def __init__( self, m ):
    if not isinstance( m, Model ): raise ValueError, 'must provide a Model
object'

    self.MODEL = m
    root = m.Map
    self.NumAt = root.numatoms
    self.ATOMS = EveryAtom( root )
    # --- for the moment we can handle only four types of interactions
    Eb = [] # EnergyID = 100001
    Ea = [] # EnergyID = 100003
    Et = [] # EnergyID = 100007
    En = [] # EnergyID = 100008
    # --- join multiple instances of each energy type

    for E in m.Energy.MEMBERS:
        Eid = E.EnergyID

        if Eid == 100001:    # bonds
            for i, j, Kb, B0 in E.termlist:
                Eb.append( ( ( i, j ), ( Kb, B0 ) ) )
        elif Eid == 100003:  # angles
            for i, j, k, Ka, A0 in E.termlist:
                Ea.append( ( ( i, j, k ), ( Ka, A0 ) ) )
        elif Eid == 100007:  # improper torsions
            for i, j, k, l, Kt, T0 in E.termlist:
                Et.append( ( ( i, j, k, l ), \
                    ( Kt, T0 ) ) )
        elif Eid == 100008:  # Soft Sphere Exclusion (SSX)
            for X in E.termlist:
                En.append( X )
```

```

        # elif Eid == 100012: # VanderWaals Exclusion (VDWX)
            # probably have to select vanderWaals or SSX not both
        elif Eid == 100012: # Electrostatics Exclusion (ELX)
            pass
        else:
            raise RuntimeError, 'cannot handle %d term' % Eid

self.EXCTUP=En

self.BLIST, self.BPARM = _inter_const( Eb )
self.ALIST, self.APARM = _inter_const( Ea )
self.TLIST, self.TPARM = _inter_const( Et )

# --- other things that we can determine for all types of model
self.AMASS = []
self.CHARG = []
for a in self.ATOMS:
    self.AMASS.append( a.mass )
    self.CHARG.append( a.charge )

def writedata(self, prefix='lammmps'):

    output=open(prefix+'_data.txt',"w")
    text=[]
    text.append("LAMMPS data file for "+prefix+"_data.txt written by mbb
\n\n" )

    text.append(`self.NumAt`+' atoms \n')
    text.append(`len(self.BLIST)`+' bonds\n')
    text.append(`len(self.ALIST)`+' angles\n')
    text.append(`len(self.TLIST)`+' impropers\n\n')
    text.append(`self.NumAt`+' atom types\n')
    text.append(`len(self.BPARM)`+' bond types\n')
    text.append(`len(self.APARM)`+' angle types\n')

```

```

text.append(`len(self.TPARM)` + ' improper types\n\n')

coords=list(self.MODEL.Coordinates.intopy())
coords.sort( lambda x, y: cmp( x[0], y[0] ) )
xlow=coords[0][0] -100.0
xhigh=coords[-1][0] + 100.0
coords.sort( lambda x, y: cmp( x[1], y[1] ) )
ylow=coords[0][1] -100.0
yhigh=coords[-1][1] +100.0
coords.sort( lambda x, y: cmp( x[2], y[2] ) )
zlow=coords[0][2] -100.0
zhigh=coords[-1][2] +100.0
del coords

text.append(`xlow`+' '+'xhigh`+' xlo xhi \n'+`ylow`+' '+'yhigh`+' ylo yhi
\n'+`zlow`+' '+'zhigh`+' zlo zhi \n\n')
text.append("\n")
text.append("Masses \n\n")
i=1
for mass in self.AMASS:
    text.append(`i`+' '+'mass`+'\n')
    i+=1
text.append("\n")
text.append("Pair Coeffs \n\n")
i=1
for mass in self.AMASS:
    if mass == 300.0:
        text.append(`i`+' 0.01 5.3\n')
    else:
        text.append(`i`+' 0.01 9.5\n')
    i+=1
text.append("\n")
text.append('Bond Coeffs \n\n')
i=1

```

```

for set in self.BPARM:
    text.append(`i`+' '+'set[0]`+' '+'set[1]`+'n' )
    i+=1
text.append(`n`)
text.append('Angle Coeffs \n\n')
i=1
for set in self.APARM:
    text.append(`i`+' '+'set[0]`+' '+'set[1]`+'n' )
    i+=1
text.append(`n`)
text.append('Improper Coeffs \n\n')
i=1
for set in self.TPARM:
    text.append(`i`+' '+'set[0]`+' '+'abs(set[1])`+'n' )
    i+=1
coords=self.MODEL.Coordinates.intopy()
text.append(`n`)
text.append('Atoms\n\n')
i=1
j=1
for coord in coords:
    text.append(`i`+' '+'j`+' '+'i`+' '+'self.CHARG[i-1]`+'
'+`coord[0]`+' '+'coord[1]`+' '+'coord[2]`+'n' )
    i+=1
text.append(`n`)
text.append('Bonds\n\n')
i=1
for set in self.BLIST:
    text.append(`i`+' '+'set[1]`+' '+'set[0][0]`+' '+'set[0][1]`+'n' )
    i+=1
text.append(`n`)
text.append('Angles\n\n')
i=1

```

```

        for set in self.ALIST:
            text.append('i`+' '+'set[1]`+' '+'set[0][0]`+' '+'set[0][1]`+'
'+'set[0][2]`+' '\n' )
            i+=1
        text.append('\n')
        text.append('Improper\n\n')
        i=1
        for set in self.TLIST:
            text.append('i`+' '+'set[1]`+' '+'set[0][0]`+' '+'set[0][1]`+'
'+'set[0][2]`+' '+'set[0][3]`+' '\n' )
            i+=1
        output.writelines(text)
        output.close()
        config=open('run_'+prefix+'_config.txt','w')
        text=[]
        text.append('units\t\t real \n')
        text.append('atom_style\t full \n')
        text.append('bond_style\t\t harmonic \n')
        text.append('angle_style\t\t harmonic\n')
        text.append('improper_style\t harmonic \n\n')
        text.append('pair_style\t lj/cut/coul/cut 50.0\n\n')
        text.append('read_data\t\t\t '+'prefix+'_data.txt \n\n')
        for pair in self.EXCTUP:
            text.append('neigh_modify exclude type '+'pair[0]`+'
'+'pair[1]`+' '\n')
            text.append('\ntimestep\t\t\t 40.0 \n')
            text.append('neigh_modify\t delay 1 \n')
            text.append('dielectric\t 80.0 \n')
            text.append('thermo_style\t multi \n')
            text.append('thermo\t\t\t 100 \n')
            text.append('velocity\t\t all create 275.0 4928459 dist gaussian \n')
            text.append('fix\t\t\t 1 all nvt 300.0 300.0 300.0 \n')
            text.append('dump\t\t\t 1 all dcd 50 dump_test.dcd \n')

```

```

text.append('min_style\t\t cg \n')
text.append('minimize\t 1.0e-3 0.001 100000 1000000000 \n')
text.append('run\t\t\t 100000')
config.writelines(text)
config.close()

```

run_rrRNA.py

```

#!/usr/bin/env python

from Yup.Models.rrRNAv1.FFA import *
from Yup.Tools.MakeGraph import *
from Yup.Methods.MolMech import MolMech
from Yup.Methods.EnerMinim import Minimizers
from Yup.Methods.MolDynam import Motors
from Yup.Models.rrRNAv1.Analyzer import *
from LAMMPS.py import *
import sys

#----- MODEL CREATION -----#
R=rrRNAFFA()
L = len( sys.argv )
if L == 2 :
    blueprint_file = sys.argv[1][:3]
    init_file = "test_king"
    min_file = "test_king_min"
    md_file = "test_king_md"
    lammmps_name="lammmps"

elif L==6 :
    blueprint_file = sys.argv[1][:3]
    init_file = sys.argv[2]
    min_file = sys.argv[3]

```

```

md_file = sys.argv[4]
lammps_name=sys.argv[5]

else:

    print '\n\tUsage: python run_rrRNA_v1.py blueprint.py or '
    print '\n\tUsage: python run_rrRNA_v1.py blueprint.py kin_file_name_1
kin_file_name_2 kin_file_name_3 lammps_file_name yammp_file_name\n'
    print '\tblueprint.py\t: name of the blueprint file'
    print '\tkin_file_name_1\t: name of kin image file taken after initilized'
    print '\tkin_file_name_2\t: name of kin image file taken after minimized'
    print '\tkin_file_name_1\t: name of kin image file taken after molecular dynamics'
    print '\tlammps_name\t: name of lammps file, data and run files will have this
prefix'

    sys.exit()

R.addRNA(blueprint(blueprint_file), modname="rna_random", randomize=1,
dimensions=(5.8, 0.2, 2., 0., 0., 0.))
M=R.finish()
#----- TAKE A SNAPSHOT -----#
Kinemage( M, init_file )
#----- MINIMIZE THE INITIAL STRUCTURE -----#
O = Minimizers( M )
O.GradientNorm = 1e-2
minimize = O.SimpleSD
minimize( 10000 )
#----- TAKE A SNAPSHOT -----#
Kinemage( M, min_file )
#----- MD SIMULATION -----#
D = Motors( M )
D.set_ThermMethod( 'BERENDSEN' )
D.ThermalizationInterval = 25
dynamics = D.Verlet

```

```
dynamics( 400000, 300.0 )  
#----- TAKE A SNAPSHOT -----#  
Kinemage( M, md_file )  
#----- CONVERT THE MODEL TO LAMMPS -----#  
lammps_model=LAMMPS(M)  
lammps_model.writedata(lammps_name)
```

VITA

Mustafa Burak Boz received his BS degree in chemistry from Koç University, Istanbul, Turkey, where he was an undergraduate research assistant of the Dean of College of Science, Prof. Dr I. Ersin Yurtsever. He worked on optimization of Lennard-Jones and positively charged helium clusters in his undergraduate years. In his PhD years, he became interested in the structure of icosahedral viruses, and he worked on modeling and simulations of RNA viruses under the supervision of Prof. Dr Stephen C. Harvey, using a combination of atomistic and coarse-grained representations.