COUNTING AND ENUMERATION OF SELF-ASSEMBLY PATHWAYS FOR SYMMETRIC MACROMOLECULAR STRUCTURES

MEERA SITHARAM[*]
CISE dept., University of Florida, Gainesville, FL 32611.
Email: sitharam@cise.ufl.edu

MIKLÓS BÓNA[†]
Department of Mathematics, University of Florida, Gainesville, FL 32611.
Email: bona@math.ufl.edu

We consider the problem of explicitly enumerating and counting the assembly pathways by which an icosahedral viral shell forms from identical constituent protein monomers. This poorly understood assembly process is a remarkable example of symmetric macromolecular self-assembly occuring in nature and possesses many features that are desirable while engineering self-assembly at the nanoscale.

We use the new model of [24,25] that employs a static geometric constraint graph to represent the driving (weak) forces that cause a viral shell to assemble and hold it together. The model was developed to answer focused questions about the structural properties of the most probable types of successful assembly pathways. Specifically, the model reduces the study of pathway types and their probabilities to the study of the orbits of the automorphism group of the underlying geometric constraint graph, acting on the set of pathways.

Since these are highly symmetric polyhedral graphs, it seems a viable approach to explicitly enumerate these orbits and count their sizes. The contribution of this paper is to isolate and simplify the core combinatorial questions, list related work and indicate the advantages of an explicit enumerative approach.

1. Introduction

Icosahedral viral shell assembly is an outstanding example of nanoscale, macromolecular self-assembly occuring in nature [15]. Mostly identical *coat protein* monomers assemble with high rate of efficacy into a closed icosahedral *capsid* or *shell*; onset and termination are spontaneous, and assembly is robust, rapid and economical. All of these requirements are both desirable and difficult to achieve when engineering macromolecular self-assembly. See Figures 1.

[*] supported in part by NSF-EIA 0218435
[†] supported in part by NSA-MDA 904-03-1-0044

However the viral assembly process - just like any other spontaneous macromolecular assembly process such as molecular crystal formation - is poorly understood. Answering questions about viral assembly pathways can help both to encourage macromolecular assemblies for engineering, biosensor and gene therapy applications, but and also discourage assembly for arresting the spread of viral infection.

We use the viral assembly pathway model of [24,25] that employs static geometric constraints to represent the driving (weak) forces that cause a viral shell to assemble and hold it together. The model avoids dynamics and as a result is both tractable and tunable. Preliminary predictions of this model consistently explain existing experimental observations about viral shell assembly. This model was developed to answer focused questions about the structural properties of the most probable types of successful assembly *pathways*, which are essentially directed acyclic graphs (dags) representing valid constructions (or decompositions) of the virus. The nodes of these dags are biochemically *stable* subassemblies of the complete assembly, partially ordered by containment. See Figures 5, 6. Specifically, the model reduces the study of pathway types and their probabilities to the study of the orbits of the automorphism group of the underlying geometric constraint graph, acting on the set of pathways. In [24,25], efficient randomized algorithms are given that sample the pathway set to provide approximate answers to these questions.

Since the underlying graphs are highly symmetric polyhedral graphs, it seems a viable approach to instead explicitly enumerate these (perhaps simplified) orbits and count their sizes. The contribution of this paper is to isolate and simplify the core combinatorial questions, list related work and indicate the advantages of an explicit enumerative approach over the random sampling approach of [24,25]. The expectation is that a hybrid of the two approaches can be developed which leverages these advantages while incorporating the full generality of the model of [24,25].

Organization

Section 1.1 discusses viral structure assembly basics; Section 1.2 discusses the current state of knowledge on viral assembly including a brief sketch of the model of [24,25] and what it achieves. Section 2 develops the appropriate definitions, states and briefly discusses approaches to the combinatorial enumeration questions that we consider here and sketches their origins from the [24,25] model.

1.1. *Virus Preliminaries*

The viral shell is important in that it packages viral "life" i.e, the genomic nucleic acid, which could be single stranded DNA (*ssDNA*), double-stranded DNA, or RNA. However, in many cases, viral shell assembly occurs with no interference from the enclosed genetic material: empty shells, or shells packaging incomplete genomic material form with equal facility [1], a fact that simplifies the modeling. A symmetric shell [11] is a consequence of its consisting of (almost) identical monomers. The predominant structure of viral shells is icosahedral since the exact five-fold, three-fold and two-fold point-group symmetry of the icosahedron permits the *quasi-equivalent* symmetry [8] required to construct structures with a large number of monomers (see Figures 1, 4, 2). The number of monomers for each vertex of each triangle of the (20-triangle) icosahedron is refered to as the (typically small) '*T*' number: a T=1 virus shell has 60 identical monomers, a T=7 virus shell has 420 monomers etc. Our focus here is mainly on ssDNA T=1 viruses. See Figures 1 4, 2.

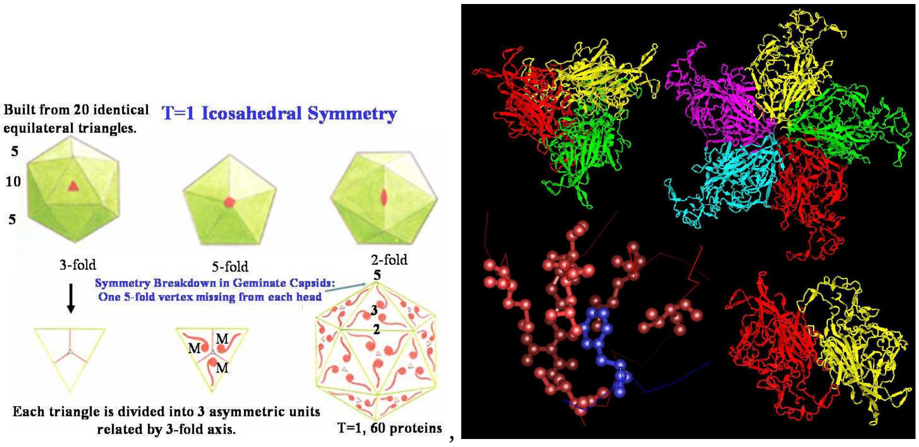

Built from 20 identical equilateral triangles.
T=1 Icosahedral Symmetry
5
10
5
3-fold
5-fold
2-fold
5
Symmetry Breakdown in Geminate Capsids: One 5-fold vertex missing from each head
3
2
M
M
M
Each triangle is divided into 3 asymmetric units related by 3-fold axis.
T=1, 60 proteins
,

Figure 1. (Left) Basic Viral Structure. (Right) Dimer, trimer and pentamer interactions and close-up of atomic interaction.

Virus assembly involves [28] highly specific monomer -monomer (protein -protein), - and possibly protein -genomic material *interactions*, all of which are governed by geometry or by weak forces that can be treated geometrically [9] (see Figure 1). More specifically, the final viral structure can be viewed formally as the solution to a system of geometric constraints that translate to algebraic equations and inequalities.

1.2. *Current state of knowledge and the New Geometric Constraint Model*

While there is a well developed structure theory of *complete* viral shells [11,8], verified by X-ray crystallography and other experimental data, the *processes* of viral shell assembly are poorly understood. From an experimental point of view, this lack of understanding is due to the extreme speed of the assembly so that wet-lab snapshots of intermediate sub-assemblies are generally unsuccessful.

From a modeling point of view, this lack of understanding is due to the fact that prior to the recent model of [24,25], previous computational models [5,6,28,27,26], [19,17,20,18] generally involve dynamics of (simplified versions) of virus assembly (further description of these approaches and comparison with the approach of [24,25], can be found in [24]). Dynamics were used previously even when the assembly models only sought to elucidate the structure of successful pathways.

Models whose output parameters are defined only as the end result of a dynamical process are computationally costly, often requiring oversimplifications to ensure tractability. In addition, such models are also not easily tunable or refinable since their input-to-output function is generally not analyzable and therefore do not provide a satisfactory conceptual explanation of the phenomenon being modeled.

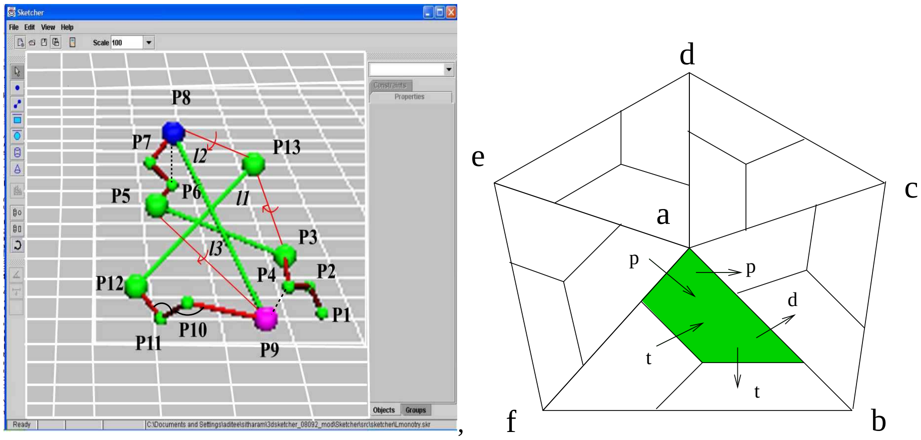
By carefully defining the probability space, using the successful assembly as a given, the static model of [24,25] gives a method to approximately compute the probabilities of successful pathway trees/dags efficiently.

Next we briefly describe the features of the [24,25] model of viral shell assembly.

The models *input parameters* are: information extracted from (a) the geometric structure of the coat protein monomer that forms the viral shell, including all relevant (rigid) conformations, Figure 2; (b) the geometric and weak-force interactions - between pairs of monomers - that drive assembly (see Figure 1). and (c) (optional) the inter-monomer contact or neighborhood structure of the complete viral shell, Figure 2.

The latter is crucial for a focused model that *only* deals with *those* pathways that are known *apriori* to lead to a complete viral shell. However, the model can be generalized to the case where (c) is not part of the input and unsuccessful assemblies are included. These input parameters are then converted into a tensegrity [14], [2], [10] and geometric constraint system [23,21] representation of the viral shell.

The output information sought from the model: first, the probability that a specific *type* of successful assembly pathway incorporates a specific *type* of subassembly, leads to the complete viral shell with bounded construction *effort*; in short, a probability distribution over successful, bounded effort assembly path-


Scale 100

Ready

C:\Documents and Settings\aditee\sitharam\3dsketcher_08092_mod\Sketcher\src\sketcher\Lmonotry.skr

Figure 2. (Left) Example monomer primitives and constraints. Balls (points) - atomic markers; Green line segments - variable length bonds; Arrows - torsion angles between green line segments; Red - distances representing fixed length bonds; Arcs – angles; Dotted lines – distances representing weak force; (screenshot using FRONTIER geometric constraint solving software). (Right) Part (c) of input to the model: Icosahedral T=1 viral shell's assembly-relevant interfaces that are used to construct the viral shell graph; the pentamer (p), trimer (t) and dimer (d) interfaces are shown for a reference monomer.

ways that incorporate certain substructures; this has a straightforward generalization ([24]) to a distribution over all possible assembly pathways (not necessarily successful) within an effort bound. The model satisfies the following requirements.

(i) The description of the model - i.e. the input-to-output function - is static, i.e. does not rely on dynamics of the assembly process. This is achieved using the state of the art theory of 3D geometric constraint decomposition [23,21] and is essential for forward analyzabilty.

(ii) The assumptions of the model are mathematically and biochemically justifiable. These justifications and rigorous comparisons of the model with existing models of viral shell assembly are given in [24].

(iii) The model is computationally tractable, i.e. *there is an efficient randomized algorithm for computing (a provably good approximation of) the pathway probability distribution.* The required algorithms are crucial modifications of state-of-the art 3d geometric constraint decomposition algorithms [23,21]. As a result, simulation software for the model is built directly upon existing opensource software for 3D geometric constraint solving [22]. See Figures 3.

Tractable computational simulation based on provably accurate algorithms is essential for backward analyzability which is needed for two reasons: first, for

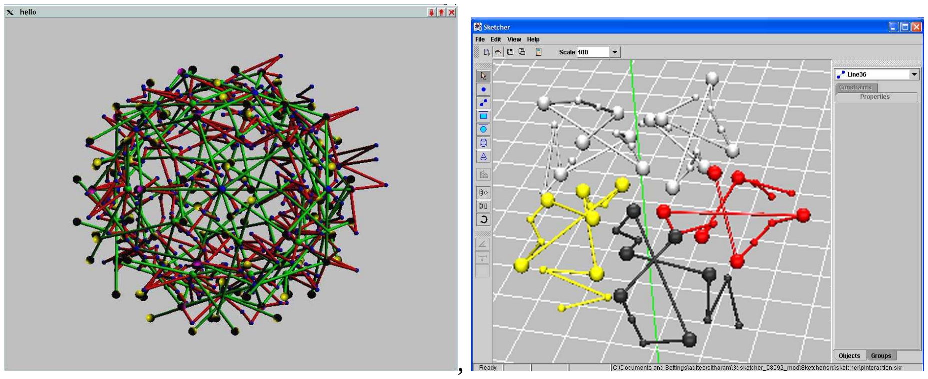

hello
Sketcher
File Edit View Help
Scale 100
Line36
Properties
Objects
Groups
Ready
C:\Documents and Settings\aditee\sitharam\3dsketcher_08092_mod\Sketcher\src\sketcher\pInteraction.skr

Figure 3. (Left)The simulated assembly of a T=1 viral shell: the solved viral geometric constraint graph of Figure 2. (Right) Pentamer subassembly of the viral geometric constraint graph of Figure 2.

iteratively refining the model so that its output matches known biochemical information or experimental results; and second, for engineering a desired output, for example engineering the monomer structure to prevent or encourage certain subassemblies, inorder to force certain pathways to become more likely than others, or to prevent successful assembly.

(iv) Preliminary simulation results (see [24], for example Figures 3) show that, in principle, the model's predictions are qualitatively consistent with known studies of viruses. More conclusive biochemical validation using 3 carefully chosen, ssDNA T=1 viruses is in process [24].

Overall, the model provides an indication of the direct, mutually beneficial interplay between (a) the concepts underlying macromolecular assembly and (b) established as well as novel concepts from combinatorial rigidity theory, geometric constraint solving, as well as polyhedral combinatorics as well as computational algebraic geometry and algebraic complexity, in general. Several promising open problems are indicated in [24].

2. Obtaining pathway type probabilities by combinatorial enumeration and counting

In this section, we first give simplified definitions of (icosahedral) viral shell graphs, valid pathways and pathway isomorphism types. These are abstracted from the geometric constraint model of Section 1.2 [24,25], although we do not explicitly discuss here the derivation process for this abstraction or simplification. For these simplified definitions, it is viable to approach the core question of estimating specific pathway type probabilities using explicit combinatorial enumeration and counting of specific types of constructions (or decompositions) of symmetric polyhedral graphs. We briefly list previous work relevant to this approach.

These probabilities are estimated for more general definitions of pathways and more general viral shell (geometric constraint) graphs that are used in the model of [24,25], using a randomized sampling method applied to a geometric constraint decomposition algorithm. The motivation of the explicit enumeration approach is to address 2 drawbacks of this method, given later in this section.

Definition 2.1. An *icosahedral $T = m$ viral shell graph* is obtained from a $T = m$ viral shell by representing each monomer as a vertex and each interface (relevant for assembly) between a pair of monomers by an edge.

The automorphism group of this graph is isomorphic to the icosahedral symmetry group of the viral shell. In fact, a more general result of [3,4] listing the possible automorphism groups of general polyhedral graphs could be useful for characterizing subgraphs of viral shell graphs that represent stable partial assemblies or subassemblies, whose significance will be clear below.

Definition 2.2. A *stable subgraph* S of a $T = m$ viral shell graph G is defined recursively. For the base case, a small set of at most k (independent of m) small *base stable* subgraphs of size at most $3m$ is specified and any subgraph S that is isomorphic to a base subgraph is stable. These constitute the *base set* B of stable subgraphs of G. For the recursion, S is stable if and only if it can be decomposed into a minimal *constituent set* Q of vertex disjoint stable subgraphs S_i such that there is a subgraph A of S that is in the base set and is not contained in the subgraph induced by any proper subset of Q.

A *stable subgraph type* is an isomorphism class obtained as the orbit of the natural action of the automorphism group of G on a stable subgraph.

For the simple T=1 viral shell graph obtained from the interfaces of Figure 2, the the two common base stable subgraphs would be 5 cycles and 3 cycles that correspond respectively to the two common stable subassemblies, pentamers and trimers, and larger stable subgraphs would be connected subgraphs built from trimers of pentamers and pentamers of trimers respectively, see Figures 4.

Definition 2.3. A *valid pathway* for a viral shell graph G is a tree where each node corresponds to a stable subgraph of G, the children of a parent form a constituent set for the parent, leaves are singleton vertices, and hence parents of leaves are subgraphs in the base set.

A *valid, successful pathway* is one whose root is the entire viral shell subgraph.

A *valid pathway type* is the isomorphism class obtained as the orbit of the natural action of the automorphism group of G on a valid pathway.

For example, Figures 5 and 6 show valid, successful pathway types and some of their representative pathways for the T=1 viral shell graph of Figure 2, in the case where the only stable subgraphs are built from trimers of pentamers and pentamers of trimers respectively (base stable set consists of trimers and pentamers).

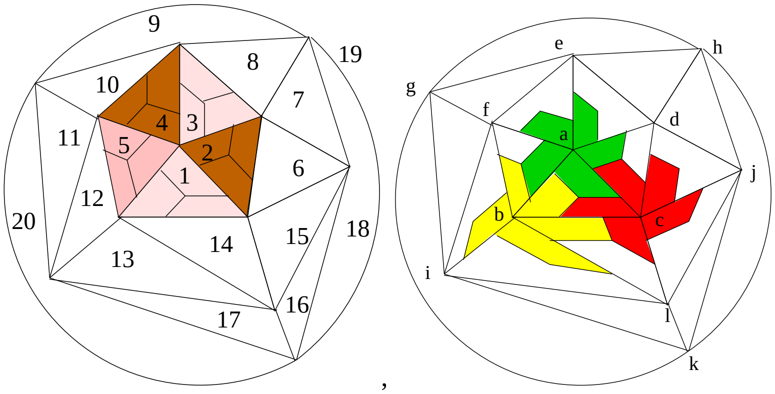


Figure 4. (Left) Facenumbers: pentamer of trimers in a T=1 shell. (Right) Vertex numbers: trimers of pentamers in a T=1 shell

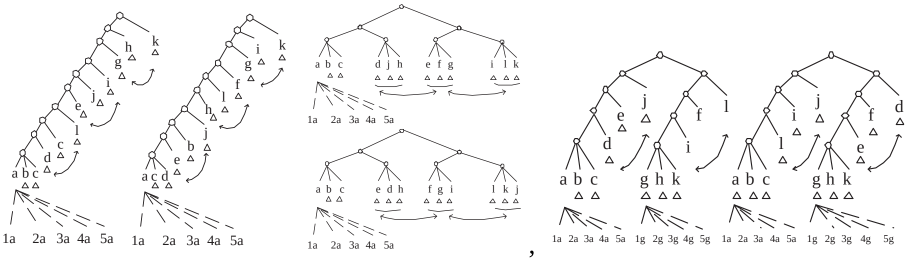


Figure 5. More T=1 Pathways based on pentameric stable subassemblies

From the model of Section 1.2 [25,24], one of the two factors that decides the probability of a successful pathway type is the size of its isomorphism class. The combinatorial enumeration and counting questions of interest are:

- to explicitly enumerate the valid (successful) pathway types of a viral

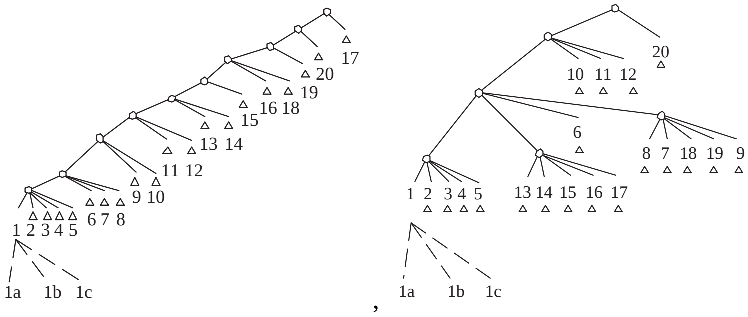


Figure 6. T=1 Pathways based on trimeric stable subassemblies

shell graph with a specified base set of stable subgraphs,

- to compute, for each valid (successful) pathway type, the size of its isomorphism class as a fraction of the total number of valid (successful) pathways,
- to answer both above questions for valid (successful) pathway types that contain a specified stable subgraph type.

By “explicit enumeration and counting” we do not preclude a clean algorithmic solution that, for instance, enumerates (or provides a clean data structure representation of) the pathway (type)s without duplication. Work on systematic constructions of Fullerenes [12,13,7,16] develop techniques that could be relevant to answering these questions. (Fullerenes are symmetric polyhedral graphs named after Buckminster Fuller who incidentally popularized tensegrity principles which inspired Caspar and Klug’s quasi-equivalence theory of viral structure).

At this point, an explicit counting approach suggested in this paper appears realistic only for the above simplified viral shell graphs - not for the full generality of viral geometric constraint graphs and pathways arising from the model of Section 1.2. However, the more general approach for estimating successful pathway type probabilities given in [25,24] - using randomized sampling of pathways and geometric decomposition algorithm - has other drawbacks. Specifically, it does not utilize the icosahedral symmetry of these viruses, potentially a powerful tool in determining the pathway isomorphism classes; secondly it does not easily extend to pathways with additional properties, such as the 3rd question above. The expectation is that a hybrid of these two approaches can be developed which addresses these drawbacks while incorporating the full generality of the model of [24,25].

References

1. M Agbandje, R McKenna, MG Rossmann, ML Strassheim, and PR Parrish. Structure determination of feline panleukopenia virus empty particles. *Proteins*, 16:155–171, 1993.
2. W Whiteley B Roth. Tensegrity frameworks. *Transactions of the AMS*, 265:419–446, 1981.
3. László Babai and Wilfried Imrich. On groups of polyhedral graphs.. *Discrete Math.*, 5:101–103, 1973.
4. László Babai and Wilfried Imrich. Sense preserving groups of polyhedral graphs. *Monatsh. Math.*, 79:1–2, 1975.
5. B. Berger, P. Shor, J. King, D. Muir, R. Schwartz, and L. Tucker-Kellogg. Local rule-based theory of virus shell assembly. *Proc. Natl. Acad. Sci. USA*, 91:7732–7736, 1994.
6. B Berger and PW Shor. Local rules switching mechanism for viral shell geometry. *Technical report, MIT-LCS-TM-527*, 1995.
7. Gunnar Brinkmann and Andreas Dress. A constructive enumeration of fullerenes. *Journal of Algorithms.*, 23:345–358, 1997.
8. D Caspar and A Klug. Physical principles in the construction of regular viruses. *Cold Spring Harbor Symp Quant Biol*, 27:1–24, 1962.
9. P Ceres and A Zlotnick. Weak protein-protein interactions are sufficient to drive assembly of hepatitis b virus capsids. *Biochemistry*, 41:11525–11531, 2002.
10. Robert Connelly. Tensegrity structures: why are they stable? *Rigidity Theory and Applications*, pages 47–54, 1998.
11. FHC Crick and JD Watson. Structure of small viruses. *Nature*, 177:473–475, 1956.
12. Antoine Deza, Michel Deza, and Viatcheslav Grishukhin. Fullerenes and coordiation polyhedra versus half-cube embeddings. *Discrete Math*, 192:41–80, 1998.
13. Tomislav Doslic. On some structural properties of fullerene graphs. *Journal of Mathematical Chemistry.*, 31(2):187–195, 2002.
14. A Edmondson. *A Fuller Explanation: The Synergetic Geometry of R. Buckminster Fuller*. Birkhauser Verlag, 1987.
15. D Brenner S Lyshefski G Iafrate WA Goddard. *Handbook of nanoscience engineering and technology*. CRC press, 2002.
16. Stanislav Jendrol and Marián Trenkler. More icosahedral fulleroids. *Journal of Mathematical Chemistry.*, 29(4):235–243, 2001.
17. J E Johnson and J A Speir. Quasi-equivalent viruses: a paradigm for protein assemblies. *J. Mol. Biol.*, 269:665–675, 1997.
18. C J Marzec and L A Day. Pattern formation in icosahedral virus capsids: the papova viruses and nudaurelia capensis β virus. *Biophys*, 65:2559–2577, 1993.
19. D Rapaport, J Johnson, and J Skolnick. Supramolecular self-assembly: molecular dynamics modeling of polyhedral shell formation. *Comp Physics Comm*, 1998.
20. V S Reddy, H A Giesing, R T Morton, A Kumar, C B Post, C L Brooks, and J E Johnson. Energetics of quasiequivalence: computational analysis of protein-protein interactions in icosahedral viruses. *Biophys*, 74:546–558, 1998.
21. M Sitharam. Frontier, an opensource 3d geometric constraint solver: algorithms and architecture. *monograph, in preparation*, 2004.
22. M. Sitharam. Frontier, opensource gnu geometric constraint solver: Version 1

(2001) for general 2d systems; version 2 (2002) for 2d and some 3d systems; version 3 (2003) for general 2d and 3d systems. In *http://www.cise.ufl.edu/∼sitharam*, *http://www.gnu.org*, 2004.

23. M Sitharam. Graph based geometric constraint solving: problems, progress and directions. In Dutta, Janardhan, and Smid, editors, *AMS-DIMACS volume on Computer Aided Design*, 2004.
24. M Sitharam and M Agbandje-Mckenna. A geometry and tensegrity based virus assembly pathway model. *submitted, available upon request*, 2004.
25. M. Sitharam and M. Agbandje-Mckenna. Modeling virus assembly pathways using computational algebra and geometry. In *Proceedings of the 10th Applications of Computer Algebra conference*, 2004.
26. A Zlotnick. To build a virus capsid: an equilibrium model of the self assembly of polyhedral protein complexes. *J. Mol. Biol.*, 241:59–67, 1994.
27. A Zlotnick, R Aldrich, J M Johnson, P Ceres, and M J Young. Mechanisms of capsid assembly for an icosahedral plant virus. *Virology*, 277:450–456, 2000.
28. A Zlotnick, JM Johnson, PW Wingfield, SJ Stahl, and D Endres. A theoretical model successfully identifies features of hepatitis b virus capsid assembly. *Biochemistry*, 38:14644–14652, 1999.