On the Mathematics of Virus Shell Assembly

Bonnie Berger* Peter W. Shor†

July 15, 1994

Abstract

A local rule theory is developed which shows that the self-assembly of icosahedral virus shells may depend on only the lower-level interactions of a protein subunit with its neighbors, i.e. *local rules,* rather than on larger structural building blocks. The local rule theory provides a framework for understanding the assembly of icosahedral viruses. These include both viruses that fall in the quasi-equivalence theory of Caspar and Klug and the polyoma virus structure, which violates quasi-equivalence and has puzzled researchers since it was first observed. Local rules are essentially templates for energetically favorable arrangements. The tolerance margins for these rules are investigated through computer simulations. When these tolerance margins are exceeded in a particular way, the result is a "spiraling" malformation that has been observed in nature.

*Mathematics Department and Laboratory for Computer Science, MIT. MIT grants AF #78181 and DARPA #N00014-92-J-1799.

†AT&T Bell Laboratories, Murray Hill, NJ.

1 Introduction

The study of virus shell structure and assembly is crucial for understanding how viruses reproduce and how anti-viral drugs might interfere with the assembly of virus shells. One of the most notable aspects of virus shells is their highly regular structure: they are generally spherical and possess strong symmetry properties. Almost all human viruses, and many plant and animal viruses, are icosahedral [2, 11]. These include the rhinovirus, poliovirus, and herpesvirus, all of which have rounded icosahedral shells. These shells are constructed of repeated protein subunits, or coat proteins, which surround their condensed DNA or RNA genomes. A given shell usually consists of hundreds of copies of one protein, but sometimes copies of two or three different proteins.

Many of these viral shells are believed to assemble with only limited aid from cellular machinery; they appear to "self-assemble," or spontaneously polymerize and take shape, in the host cell environment. This was originally established for the rod shaped tobacco mosaic virus [12] and has since been shown for many spherical viruses [6]. Sometimes assembly is assisted by scaffolding proteins, which assemble with the coat proteins to form a precursor shell, but are removed before the shell matures. At first glance, the assembly of the shells seems easy to understand, because the structure is so regular. In fact, it has been difficult to determine the actual pathway through which the subunits interact to form a closed shell composed of hundreds of subunits [25]. In icosahedral viruses this has been particularly difficult to explain because very often the same protein occurs in non-symmetric positions [25].

Previous attempts at explaining the formation of closed icosahedral shells from subunits have focused on the icosahedral symmetry, through the Caspar and Klug theory of *quasi-equivalence* [8]. This theory classifies icosahedral shells whose protein subunits all have very similar (quasi-equivalent) neighborhoods and form hexamers and pentamers in the virus shell. The general belief was that shells were formed by assembly of these pentamer and hexamer building blocks. However, in the most closely analyzed experimental system for studying the subunit assembly process, the bacterial virus P22, closed icosahedral shells assemble efficiently from purified monomeric protein subunits, even though the subunits are arranged as pentamers and hexamers in the final shell structures [13, 22, 23]. This suggests that the emphasis on the final symmetry of the structure has been a barrier to understanding how these proteins assemble into such complex structures.

It was also generally believed that proteins took on only one conformation, particularly very stable proteins such as those that form virus shells. Recent evidence indicates that virus shell proteins in fact take on several conformations [14, 19, 21], as has been proposed by [7, 25]. This important observation informs the approach to virus shell assembly presented below.

It is useful to consider what an individual protein "knows," from an information-theoretic point of view. It might appear that each protein needs to know something about the global structure. In fact, if the proteins assume different conformations during the assembly process depending on their relative positions, each protein has enough local information to "know" where to bind. Thus, a protein needs to know nothing about what is going on in the rest of the structure in order to form icosahedral shells.

In particular, the local rule-based theory introduced by Berger *et al.* [3] showed that possible assembly pathways can be given that depend on only the interactions of a protein with its immediate neighbors, rather than on larger structural building blocks. This leads to simple, local rule theories that suffice to explain assembly. The shell proteins of each virus obey a set of local rules which direct assembly. A local rule theory can be divided into a combinatorial part, which says which conformations can bind to each other, and a numerical part, which gives the relative angles and

the lengths of these bonds. In [3], two combinatorially different sets of local rules for the $T = 7$ shell are given (the T number classifies the combinatorial structure of the shell). One uses seven conformations of the coat protein; the other uses only four. Currently there is not adequate experimental evidence to say that one of these sets directs assembly in any particular virus.

In this paper, the mathematical theory behind the local rules is presented for the first time. We focus on the combinatorics of the local rules, giving several alternative sets of rules for each combinatorial shell structure. These combinatorial sets of rules may help in the determination of virus structures: a virus with unknown structure might be hypothesized to obey a given combinatorially set of local rules for assembly, possibly by analogy with a related virus. We also give a proof that a set of local rules for each icosahedral structure guarantees the final form.

A complete set of local rules for a virus, in addition to the combinatorial structure of the rules, must also specify interaction angles in 3D space, torsional angles, and interaction lengths. These will determine the exact shape of the assembled structure. Two sets of rules that are approximately the same tend to produce nearly identical final shapes, whereas those that are very different from each other produce different shapes. Computer simulations can show the relationship between the interaction angles and lengths of the local rules and the consequent virus shape. For the $T = 7$ shell, computer investigations of the interaction angles and lengths indicate that if these are changed by less than 8%, the final shapes will be nearly identical. In this paper, we further elaborate on the numerical parts of the local rules through computer experimentation.

Local rule theories are not limited to viruses that conform to the theory of quasi-equivalence. Recently, investigators have discovered viruses with unusual icosahedral symmetries [19, 24, 20], such as the polyomavirus, which causes tumors in mice. These have five-sided building blocks surrounded by *six* neighboring building blocks. Little is understood about the assembly of these "non-quasi-equivalent" viruses. The distinction between quasi-equivalence and non-quasi-equivalence is not so important if viewed in terms of local rules for assembly. Because local rules break the quasi-symmetry during the assembly process in any case, the fact that this symmetry is broken in non-quasi-equivalent viruses does not affect the local rule hypothesis. In particular, a set of local rules is presented that completely and uniquely determines the final conformation of the polyomavirus.

For a number of viruses, incorrect or malformed assemblages of coat protein subunits have been described, including tubular and spiral variants. The coat subunits in these structures are normal, but there have been errors in their interactions with each other [10, 18]. We have a computer simulation (see also [3]) which shows that if the local rules are distorted in certain ways, or if certain mistakes are made in the assembly process, spiral structures can form.

Previous attempts at interfering with the infection process have mainly focused on interrupting infection by a fully-formed shell at the binding site. The local rules tell us that if we can interfere with a single binding interaction, the shells may not close. Recent experiments indicate that the subunit assembly process may be a sensitive locus for inhibitors of virus assembly [29].

2 Icosahedral Structure

One striking aspect of virus shells is their strong symmetry properties. All the viruses discussed in this paper have what is called *icosahedral structure.* They are based on the icosahedron (figure 1a), a mathematical solid, and they have the same symmetry properties as the icosahedron.

Caspar and Klug [8] pointed out the link between icosahedra and icosahedral virus shells in

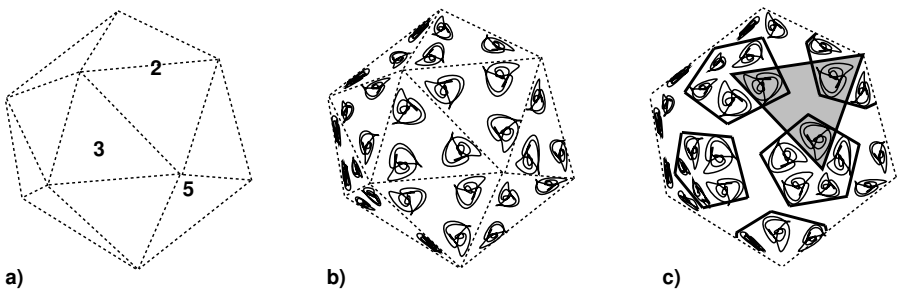

Figure 1: a) An icosahedron has 5-fold rotational symmetry at its 12 vertices, 3-fold rotational symmetry at its 20 triangular faces, and 2-fold rotational symmetry at its 30 edges. There are 60 symmetric regions in an icosahedron, each one lying in a third of a triangular face. b) Each triangular face has three proteins, one in each symmetric region. This icosahedral virus shell is made of 20 triangular faces and 60 identical proteins. c) The same icosahedral structure as in (b) but with pentameric clustering. Heavy lines group the proteins into pentamers around the 5-fold axes of symmetry. One triangular group is shaded for contrast.

their theory of quasi-equivalence. They classify icosahedral shells according to their *T number.* Their definition of T number is equivalent to the following: The T number of an icosahedral shell is the number of subunits per corner of each triangular face.

The theory of quasi-equivalence classifies icosahedral shells whose protein subunits all have very similar (quasi-equivalent) neighborhoods. Caspar and Klug assume that every protein subunit is a member of a hexamer or a pentamer, that the hexamers are arranged in a hexagonal lattice, and that there are exactly 12 pentamers, which lie on the five-fold axes of symmetry of the icosahedral shell. These assumptions are derived from the hypothesis that all subunits lie in nearly identical neighborhoods. Given these assumptions, it follows that there are only a limited number of possibilities for shell structures. A mathematical consequence is the restriction of the possible set of T numbers: the only ones allowed are $1, 3, 4, 7, 9, 12, 13, 16, 19, 21, 25, \ldots$. Although the following theorem has been cited [8, 5, 4], a proof has never appeared in the literature to the best of our knowledge.

Theorem 2.1 *Given the above hypothesis, the possible T numbers are of the form $a^2 + ab + b^2$, where a and b are non-negative integers.*

Proof: We start by embedding the icosahedron on a hexagonal lattice where each pair of pentagonal vertices v and w are (a, b) apart; that is, we can get from v to w in the lattice by going in a straight line distance a, making a 120° turn, and going in a straight line distance b. Thus, a and b form two sides of a triangle lying on the lattice and (v, w) can be thought of as the third side, not necessarily on the lattice. By the law of cosines, we have

$$dist(v, w) = \sqrt{a^2 + b^2 - 2ab\cos(120°)} = \sqrt{a^2 + b^2 + ab}.$$

Consequently, the icosahedron is composed of 20 equilateral triangles, each with sides of length $dist(v, w)$. Since the area of an equilateral triangle is $\sqrt{3}/4$ times the length of a side squared, the surface area of the icosahedron is $20 * (a^2 + ab + b^2) * (\sqrt{3}/4)$.

Next we count the number of proteins that can pack on the icosahedron by counting the number of unit triangles that pack on it. Since the unit triangles cover the entire icosahedron, the number

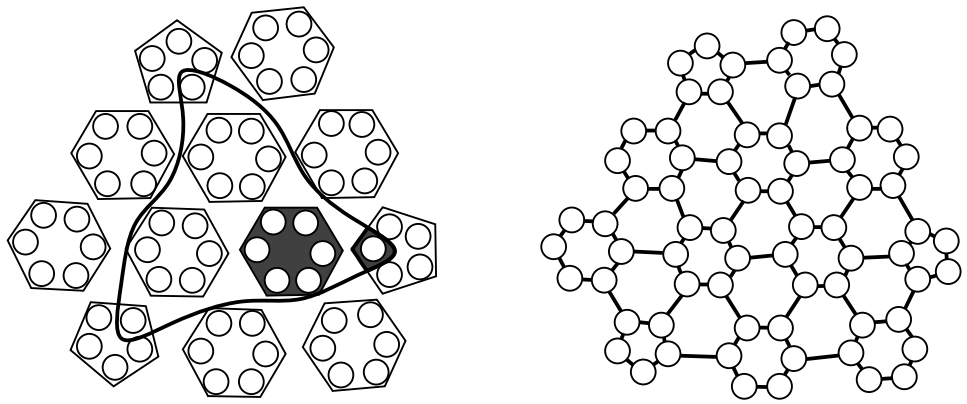

Figure 2: a) A portion of a $T = 7$ virus shell with the seven subunits in a corner unit shaded, the pentamers and hexamers drawn in light lines, and the triangular face in a heavy curved line. The protein subunits are depicted as circles. b) The same overall structure as (a), but redrawn in a graph representation to emphasize binding interactions, rather than the pentameric and hexameric building blocks. Every protein is a vertex and every binding interaction is an edge.

of unit triangles is the surface area divided by the area of a unit triangle, which is

$$\frac{20(a^2 + ab + b^2)(\sqrt{3}/4)}{\sqrt{3}/4} \;=\; 20(a^2 + ab + b^2).$$

Since each unit triangle can have three proteins, one at each corner, there are $60(a^2 + ab + b^2)$ proteins overall. Consequently, there are $T = a^2 + ab + b^2$ proteins per corner of each triangular face. □

This paper represents these shells in a way that better illustrates local rules. For example, the $T = 1$ shell of satellite tobacco necrosis virus [5] is typically viewed as an icosahedron, except that instead of having one protein at each vertex, it has a protein at each corner of each triangular face (figure 1b). The same structure can be redrawn by grouping the proteins at each vertex into pentamers, as in Figure 1c. See Figure 2 for an example of a $T = 7$ icosahedral shell. A graph representation of an icosahedral structure can be obtained by replacing the proteins with vertices and drawing an edge between two vertices when there is a binding interaction[1] between the two proteins (Figure 2b). All these structures still have icosahedral symmetry.

The focus on quasi-equivalence has led to the restriction on T numbers, which may be somewhat artificial. These constraints have grounding in mathematical symmetry, but they are not required by nature, because not all viruses satisfy Caspar and Klug's hypotheses. There is no physical reason for ruling out $T = 2$, for example. Indeed, $T = 6$, which will be discussed later, occurs in nature but does not fit the quasi-equivalent format. However, most viruses do fit into this format.

3 Local Rules

An alternative hypothesis is described for how icosahedral structures form, based solely on simple local rules for determining how proteins interact. For simplicity, we will assume virus shells contain

[1] For the purposes of abstraction, we refer to the interactions between two proteins, comprised of electrostatic, van der Waals, or other non-covalent chemical interactions, as a single binding interaction.

a single type of coat protein; the theory of assembly presented here works in all cases.

For each possible T number or shell size of an icosahedral virus, a set (or several alternative possible sets) of local rules exist that build the corresponding shell. These local rules are all of the following form: We assume that identical protein subunits take on a small number of distinct conformations. The local rules then specify, for each conformation, which other conformations it can bind to and the approximate interaction angles, interaction lengths, and torsional angles that will then occur between them. We are able to show that solely by following this local information, a closed, icosahedral shell will form with the desired T number. Some, but not all, sets of local rules require the building process to start with a given initiation complex to guarantee formation of the desired structure. We first give an example set of local rules where the assumed number of conformations is the same as the T number. We then give example sets of local rules where the assumed number of conformations is less than the T number, but the rules are somewhat more complex.

In a local rule theory, a protein "knows" where it is in the shell by looking at the conformations of its neighbors. Thus, any set of local rules for structures with $T > 1$ requires a number of different conformations; otherwise, a protein binding to the structure will not "know" in which of the various non-equivalent positions it will be and thus whether it should be part of a pentamer or a hexamer. These different conformations need not be maintained in the mature form of the virus. For P22, the differences between the protein conformations in non-equivalent positions are noticeably less in the mature form than in the precursor form [21]. Perhaps functionally different protein conformations are required for assembly, while in the mature form all the proteins assume the same functionality and need only be different enough to hold the shell together stably. This may also be the reason for the substantial changes that other bacteriophages undergo between their precursor and mature forms [15].

Local rule theories can be constructed for all T numbers. There is always a set of local rules with the number of conformations equal to the T number. However, in the case of higher T number structures, this many conformations would be infeasibly large. The number of conformations can sometimes be reduced by using rules that assign the same conformation to non-equivalent positions.

3.1 Local Rules for Quasi-Equivalent Viruses

3.1.1 Example: $T = 7$ Rules

The local rule theory can be illustrated through the example of the bacteriophage P22 virus shell, which is a $T = 7$ virus; that is, there are seven proteins in positions that are not equivalent under icosahedral symmetry, giving 420 proteins overall.

Seven conformations of the coat protein, or shapes, have been observed in the P22 precursor capsid [21]; however, it is not clear these are all truly distinct. First, let us suppose there are seven conformations. One of these seven conformations will be considered first. Figure 3 gives the rules for how the first conformation chemically binds in 3D. The type 1 conformation is in the center. Given the binding interaction to the type 2 neighbor, then, at a position clockwise from this at about an angle of 130°, only a type 1 conformation can attach. Similarly, only a type 1 conformation can attach at an angle of about 108° from this latter binding interaction. We call this representation the type 1 local rule.

Similar local rules can be constructed for all the seven conformations in bacteriophage P22 (figure 4). For these rules, all subunits with the same conformation will have the same binding interactions. The binding interactions in the local rules are present in micrographs of the shell;

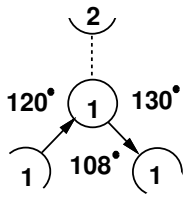

Figure 3: A representation of a local rule for the type 1 conformation of P22. Each protein subunit is represented as a circle or a part of a circle labeled with its conformation. Angles between binding interactions are the approximate number of degrees between the centers of the protein subunits, obtained from computer simulations. Binding interactions are represented as unit length, possibly with an associated direction. One way to think about a binding interaction is that the protein is sticking its arm out into another protein, like a key into a lock. For many viruses, this interpretation is consistent with experimental observations.

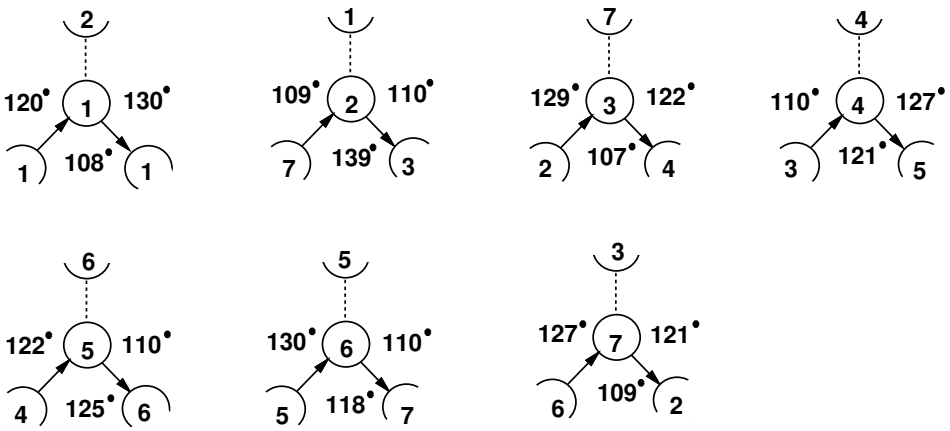

Figure 4: Possible local rules for a left-handed $T = 7$ virus. Angles are not based on any particular virus, but are derived from a computer simulation.

however, additional interactions may also be present [21]. These could be having a secondary effect which can be accounted for by the 8% tolerance margins of the local rule theory, described below. For these rules, the angles were at first derived from a physical model of a spherical $T = 7$ structure and subsequently refined by using the results of an initial computer simulation. The rule angles and lengths could likewise be derived by first guessing likely values, simulating on the computer to see where, if at all, the structure breaks down, and modifying the guess accordingly. This process is repeated until a guess is found for which the structure closes.

As soon as a subunit has at least one binding interaction, these rules can be applied unambiguously to determine the subunit's remaining neighbors. The different orders in which local rules can be applied give all the possible ways in which the assembly process might proceed. It is also conceivable that the rules can be applied simultaneously. While it would be consistent with the local rules that pentamers and hexamers initially form and later bind together, as previously believed, this is not required by the theory.

Chemically speaking, the local rules do not dictate which event comes first: a protein adopting

a conformation and then forming binding interactions as specified, a protein acquiring a binding interaction and then being forced into the corresponding conformation, or some cooperative combination. It is also possible that a protein does not adopt the conformation specified by the rules until after several of its neighbors have arrived.

The question remains, what structures can be built if these local rules must be respected? Applying the local rules to an arbitrary starting protein can result in a structure resembling the $T=7$ shell, or some subset of the shell, but nothing else (figure 5).

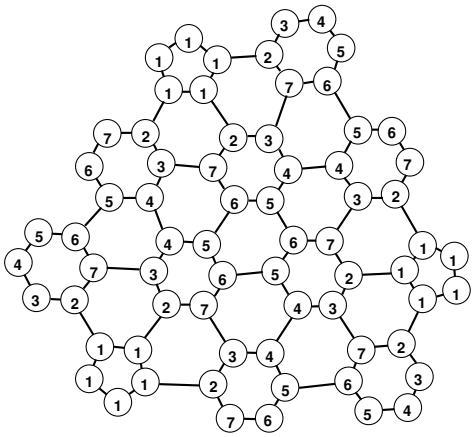

Figure 5: The same overall structure as in figure 2b, but redrawn to emphasize local rules.

Theorem 3.1 *The only structures consistent with those produced by the $T = 7$ rules in Figure 4, allowing sufficiently small tolerances, are subsets of the final $T = 7$ structure.*

Proof Sketch: Consider the finished shell, which by assumption has icosahedral symmetry. Suppose the angles and lengths of the rules are exactly those of the final shell. Since the rules give unique positions and conformations for the neighbors of any subunit, any subunit binding to a site in a partially completed shell must occupy the proper position and take the proper conformation of the corresponding subunit in the finished shell. Thus, the only possible structures that can be built respecting the local rules are subsets of the shell.

Next, we show that there is a length δ and an angle θ such that if all the interaction angles are perturbed by θ and the interaction lengths by δ, the only structures allowed are still subsets of the shell. Consider a partially completed shell. Any two subunits are connected by a path of at most length 420, since there are only 420 subunits in the completed shell. Thus, if we perturb every angle by θ, the direction of a bond between a given pair of proteins can only be perturbed from its position in the final shell by 420θ. The the vector between two adjacent proteins is thus perturbed by at most $420\theta l + \delta$, where l is the interaction length between these two proteins. Summing over all edges between a pair of proteins in a partially completed shell, we find that the distance between these proteins changes by at most $420^2\theta l + 420\delta$, so if θ and δ are chosen sufficiently small, the only possible resulting structure is the desired shell. □

In actuality, if a protein binding to a growing shell chooses a random growth site, the probability of getting a long path at any point during the assembly process is fairly small, since subunits would likely attach at sites making a "shortcut" on this path. Thus, actual tolerances in the rules which produce $T = 7$ shells with high probability are much larger than the ones in the above theorem.

We have done computer experiments investigating these tolerances, which will be discussed later in this paper.

An alternate set of local rules for $T = 7$, using only four conformations, is given in figure 6. There is a trade-off for using fewer conformations in that this set of rules is not as robust as the set given in Figure 4, and the last rule (the forbidden hexagon) may require a more complicated control mechanism.

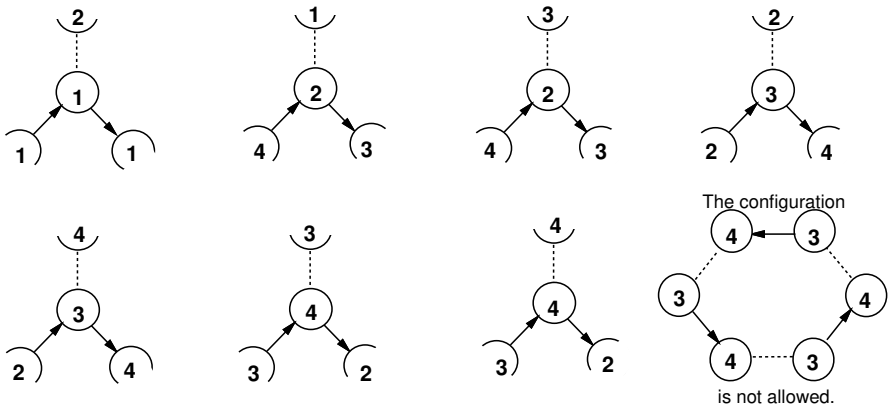

Figure 6: A second set of local rules for assembly of a left-handed $T = 7$ virus. Solid lines are binding interactions within capsomeres; dotted lines are binding interactions between capsomeres. We assume that the shell is initiated at a pentamer and that a protein does not assume its final configuration until there is at least one other protein in the same capsomere. The disallowed configuration might be impossible because the sum of the angles around the hexagon is too large or too small to allow it to close. Alternatively, since the three type 4 conformations in the hexagon are spatially adjacent, they may form a trimer that has higher energy than the 4-4-2 trimer. These rules produce the structure in Figure 6 but with conformations 5, 6, and 7 replaced by 2, 3, and 4, respectively.

In the alternate set of $T = 7$ rules, the hexamers are symmetric under rotations of 180°. This is intriguing because the micrographs of P22 show near-symmetry of the hexamers under 180° rotations [21]. It is possible that this alternate set of rules is responsible for determining the assembly of P22, and that the reason seven different conformations of the protein are observed is that functionally equivalent conformations in non-equivalent positions are responding to different stresses.

This alternate set of local rules for $T = 7$ is nearly the same as a set of local rules for $T = 4$: the first, second, fifth, and sixth rules in Figure 6 are the same as the rules in Figure 13. By changing the forbidden-hexagon rule (which is what prevents this set of rules from forming $T = 4$ shells) we can obtain a set of rules for a $T = 4$ shell. In fact, the coat proteins of three $T = 7$ bacteriophages can also form $T = 4$ shells. One of the mistakes observed in the assembly of P22 in the absence of scaffolding proteins is the formation of a $T = 4$ shell instead of a $T = 7$ structure [10]. This would seem to indicate that the scaffolding protein is involved in enforcing the forbidden-hexagon rule in P22. In bacteriophage λ, mutations of the coat protein exist that form functional $T = 4$ shells [17]. Finally, the $T = 7$ bacteriophage P2 has a satellite virus P4 that forms $T = 4$ shells using the P2 coat protein by substituting a different scaffolding protein [9, 1].

3.1.2 Example: $T = 3$ Rules

Many viruses have $T = 3$ shells. Because of the small size of these viruses, the determination of their atomic structure is in many cases feasible, and the atomic structure for several of these viruses is known. Thus, $T = 3$ shells make an excellent testing ground for exploration of local rule theories. We have been able to find three combinatorially different sets of local rule theories, and for each of these sets there are viruses whose structures suggest that that set directs the assembly of the virus.

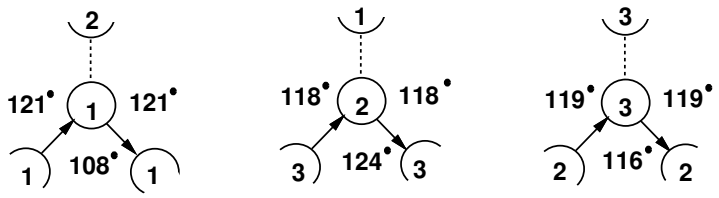

Figure 8: Possible local rules for a $T = 3$ virus. Angles are not based on any particular virus, but are derived from a computer simulation.

The simplest set of local rules for a $T = 3$ shell, with three comformations of the coat protein, is given in figure 8. The picornaviruses, which include rhinovirus and poliovirus have three different proteins which make up the shell [4]. These proteins appear to be designed so that they can fit together in only one way to form a virus shell, thus corresponding to the set of rules in 8.

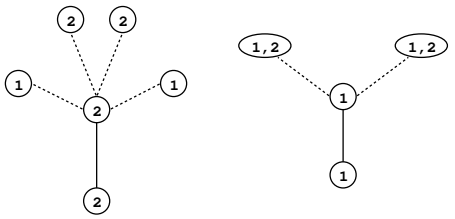

Figure 9: A second set of local rules for assembly of a $T = 3$ virus.

Another well-studied class of icosahedral viruses are the $T = 3$ plant viruses [14, 16, 27]. Several theories for their assembly have been advanced [26, 28]. Although these $T = 3$ virus shells have three non-equivalent positions, the proteins in two of these positions assume quite similar conformations [28, 27]. These are labeled 1 in the graph representation in Figure 10, while proteins in the third position are labeled 2. A set of rules can be extracted from this representation that permits both $T = 3$ and $T = 1$ shells (Figure 9). The coat proteins of many of these viruses can in fact form $T = 1$ shells [26]. However, as similarly noted by [28], if assembly is initiated by a structure containing a type 2 conformation, these will propagate during assembly to uniquely determine the $T = 3$ structure.

The final set of local rules for $T = 3$ is given in Figure 11. This set of rules has two conformations, 1 and 2, and determines the $T = 3$ structure given in Figure 12. The structure of the bacteriophage MS2 was recently determined [30]. This bacteriophage has a coat protein which is different structurally from any previously known viral coat protein. It also appears to use this combinatorial set of rules for assembly, again unlike any previously known virus. Although this bacteriophage has three non-equivalent positions, the proteins in two of these positions (2 and 2'

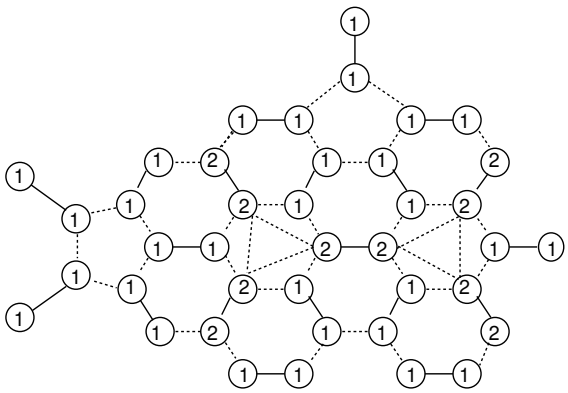

Figure 10: A graph representation for $T = 3$ plant viruses. Not all binding interactions are shown, but the binding interactions shown are sufficient to abstract a set of local rules (see Figure 9) that direct assembly of virus shells. The shell proteins are believed to form dimers in solution, which are represented by solid lines in the figure.

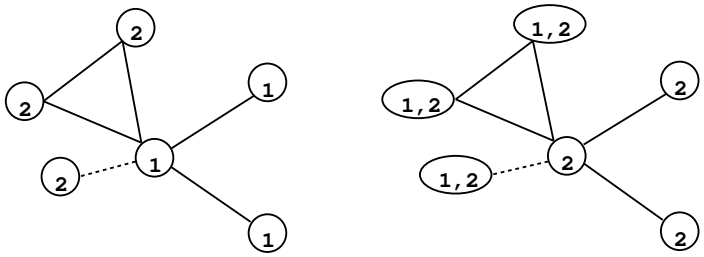

Figure 11: A third set of local rules for assembly of a $T = 3$ virus. The three-way interactions represented by triangles with edges are required to have one conformation 1 and two conformation 2's.

in Figure 12) are in quite similar conformations. The third protein (1 in Figure 12 has a loop that is configured quite differently near where the five proteins in conformation 1 meet. Also, as in the rules for Figure 11, the three-way interaction shown by curved triangles in Figure 12 is not symmetrical; the spatial relationship between conformations $2'$ and 1 is different from the relationships between conformations 2 and $2'$ and conformations 1 and 2.

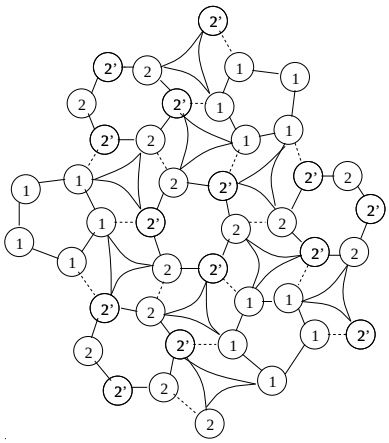

Figure 12: The graph representation of the $T = 3$ bacteriophage MS2. This structure is produced by the set of rules in Figure 11.

These sets of rules correspond to three of the five set partitions of the three non-equivalent positions on a $T = 3$ shell. The set partition taking all proteins to the same conformation cannot give rise to a local rule theory, because a protein will not have enough information to determine whether it should be in a pentamer or a hexamer. The remaining set partition, which takes two of the three positions to conformation 1's and the other two conformation 2's, but in a different manner than in Figures 11 and 9 also does not appear to have any set of local rules. In particular, there does not appear to be any way of drawing local interactions between neighboring proteins which permits an assembly path which uniquely determines the desired shell.

3.1.3 Example: $T = 4$ Rules

Simple rules for a $T = 4$ shell are given in figure 13. There are other sets of possible rules for $T = 4$ shells with both two and three distinct conformations. One of these is given in Figure 14. We do not know of any cases where enough information is known about a $T = 4$ virus to form a hypothesis that a particular set of rules directs the assembly of any particular virus.

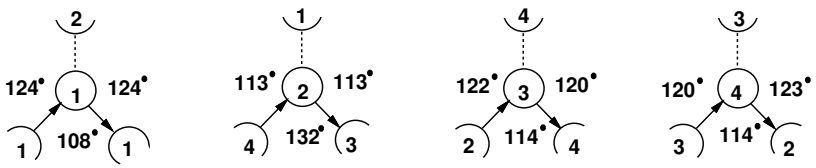

Figure 13: Possible local rules for a $T = 4$ virus. Angles are not based on any particular virus, but are derived from a computer simulation.

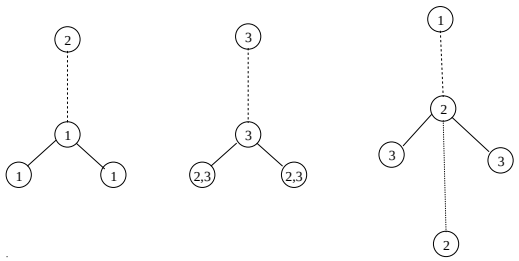

Figure 14: A second set of local rules for assembly of a $T = 4$ virus.

3.2 Local Rules for a Non-Quasi-Equivalent Virus.

In what follows, we apply the local rule theory to a polyomavirus, simian virus 40 (SV40), to produce a new hypothesis for its assembly. SV40 is one of a class of polymaviruses which all have similar structures and cause cancer in various species [19]. SV40 is a 360-subunit spherical virus with a shell consisting entirely of pentamers, some of which contact five other pentamers and some of which contact *six* other pentamers (figure 15). This structure still has 5-fold, 3-fold, and 2-fold

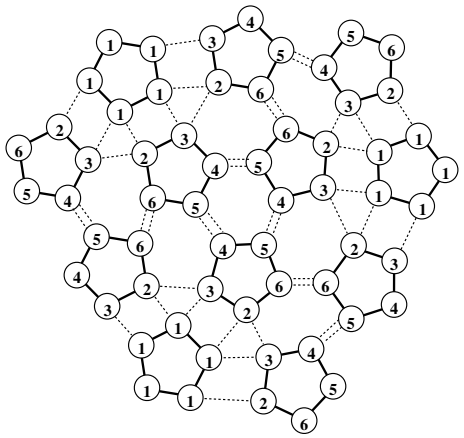

Figure 15: A simplified diagram of how the coat proteins in the SV40 shell connect with each other.

symmetry and is therefore icosahedral.

SV40 has been considered an anomaly, because the theory of quasi-equivalence assumes that all icosahedral shells have exactly 12 equally-spaced *penta*valent (i.e. have five neighboring capsomeres) pentamers and that all other subunits are packed into hexavalent hexamers. Clearly this assumption does not hold for SV40. Research [19, 24, 20] on the structure of SV40 has focused on how pentamers could be hexavalent and on how the same protein could occupy very asymmetric environments.

Another way to describe this anomaly is in terms of T numbers; SV40 would correspond to a $T = 6$ in that there are six proteins per corner of each triangular face. However, as the theory of quasi-equivalence does not allow 6 as a possible T number, it was classified as an anomalous $T = 7$. The capsomeres are indeed arranged in much the same pattern as for a $T = 7$ virus.

Local rules for SV40 can be constructed that are not substantially different than for other icosahedral viruses. It could simply have six local rules (figure 16), one for each of its conformations, as described in Berger *et al.* [3]. For SV40, six protein conformations have been confirmed, but the binding interactions are more complicated than as indicated by the local rule theory [19].

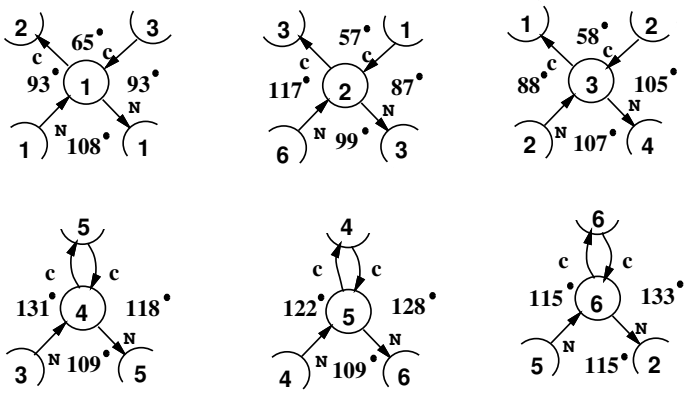

Figure 16: Local rules for the SV40 virus. Each protein subunit is labeled with the type of its conformation. The double, directed edges could be simplified to either a single, directed or undirected, edge; we have drawn them as double edges so as to correspond to the known biological structure [21]. Each inter-pentameric binding interaction is a C–terminal arm of the protein subunit, labeled with a direction to indicate which subunit it is from and a "C" to indicate which arm. Each intra-pentameric binding interaction is an N–terminal arm, labeled also with a direction and "N."

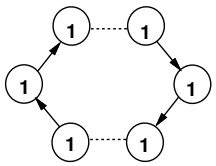

Figure 17: An example of how the local rules can be "broken": the connected components with solid directed edges between them are assumed to form before the dashed edges arise.

These simple rules for six conformations guarantee the final form: applying the rules in figure 16 in random order on the computer resulted in the same pattern of interconnectivity as in figure 15 with high probability. It is true that the rules are somewhat different than those in figure 4, since a protein can have four binding interactions instead of three.

4 Closure and Malformation

Although the above discussion might suggest that closure is easily assured, simulations show that a spiraling malformation can occur if the local rules are "broken" just once. Such incorrectly polymerized spiral structures have been observed with P22 and other viruses [18, 10, 15].

Suppose the two distinct connected components in figure 17, drawn as solid lines, were formed independently. When they encounter each other (i.e. the dashed lines are added), the local energy may be minimized by their forming a hexamer, when they really should form a pentamer. This distortion of the rules allows six type 1 subunits instead of five to fit together to form a capsomere at what "should" be a 5-fold axis of symmetry. If the local rules are correctly followed thereafter,

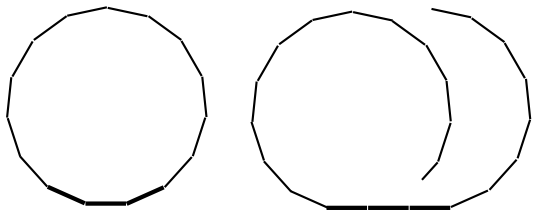

Figure 18: A cross-sectional diagram of spiraling. On the left, a spherical shape is constructed from segments with regularly-spaced curvature. On the right, a region without curvature is created at the bottom of the sphere, but subsequent growth retains its regularly-spaced curvature. The resulting structure does not achieve closure.

this hexamer would next be surrounded with six capsomeres instead of five. Since hexagons tile a plane, a region of an icosahedral surface having a hexamer in place of a pentamer will be relatively planar, but the regions growing around it will have the normal radius of curvature. When the sides have curved 180°, they will not be near enough to close because they will be separated by the length of the planar region (figure 18). One side may curl inward, and the second may form another shell layer around it. Computer experiments show that if local rules are broken in this way, spiraling can indeed occur. This exact malformation may not occur in nature, but other mistakes in formation could result in large planar sections of a shell which would likewise spiral.

Another common malformation of virus shells is the formation of open tubes. These can be viewed as planar sheets of hexamers which have been curved until two of their edges meet. This kind of malformation has been observed in many viruses, including SV40 [2] and bacteriophages λ and T4 [15]. The formation of these tubes may require some binding interactions between two conformations which are not allowed in the local rules. Possibly these binding interactions have higher energy than those in closed shells, but once a tube has begun to form, the lowest energy additions to this structure are a continuation of the tube. The structure of these tubes might further illuminate the rules for assembly.

5 Implementation

Local rules are essentially templates for energetically favorable arrangements. An individual protein can adhere to a slightly asymmetric location, and then the surrounding structure would readjust to find a low energy arrangement. Implementing these flexible templates was the main difficulty in doing a computer simulation of the assembly process. We used computer simulations to explore which rules result in closed shells; for example, starting with the local rules in Figure 4, a closed $T = 7$ shell can be built.

The computer simulations worked as follows: An energy model was set up assuming a quadratic penalty for deviations from the interaction angles, torsional angles, and interaction lengths given in the rules. The proteins were added to existing binding sites; if there were no candidates able to attach in the existing structure within one protein diameter of the binding site, a new protein was added. The local rules were used to determine the conformation and location of each new protein. After a protein was added, the resulting structure was optimized to minimize energy by iterating optimization steps. In each step, all the proteins were moved in accordance with the forces and torques computed from the energy model. The binding sites were examined both in random and breadth-first orders, in each case resulting in the formation of a closed shell.

Computer simulations show that the local rules are relatively robust. Even initial rules offset from the rules in figure 4 by a randomly selected amount of up to 9.6° (about 8%) for each rule angle and 8% for each interaction length lead to the formation of a closed shell in 3D space (figure 7b), which looks nearly identical to the one formed by the original rules (figure 7a). If the angles were changed by up to 10%, the shell failed to close in approximately half the trials; but when it closed, it still looked very similar to the original shell. (Clearly, these numbers depend on the underlying assumptions of the algorithm.) Through more substantial (non-random) changes in the local rules, a virus' shell can vary between spherical and polyhedral shapes.

As discussed above, we have also performed computer simulations on all sets of simple local rules up to $T = 16$, the $T = 6$ shell, and spiraling malformations. We even have produced in our simulations $T = 2$, $T = 5$, and $T = 8$ shells, which are not "allowable" T numbers. Furthermore, we have shown through computer simulation that by allowing a protein to make only one of its bonds, given by the simple local rules, with a growing shell structure, the shell will not form correctly. This suggests that biological experiments which add a mutant protein to a growing shell may be fruitful.

Acknowledgements

We are grateful to Jonathan King for his assistance with this work.

References

[1] M. Agarwal, M. Arthur, R. D. Arbeit, and R. Goldstein. Regulation of icosahedral virion capsid size by the *in vivo* activity of a cloned gene product. *Proc. Natl. Acad. Sci. U.S.A.*, 87:2428–2432, 1990.

[2] T. S. Baker, D. L. D. Caspar, and W. T. Murakami. Polyoma virus "hexamer" tubes consist of paired pentamers. *Nature*, 303:446–448, 1983.

[3] B. Berger, P. W. Shor, L. Tucker-Kellogg, and J. King. Local rule-based theory of virus assembly. *Proc. of the Natl. Academy of Sci.*, 91(16), aug 1994.

[4] C. Branden and J. Tooze. *Introduction to Protein Structure*. Garland Publishing, New York, 1991. chapter 11.

[5] S. Casjens. An introduction to virus structure and assembly. In S. Casjens, editor, *Virus Structure and Assembly*, pages 1–28. Jones and Bartlett, Boston, MA, 1985.

[6] S. Casjens and J. King. Virus assembly. *Annu. Rev. Biochem.*, 44:555–611, 1975.

[7] D. L. D. Caspar. Movement and self-control in protein assemblies: Quasi-equivalence revisited. *Biophys. J.*, 232:103–138, 1980.

[8] D. L. D. Caspar and A. Klug. Physical principles in the construction of regular viruses. *Cold Spring Harbor Symp. Quant. Biol.*, 27:1–24, 1962.

[9] T. Dokland, B. H. Lindqvist, and S. D. Fuller. Image reconstruction from cryo-electron micrographs reveals the morphopoietic mechanism in the P2–P4 bacteriophage system. *EMBO J.*, 11:839–846, 1992.

[10] W. Earnshaw and J. King. Structure of phage P22 coat protein aggregates formed in the absence of the scaffolding protein. *J. Mol. Biol.*, 126:721–747, 1978.

[11] B. N. Fields, editor. *Virology*. Raven Press, New York, 1985.

[12] H. Fraenkel-Conrat and R. C. Williams. Reconstitution of active tobacco mosaic virus from its inactive protein and nucleic acid components. *Proc. Natl. Acad. Sci. U.S.A.*, 41:690–698, 1955.

[13] M. T. Fuller and J. King. Assembly *in vitro* of bacteriophage P22 procapsids from purified coat and scaffolding subunits. *J. Mol. Biol.*, 156:633–665, 1982.

[14] S. C. Harrison, A. J. Olson, C. E. Schutt, F. K. Winkler, and G. Bricogne. Tomato bushy stunt virus at 2.9 Å resolution. *Nature*, 276:368–373, 1978.

[15] R. W. Hendrix. Shape determination in virus assembly: The bacteriophage example. In S. Casjens, editor, *Virus Structure and Assembly*, pages 169–204. Jones and Bartlett, Boston, MA, 1985.

[16] J. M. Hogle, A. Maeda, and S. C. Harrison. Structure and assembly of turnip crinkle virus I: X-ray crystallographic structure analysis at 3.2 Å resolution. *J. Mol. Biol.*, 191:625–638, 1986.

[17] I. Katsura. Structure and inherent properties of the bacteriophage lambda head shell IV: Small-head mutants. *J. Mol. Biol.*, 171:297–317, 1983.

[18] J. King, R. Griffin-Shea, and M. T. Fuller. Scaffolding proteins and the genetic control of virus shell assembly. *Quart. Rev. Biol.*, 55:369–393, 1980.

[19] R. C. Liddington, Y. Yan, J. Moulai, R. Sahli, T. L. Benjamin, and S. C. Harrison. Structure of simian virus 40 at 3.8-Å resolution. *Nature*, 354:278–284, 1991.

[20] C. T. Patch, A. S. Levine, and A. M. Lewis, Jr. The adenovirus-SV40 hybrid viruses. In H. Fraenkel-Conrat and R. C. Wagner, editors, *Comprehensive Virology*, volume 13, pages 459–541. Plenum Press, 1979.

[21] B. V. V. Prasad, P. E. Prevelige, Jr., E. Marietta, R. O. Chen, D. Thomas, J. King, and W. Chiu. Three-dimensional transformation of capsids associated with genome packing in a bacterial virus. *J. Mol. Biol.*, 231, 1993.

[22] P. Prevelige, D. Thomas, and J. King. Scaffolding protein regulates the polymerization of P22 coat subunits into icosahedral shells *in vitro*. *J. Mol. Biol.*, 202:743–757, 1988.

[23] P. E. Prevelige, Jr., D. Thomas, and J. King. Nucleation and growth phases in the polymerization of coat and scaffolding subunits into icosahedral procapsid shells. *Biophys. J.*, 64:824–835, March 1993.

[24] I. Rayment, T. S. Baker, and D. L. D. Caspar. Polyoma virus capsid structure at 22.5 Å resolution. *Nature*, 295:110–115, 1982.

[25] M. G. Rossmann. Constraints on the assembly of spherical virus particles. *Virology*, 134:1–13, 1984.

[26] M. G. Rossmann and J. W. Erickson. Structure and assembly of icosahedral shells. In S. Casjens, editor, *Virus Structure and Assembly*, pages 29–74. Jones and Bartlett, Boston, MA, 1985.

[27] A. M. Silva and M. G. Rossmann. Refines structure of southern bean mosaic virus at 2.9 Å resolution. *J. Mol. Biol.*, 197:69–87, 1987.

[28] P. K. Sorger, P. G. Stockley, and S. C. Harrison. Structure and assembly of turnip crinkle virus II: Mechanism of reassembly *in vitro*. *J. Mol. Biol.*, 191:639–658, 1986.

[29] C. M. Teschke, J. King, and P. Prevelige. *Biochem.*, 132:10658–10665, 1993.

[30] K. Valegard. The 3D structure of the bacterial virus MS2. *Nature*, 344:36–41, 1990.

Figure 7: a) *Silicon Graphics Indigo 2* computer graphics image of the shell resulting from the rules in Figure 4. b) The same figure as in (a), except formed from randomly selected rules, offset up to 8% from the rules that formed (a).