Geometry as a Weapon in the Fight Against Viruses

Reidun Twarock FIMA, University of York

This article is based on the IMA Gold Medal Lecture that Professor Reidun Twarock gave on 26 June at the Royal Society.

Viruses package their genetic material into protein containers that adopt polyhedral shapes. For over half a century, such virus architectures have been classified in terms of the Goldberg polyhedra and their dual triangulations in the seminal Caspar–Klug theory (CKT). However, following developments in our ability to image viral particles, it has become apparent that many virus structures do not conform to these blueprints.

The virus tilings described here simultaneously cover the classical viral architectures and solve open problems in structural virology. Introduction of a radial dimension into the models of virus architecture via extension of the symmetry group provides the basis for a graph theoretical approach describing the formation of viruses as a travelling salesman problem. This was instrumental in the discovery of virus assembly instructions – a virus assembly code – embedded within the genetic message of many viral pathogens.

Insights into viral geometry have thus played a key role in understanding how viruses form, evolve and infect their hosts, and have opened up new avenues in anti-viral therapy.

Symmetry in virology

Symmetry is ubiquitous in the natural world. It occurs at all scales, from particle physics to chemistry and cosmology. Mathematics describing symmetry in its various forms is therefore essential for our understanding of nature. Virology is no exception.

The protein containers (capsids) encapsulating and protecting the viral genomes resemble tiny footballs, which have the same rotational symmetries as an icosahedron. The locations of a 5-fold, a 3-fold and two 2-fold symmetry axes of icosahedral symmetry are indicated for a virus with 60 identical capsid proteins with respect to an icosahedral reference frame in Figure 1(a). The reason for this is known as the *principle of genetic economy* [1].

By repeatedly synthesising capsid building blocks from the same genomic segment, viruses minimise the portion of their genomes required for coding of the capsid. As icosahedral symmetry corresponds to the largest rotational symmetry group in three dimensions, it guarantees the largest possible number of repeats of the basic capsid building block in the capsid, thus optimising container volume. This facilitates the packaging of the genomic cargoes into their capsids, providing an explanation for the prevalence of icosahedral symmetry in virology.

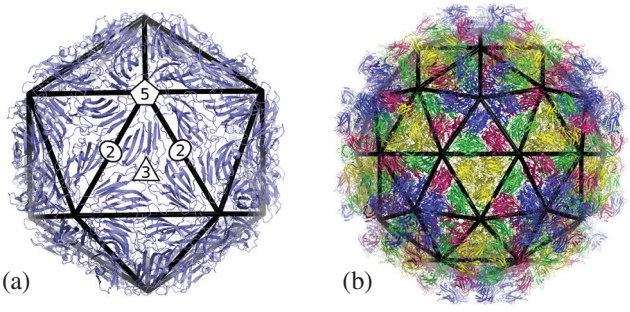

Figure 1: Viruses and icosahedral symmetry. (a) The icosahedral symmetry axes for a virus with 60 identical capsid proteins. (b) The $T = 4$ capsid structure of Providence virus.

Viral geometry

Icosahedral viruses come in different shapes and sizes. Most of them have capsids formed from more than 60 protein units, implying that capsid protein positions cannot correspond to a single orbit of the icosahedral group. Icosahedral symmetry by itself is therefore not sufficient to fully characterise viral geometries, and other mathematical approaches are required for a deeper understanding of virus architecture.

Caspar–Klug theory

In their seminal quasi-equivalence theory [2], Caspar and Klug posit that the protein subunits of larger capsids must be located in similar local environments, thus forming the same types of local interactions with neighbouring protein subunits. They model capsid architecture with reference to hexagonal surface lattices with icosahedral symmetry known as Goldberg polyhedra. Apart from hexagonal faces, these have 12 pentagonal faces as required by Euler's theorem in order to create a closed shell.

Icosahedral symmetry implies that there must be precisely $10(T-1)$ hexagonal faces, where $T = n^2 + nk + k^2$ with n and k positive integers (or zero in at most one case) is called the T-number. The case $T = 1$ corresponds to the icosahedron itself, and larger values describe triangulations of the icosahedral surface.

The insect-infecting Providence virus in Figure 1(b) is an example of a virus that can be modelled as a $T = 4$ capsid in CKT. Its Caspar–Klug surface lattice is shown superimposed on the capsid formed from 240 capsid proteins, with symmetry-equivalent capsid proteins shown in identical colours. Each triangular face indicates the positions of three capsid proteins.

The T-number has a geometric interpretation in terms of the dual triangulations called geodesic polyhedra. It indicates the number of triangular faces that, by area, cover each icosahedral face. Assigning a protein unit to every corner of these T triangles per icosahedral face, and noting that there are 20 such faces in an icosahedron, these polyhedral blueprints accommodate precisely $60T$ proteins. This means that the protein numbers in the viral capsid models in CKT are quantised. This astonishing prediction was long thought to be universally true. However, with the development of refined imaging techniques, many outliers to quasi-equivalence theory were discovered with protein numbers violating this restriction, instigating the development of new mathematical approaches.

Viral tiling theory

Prominent examples of such viruses are the cancer-causing polyoma- and papillomaviruses. Their capsids are formed from 72 pentamers (clusters of five proteins), rather than the characteristic combination of 12 pentamers and otherwise hexamers (clusters of six proteins) as in Caspar and Klug's approach. As planar

lattices formed from pentagons do not exist without gaps – a result known as the crystallographic restriction – a simple adaptation of the Caspar–Klug construction is not possible in these cases.

This is reminiscent of the mathematical challenges faced in the modelling of quasicrystals, alloys exhibiting long-range order but lacking periodicity, that were discovered in 1984 and won Dan Shechtman the Nobel Prize in Chemistry in 2011.

Adapting techniques from tiling theory used in the modelling of quasicrystals (in particular Penrose tilings), a tiling approach for the modelling of virus architecture has been developed that accounts for the surface structures of the polyoma- and papillomaviruses, see Figure 2(a). Symmetry equivalent proteins are shown colour-coded, and are organised into 72 pentamers (clusters of five) in Figure 2(a). This is the first example [3] of what is now known as viral tiling theory (VTT).

In VTT, capsids are represented by icosahedral tilings in which tiles have a dual role. As in CKT, individual tiles can correspond to clusters of protein subunits. In these cases, tiles represent capsid building blocks (capsomers) that form in solution, before multiple capsomers associate to assemble the capsid in the next step. Such tiles can be identified with interactions *within* capsomers.

In contrast to CKT, VTT also accommodates shapes other than the pentagons and hexagons representing pentamers and hexamers. For example, it includes rhomb tilings representing dimers (clusters of two proteins), which better describe the surface architectures of viruses such as bacteriophage MS2 (Figure 2(b)), a bacterial virus, which is formed from 90 protein dimers. VTT thus distinguishes MS2's architecture from the Caspar–Klug capsid structures with the same numbers of protein subunits. Hence, VTT also discriminates between different capsid layouts with the same number of capsid proteins, which would be represented by the same surface lattice in CKT.

Moreover, VTT also considers tilings in which tiles represent interactions *between* capsomers, as in the tiling in Figure 2(a). Here, different types of tiles correspond to different types of interactions between proteins in neighbouring pentamers. Interactions between pentamers occur in groups of two (dimer) and three (trimer) interactions, and are represented by rhombs and kites, respectively. Such tilings also cover capsid architectures with protein numbers excluded by CKT, such as the 360 proteins in Figure 2(a), which correspond to a *forbidden* T-number of 6.

VTT was designed to remedy the shortcomings of CKT, with the primary goal of explaining capsid architectures assembled from (nearly) identical protein subunits. However, increasing numbers of larger viruses have been discovered that assemble from a combination of different types of proteins, such as major and minor capsid proteins. For such capsids, the assumption of quasi-equivalence no longer holds, because only (nearly) identical protein subunits can be expected to occupy similar local environments, or, from a mathematical point of view, occupy positions in a lattice formed from a single type of building block.

A natural generalisation of quasi-equivalence is therefore to assume that capsids are built from a combination of different polygons, each of which are specific to a given type of capsid protein building block and are such that their relative sizes reflect their footprints on the capsid surface. The local rules according to which different types of proteins interact with each other should be universal across the capsid, and such capsid architectures can therefore be modelled by uniform lattices, i.e. lattices with one vertex type.

The planar uniform tilings were classified by Kepler in his *Harmonices Mundi* in 1619, and are also known as Archimedean lattices. As in the Caspar–Klug construction, we used these lattices to construct infinite series of polyhedra as models of capsid architecture, which contain the Caspar–Klug series of T-number geometries as a special case. Together with their duals, the Laves lattices, they provide blueprints for virus architectures that fall out of CKT/ VTT, as in the case of herpes simplex virus (Figure 2(c)) that follows the architecture of one of the new lattice series. The new polyhedral models of virus architecture thus have expanded the repertoire of allowed capsid protein numbers and have solved open problems in structural virology [4].

Applications of viral tiling theory

The tiling models of virus architecture are important for understanding the biophysical properties of viruses, such as their stability. They also provide novel insights into viral evolution. For example, an analysis of a wide range of capsids revealed that viruses in the same family follow similar geometric layouts. Even viruses from different families can exhibit similar capsid protein folds and surface lattice types despite a lack of significant sequence similarity, which is the usual measure of evolutionary relatedness. This suggests that the limited spectrum of geometrically possible lattice types may act as a driver of convergent evolution.

The tiling models also have many practical applications in nanotechnology. For example, in an adaptation of VTT to the modelling of self-assembling protein nanoparticles (SAPNs) [5], edges represent the positions of the protein building

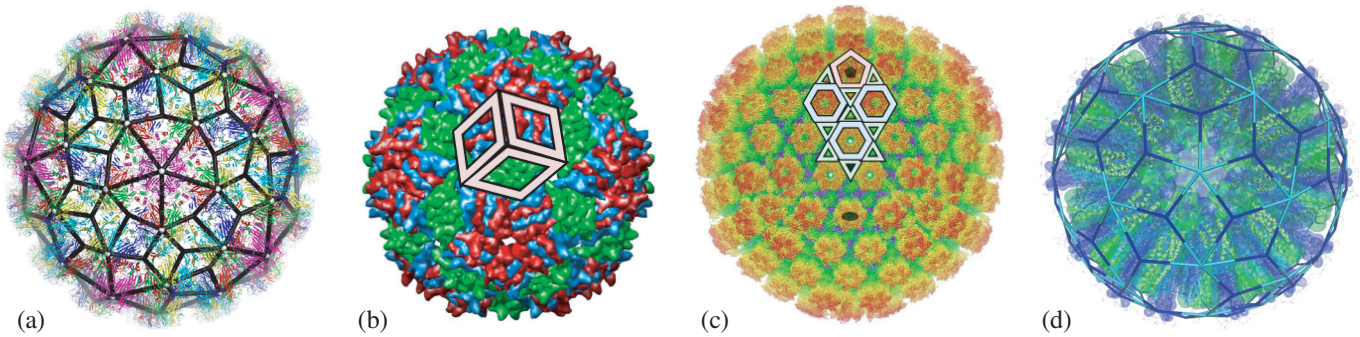

Figure 2: Viral tiling theory. (a) Kite-and-rhomb tiling encoding the surface of papilloma- and polyomaviruses. (b) Capsid of bacteriophage MS2 is represented by a rhomb tiling in VTT. (c) Large viruses, like this herpes simplex virus, require tilings with more than one type of polygon. (d) An adaptation of VTT to the modelling of SAPNs that are used to make malaria vaccines.

blocks, which are a pair of fused helices that form contacts with other copies in groups of five (green helices) and three (blue helices). Such particles (Figure 2(d)) are used in the design of malaria vaccines.

Symmetry is more than skin deep

CKT, VTT and its extensions described above, model virus architecture in terms of surface lattices that indicate protein positions, protein cluster types and their relative orientations. Whilst this can be used to address many fundamental structural questions in virology, it cannot provide any insights into the organisation of material at different radial levels in a virus. To obtain this, mathematical techniques are required that go beyond the description of surface lattices.

One option is to view the surface lattices as subsets of 3D tilings. Since icosahedral symmetry is non-crystallographic, such tilings must be aperiodic as in the case of quasicrystals. Such aperiodic tilings can be constructed from periodic tilings (lattices) in a higher dimensional space via projection onto an invariant subspace, akin to Plato's 'Allegory of the Cave', where puppets are only seen via their projections – shadows – on the wall of the cave. Typically, the dimension of this higher dimensional lattice (the minimal embedding dimension) is chosen such that: (i) the symmetry group is crystallographic in that dimension, and (ii) there exists a 3D subspace that is invariant under the action of the symmetry group and can therefore serve as a space in which the objects of interests can be modelled.

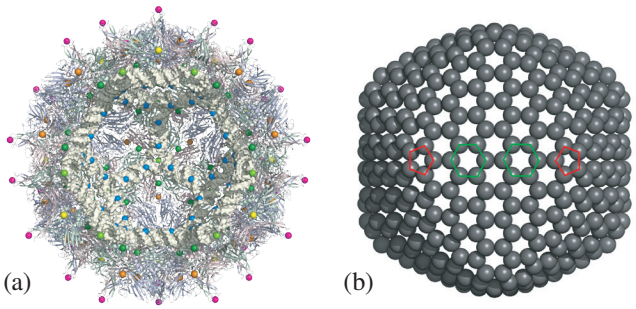

Figure 3: Affine extensions of the non-crystallographic symmetry group in virology and carbon chemistry. (a) Point arrays map around the material boundaries of viruses, here shown for the Pariacoto virus and (b) model the atomic positions of nested fullerene cages called carbon onions.

Modelling viruses with icosahedral symmetry in 3D requires a minimal embedding dimension of 6. There are three 6D lattices with icosahedral symmetry: the simple cubic (SC), the face-centred cubic (FCC) and the body-centred cubic (BCC) lattice. Working either with projections of orbits of their lattice groups into the 3D invariant subspace, or using projections of their 6D lattice bases in order to construct affine extensions of the icosahedral group in 3D, we derived and classified 3D point arrays with icosahedral symmetry. We demonstrated that elements of the resulting library of point arrays map around material boundaries of viral capsids, as shown for the Pariacoto virus in Figure 3(a) [6].

The multi-shell models also apply more widely to icosahedral multi-shell structures in science, such as the atomic positions of nested carbon cage structures called carbon onions [7]. The atomic positions of one shell are shown as grey dots in Figure 3(b), with two 5- and 6-fold faces of the corresponding polyhedron indicated in red and green, respectively.

Viral geometry and code breaking

The deeper understanding of viral geometry enabled by VTT and affine extended icosahedral symmetry has been a driver of discovery in virology. For example, it has provided a novel way of interrogating viral genomic RNAs for sequence/structure motifs in contact with the capsid shell.

For this, the best-fitting point array is selected from the library generated via affine extensions of the icosahedral group based on the outermost features of a virus [7]. The point array for bacteriophage MS2 is shown superimposed on a cross-sectional view of the capsid and packaged genome, obtained via cryo-electron microscopy (Figure 4(a)). There are points in the array at the binding sites between capsid protein and RNA, as shown in a magnified view of the vertices in yellow and orange in Figure 4(b), which mark the contact points between genomic RNA (bottom; two RNA structures called stem-loops are shown in blue and orange) and capsid protein (top; rendered as a cartoon showing protein sheets and helices). These points can be connected into a polyhedron (Figure 4(c)).

In any given virus, each vertex can be occupied by at most one stem-loop. Connecting vertices in the order in which they are occupied by sites in the viral genome, from its $5'$ to its $3'$ end, results in a Hamiltonian path (shown in yellow) on this polyhedron (Figure 4(d)). This is a path where each vertex is visited exactly once, similar to the travelling salesman problem.

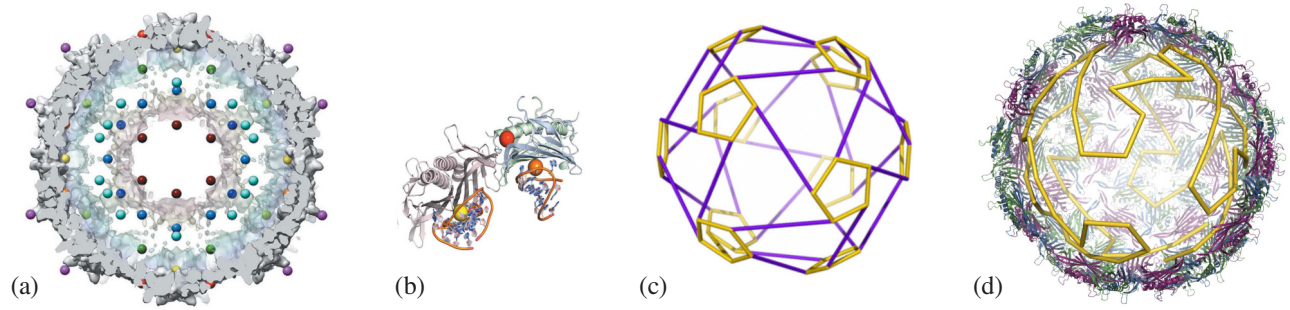

Figure 4: Hamiltonian paths and code breaking. (a) The multi-shell model for bacteriophage MS2. (b) A magnified view of the contact points between genomic RNA (bottom) and capsid protein (top). (c) Connecting points corresponding to neighbouring contact sites at the inner capsid surface results in a polyhedral shell. (d) A Hamiltonian path (yellow) on this polyhedron indicates the positions of the contact sites in the viral genome, a vital step in deciphering the PS-mediated assembly code.

It provides a mathematical bookkeeping device for the positions of successive capsid protein contact sites along the viral genome [8]. It also encodes the geometries of the assembly intermediates that occur during formation of the capsid, that is, Hamiltonian paths are in a one-to-one correspondence with virus assembly pathways.

Note, however, that Hamiltonian paths do not fully describe the actual conformation of the genomic RNA within the capsid, because portions of the genome extend into the capsid interior. Using the Hamiltonian path concept in combination with bioinformatics, we discovered that there are multiple dispersed sites within the genomic RNA with affinity for capsid protein mediated by a shared consensus sequence/structure motif [9]. We termed these motifs *packaging signals* (PSs) due to their roles in genome packaging and capsid assembly.

We thus discovered an assembly code embedded within the genetic code of a virus. Together with experimental collaborators, we developed novel analysis strategies to identify this code in a number of viruses, including major human pathogens, and jointly hold patents exploiting this discovery for anti-viral therapy and the design of virus-like particles [10].

Using geometry to understand viral life cycles

The Hamiltonian paths approach has provided a novel perspective on a fundamental question in virology. How do viruses assemble their capsids efficiently in light of the vast number of possible assembly pathways? This search for the drivers of efficient capsid assembly is akin to Levinthal's paradox in protein folding – the conundrum of how proteins achieve their biologically functional (native) state swiftly via specific folding pathways, rather than a random exploration of all possible pathways.

Teaming up stochastic simulations of capsid assembly based on Gillespie-type algorithms with the geometric understanding of capsid formation in terms of Hamiltonian paths, we were able to uncover the mechanism by which viruses solve this paradox [11]. Varying the PS sequences around a consensus motif results in a distribution of PSs along the genomic RNA with different affinities for capsid protein. In regimes where capsid protein concentration is small, as is the case at the start of an infection, different affinity distributions result in distinct assembly scenarios, and evolution has tuned the PS sequences akin to knobs on a radio to optimise assembly efficiency via mutation and selection. This explains why PS-mediated assembly can be observed only under the condition of protein concentrations consistent with a gradual build-up of capsid protein concentration, and would be obscured in experiments of virus assembly in which the full aliquot of capsid protein is added at the start of the assembly reaction, as is typically the case.

The strong variation of PSs around a consensus motif, which we now understand is an integral part of the mechanism of PS-mediated assembly, is also the reason why PSs were so difficult to detect without the insights from geometry.

The journey has just begun . . .

The discovery of PS-mediated assembly has overturned the existing paradigm in virus assembly, and has opened up novel routes for anti-viral therapy that we have only just started to explore in the context of viral infection dynamics at the intra- and intercellular level [12]. First results show that drugs targeting this mechanism are less likely to elicit therapy resistance through mutation than existing forms of therapy.

Many fascinating biological and mathematical questions are still open. For example, a large number of viruses have asymmetric components that distort the icosahedral lattice symmetry. The analysis of their various functional roles, for example in capsid assembly and disassembly, is still in its infancy. Moreover, the fascinating implications of PS-mediated assembly for viral evolution are shedding new light on the fundamental principles underpinning viral evolution.

Finally, the discovery of PS-mediated assembly is an invitation to turn the table on viruses, and abstract and optimise the assembly code for applications in nanotechnology, for example for the design of gene-delivery vehicles and vaccination.

For teachers:

Many of the mathematical concepts described here lend themselves for exploration in a classroom environment. Material for this is available at: `www-users.york.ac.uk/~rt507/teaching_resources.html`.

References

1 Crick, F. and Watson, J. (1956) Structure of small viruses, *Nature*, vol. 177, pp. 473–475.

2 Caspar, D. and Klug, A. (1962) Physical principles in the construction of regular viruses, *Cold Spring Harbor Symp. Quant. Biol.*, vol. 27, pp. 1–24.

3 Twarock, R. (2004) A tiling approach to virus capsid assembly explaining a structural puzzle in virology, *J. Theor. Biol.*, vol. 226, pp. 477–482.

4 Twarock, R. and Luque, A. Structural puzzles in virology solved by novel icosahedral designs, *Nature Commun.* (in press).

5 Indelicato, G. et al. (2016) Principles governing the self-assembly of coiled-coil protein nanoparticles, *Biophys J.*, vol. 110, pp. 646–660.

6 Keef, T. et al. (2013) Structural constraints on the three-dimensional geometry of simple viruses: case studies of a new predictive tool, *Acta Crystallogr. A*, vol. 69, pp. 140–150.

7 Dechant, P., et al. (2014) Viruses and fullerenes – symmetry as a common thread? *Acta Crystallogr. A*, vol. 70, pp. 162–167.

8 Twarock, R., Leonov, G. and Stockley, P.G. (2018) Hamiltonian path analysis of viral genomes, *Nature Commun.*, vol. 9, p. 2021.

9 Dykeman, E.C., Stockley, P.G. and Twarock, R. (2013) Packaging signals in two single-stranded RNA viruses imply a conserved assembly mechanism and geometry of the packaged genome, *J. Mol. Biol.*, vol. 425, pp. 3235–3249.

10 Twarock, R. and Stockley, P.G. (2019) RNA-mediated virus assembly: mechanisms and consequences for viral evolution and therapy, *Annu. Rev. Biophys.*, vol. 48, pp. 495–514.

11 Dykeman, E.C., Stockley, P.G. and Twarock, R. (2014) Solving a Levinthal's paradox for virus assembly suggests a novel anti-viral therapy, *PNAS*, vol. 111, pp. 5361–5366.

12 Bingham, R.J. et al. (2017) RNA virus evolution via a quasispecies-based model reveals a drug target with a high barrier to resistance, *Viruses*, vol. 9, p. 347.