Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 17 July 2012
Accepted 16 November 2012

Structural constraints on the three-dimensional geometry of simple viruses: case studies of a new predictive tool

Thomas Keef,[a] Jessica P. Wardman,[a] Neil A. Ranson,[b] Peter G. Stockley[b] and Reidun Twarock[a]*

[a]York Centre for Complex Systems Analysis, Departments of Mathematics and Biology, University of York, York, England, and [b]Astbury Centre for Structural Molecular Biology, University of Leeds, Leeds LS2 9JT, England. Correspondence e-mail: reidun.twarock@york.ac.uk

Understanding the fundamental principles of virus architecture is one of the most important challenges in biology and medicine. Crick and Watson were the first to propose that viruses exhibit symmetry in the organization of their protein containers for reasons of genetic economy. Based on this, Caspar and Klug introduced quasi-equivalence theory to predict the relative locations of the coat proteins within these containers and classified virus structure in terms of *T*-numbers. Here it is shown that quasi-equivalence is part of a wider set of structural constraints on virus structure. These constraints can be formulated using an extension of the underlying symmetry group and this is demonstrated with a number of case studies. This new concept in virus biology provides for the first time predictive information on the structural constraints on coat protein and genome topography, and reveals a previously unrecognized structural interdependence of the shapes and sizes of different viral components. It opens up the possibility of distinguishing the structures of different viruses with the same *T*-number, suggesting a refined viral structure classification scheme. It can moreover be used as a basis for models of virus function, *e.g.* to characterize the start and end configurations of a structural transition important for infection.

1. Introduction

Viruses are striking examples of order at the nanoscale. In many viruses the protein-based containers that package their genomic nucleic acids exhibit icosahedral symmetry (see Fig. 1*a*). As Crick & Watson (1956) argued, this is for reasons of genetic economy, because viral capsids with this symmetry can be formed from the maximal possible number of coat protein subunits for the least genetic information, hence providing the largest genome packaging volume. The first mathematical models of virus structure used this fact to represent viral capsids *via* icosahedrally symmetric surface lattices or tilings. In particular, Caspar and Klug showed that the positions and relative orientations of the capsid proteins of most viruses follow triangulations of an icosahedral surface (Caspar & Klug, 1962; Coxeter, 1972) such as the one shown in Fig. 1(*b*). These triangulation-based models apply for scenarios in which the proteins are organized in quasi-equivalent local environments, *i.e.* for the cases in which the local bonding environments of the capsid proteins are similar. An extension of Caspar–Klug theory has been formulated for the non-quasi-equivalent cases, using tessellations in terms of different types of shapes to model the different types of local bonding environments. For example, the tiling in Twarock (2004) represents the protein organization in the capsids of the cancer-causing polyomaviridae *via* two different shapes, a rhomb and a kite, representing the dimer (respectively, trimer) interactions between capsid proteins.

A common property of these approaches is the fact that, being based on surface lattices, they do not provide any radial information that could account for, *e.g.*, the thickness of the capsid or details of the capsid proteins other than their relative positions and orientations. Moreover, they provide no information on features of the organization of the packaged genomes. The latter is of particular interest as many viruses exhibit ordered features in their genomes in structures determined at moderate resolution (van den Worm *et al.*, 2006).

Striking examples of this are the dodecahedral cage of dsRNA seen in Pariacoto virus (PaV) (Tang *et al.*, 2001) and the double-shell architecture of the viral RNA observed in bacteriophage MS2 (Toropova *et al.*, 2008). These results suggest that there are further constraints on the three-dimensional structures of these highly ordered capsids and genomes, *i.e.* that there should exist a wider set of constraints on virus architecture than those formulated in Caspar–Klug theory and viral tiling theory.

These theories demonstrate that important information on the structural organization of viruses can be derived from geometric considerations alone. This does not mean that the physical and chemical principles underlying virus structures are not important, but it does suggest that there is a strong geometric constraint on the ways in which these principles can manifest themselves on a structural level. This view is supported further by the observation that there are only a limited number of different capsid protein folds (Bamford *et al.*, 2005), which occur repeatedly across different viral families with a surprising lack of any statistically significant amino-acid-sequence similarity. This suggests that there should be strong geometric constraints on the evolution of capsid protein geometry that should be classifiable using geometric tools.

Janner has explored the use of lattices to explain features of the three-dimensional geometry and genome organization in viruses. In a series of papers (Janner, 2010*a*,*b*, 2011*a*,*b*,*c*) he has shown that encasing forms can be constructed for viral components at different radial levels by embedding virus structure into lattices. This approach is by construction descriptive as it is not *a priori* clear which subset of a lattice is important for a virus of interest, and the fitting has been carried out *via* visual inspection, allowing for the conclusion that the model is a good approximation, without being able to quantify this further.

Here we introduce a method that focuses on the symmetry group of the underlying lattice. Indeed, as lattices with icosahedral symmetry do not exist in three dimensions due to the crystallographic restriction (Scherrer, 1946), we work with quasi-lattices, *i.e.* structures with long-range order lacking periodicity. Such structures are known to occur in physics in the form of quasicrystals, alloys with atomic positions organized according to quasi-lattices (de Bruijn, 1981*a*,*b*; Senechal, 1996). In this paper, we consider Janner's lattices as approximations of quasi-lattices with icosahedral symmetry. Finite subsets of the vertex sets of such quasi-lattices can be constructed iteratively from affine extensions of the icosahedral group as we demonstrate in the next section. Since viruses are finite objects, we use these finite point arrays as discrete models for virus structure. Figs. 1(*c*) and 1(*d*) demonstrate how such point arrays relate to surface tessellations in Caspar–Klug theory and viral tiling theory. Indeed, the vertices of the triangulation in Fig. 1(*b*) form a subset of the vertices of the three-coloured icosahedra in Fig. 1(*c*), which are related to each other *via* generators of an affine extension of the icosahedral group. By considering all vertices of these translated and rotated copies, as shown in Fig. 1(*d*), correlated structural constraints at different radial levels are obtained. Fig. 1 shows one example of a point array encoding the action of an affine-extended symmetry group. According to the classification in Wardman (2012), there exists a finite set of such structures, the library of point arrays. We introduce here a best-fit algorithm that compares all structures in this library against structural data provided in the form of a PDB file and demonstrate the predictive power of our method for a number of viruses. The implication of this study is that icosahedral symmetry and quasi-equivalence are part of a wider set of structural constraints on virus structures.

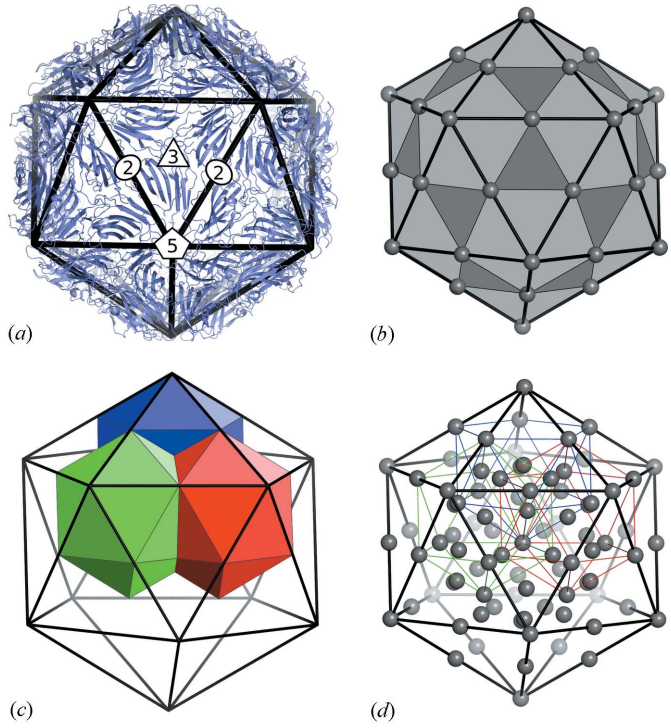

Figure 1
Quasi-equivalence is part of a wider set of structural constraints on virus architecture. (*a*) A large number of viruses exhibit icosahedral symmetry in the organization of their protein containers, *i.e.* they share a set of five-, three- and twofold axes with an icosahedron. Here, four example symmetry axes are shown on a viral protein container. (*b*) The $T = 4$ surface lattice from quasi-equivalence theory that encodes the protein organizations in capsids composed of 240 coat proteins. (*c*), (*d*) A packing of overlapping icosahedra generates a partition of the icosahedral face akin to the $T = 4$ structure: (*c*) shows three of the 60 translated and rotated icosahedra as solids; (*d*) shows the edges of these icosahedra, together with the vertices of all 60 icosahedra. This is an example of the structural constraints implied by our theory.

2. Three-dimensional constraints on virus architecture from affine-extended symmetry groups

The surface lattices and tilings in Caspar–Klug and viral tiling theory can be viewed as subsets of spatially extended structures in three dimensions, as illustrated for the icosahedral triangulation (a $T = 4$ structure in the Caspar–Klug classification) in Fig. 1. In this section, we discuss how point arrays such as that in Fig. 1(*d*) can be systematically constructed from affine extensions of the icosahedral group. For this, note that point arrays obtained *via* affine extensions of non-crystallographic groups form subsets of the vertex sets of quasi-lattices. We demonstrate this in Fig. 2 for a planar symmetry group, as graphical representation is simpler in this case (see also supplementary movie 1).[1] In this figure, we

[1] Supplementary movies are available from the IUCr electronic archives (Reference: WX5023). Services for accessing these movies are described at the back of the journal.

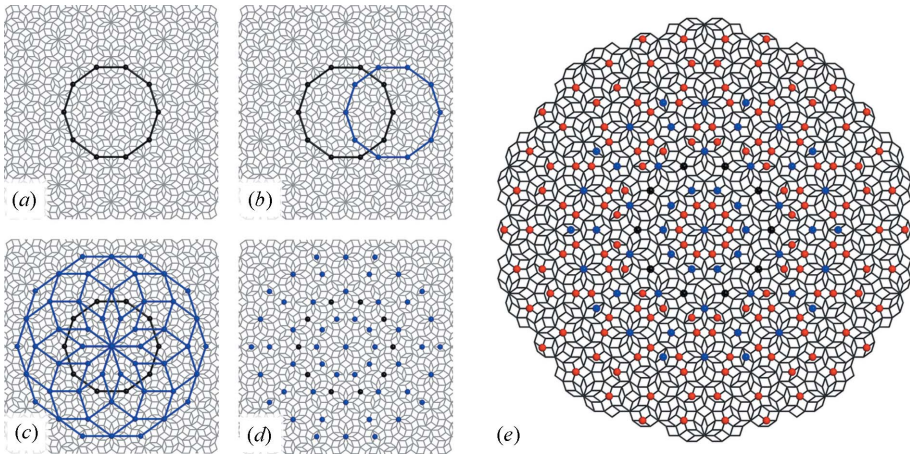

Figure 2
The geometric principle can be encoded in a tiling. The figure demonstrates the relation of extended symmetry groups with tilings. For simplicity, the principle is demonstrated for the two-dimensional case of tenfold symmetry; the same principle has been applied to icosahedral symmetry in three dimensions and has resulted in the library of point arrays used here (Keef & Twarock, 2009; Wardman, 2012). (*a*) A decagon, a geometric representation of tenfold symmetry, superimposed on a Penrose-type tiling such that its corners coincide with vertices of the tiling. (*b*) Addition of a translation to the rotational symmetries of the decagon (*i.e.* an affine extension of tenfold rotational symmetry) results in translated copies of the decagon (shown in blue) with corners also coinciding with vertices of the tiling. (*c*) Subsequent rotations about the tenfold axis at the centre of the original decagon result in ten copies of the translated decagon. (*d*) Since corner points of the decagon are geometric representations of the (rotational) symmetries, the (artificial) decagonal edges are faded away. An iterative process of translation and rotation leads to the addition of further points. For example, in the second iteration step the red points in (*e*) are added and more vertices of the tiling are covered.

consider the rotational symmetry group of a decagon. *Via* an affine extension of this group by the translation mapping the black onto the blue decagon in Fig. 2(*b*), and subsequent application of all rotational symmetries in Fig. 2(*c*), one obtains the point array in Fig. 2(*d*) that forms a subset of the vertex set of a tiling. This tiling has been obtained *via* the projection method (Kramer & Shlottmann, 1989) from a five-dimensional lattice with decagonal symmetry. Fig. 2(*e*) demonstrates that iterating the action of the translation operation further, again with subsequent application of the rotational symmetries, results in a larger point array that contains in addition the vertices shown in red in Fig. 2(*e*). With increasingly higher iterations, the point array would become denser and more extended in space. Since viruses are finite objects, we therefore use a cutoff in the number of iterations employed. For smaller viruses (*i.e.* viruses up to $T = 4$ in the Caspar–Klug classification), this is typically after the first iteration step.

This example demonstrates how point arrays constructed *via* an affine extension of a symmetry group relate to the vertex sets of tilings obtained *via* the projection method. Note that the geometry of the shape representing the symmetry group (here the decagon as a geometrical representation of decagonal symmetry) is related to the basis of the higher-dimensional lattice from which the tiling is obtained *via* projection. In the case of icosahedral symmetry, the minimal dimension (minimal embedding dimension) in which a lattice with icosahedral symmetry exists is six dimensional. In order to construct all affine extensions of the icosahedral group that can give rise to vertex sets of quasi-lattices in this way, one therefore needs to apply the procedure to the projections of all six-dimensional Bravais-lattice types with icosahedral symmetry. There are three such lattice types: the simple cubic (s.c.), the face-centred cubic (f.c.c.) and the body-centred cubic (b.c.c.) lattice. According to Table 1 in Indelicato, Cermelli *et al.* (2012), the projections of their basis vectors into one of the two three-dimensional subspaces invariant under the icosahedral group correspond to an icosahedron (vertices on the five-fold axes of icosahedral symmetry), an icosidodecahedron (vertices on the twofolds) and a dodecahedron (vertices on the threefolds). The construction and classification of affine extensions of the icosahedral group have therefore been based on these polyhedra (Keef & Twarock, 2009; Wardman, 2012). In particular, each of the three polyhedra is used as a *start configuration* in the terminology of that paper, *i.e.* as the shape representing the symmetry group from which all translations extending the icosahedral group are determined as those operations that result in the translated and rotated copies sharing vertices. From a group-theoretical point of view, this implies that the resulting affine extensions have non-trivial group relations (*i.e.* do not correspond to the free group). This is because vertices correspond to words in the group generators in this context, and coinciding vertices hence define non-trivial relations between words of generators. The result of the classification, adapted from Keef & Twarock (2009) and Wardman (2012), is given in Table 1.

As described in these references, since start configurations are auxiliary objects containing information regarding the six-dimensional basis from which they have been obtained *via* projection, two start configurations, and hence also their point arrays, can be combined provided that they have been derived *via* the same translation operation. In particular, the entries in the table are given for start configurations normalized as follows: vectors pointing to the vertices of the icosahedral, dodecahedral and icosidodecahedral (IDD in Table 1) start configurations have length $(2+\tau)^{1/2}$, $3^{1/2}$ and τ, respectively, where $\tau = (1+5^{1/2})/2 \simeq 1.618$. Moreover, translation lengths along a fivefold ($\mathbf{T}_5$), threefold ($\mathbf{T}_3$) and twofold ($\mathbf{T}_2$) direction are indicated in multiples of vectors of length $(2+\tau)^{1/2}$, $3^{1/2}$ and τ, respectively.

Since start configurations and their associated translations scale simultaneously (as only their relative sizes matter in the construction of the affine extensions), it is possible to rescale the entries for different start configurations in the table such

Table 1
The cardinalities of the sets for the 55 extended symmetry systems calculated in Wardman (2012).

Icosahedron		Dodecahedron		IDD	
Translation	Size	Translation	Size	Translation	Size
$(-1+\tau)\mathbf{T}_5$	116	$(2-\tau)\mathbf{T}_5$	200	$\frac{1}{2}(-1+\tau)\mathbf{T}_5$	290
$\mathbf{T}_5$	85	$(-1+\tau)\mathbf{T}_5$	172	$\frac{1}{2}\mathbf{T}_5$	242
$\tau\mathbf{T}_5$	116	$\mathbf{T}_5$	172	$\frac{1}{2}\tau\mathbf{T}_5$	242
		$\tau\mathbf{T}_5$	200	$\mathbf{T}_5$	360
				$\frac{1}{2}(1+\tau)\mathbf{T}_5$	290
				$\tau\mathbf{T}_5$	360
$(-1+\tau)\mathbf{T}_3$	192	$(2-\tau)\mathbf{T}_3$	360	$\frac{1}{2}(-1+\tau)\mathbf{T}_3$	510
$\mathbf{T}_3$	164	$(-1+\tau)\mathbf{T}_3$	252	$\frac{1}{2}\mathbf{T}_3$	362
$\tau\mathbf{T}_3$	164	$\mathbf{T}_3$	191	$\frac{1}{2}\tau\mathbf{T}_3$	374
$(1+\tau)\mathbf{T}_3$	192	$\tau\mathbf{T}_3$	252	$\mathbf{T}_3$	600
		$(1+\tau)\mathbf{T}_3$	360	$\frac{1}{2}(1+\tau)\mathbf{T}_3$	362
				$\tau\mathbf{T}_3$	570
				$\frac{1}{2}(1+2\tau)\mathbf{T}_3$	510
				$(1+\tau)\mathbf{T}_3$	600
$(-1+\tau)\mathbf{T}_2$	342	$(2-\tau)\mathbf{T}_2$	590	$\frac{1}{2}(-1+\tau)\mathbf{T}_2$	870
$2(2-\tau)\mathbf{T}_2$	272	$2(-3+2\tau)\mathbf{T}_2$	500	$(2-\tau)\mathbf{T}_2$	710
$\mathbf{T}_2$	342	$(-1+\tau)\mathbf{T}_2$	560	$\frac{1}{2}\mathbf{T}_2$	870
$2(-1+\tau)\mathbf{T}_2$	212	$2(2-\tau)\mathbf{T}_2$	332	$(-1+\tau)\mathbf{T}_2$	552
$2\mathbf{T}_2$	212	$\mathbf{T}_2$	590	$\frac{1}{2}\tau\mathbf{T}_2$	870
$2\tau\mathbf{T}_2$	272	$2(-1+\tau)\mathbf{T}_2$	344	$\mathbf{T}_2$	361
		$2\mathbf{T}_2$	332	$2(-1+\tau)\mathbf{T}_2$	870
		$2\tau\mathbf{T}_2$	500	$\tau\mathbf{T}_2$	552
				$2\mathbf{T}_2$	840
				$(1+\tau)\mathbf{T}_2$	710
				$2\tau\mathbf{T}_2$	870

that their translation lengths match. This results in 569 pairings of start configurations and we call the associated point arrays the *library* of point arrays. It would potentially be possible to pair more than two entries in this way; however, this would lead to increasingly extended and dense arrays similar to those obtained *via* higher iterations of the symmetry group. These may be relevant for larger viruses, but for smaller viruses ($T = 1$ to $T = 7$ in the Caspar–Klug nomenclature) these would be too dense. In the following, we will demonstrate that this library of point arrays provides information on the structures of a wide range of simple viruses covering all T-numbers up to $T = 7$.

3. The best-fit algorithm

The affine extension procedure has resulted in a finite library of point arrays with the property that the points of each individual array are related to each other by elements of an affine-extended icosahedral group (see also supplementary movie 2). This interdependence of points in any given array makes it possible to use these arrays as predictive tools: by designing a best-fit algorithm that selects a point array in the library based on the fit of a subset of its points (*e.g.* the exterior points) to part of the virus (*e.g.* the capsid), the remaining points of the array (*e.g.* points at radial distances overlapping with the interior of the capsid) can then be used to infer information regarding the organization of other components of the virus (*e.g.* the genome). This is of particular interest as some structural information can be more easily obtained at higher resolution than other information with current experimental capabilities. Indeed, whilst capsid structure determination by crystallography or cryo-electron microscopy (cryo-EM) typically achieves high resolutions, cryo-EM data regarding genome organization are currently limited to about 8 Å resolution. Therefore, we have designed the best-fit algorithm to select the best-fit point array(s) for any given virus based on the PDB file of the viral capsid (coordinates of the atomic positions of the capsid proteins), which is readily available for a large number of viruses, *e.g.* from VIPER (Carrillo-Tripp *et al.*, 2009). We then infer information regarding genome organization within the capsid from the points in the best-fit array overlapping with the interior of the capsid.

In this section, we provide details of the best-fit algorithm. In particular, the following outlines the procedure according to which the best-fit algorithm determines which point array (best-fit point array) in the library best describes the surface structure and topography of a virus based on its PDB file:

(*a*) Sample preparation. The coordinates of the test viral proteins in the fundamental domain of the icosahedral group (also called asymmetric unit in the biological literature) are retrieved from a PDB data bank and icosahedral symmetry operations used to generate the entire capsid. The proteins are then represented by spheres of radius 1.9 Å around each atomic position, which corresponds to the maximal van der Waals radius of all atoms in the PDB file, calculated using *CHARMM* (Brooks *et al.*, 2009).

(*b*) Alignment and scaling. Our classification contains 569 point arrays. These are aligned with the modelled viral surface *via* their common symmetry axes. In order to scale each array to the test virus, the smallest possible scaling of the point array is determined such that the modelled capsid surface is entirely contained in the convex hull of the array.

(*c*) Sifting. The aligned and scaled point arrays contain points at different radial levels. For some arrays which are not good representations of the viral architecture, some of these points will fall within the van der Waals radius of protein atoms; we exclude all such point arrays from further consideration.

(*d*) Goodness of fit. In order to quantify the goodness of fit of these arrays to the protein container, two values are calculated: a root-mean-square deviation (RMSD) score $S_{\rm RMSD}$ and a topography score $S_{\rm Top}$. These measure how well the points in an array match the van der Waals radii of the atoms in the capsid proteins ($S_{\rm RMSD}$) and how well they represent the overall surface topography of the virus ($S_{\rm Top}$). In order to reduce computational complexity, we only consider a single asymmetric unit with its neighbouring proteins in this calculation and truncate all point arrays to the same copy of the asymmetric unit.

(*d*1) RMSD score $S_{\rm RMSD}$. The RMSD score examines the closeness of the fit of the point array around capsid proteins. Therefore, only points encompassing the volume occupied by the capsid proteins are considered for this analysis. In order to

ensure that all features of the inner surface are represented, we apply an (arbitrary) cutoff to include only the points with radii within 4 Å from the inner capsid surface and above. Since some of these points may be located between different capsid proteins, we compute the distances of each such point from the van der Waals radii of all protein atoms in its vicinity. For all T (as in T-number in the Caspar–Klug classification) different quasi-equivalent conformers and all points i ($i = 1, \ldots, M$) in the asymmetric unit under consideration, the minimal distances $R_{i,j}$ ($j = 1, \ldots, T$) are determined. The RMSD score is then computed over all N points in the asymmetric unit as follows:

$$S_{\rm RMSD} = \left[\frac{\sum_{i=1}^{N}\left(m_i \sum_{s=1}^{K_i} R_{i,s}^2\right)}{\sum_{i=1}^{N} m_i K_i}\right]^{1/2}. \qquad (1)$$

Here, m_i denotes the multiplicity with which each point i occurs in the point array after application of icosahedral symmetry and, for each point i, $K_i \leq T$ denotes the number of distances $R_{i,s}$ ($s = 1, \ldots, K_i$) that are smaller than or equal to an arbitrarily chosen cutoff of 2 Å. The latter is due to the fact that such points are likely to be located on boundaries of adjacent capsid proteins, and we therefore consider their fit to all these capsid proteins in the formula.

(*d*2) Topography score $S_{\rm Top}$. The topography score determines which point array best matches the overall surface topography of the virus. To generate a quantitative score, we have used a clustering algorithm that approximates these external features as follows: we locate the most radially distal 5% (by distance) of Cα atoms in the viral capsid. These are then clustered using the *hclust* algorithm in R (R Development Core Team, 2008) with a *cutree* height threshold of 20 Å. The mean of all atoms within each cluster is then calculated and scaled to the outermost surface. The shortest distance of this mean (*e.g.* the midpoint of the towers in PaV, see Fig. 4) from any point in the array yields the topography score that measures how well external features of the virus are captured in our coarse-grained surface representation.

(*d*3) Combined score. The combined score $S_{\rm com} = (S_{\rm RMSD}^2 + S_{\rm Top}^2)^{1/2}$ simultaneously optimizes the two independent values $S_{\rm RMSD}$ and $S_{\rm Top}$, and the point array with the lowest combined score is considered the best match for the test virus.

(*e*) Evaluation of scores. The values for the viruses considered in this paper are listed in Table 2. Note that since we are not comparing like with like, but the fit of the point sets around a surface, the RMSD scores are higher than in crystallographic studies. Therefore, their relative, rather than absolute, values are important here. Only points of the best-fit array within the volume encompassing the viral proteins are used for calculating $S_{\rm com}$. Any inferences regarding points overlapping with the area occupied by the viral genome are hence predictions of this approach. For example, the vertices in the minor grooves of the dodecahedral cage in PaV or those encasing the two-shell RNA architecture in MS2 are predictions of this algorithm.

Table 2
The RMSD score, topography score and combined score in Å for the test viruses discussed in this paper.

TNV is tobacco necrosis virus; DYMV is desmodium yellow mottle virus.

Virus name	RMSD score ($S_{\rm RMSD}$)	Topography score ($S_{\rm Top}$)	Combined score ($S_{\rm com}$)
MS2	2.53	5.83	6.35
Pariacoto	4.7	0.89	4.78
SV40	1.48	2.39	2.81
GA	2.17	4.24	4.76
Hepatitis B	4.06	1.62	4.38
TBSV	1.92	5.92	6.23
CCMV	0.94	6.97	7.04
CCMV swollen	2.17	3.93	4.48
STMV	1.19	2.78	3.02
TNV	4.63	12.64	13.46
DYMV	1.25	1.37	1.85

(*f*) Remarks regarding robustness of the algorithm. Data from PDB files indicate atomic positions only up to a certain resolution and it is therefore important to ensure that the best-fit array is not an artefact of this intrinsic uncertainty. For this, the algorithm monitors robustness of the best-fit array against small incremental rescalings of the PDB data with respect to the array. The *prevalence score* measures for how many rescalings, each moving point in increments of 0.1 Å, the best-fit point array prevails as optimal fit. All best-fit arrays determined for the applications presented here have prevalence scores of at least 4 and hence perform positively with respect to this test.

Finally, note that array points are not one-to-one with atomic positions of capsid protein or the packaged genomes, *i.e.* the RNA molecules in the cases discussed here. Rather, they are mapping onto material boundaries, which correspond to the surface representations that can be obtained from the atomic positions, *e.g. via* the software package *PyMol* (http://www.pymol.org/). Therefore, proximity of array points to material boundaries is computed using the RMSD to atoms in the surface; arrays with points within material are not considered a good fit to the structure and are hence excluded from the procedure that determines the best-fit point array. The reasons for this are twofold. First, point arrays are much too sparse to account for all atomic positions. Second, this choice is inspired by Janner's work, which uses lattice techniques to model material boundaries in viruses. Our approach is a natural extension of this, but uses quasi-lattices as opposed to lattices, and provides a library of finite point arrays which are subsets of quasi-lattices and describe the structure of the virus as a whole. As a result of our procedure, the best-fit algorithm selects that point array from among the 569 that best represents the material boundaries of the capsid in this sense.

4. Applications to test cases

Viruses with the same number of capsid proteins are described by the same element in the Caspar–Klug classification (same T-number). We therefore analyse here two viruses of the same

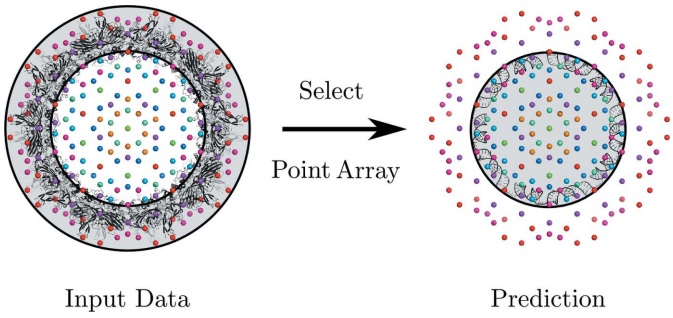

Figure 3
An illustration of the procedure underlying the best-fit algorithm. Only those points in each array in the library of 569 point arrays are used as input that overlap with the capsid, here shown as a shaded ring structure. Once the best-fit point array has been selected, its remaining points overlapping with the capsid interior must also be descriptors of capsid geometry. They therefore predict additional structural constraints. In our case studies below we show that these predictions provide additional insights into the genome packaging structure that agree well with experimental data.

T-number (Pariacoto virus and bacteriophage MS2, both $T = 3$) to demonstrate that our procedure provides additional information that distinguishes between their structures.

The best-fit algorithm uses as input only those points in each of the 569 arrays that overlap with the capsid and selects the best-fit point array on that basis. However, from a mathematical point of view, all points in the array are collectively determined by an application of symmetry generators. It is therefore not possible to ignore any array points in the mathematical structure and the entire point array implies constraints on the structure of the virus. In particular, the array points of the best-fit array that have not been used as input for the best-fit algorithm provide a prediction for genome organization, as illustrated schematically in Fig. 3.

4.1. Pariacoto virus (PaV)

We first examine PaV (Tang *et al.*, 2001), a $T = 3$ virus that infects insects. The PaV capsid is formed from 180 identical coat proteins clustered as 60 trimers. As well as revealing the structure of the protein shell, the X-ray crystal structure shows that ∼35% of the genomic RNA is also icosahedrally ordered. This portion of the genome has been modelled as a dodecahedral cage of *A*-type duplexes that meet at three-way junctions. Quasi-equivalence theory precisely predicts the positions of the three distinct conformers required to build the $T = 3$ PaV capsid. However, since it only deals with the positions of proteins within the surface (two-dimensional) lattice that tessellates a sphere, it makes no predictions about the topography of those virus coat protein conformers, or about the position of the genomic RNA with respect to the overlying protein shell. For particles with extended symmetry, however, the positions and shapes of all viral components must be correlated.

Fig. 4(*a*) shows the structure of the PaV protein shell viewed from the outside, with an icosahedron superimposed to highlight the local symmetry of the protein lattice. The match to symmetry elements such as the two-, three- and fivefold axes in this case is obvious. Extended symmetries are characterized by arrays of points in three dimensions that are both difficult to visualize and whose locations are non-intuitive (Fig. 4*b*). The best-fit algorithm scans the library of point arrays by scaling them and localizing these points to defined positions relative to the PaV structure. In this case a clear best match was found between the most radially extended points of one array and the tops of the trimeric protein spikes (see supplementary movies 2 and 3). These matches are highlighted in Fig. 4(*a*). Note that they do not lie on a global symmetry axis of the icosahedron, implying that the array library contains information about protein topography as well as geometry. These external scaling points are shown as magenta spheres. Further points in this array delineate the outer (red points in Figs. 4*b*, 4*c*) and inner surfaces (orange and yellow points) of the protein shell. The positions of these additional points are fixed by the scaling of the entire array to the spikes on the protein capsid surface. No points occur within the volume of the protein subunits. In other words, this array of points describing one of the allowed forms of extended icosahedral symmetry accurately maps both the major surface features and the thickness of the PaV coat protein shell.

As a test of our procedure, we have inverted the procedure in Fig. 3 and have ranked all 196 point arrays that have been identified as potential candidates by the best-fit algorithm solely according to their fit to the RNA cage in Pariacoto virus. The best-fit array determined previously ranks 19th based on RMSD alone, but if robustness against small deviations (the prevalence score) is taken into account, it ranks first. This suggests that this algorithm indeed identifies the point array in the library that best captures its overall structure.

4.2. Bacteriophage MS2

To see if viruses other than PaV show evidence of extended symmetry we applied the same matching algorithm to MS2 (Valegård *et al.*, 1990), a bacteriophage infecting *Escherichia coli*. The matching algorithm again returned a single best-fit point array, scaling to the most radially distant features of the MS2 capsid, namely a subset of the N-terminal β-hairpins of the capsid proteins (see magenta points in Fig. 5*a* and supplementary movie 4). Note, in this $T = 3$ shell only the hairpins on the *B*-type subunits are associated with array points. These are at a slightly higher radius than the equivalent structures on *A*- and *C*-type subunits. The best-fit array has further points that mark the approximate positions of the outer (red points) and inner surfaces (orange and yellow points) of the protein shell (Fig. 5*b*). Interestingly, the orange and yellow points locate the bottom surfaces of both types of quasi-equivalent protein dimers (*A*/*B* and *C*/*C*, shown in blue/green and pink, respectively) (Valegård *et al.*, 1990) required to form the $T = 3$ surface lattice, even though they are at different radial levels. As for PaV, the extended symmetry point array correctly delimits the MS2 capsid. Both PaV and MS2 are $T = 3$ capsids and are equivalent as far as quasi-equivalence theory is concerned, although we can distinguish

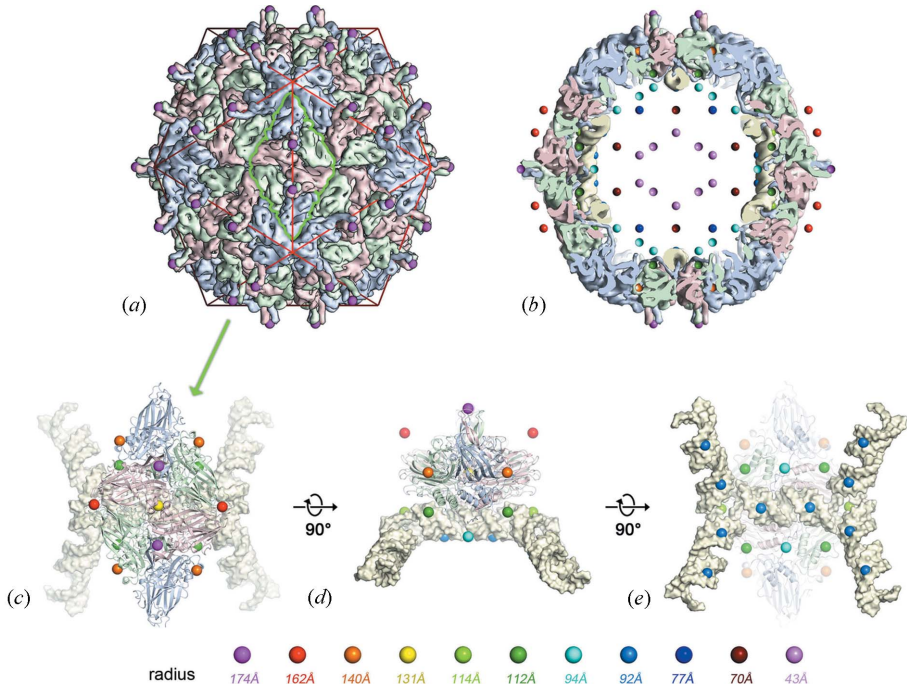

Figure 4
Structural constraints encode protein topography and genome organization in PaV. (*a*) The capsid of PaV is organized according to icosahedral symmetry as illustrated by its match to the superimposed icosahedron (red). The magenta points are additional constraints encoded by our theory and correspond to the outermost points of the best-fit array. They match to the tops of the trimeric protein spikes, which is striking given that these are not located on axes of icosahedral symmetry. (*b*) A cross-sectional view (52 Å thick) of the capsid, showing the locations of the best-fit array points relative to the protein container and its closely associated dsRNA cage (light yellow). (*c*)–(*e*) Close-ups of the two trimers bounded by the green rhomb in (*a*) together with an associated portion of the dodecahedral RNA cage viewed from (*c*) outside the particle, (*d*) the side and (*e*) inside the particle. The point array encodes constraints on the trimeric protein complex (orange and yellow points) and the relative sizes of the capsid and RNA cage. Strikingly, (predictive) green points map on the three-way junctions of this cage, and (predictive) blue points fit snugly into the minor grooves of the *A*-type RNA duplexes. Since the locations of all points are fixed by extended symmetry with respect to the outermost array points (magenta), this implies that protein topography and RNA organization are correlated by a geometric scaling principle that is encoded by extended icosahedral symmetry. For clarity, array points are shown here and throughout as spheres of 4.5 Å radius, colour coded by their radial positions. Note, the PaV crystal structure is the result of icosahedral symmetry averaging. This procedure does not assume any interdependence of molecules at different radial levels and does not alter the conclusions from the matching to the array described above.

them by structure determination. The new geometric principle described here shows that they are both distinct solutions for particles having extended symmetry.

4.3. Genome organization is also predictable

Points in the best-fit array for PaV overlapping with the volume occupied by the RNA genome (see Fig. 4*b* and also supplementary movie 3) provide predictions on genome organization. Strikingly, the mid-blue points at a radius of 92 Å are located in the minor grooves of the *A*-type duplex RNA, and green points at a radius of 114 Å mark the positions of the threefold junctions (see their locations in relation to a portion of the dodecahedral RNA cage in Fig. 4*c*). This is a remarkable result since the scaling of the entire point array was to a feature on the outer surface of the protein capsid. More significant than the placement of individual points adjacent to the genomic RNA is the fact that they define a set of five intersecting lines about each pair of three-way junctions, albeit with the lines defined by just two array points (Figs. 4*d*, 4*e*). These lines are parallel to the helical axes of the modelled genomic RNA, implying that the virus has evolved to maximize its symmetry in three dimensions. This result could only occur if there is an intrinsic geometric relationship between the shapes and structures of the coat protein layer and of the packaged genome, suggesting a completely new concept in virus biology. The obvious conclusion is that viruses show evidence of extended icosahedral symmetry.

Very similar results were obtained by applying the same analysis to the RNA bacteriophage MS2 (see Figs. 5*c*, 5*d*). The high-resolution X-ray structure of the MS2 virion, like many ssRNA viruses, does not show any density that can be ascribed to the viral RNA (Valegård *et al.*, 1990). However, an icosahedrally averaged cryo-EM structure at intermediate (∼9 Å) resolution reveals a double shell of RNA packaged with approximate order (Toropova *et al.*, 2008). Strikingly, once again, array points overlapping with the volume occupied by the RNA map around the contours of this RNA density (Figs. 5*c*, 5*d*). Again, the positions of these points were fixed by scaling to the outside of the protein shell. In MS2, we have shown that the contacts between the genome and the coat proteins in the shell play significant roles in virus assembly (Basnak *et al.*, 2010; Dykeman & Sankey, 2010; Stockley *et al.*, 2007; Dykeman *et al.*, 2011; Morton *et al.*, 2010; Rolfsson *et al.*, 2010; Toropova *et al.*, 2011). We have argued that multiple RNA stem-loop coat protein dimer interactions are required to determine the positions of the *A*/*B* quasi-equivalent dimers in the final capsid. The locations of these interactions have been determined using a defined, high-affinity stem-loop, the translational repressor TR (that is, the stem-loop formed by the sequence 5′-ACAUGAGGAUUACCCAUGU-3′), which has allowed determination of X-ray structures for this complex in the context of the capsid (Valegård *et al.*, 1990, 1997). Fig. 5(*d*) shows that the array points map neatly to the bound RNA fragments at the different quasi-equivalent locations in the $T = 3$ shell, suggesting that the intrinsic scaling implied by the PaV result also applies to MS2.

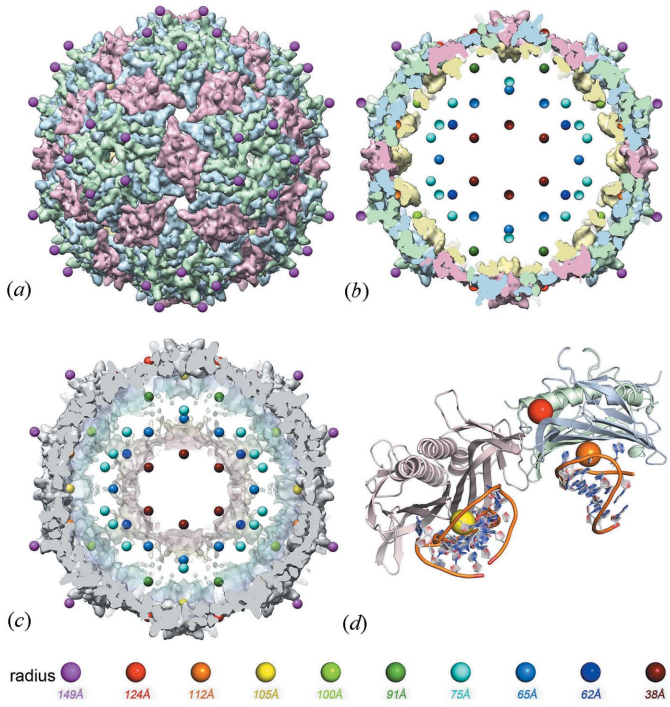

Figure 5
The structural constraints predict a two-shell genome organization in bacteriophage MS2. (*a*) The outermost points of the best-fit array scale to the N-terminal β-hairpins of the capsid proteins in MS2 (magenta). (*b*), (*c*) Central sections through the particle; (*b*) illustrates the match to the crystal structure in surface representation (Valegård *et al.*, 1990), with TR stem-loops shown in yellow. As can be seen in the close-up in (*d*), array points are located at the contact points between stem-loops and protein. This is even more striking given that yellow and orange points are predicted to be located at different radial levels, corresponding to the contacts with the two different types of dimeric building blocks (*A*/*B* and *C*/*C*) of the capsid. (*c*) An illustration of the match with the cryo-EM RNA density (Toropova *et al.*, 2008), shown here as a radially coloured transparent surface. Array points map the inside (maroon points, radius of 38 Å) and outside (blue and mid-blue points, 62 and 65 Å, respectively) surfaces of the inner RNA shell, and also mark the density connecting the inner and outer RNA shells (cyan, 75 Å). Strikingly, magenta and maroon points together define the spatial extent of material in this particle.

A comparison of the PaV and MS2 results shows that, even though these viruses are both $T = 3$ structures, they are represented by different elements in the library of point arrays. Consequently, our algorithm implies different predictions for the genome organization inside their capsids.

4.4. Application to a non-quasi-equivalent virus

As much as being able to provide additional information for the quasi-equivalent cases in Caspar–Klug theory, our approach complements viral tiling theory. The polyomaviridae include cancer-causing viruses in humans and have non-quasi-equivalent capsids, *i.e.* capsids built entirely from pentamers, rather than the mixture of pentamers and hexamers of a capsid built according to the rules of quasi-equivalence, and their surface structures therefore follow viral tiling theory (Twarock, 2004). Simian virus 40 (SV40) (Liddington *et al.*, 1991) is an example of such non-quasi-equivalent capsid

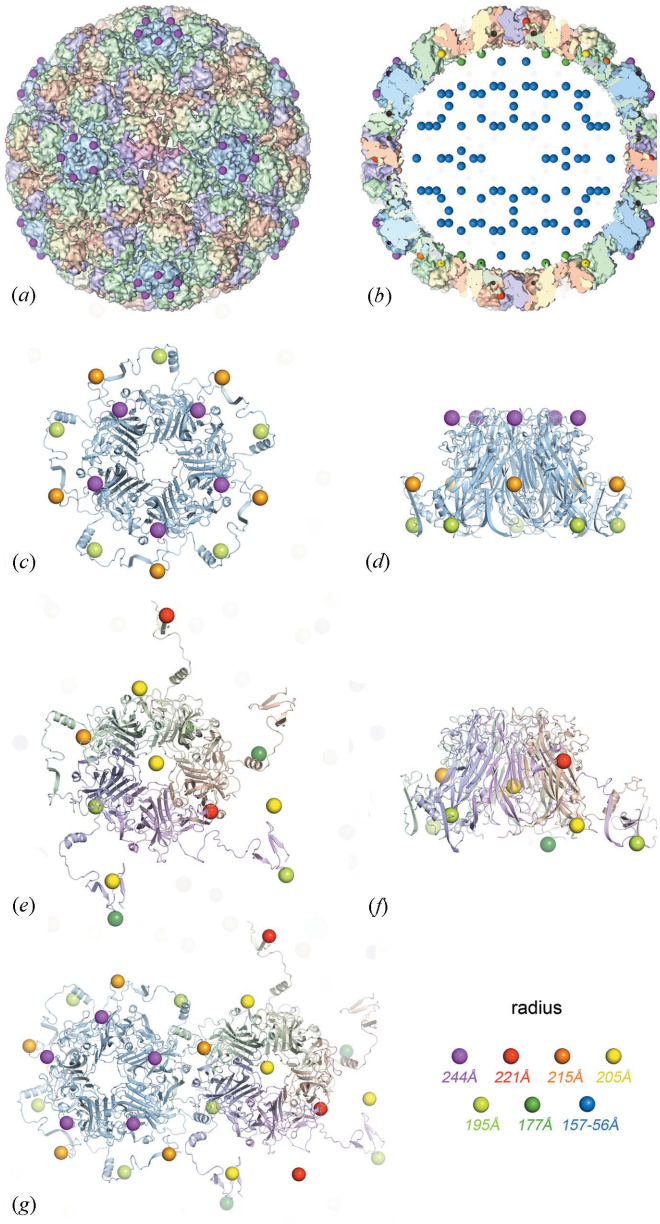

Figure 6
Non-quasi-equivalent configurations can also be predicted. The viral capsid of SV40 contains 360 identical coat protein subunits arranged as 72 pentamers, an example of a non-quasi-equivalent capsid organization. (*a*) The outermost constraints (magenta) are grouped around the 12 clusters of five proteins (pentamers) at the particle fivefold axes of icosahedral symmetry. (*b*) Cross-sectional view of the capsid, showing the locations of points in the array relative to protein. (*c*)–(*g*) Ribbon representations of the two different types of pentamers in the capsid; (*c*), (*d*) top and side view of the 12 pentamers at the fivefold axes, viewed from outside the capsid; (*e*), (*f*) show the corresponding views for the 60 pentamers off the symmetry axes; (*g*) shows both pentamer environments simultaneously as situated in the capsid. The new geometric principle of virus architecture distinguishes between the two types of pentamer environments and incorporates this viral geometry that cannot be modelled in quasi-equivalence theory.

architecture (see Fig. 6*a*). Its capsid is composed of 72 coat protein pentamers. This deviation from the standard pattern is possible because the coat proteins in the pentamers exhibit two distinct types of bonding interactions with proteins in

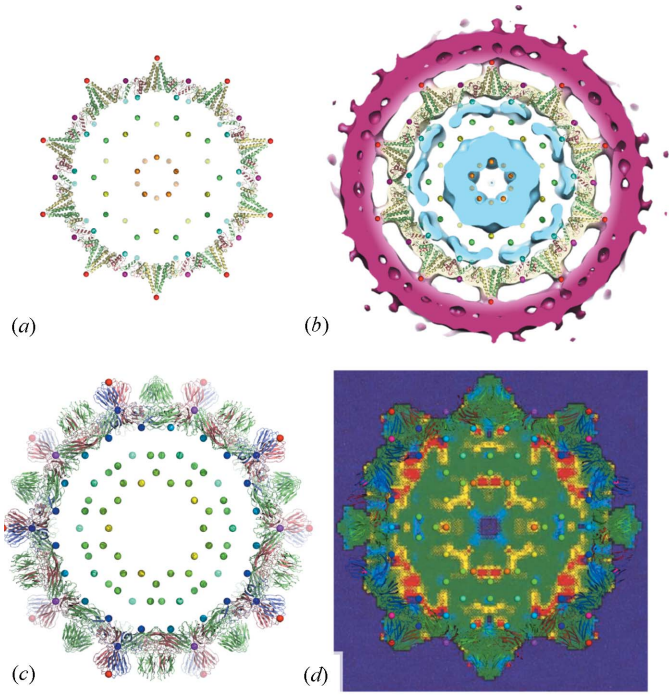

Figure 7
(*a*) The best-fit point array for hepatitis B viewed down a fivefold axis and (*b*) overlaid on the cryo-EM data of the genomic material, adapted from Wynne *et al.* (1999). (*c*) The best-fit point array for TBSV viewed down a fivefold axis and (*d*) overlaid on neutron scattering data, adapted from Hopper *et al.* (1984), Olson *et al.* (1983) and Hogle *et al.* (1983).

neighbouring pentamers defined by differing conformations of their extended C-terminal arms. We used SV40 to see if our extended-symmetry approach would also shed light on such non-quasi-equivalent structures.

The matching algorithm again unambiguously identified a single member of the library of point arrays for SV40, which scales to the outer surface of the protein capsid (magenta points in Figs. 6*a*–6*g*). This array also has points located on the inner surface (green points in Figs. 6*b*–6*g* and supplementary movie 5), *i.e.* the array of points matches the protein topography and defines the capsid thickness, as for the other test viruses. Further points overlap with the volume occupied by the genome, although there are no structural data for this component in this case (Fig. 6*b*). The distribution of array points differs for the two types of pentamers. Strikingly, a subset of the points discriminates the positions of the two different types of C-terminal arm conformation (Figs. 6*c*–6*g*). In other words, SV40 follows the predictions of extended symmetry, *i.e.* it is not an anomaly. This resolves a long-standing structural puzzle in virology.

4.5. Application to a wide range of viruses

The test cases in §§4.1–4.4 demonstrate that the library of point arrays can distinguish between two $T = 3$ viruses and also applies to a non-quasi-equivalent case, a $T = 7$ virus. In order to demonstrate that the new approach can also account for other virus structures, we have performed the analysis for a wide range of viruses (see Table 2). We used hepatitis B virus (PDB id 1qgt) shown in Figs. 7(*a*) and 7(*b*), which also exhibits a cage structure in its packaged genome (Wynne *et al.*, 1999), to demonstrate that the library of point arrays also applies to a $T = 4$ structure. STMV (satellite tobacco mosaic virus) has been used as an example of a $T = 1$ virus, and our results are in good agreement with the RNA fragments represented in its PDB file (PDB id 1a34; Larson *et al.*, 1998). We have shown that bacteriophage GA (PDB id 1gav; Tars *et al.*, 1997), which is in a different group in the same family as MS2, has the same best-fit point array, showing that the structures of two evolutionarily related viruses are represented by the same point array in our classification.

Moreover, the point arrays determined as the best fit for tomato bushy stunt virus (TBSV) (PDB id 2tbv) (Hopper *et al.*, 1984; Olson *et al.*, 1983; Hogle *et al.*, 1983) show that the two-domain architecture of its capsid protein is reflected in the array (Figs. 7*c*, 7*d*). Different features in the genome organization, such as the polymorphic organization of the Seneca Valley genome, are also accounted for by best-fit point arrays. Finally, we interrogated the capsid structure of CCMV (cowpea chlorotic mottle virus) before (Speir *et al.*, 1995) and after (Tama & Brooks, 2002; Liu *et al.*, 2003) expansion. Best-fit point arrays provided coarse-grained models that were used to determine features of the structural transition *via* a lattice approach (Indelicato, Keef *et al.*, 2012).

5. Discussion

The majority of viruses use icosahedral symmetry to build their capsids because of genetic economy. A container built from multiple copies of a single coat protein subunit has the largest volume if organized according to icosahedral symmetry. This concept is the same as the reason why bees build hives with hexagonal lattices minimizing the amount of wax needed, and intersecting soap bubbles minimize their surface areas. If viral subunits can adopt multiple, quasi-equivalent conformations then much larger capsids, able to package even larger genomes, can be constructed from this single gene product, and the vast majority of known capsids are of this form. They are currently classified in terms of their T-numbers (Caspar & Klug, 1962), which can be very large for large viruses and imply that capsids are built from $60T$ copies of a coat protein subunit. The T-number predicts the locations of the coat proteins in the capsid relative to a tessellation of the surface of a sphere that encodes the structural organization of the capsid. Although coat protein subunits are three-dimensional objects, their locations in quasi-equivalence theory are described in terms of a two-dimensional surface. In this study, we have applied the mathematics of extended symmetry to virus structures and generated results that go beyond quasi-equivalence into three dimensions. Unlike quasi-equivalence, our theory can distinguish between the architectures of different viruses with the same T-number and, strikingly, predicts aspects of coat protein and genome topography. It also incorporates into a single scheme previously anomalous virus architectures. Our results imply that all parts

of a virus are structurally related *via* geometric constraints that are an implicit property of the symmetry of its capsid. Viruses are therefore even more strongly constrained by symmetry than previously realised.

An obvious question is: why is this the case? Zandi *et al.* (2004) have shown that viral protein containers organized according to the principle of quasi-equivalence correspond to local minima in an energy landscape. By analogy, perhaps entire viral particles organized according to extended symmetry also represent such local free-energy minima, albeit in a more complex energy landscape? The new principle of extended symmetry would then help to explain the inherent stability of viral particles, which is important between rounds of infection. This would provide a plausible evolutionary mechanism by which extended symmetry could be selected for during viral evolution. At first glance this seems improbable but the same type of interaction between evolution and the constraints of three-dimensional geometry gave us bees that understand how to make a hexagonal lattice and hence store the most honey for the least amount of beeswax.

Extended symmetry has far-reaching implications for our understanding of virus biology, because it reveals for the first time the interdependence of the shapes and sizes of all viral components. For example, virus particles commonly undergo large conformational changes during maturation or infection. The new geometric principles revealed by extended symmetry must apply to all these metastable conformational states. The structural transitions that occur during maturation and infection are therefore inherently predictable based on an X-ray or cryo-EM structure of the native capsid (Indelicato, Cermelli *et al.*, 2012). These insights are extremely timely, as we are just starting to understand the detailed molecular mechanisms that underlie virus assembly (Basnak *et al.*, 2010; Dykeman & Sankey, 2010; Stockley *et al.*, 2007; Dykeman *et al.*, 2011; Morton *et al.*, 2010; Rolfsson *et al.*, 2010; Toropova *et al.*, 2011) and exploit them. Potential applications include the creation of targeted drug-delivery vehicles and imaging contrast agents (Wu *et al.*, 1995; Lewis *et al.*, 2006), as well as the use of viruses and virus-like particles for vaccine development (Jagu *et al.*, 2010).

Viral coat proteins are striking in their levels of topographical conservation despite very low levels of primary sequence identity. This is usually ascribed to the existence of a common ancestor, *i.e.* to a divergent evolutionary process (Bamford *et al.*, 2005). Since evolution acts at the level of the phenotype, the implication is that only very few protein folds can accommodate the requirements of viral coat protein subunits. From quasi-equivalence these are relatively modest, and include an ability to assemble into an icosahedral surface lattice and be flexible enough to create quasi-conformers, as well as producing viable virions. Many other multi-protein complexes are known whose proteins seem able to fulfil many of these criteria, leaving a puzzle. Our results suggest an alternative explanation for viral protein folds based on the competitive advantage of virion stability implied by highly symmetric capsids in three dimensions. Even if such a phenotype conferred a minimal advantage compared to the bulk population, the trait of favouring extended capsid symmetry would rapidly become fixed in a viral population. This additional constraint would dramatically reduce the numbers of protein folds capable of achieving maximal stability, partially explaining the observed level of conservation. The approach presented here provides a tool for the prediction of the structural constraints that a virus organized with highest symmetry should obey. This does not completely constrain its structure, but rather provides important insights into those structural features having the highest level of symmetry and which are therefore less likely to change in an evolutionary cycle. As we have demonstrated with this analysis, the best-fit point arrays of MS2 and the evolutionarily related GA are identical. Since the outermost points in the best-fit array map around the immuno-dominant epitopes, usually the most radially distant features of a virus, this has implications for features that are preserved when viruses mutate. An appreciation of extended symmetry will therefore underpin our attempts to develop new therapies, *e.g.* by development of vaccines and agents that target viral assembly. This includes those oncogenic viruses, such as the human papillomaviruses discussed here, that underlie cervical cancers, the structures of which have not previously been explained by quasi-equivalence theory alone (Jagu *et al.*, 2010).

We thank Professor Simon Phillips, Research Complex at Harwell, for helpful discussions, and Professors Alasdair Steven, NIH, and Colin Kleanthous, York, for comments on the manuscript. We are grateful for financial support *via* a Leverhulme Trust Research Leadership Award (to RT), the UK BBSRC and especially the Leverhulme Trust (PGS and NAR), as well as the University of Leeds for its support of research within the Astbury Centre.

References

Bamford, D. H., Grimes, J. M. & Stuart, D. I. (2005). *Curr. Opin. Struct. Biol.* **15**, 655–663.
Basnak, G., Morton, V. L., Rolfsson, O., Stonehouse, N. J., Ashcroft, A. E. & Stockley, P. G. (2010). *J. Mol. Biol.* **395**, 924–936.
Brooks, B. R. *et al.* (2009). *J. Comput. Chem.* **30**, 1545–1614.
Bruijn, N. G. de (1981*a*). *Nederl. Akad. Wetensch. Indag. Math.* **43**, 39–52.
Bruijn, N. G. de (1981*b*). *Nederl. Akad. Wetensch. Indag. Math.* **43**, 53–66.
Carrillo-Tripp, M., Shepherd, C. M., Borelli, I. A., Venkataraman, S., Lander, G., Natarajan, P., Johnson, J. E., Brooks, C. L. & Reddy, V. S. (2009). *Nucleic Acids Res.* **37**, D436–D442.
Caspar, D. L. & Klug, A. (1962). *Cold Spring Harb. Symp. Quant. Biol.* **27**, 1–24.
Coxeter, H. S. M. (1972). In *A Spectrum of Mathematics*, edited by J. C. Butcher, pp. 98–107. Oxford University Press.
Crick, F. H. & Watson, J. D. (1956). *Nature (London)*, **177**, 473–475.
Dykeman, E. C., Grayson, N. E., Toropova, K., Ranson, N. A., Stockley, P. G. & Twarock, R. (2011). *J. Mol. Biol.* **408**, 399–407.
Dykeman, E. C. & Sankey, O. F. (2010). *Phys. Rev. E*, **81**, 021918.
Hogle, J., Kirchhausen, T. & Harrison, S. C. (1983). *J. Mol. Biol.* **171**, 95–100.
Hopper, P., Harrison, S. C. & Sauer, R. T. (1984). *J. Mol. Biol.* **177**, 701–713.

Indelicato, G., Cermelli, P., Salthouse, D. G., Racca, S., Zanzotto, G. & Twarock, R. (2012). *J. Math. Biol.* **64**, 745–773.
Indelicato, G., Keef, T., Cermelli, P., Salthouse, D. G., Twarock, R. & Zanzotto, G. (2012). *Proc. R. Soc. Lond. Ser. A*, **468**, 1452–1471.
Jagu, S., Kwak, K., Garcea, R. L. & Roden, R. B. (2010). *Vaccine*, **28**, 4478–4486.
Janner, A. (2010*a*). *Acta Cryst.* A**66**, 301–311.
Janner, A. (2010*b*). *Acta Cryst.* A**66**, 312–326.
Janner, A. (2011*a*). *Acta Cryst.* A**67**, 174–189.
Janner, A. (2011*b*). *Acta Cryst.* A**67**, 517–520.
Janner, A. (2011*c*). *Acta Cryst.* A**67**, 521–532.
Keef, T. & Twarock, R. (2009). *J. Math. Biol.* **59**, 287–313.
Kramer, P. & Shlottmann, M. (1989). *J. Phys. A*, **22**, L1097–L1102.
Larson, S. B., Day, J., Greenwood, A. & McPherson, A. (1998). *J. Mol. Biol.* **277**, 37–59.
Lewis, J. D., Destito, G., Zijlstra, A., Gonzalez, M. J., Quigley, J. P., Manchester, M. & Stuhlmann, H. (2006). *Nat. Med.* **12**, 354–360.
Liddington, R. C., Yan, Y., Moulai, J., Sahli, R., Benjamin, T. L. & Harrison, S. C. (1991). *Nature (London)*, **354**, 278–284.
Liu, H., Qu, C., Johnson, J. E. & Case, D. A. (2003). *J. Struct. Biol.* **142**, 356–363.
Morton, V. L., Dykeman, E. C., Stonehouse, N. J., Ashcroft, A. E., Twarock, R. & Stockley, P. G. (2010). *J. Mol. Biol.* **401**, 298–308.
Olson, A. J., Bricogne, G. & Harrison, S. C. (1983). *J. Mol. Biol.* **171**, 61–93.
R Development Core Team (2008). *R: a Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
Rolfsson, O., Toropova, K., Ranson, N. A. & Stockley, P. G. (2010). *J. Mol. Biol.* **401**, 309–322.
Scherrer, W. (1946). *Elemente der Mathematik*, **1**, 97–98.
Senechal, M. (1996). *Quasicrystals and Geometry*. Cambridge University Press.
Speir, J. A., Munshi, S., Wang, G., Baker, T. S. & Johnson, J. E. (1995). *Structure*, **3**, 63–78.
Stockley, P. G., Rolfsson, O., Thompson, G. S., Basnak, G., Francese, S., Stonehouse, N. J., Homans, S. W. & Ashcroft, A. E. (2007). *J. Mol. Biol.* **369**, 541–552.
Tama, F. & Brooks, C. L. III (2002). *J. Mol. Biol.* **318**, 733–747.
Tang, L., Johnson, K., Ball, L., Lin, T., Yeager, M. & Johnson, J. E. (2001). *Nat. Struct. Biol.* **1**, 77–83.
Tars, K., Bundule, M., Fridborg, K. & Liljas, L. (1997). *J. Mol. Biol.* **271**, 759–773.
Toropova, K., Basnak, G., Twarock, R., Stockley, P. G. & Ranson, N. A. (2008). *J. Mol. Biol.* **375**, 824–836.
Toropova, K., Stockley, P. G. & Ranson, N. A. (2011). *J. Mol. Biol.* **408**, 408–419.
Twarock, R. (2004). *J. Theor. Biol.* **226**, 477–482.
Valegård, K., Liljas, L., Fridborg, K. & Unge, T. (1990). *Nature (London)*, **345**, 36–41.
Valegård, K., Murray, J. B., Stonehouse, N. J., van den Worm, S., Stockley, P. G. & Liljas, L. (1997). *J. Mol. Biol.* **5**, 724–738.
Wardman, J. (2012). PhD thesis, University of York, England.
Worm, S. H. van den, Koning, R. I., Warmenhoven, H. J., Koerten, H. K. & van Duin, J. (2006). *J. Mol. Biol.* **363**, 858–865.
Wu, M., Brown, W. L. & Stockley, P. (1995). *Bioconjugate Chem.* **6**, 587–595.
Wynne, S. A., Crowther, R. A. & Leslie, A. G. (1999). *Mol. Cell*, **3**, 771–780.
Zandi, R., Reguera, D., Bruinsma, R. F., Gelbart, W. M. & Rudnick, J. (2004). *Proc. Natl Acad. Sci. USA*, **101**, 15556–15560.