Architecture and Assembly of Maize streak virus:

Insights from 3D Electron Microscopy

Kyle Clayton Dent

A thesis submitted in partial fulfilment of the requirements of the degree of Master of Science (Structural Biology) at the University of the Cape Town

Supervisors:

Dr Arvind Varsani
Prof. Trevor Sewell

October 2013

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

Keywords

Maize streak virus
Geminiviridae
ssDNA virus
Virus structure
Macromolecular assembly
Quasi-equivalence
Electron cryo-microscopy
Single-particle 3D reconstruction

Abstract

Maize streak virus (MSV), circular single stranded DNA (ssDNA) virus (~2.7kb), is the causative agent of Maize streak disease, and is a devastating pathogen that causes severe crop losses to subsistence farmers in sub-Saharan Africa. MSV is transmitted by the leafhopper *Cicadulina mbila*, and is the type member of the Mastrevirus genus (family *Geminiviridae).* MSV shares a unique twinned icosahedral ("geminate") virion architecture (22 x 38 nm) with all other family members. Geminate particles consist of 110 coat protein (CP) subunits that assemble onto a circular single-stranded DNA (ssDNA) genome. Each *T = 1* unit is an incomplete icosahedron assembled from 55 CPs. The structures of MSV and *African cassava mosaic virus* (ACMV, genus Begomovirus) have been studied by electron cryo-microscopy (cryo-EM) previously. While these investigations revealed some details about the geminate architecture, the interactions of capsid components have not yet been adequately modelled. The two incomplete icosahedral "heads" of the geminate particle are offset from one another and apparently make distinct CP:CP contacts at this region of the virion. Information regarding the nature of quasi-equivalent CP conformers or the sets of amino acid residues that mediate these interactions has not been forthcoming. Since the experimental results of these previous studies are not available in a public database, we were motivated to revisit the structure of MSV in order to obtain a 3D experimental density that might aid pseudo-atomic modelling. The MSV CP:ssDNA interaction has also been shown to be crucial for systemic movement through the host. Hence, quasi-atomic modelling may inform development of antiviral strategies which aim to interfere with virion assembly.

MSV virions were isolated from the leaves of maize plants infected by agro-inoculation and visualized in both heavy metal stain and vitreous ice after they had been adsorbed to a thin-layer of continuous carbon to prevent virion aggregation. Virus preparations consisted of distinct CP assemblies consisting of multiples of the incomplete *T = 1* icosahedral unit. Monopartite (icosahedral), bipartite (geminate), tripartite, and higher assemblies were observed suggesting the MSV CP is not only multifunctional but also structurally versatile being able to package ssDNA of variable sizes. Low-dose images were recorded on film, and 3D reconstruction of both monopartite and bipartite capsid species carried out using standard single-particle image processing methodology. The resolution of the bipartite reconstructions was 26 Å for the negative-stain dataset, and 23 Å for cryo-EM dataset, while the resolution of the monopartite reconstruction was estimated to be ~ 15 Å. Comparative modelling of the MSV CP was undertaken using the pentamer (CP_5) of *Satellite tobacco necrosis virus* (STNV) as a structural template. Correlation-based fitting of icosahedral and geminate atomic models that varied in geometric arrangement of MSV CP_5 allowed the geometric parameters of the bipartite capsid to be determined.

Fitting of MSV CP_5 into the EM densities informed our understanding of interfaces which allow the CP to self-associate, and showed that CP_5 is in fact displaced within the icosahedral geometry of the heads by a 10° rotation about the 5-fold axes of symmetry in comparison to STNV; hence, while quaternary structure of the pentameric capsomer is conserved between these viruses, the quaternary interactions between capsomers of the *T = 1* unit has diverged considerably. This study shows that the offset between the geminate heads of the MSV virion is ~ -11°, and that this geometry appears to arise owing to a distinct set of CP:CP interfaces which occur across the equator between two quasi-icosahedral heads and involve regions that would interact to form the CP_5: CP_5 interfaces within each of the heads (across 2-fold and 3-fold symmetry axes). Notably this offset differs from that reported for ACMV, which has a reported offset of 20°. Additionally, the resolution afforded by the icosahedral monopartite reconstruction provided the first structural evidence to suggest that the calcium ion binding site of the STNV CP_5 (located on the C5 axis) is likely to be conserved in MSV. This result suggests that in common with other plant viruses, depletion of calcium ions may be required for genome egress in a newly infected host cell. This study highlights the importance of future high-resolution studies of this unique virion morphology by both X-ray crystallography and cryo-EM.

Declaration

I declare that ***"Architecture and Assembly of Maize streak virus: Insights from 3-D Electron Microscopy"*** is my own work, and that virus isolations, sample preparation, and electron microscopy were performed by my hand under supervision gratefully acknowledged. This dissertation has not been submitted towards any other degree at any other university. All sources of information I have used or quoted have been indicated and acknowledged by complete references.

Kyle C. Dent October 2013

Signed:

Acknowledgements

I thank Trevor Sewell, Arvind Varsani, Darren Martin, and Dionne Shepherd for allowing me to participate in their research group.

Trevor, thank you for your vision of Structural Biology in South Africa from which I benefited greatly.

Thanks to Mohammed Jaffer, Sean Karrim, Miranda Waldron and Brandon Weber for all of your help in so many ways. This work would not have been possible without you.

I would like to acknowledge the financial support of the *Harry Crossley Foundation*, the *National Research Foundation* as well as the *Carnegie Corporation of New York*. In addition, I thank the *Postgraduate Student and Funding Office* of the University of Cape Town for making my time at the University of Cape Town possible, and for giving me a start at a career.

In particular I would like to mention Irene Petrony, Linda Whitfield, and Linda Vranas for the work they do for students.

QE	Quasi-equivalence
RCR	Rolling circle replication
SDS-PAGE	sodium dodecyl sulphate polyacrylamide gel electrophoresis
SEM	Scanning electron microscopy
SNR	Signal-to-noise ratio
STNV	*Satellite tobacco necrosis virus*
SIR	Short intergenic region
ssRNA	Single-stranded ribonucleic acid
ssDNA	Single-stranded deoxyribonucleic acid
TBSV	*Tomato bushy stunt virus*
TEM	Transmission electron microscope
TMV	*Tobacco mosaic virus*
TNV	*Tobacco necrosis virus*
TNV	*Tobacco necrosis virus*
ToLCV	*Tomato leaf curl virus*
TpCTV	*Tomato peudo-curly top virus*

Table Contents

Abbreviations

2D	Two-dimensions
3D	Three-dimensions
ACMV	*African cassava mosaic virus*
BTCV	*Beet curly top virus*
BGYMV	*Bean golden yellow mosaic virus*
C5	5-fold cyclic point-group symmetry
CCMV	*Cowpea chlorotic mottle virus*
Cryo-EM	Electron cryo-microscopy
CP	Coat protein
CP_5	Coat protein pentamer
CPMV	*Cowpea mosaic virus*
CSMV	*Chloris striate mosaic virus*
D5	5-fold dihedral point-group symmetry
DED	Direct electron detector
EM	Electron microscopy
EDTA	Ethylene-diamine-tetraacetic acid
FSC	Fourier shell correlation
LIR	Long intergenic region
LN	Liquid nitrogen
MSV	Maize streak virus
NCS	Non-crystallographic symmetry
NLS	Nuclear localization signal
ORF	Open reading frame
PDB	Protein Data Bank
PSIPRED	Protein structure prediction webserver

1 Structural biology of viruses

1.1 An overview of virus structure and function

Viruses are macromolecular assemblies made up of proteins and nucleic acids that may or may not also include a lipid component. These assemblies have the potential to infect the cells of a host organism, and by harnessing the cells biosynthetic machinery, are then able to synthesize additional protein components as well as replicate their nucleic acid genomes. The protein components are able to assemble to encapsidate new copies of the viral genome, and hence produce more infectious virus particles (termed virions) that are able to spread to repeat the cycle of infection. Virions consist of a protein shell, the capsid that protects the encapsidated genome (together termed the nucleocapsid) from degradation by nucleases that may be present within the host cell, or in the environment. Lipid enveloped viruses derive their lipids from the host cell during a process of budding, that may be from the cell membrane or from another organelle inside the cell often related to secretory functions. Lipid enveloped viruses often include glycoproteins in their lipid bilayers, and these components also mediate interactions with the host which may allow the virus to enter and express its genes. For non-enveloped mammalian viruses it is the capsid of the virus that will mediate these interactions with the cell boundary, whereas for plant viruses, the virus will often rely on mechanical disruption of the plants thick cell walls to enter the host cells cytoplasm.

Viruses are classified according to their different properties such as genome (nucleic acid) type (DNA/RNA), genome architecture, serology, and capsid morphologies. The Baltimore classification (Baltimore, 1971) focuses on properties of the viral genome, whether either single-stranded (ss) or double-stranded (ds) and whether it is either RNA or DNA, as it is these properties that determine a viruses replication strategy. Additionally, viral genomes may be monopartite, or multipartite, meaning that they are encoded by only a single strand, or multiple strands of nucleic acid, respectively. Whether the genomic segments are packaged into a single virion, or travel separately varies from virus to virus. Viruses also vary in size dramatically, from simple viruses with diameters of 180 Å (e.g. *Porcine circovirus* (*Circoviridae*) with a single-stranded RNA (ssDNA) genome ~2000 nt long (Biagini, 2004)). Conversely, *Acanthamoeba polyphaga mimivirus*, one of the largest viruses known, has a capsid diameter of 400 nm (4000 Å) and double-stranded DNA genome 1,181,000 bp long (Scola *et al.*, 2003).

An emerging classification system for viruses relies on atomic resolution information about the virus capsid structure derived from X-ray crystallography. It is argued that owing to the high rates of viral divergence, especially for RNA viruses which have low fidelity polymerases lacking 3'

to 5' exonuclease activity, that the only reliable means for truly identifying distant evolutionary relationships is to rely on observing conservation of the capsid fold and architecture, which is less prone to change with divergence because of the large redundancy in sequence-structure space (many sequences fold to assume the same 3D shape). This implies that even with great sequence variation, whether at the protein or nucleic acid sequence level, the structural properties of the virus capsid will continue to show significant similarities, which will in turn implicate two viruses as having diverged from a common ancestor (Bamford *et al.*, 2005).

Plant viruses, one of which is the subject of this study, are often considered simple viruses for a number of reasons. Many have relatively small genomes, and consequently also have small regular protein shells (capsids) that are not enveloped by a lipid bilayer. Exceptions do exists, such as the Tospoviruses; (-) ssRNA viruses of the family *Bunyaviridae*. Many plant RNA viruses are helical (e.g. *Tobacco mosaic virus*), however spherical viruses are most relevant to this thesis, and will be discussed in the most detail.

Most of the spherical capsids are $T = 3$ capsids that are roughly 30 nm in diameter; refer to Section 1.3.1.4 for a description of conventions for describing the spherical architecture of viruses. The size of the capsid imposes limitations on the size of the genome, and the maximum genomic length of viruses with capsids of this size is roughly 9000 nucleotides (nt), or ~1.6 MDa worth of genomic material. Virus "satellites" also exist, these are small segments of nucleic acid that rely on functionality of another ("helper") virus for either replication or transmission. Most of these viruses are packaged by coat protein subunits (CPs) that they encode themselves, and are often heavily diverged at a sequence level from that of their "helper" virus. Important examples of these are *Satellite panicum mosaic virus* (SPMV) and *Satellite tobacco necrosis virus* (STNV), which are $T = 1$ viruses, possessing the simplest architecture of units corresponding to icosahedral symmetry (Figure 1-1). Other satellite viruses encode their own replicases, while relying on the CPs of their helper viruses for transmission; examples of these include the alpha and beta satellites of the geminiviruses. Satellite viruses typically have small genomes ± 1500-2000 nt, and are often only large enough (genetically) to encode a single gene, and are not genetically large enough to encode all functions required for a whole infectious- (or life-) cycle.

As mentioned previously, plant viruses rely on vectors (e.g. insects) to penetrate the thick cell boundary of a typical plant cell. Consequently, they do not recognise cell surface specific receptors (as mammalian viruses do) and hence do not have sophisticated injection machinery (e.g. as that of the bacteriophages which must also bypass a thick cell boundary). For this reason the surfaces of their capsids are typically inert, unreactive and often quite sturdy.

Plant viruses are in many cases relatively "simple" organisms, and are often amenable to inexpensive large scale production. These properties make them suitable for application to real world problems such as vehicles for drug-delivery, epitope presentation (i.e. the production of vaccines), as well as templates for materials synthesis (Douglas and Young, 1998). Consequently, much effort has been invested in developing them as a "toolbox" applicable to a variety of biotechnological applications (Steinmetz and Evans, 2007).

Many such technological applications require an understanding of the assembly behaviour of their CP subunits, which vary dramatically depending on how the virus particle is "constructed" or self-assembled in the biological context. Additionally, the production of empty capsids is often favoured as the infectious genome is a biosafety concern in agricultural settings. Examples of plant viruses developed toward biotechnological applications include *Cowpea mosaic virus* (CPMV) (Sainsbury *et al.*, 2010) and *Cowpea chlorotic mottle virus* (CCMV) (Brumfield, 2004). It is possible that the CPs of geminiviruses could be developed toward such specific applications as well, and it may be that the unique geminate architecture may possess useful properties towards these ends. This structural study was motivated by the understanding that information about these structures, as well as their assembly, will be required if they are to be successfully modified for such applications.

1.2 Why study the structure of viruses?

In the context of molecular biology, *function follows form*. It is the structure of a macromolecule (or complex), i.e. its shape and electrostatic properties that will determine how it interacts with other molecules in its environment. It is for this reasons that structural biology attempts to inform our understanding of molecular function by providing visual information of macromolecules, sub-cellular structures and entire cells. As complex macromolecular systems, viruses lend themselves well to study by structural techniques, and are attractive specimens because they possess qualities which facilitate their study (they are discrete structures, which can often be isolated in large quantities, or identified by direct observation inside cells within their functional environment). Indeed, owing to high rates of replication, and divergence, viruses can explore vast sequence and therefore structure spaces, and hence viruses can be informative model systems of molecular evolution and have developed many unique "machines" to facilitate progression of their life-cycles. It is for this reason that viruses may also be considered modular; each machine (e.g. replication protein, coat protein, or packaging motor) will fulfil a role largely independently of the others at each stage of the virus life cycle.

The virions of simple viruses are considerably less complex than their larger counterparts (often only encoding less than 10 functional proteins), however, these entities are not static containers

and must encode subtle mechanisms to allow the typical stages of a virus life-cycle to proceed (e.g. cell recognition, uncoating or disassembly, and subsequent assembly to package progeny viral genomes). These mechanisms are challenging topics of study, and little detail is known (with any certainty) about how these viruses go about their life-cycles.

Macromolecular self-assembly is a topic of broad implication, and is a phenomenon important to our understanding of the molecular foundations of all biological mechanisms and processes; including signal transduction (Johnson, 1996). Simple viruses have long been used as model systems for this phenomenon as they can be reassembled and in some cases reconstituted, i.e. infectious virions produced from a series of purified components in a test tube (Bancroft and Hiebert, 1967; Fraenkel-Conrat and Williams, 1955). Previously, assembly was studied using recombinant coat protein alone, however, more recently studies carried out in the presence of the viral genome, or derivate thereof, have indicated an import role of the genome in the particle formation (Stockley *et al.*, 2013). In these cases the genome participates in the assembly in a "concerted" or simultaneous mode of packaging and assembly. This differs from more complex viruses, which typically produce procapsids and progeny genomes, before a specialized motor translocates the genome into the capsid by an ATP dependent process. Assembly and packaging of viruses is discussed in further detail in Section 1.3.1.1.

1.3 Structural biology of viruses

Structural studies of viruses began with the plant viruses *Tobacco mosaic virus* (TMV) and *Tomato bushy stunt virus* (TBSV). In both cases X-ray diffraction studies indicated that that these viruses were symmetric. The observation of layer lines in the diffraction pattern of TMV (Franklin, 1956), similar to those seen for double-stranded DNA (Watson and Crick, 1953), strongly suggested that TMV was helically symmetric. While the symmetry present in the diffraction patterns of TBSV indicated 5-fold rotational symmetry axes, lending support to the idea that the virus might be spherical, possessing cubic (icosahedral) symmetry (Caspar, 1956).

While these X-ray diffraction studies did not allow the specimen to be viewed directly (as the diffraction pattern recorded only the intensities of the diffracted X-rays), subsequent visualization using electron microscopy did – albeit at considerably lower resolution. Owing to the charge of the electron, the electron microscope (EM) is able to preserve the phases of the scattered electrons by recombining them using a powerful electromagnetic objective lens to form an image. The EM observations confirmed that two primary nucleocapsid morphologies exist. TMV was seen to be rod shaped, while TBSV is spherical or isometric.

Due to the overall size of virus particles (virion), both X-ray crystallography and electron microscopy have become the primary techniques used by structural virologists. Nuclear magnetic resonance spectroscopy (NMR) is generally considered inapplicable owing to the relatively large size of virion macromolecular assemblies, although the method has been used to study of the structure and interaction of isolated viral components or intermediates.

Both of these principal methods have their benefits, challenges, and drawbacks. X-ray crystallography relies on the availability of a high concentration of very pure sample, which must then be crystallized; often the most challenging stage of any investigation. These conditions being met, X-ray diffraction has the potential to routinely provide atomic resolution information, allowing atomic models of the structures to be created directly from experimental data; first achieved for TBSV (Harrison *et al.*, 1978). Another challenge associated with X-ray crystallography is the "phase problem", and describes that the success of a X-ray study is dependent on whether the phases associated with the diffraction data can be solved either empirically, or by following a computational strategy.

EM has historically provided lower resolution 2D view of the specimen, but because the sample needn't be crystallized (the sample is viewed freely dispersed in an immobilized state) studies often proceed more quickly, and do not require such a large amount of material. Electrons interact more strongly with matter than do X-ray's, and this means that radiation damage can be a severe problem, but the strength of the interaction means that structural information can be gleaned from single instances of a macromolecular complex. Being radiation-dose limited, EM data suffers from significantly low signal-to-noise ratio; essentially, there is more noise than signal in any given image, a problem that must be overcome computationally. These challenges, which directly limit resolution, have been addressed with technological advances such as improved detectors, improvements in stability of specimen stages, as well as quality of electron sources and electromagnetic lenses. Furthermore, to achieve a 3D representation of the specimen, computational image processing techniques have been developed to carry out "3D reconstruction" of the 2D EM data. Some of these procedures have been applied in this study, and are described further in Section 1.3.5.

1.3.1 Symmetry arising from assembly

1.3.1.1 Virus assembly

The assembly of virus nucleocapsids is a broad topic, but is overviewed here with an emphasis on simple viruses. Viral genomes (by definition) must always encode at least one coat protein gene that is ultimately expressed to form structural units that assemble to form a protective layer around progeny genomes. Packaging of viral nucleic acid generally follows one of two broad

schemes. The first involves self-assembly of the protein and nucleic acid components to form the complete virion (a 'concerted' mechanism of folding and association of protein and nucleic acid), and is a strategy often used by simple viruses. The second involves active packaging of viral DNA using a motor protein complex into a preformed coat protein assembly, often referred to as the procapsid. Examples of viruses which utilize this sequential mechanism include dsDNA phages such as bacteriophage P22 (family *Podoviridae*), as well as the herpesviruses which are built from a $T = 7$ icosahedral architecture and a considerably larger than simple viruses.

Small or relatively simple (+) ssRNA viruses (also ssDNA viruses such as gemini- and parvoviruses) viruses are thought to take advantage of a concerted self-assembly process as these do not require the incorporation of other packaging genes into their genomes (a more economic solution in regard to the size of their genomes). Picornaviruses are somewhat of an exception, and may be considered simple viruses with a more complicated mechanism of packaging and assembly. These viruses have coupled translation and replication of the viral genome with assembly in a mechanism that is yet to be fully elucidated (Nugent *et al.*, 1999). A feature or by-product of this mechanism is that empty capsids containing absolutely no genetic material are also assembled. This is in contrast to other simple viruses e.g. the *Nodaviridae* (ssRNA viruses infecting insects) that cannot form empty capsids. Other viruses such as Hepatitis B virus, are able to assemble empty capsids, but only at high CP concentrations. The *Picornaviridae* and *Nodaviridae* demonstrate auto-catalytic proteolysis of structural proteins in order to "mature" to an infectious form. Remarkably, the RNA genome is thought to participate in these proteolytic reactions (Basavappa *et al.*, 1994).

The concerted strategy of assembly has been revealed to be a multi-stage process. Protein subunits are produced by translation of mRNA, subsequently adopt a more defined conformation or tertiary structure via a process of folding, and aggregate specifically to form assembled multimers. These multimers may be either dimers, as in the case of bacteriophage MS2 (Stockley *et al.*, 2007), CCMV or Hepatitis B virus (Zhou and Standring, 1992), or pentameric (as is thought for Satellite tobacco necrosis virus, STNV (Jones and Liljas, 1984a)). These multimers act as the 'units of assembly' and further interact with other instances of themselves to result in the fully assembled capsid. In cases where the genome is known to interact with the capsid subunits, it is thought to have an important, but poorly understood role in this process of determining the assembly pathway, however, the transient nature of assembly intermediates makes their observation challenging. Techniques such as mass spectroscopy, analytical ultra-centrifugation, and size-exclusion chromatography have been useful when used in conjunction with *in vitro* assembly systems.

Multiple studies on different viruses have shown that a specific region of the genome will nucleate assembly by recruiting coat protein subunits by virtue of a sequence specific recognition event, and that multiple regions throughout the genome may play a role to a greater or lesser degree depending on their overall affinity with the viral coat protein (Borodavka *et al.*, 2012; Ford *et al.*, 2013), observations which have been formulated into the "packaging signal" hypothesis (Stockley *et al.*, 2013). For viruses following the sequential strategy (i.e. more complex viruses with larger genomes), the assembly of the procapsid is thought to be nucleated by the portal complex itself (a protein-protein interaction rather than a specific interaction with the viral genome), and selectivity of packaging follows a sequence specific recognition of the genome by the portal complex itself.

An interesting implication of the packaging mechanism is the density of genetic material that can be incorporated into the closed shell of the capsid. For the dsDNA viruses, the nucleic acid genome does not interact with the coat protein layer strongly, but is packaged to very high pressures (e.g. 100 atm) under force of ATP hydrolysis. For simple viruses the packaging density is considerably lower, as virion assembly requires on spontaneous folding and interaction of both the single-stranded genome and the protein subunits (Speir and Johnson, 2012).

1.3.1.2 Symmetric properties of viral nucleocapsids

Biological macromolecules self-assemble owing to non-covalent interactions that are highly specific. Evolutionary (selective) pressures may induce interacting surfaces to become structurally and energetically complementary, and it is this property that allows for specificity and as a result causes protein or nucleic acids to "fit together" spontaneously to make complex assemblies such as the ribosome.

The ribosome, an example of a biological assembly essential to life, is termed an "asymmetric" structure because each view of the structure is unique. But if identical molecules (e.g. protein subunits) interact with each other in identical ways, the repeated association will results in assemblies (or aggregates) with defined symmetry. An object may be described as possessing symmetry if an operation (also known as an affine transformation) such as rotation about an axis, or translation along an axis, is able to bring the object into "self-coincidence"; i.e. the object appears the same after the operation has been performed. Symmetric objects are often characterized by a number of symmetric axes. For example, a cube has three 4-fold axes, six 2-fold, and four 3-fold axes of rotational symmetry. The total number of symmetry operations corresponds to the overall symmetry of the object, for a cube this would be 12-fold symmetry. The construction of a symmetric object can be described in relation to the asymmetric unit, which is essentially the "local" building block of the object with global symmetry. For an object such as

a cube with 12-fold symmetry, 12 asymmetric units are present, and the cube can be "reconstructed" if the symmetry operations of the cube are applied to a single instance of the asymmetric unit; we might say that the cube displays 12-fold redundancy.

The most basic symmetric assembly is that obeying helical symmetry. Nucleocapsids corresponding to this arrangement are exhibited by some plant viruses (e.g. tobacco mosaic virus or potato virus Y), as well as some (-) ssRNA enveloped viruses that infect mammals (e.g. members of the *Paramyxoviridae* such as measles virus). In such cases self-assembly across two interacting surfaces (interfaces) will result, geometrically speaking, in a displacement along a screw axis. In principle this mode of interaction can be repeated indefinitely, producing a fibre of potentially infinite length. In most cases the viral genome acts as a "ruler" defining the size of the assembly, as protein-nucleic acid interactions play an integral role in the stability of the complex. Helical assembles are described by a number of parameters (e.g. the number of subunits per 360° turn, and the axial displacement between subunits). TMV is considered the classic example of such viruses, producing particles which are ~ 300 nm long with a cylindrical diameter of 18 nm, and is assembled from 2000 coat protein subunits to form right-handed geometry with a pitch of 2.3 nm and exhibiting 16.34 subunits per turn (Namba *et al.*, 1989). The coat protein subunit of TMV consists of a four-helix-bundle, which arranges to form a cylindrical shell with the RNA interacting along a helical path internally, protected from the environment. Isometric viruses assemble using more than two interactions to produce closed isometric (i.e. spherical) shells that exhibit icosahedral symmetry. Geminiviruses capsids assemble according to similar rules to those of spherical viruses that obey strict icosahedral symmetry, and consequently these principles will be discussed in detail in the following sections.

1.3.1.3 Icosahedral symmetry

As already mentioned, the capsids of isometric or spherical viruses possess icosahedral symmetry. An icosahedron is a regular polyhedron that consists of 20 equilateral triangular faces. The structure consists of 30 edges and 12 vertices related by six 5-fold, ten 3-fold and fifteen 2-fold axes of rotational point-group symmetry (Figure 1, a). While the 5-fold axes intercept the vertices, the 3-fold axes pass through the centre of the triangular faces, and the 2-fold axes intercept the edges (Baker *et al.*, 1999). This closed structure is described as possessing 532 point group symmetry with a total of 60 symmetry elements. The 60-fold symmetry of the structure indicates that the asymmetric unit may be described using $1/60^{th}$ of the structure. The dodecahedron is similar to the icosahedron in that it possesses the same symmetry elements, but the structure has complementary properties. The most distinguishing feature is that dodecahedra consist of pentagonal faces, rather than triangular faces.

In the biological context, self-association of the viral coat protein requires that suitable interfaces are presented by each subunit. Even though icosahedral (532) symmetry is common to these assemblies, the nature of the protein-protein contacts may vary, and a main benefit of structural virology has been to characterize these architectural differences. That all spherical virus capsids are icosahedral is significant. Icosahedral shells provide the lowest energy solution to the construction of a closed shell from subunits of the smallest size (Caspar and Klug, 1962). That a complex and large structure such as an isometric capsid can be encoded by a single protein is an efficient solution to the problem of encapsidating a genome which encodes for its own packaging machine (Crick and Watson, 1956). Viruses must construct their protein shells from protein subunits that are asymmetric. The simplest way to achieve this is by assembly of 60 chemically identical protein subunits. Because of the symmetry constraints, each subunit is expected to be in an equivalent environment, stabilized by equivalent interactions with its neighbouring subunits. Because of this essentially crystalline symmetry, Watson and Crick had originally noted that virus capsids might be considered surface crystals. Views of the capsid of STNV are shown in Figure 1-2, as an example of a virus that is constructed from 60 subunits. However, electron microscopy and biochemical studies revealed that virus capsids more often than not consisted of more than 60 protein subunits. How large capsids are constructed in such instances is described in the following section.

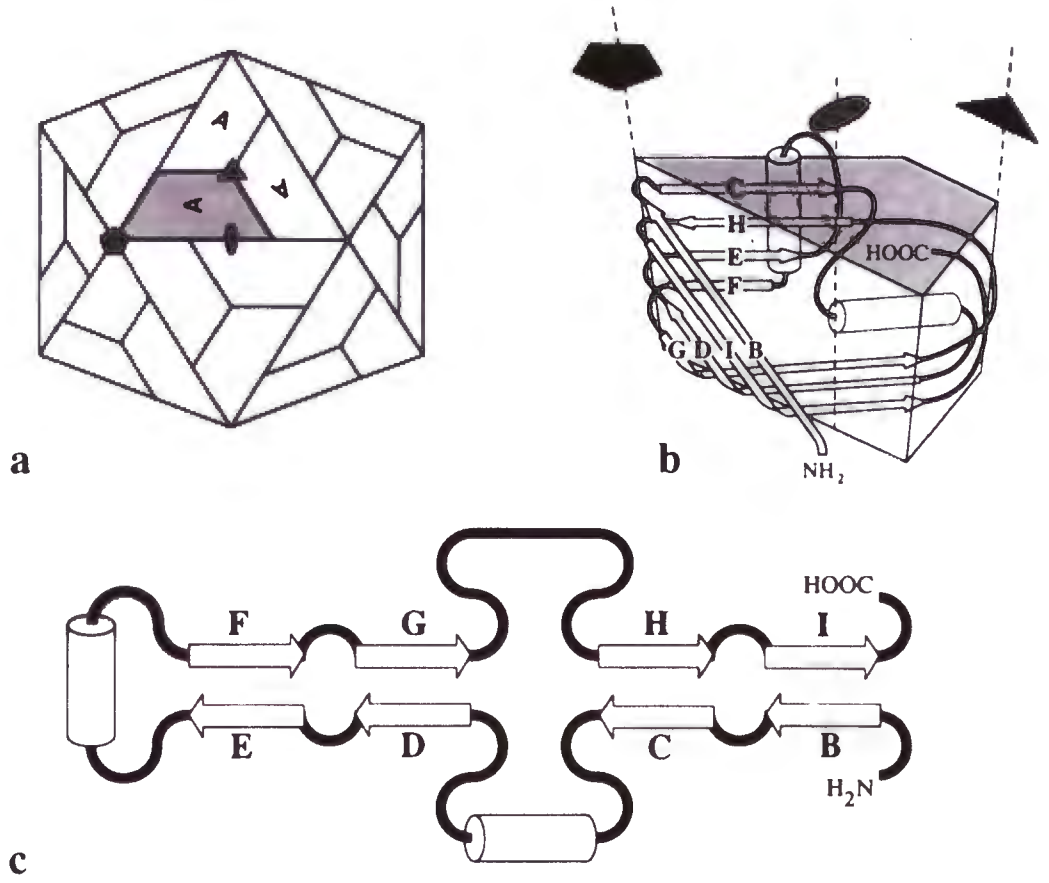

Figure 1-1 Structure and architecture of simple T = 1 icosahedral viruses. a) Schematic of a *T = 1* icosahedral capsid built from 60 units (labelled 'A') in which one asymmetric unit is shaded grey. Axes of 5-, 3-, and 2-fold rotational symmetry are indicated by relevant geometry shape (shaded black). b) The architecture, and c) secondary structure topology of the β-barrel "jelly-roll" fold which is shared by *T = 1* viruses, and prevalent in the structural components of viruses across many lineages including predominantly RNA viruses, but also some DNA viruses is shown by schematic. The jelly-roll fold, also known as the β-sandwich, consists of eight anti-parallel B-strands (B-I) and is wedge shaped, as shown in b). The coat protein subunits that fold to adopt this shape, are typically between 20 and 40 kDa in size, for ssRNA viruses, but are often larger for ssDNA viruses. The viral CPs vary based on the lengths of the N- and C-termini, as well as the length of loop insertions between the β-strand structural elements. Figure reproduced from (Johnson, 1996).

1.3.1.4 Macromolecular quasi-equivalence

The theory of quasi-equivalence (QE) is fundamental to our understanding of the structure of isometric viruses, and has broader implications applicable to our understanding of the functionality of macromolecular assemblies in general (Bamford *et al.*, 2001). QE was developed by Donald Casper and Aaron Klug (Caspar and Klug, 1962) to account for viral capsids which consist of more 60 subunits, and defines the triangulation (T) number to relate the number of alternative modes of packing a coat protein subunit may adopt within a capsid to form icosahedral architectures of different sizes. The consequence of using multiple conformations to construct a closed shell is that the assembled capsid is larger, hence able to package more nucleic acid. Virus coat proteins capable of such behaviour typically have rigid core domains, e.g. the β-barrel (jelly-roll) domain (discussed in the next Section), which interacts via flexible protein segments such as loops. The original prediction of quasi-equivalence was that coat protein subunits adopt different non-symmetry related conformers which preserve bonding patterns but with slightly different geometries, and that it is in principle the flexible regions which allow for quasi-equivalent interactions to take place at all. Experimental observations allowed by crystallography showed that in many cases strictly equivalent interactions are not preserved, but rather that the different conformations allow different sets of interacting residues to meet within in each of the unique environments (Harrison, 1980).

Geometrically, the *T* number also describes how the triangular facets are constructed by sub-triangulation of a hexagonal lattice (wallpaper group p6, refer Figure 1-3). A hexagonal lattice can be folded to create an icosahedron if hexagons are replaced with pentagons at twelve positions. The equation $T = h^2 + hk + k^2$, describes the magnitude of vectors *h* and *k*, which determine the displacement of pentagonal vertices from an arbitrarily specified centre of a hexagon on the hexagonal lattice (refer Figure 1-3). As stated in the previous section, the simplest arrangement consists of 60 subunits occupying identical environments (referred to as $T = 1$, i.e. all CPs existing in the same conformation, and all hexagon positions replaced with pentagons). A $T = 3$ virus, however, will exhibit **3** x 60 = 180 subunits in three alternative conformations (generally termed A, B and C) in its asymmetric unit, and this architecture is exhibited by bacteriophage MS2, CCMV, and Hepatitis B virus.

The theory of quasi-equivalence isn't always followed in nature. An important example is the structural architecture of the 'papovaviruses' (consisting of the *Papillomaviridae* and *Polyomaviridae* families). These dsDNA viruses are assembled from 72 pentameric capsomers (composed of a protein termed VP1), which interact with neighbouring pentamers via flexible C-terminal arms to form $T = 7$ architecture. Instead of being composed of 420 subunits of the VP1

structural protein (as QE would predict), these viruses are in fact composed of only 360 copies. Despite the mathematical disagreement, the basic structural and molecular principals (of quasi-equivalent interactions) still hold. Flexible regions allow for lattice arrangements of structural units to interact in non-equivalent ways, allowing large geometric lattices to be constructed from asymmetric structural units.

Geminiviruses must be another instance in which conformational switching mediates assembly of the virion, although the geminivirus behaviour (being essentially a composite of two $T = 1$ assemblies; i.e. 'bipartite') cannot be expressed in terms of the traditional theory of quasi-equivalence, and warrants further investigation for this reason alone.

1.3.2 Architecture of spherical plant viruses

Many spherical plant and animal viruses share common structural features at both a tertiary and quaternary level. This was first highlighted after the crystal structure of poliovirus (family *Picornaviridae*, infecting animals) was seen to have a very similar tertiary structure to the plant viruses that had been solved by X-ray crystallography (Hogle *et al.*, 1985). At the level of tertiary structure, the coat proteins of these viruses share a fold referred to as the anti-parallel β-barrel, or jelly-roll (Rossmann and Johnson, 1989). The biological prevalence of the β-barrel fold suggested that these viruses might share a common ancestor, and hinted that the great diversity of viruses might be reduced by a classification scheme relying on structural methods (Abrescia *et al.*, 2012; Bamford, 2002; Bamford *et al.*, 2005; Krupovic and Bamford, 2010). Five different viral coat protein folds have been identified, and are reviewed in (Liljas, 2004), however this discussion will focus on the β-barrel fold, as it is the most relevant to our understanding of geminivirus structure and evolution.

The β-barrel fold consists of eight anti-parallel β-strands that form two four-stranded β-sheets that are sandwiched via a hydrophobic core that forms a rigid domain. The β-strands are labelled in alphabetical order from B to I by convention, and typically show some degree of conservation at the sequence level owing to the number of hydrophobic residues that stabilize the core of the domain. The loops between β-strands show high rates of divergence (varying in length and sequence) and are considered more flexible than the core of the domain. For many viruses, this domain will in many cases present a N-terminal "arm" that will usually extend into the core of the assembled virus particle. Numerous positive charges on the arm are thought to mediate stabilizing interactions of the coat protein with regions of the packaged viral genome, although these interactions are poorly understood because the arms are generally disordered and not visible in electron density maps provided by X-ray crystallography or cryo-EM studies.

In the following sections, the architecture of *Satellite tobacco necrosis virus* (STNV), Cowpea chlorotic mosaic virus (CCMV) and other viruses are described in brief detail. STNV is particularly relevant because an apparent evolutionary connection between STNV and MSV CPs that is revealed by sequence alignment (Zhang *et al.*, 2001). CCMV (family *Bromoviridae*) is discussed as an example of a virus with a $T = 3$ architecture with similar (but also evidently different) properties to the geminate architecture.

1.3.2.1 Satellite tobacco necrosis virus (STNV)

STNV is a small isometric virus with a 1200 nt positive-sense single-stranded RNA genome. The virus relies on *Tobacco necrosis virus* (TNV) for replication; and is as a consequence referred to as a "satellite" of TNV. The genome encodes as single coat protein (195 amino acids) that self-assembles to form a $T = 1$ icosahedral shell consisting of 60 subunits of identical conformation (Jones and Liljas, 1984a). The assembly is also able to coordinate 92 calcium ions at interfaces between the coat protein subunits. The coat protein adopts the frequently observed jelly-roll fold, and assembles such that ten amino acids of each N-terminal arm adopt an α-helical conformation that is presented onto the inside of the capsid along the axes of 3-fold symmetry. These α-helices appear to contribute an essential interactions across these axes of the capsid. The tertiary structure, architecture and CP:CP interactions of STNV are depicted in Figure 1-2, while the overall capsid architecture is shown in Figure 1-4. The first ten residues of the STNV CP N-terminal arms are disordered and not visible to structural techniques but nevertheless these present positive charges which are thought to neutralise the charge density of the genome allowing compaction during assembly. Recent studies have suggested that multiple regions in the STNV genome adopt stem-loop secondary structure, and that it is these stem-loops that bind the N-terminal α-helices in a sequence dependent process inducing assembly of the intact virion (Ford *et al.*, 2013).

The STNV capsid coordinates calcium ions at 3 distinct sites that involve oxygen ligands contributed from amino acid carbonyl atoms, carboxylate side charges (of aspartate and glutamate residues), as well as structured water molecules (Jones and Liljas, 1984a). Calcium ions may bind at sites on the 5-fold vertices, 3-fold interfaces, or at a general site which is located at site between two CPs displaced by 7 Å from the 3-fold axes of symmetry. Incorporation of calcium ions into plant virus capsids is commonly observed and thought to be involved in a mechanism which contextually modulates the meta-stability of the virion such that uncoating of the genome (requiring rearrangements and possibly dissociation of the CPs) is possible only once the virions are introduced into uninfected host cells (Durham *et al.*, 1977). A number of simple viruses incorporate metal ions into their capsids, and examples include Tomato bushy stunt virus (Robinson and Harrison, 1982), Turnip crinkle virus

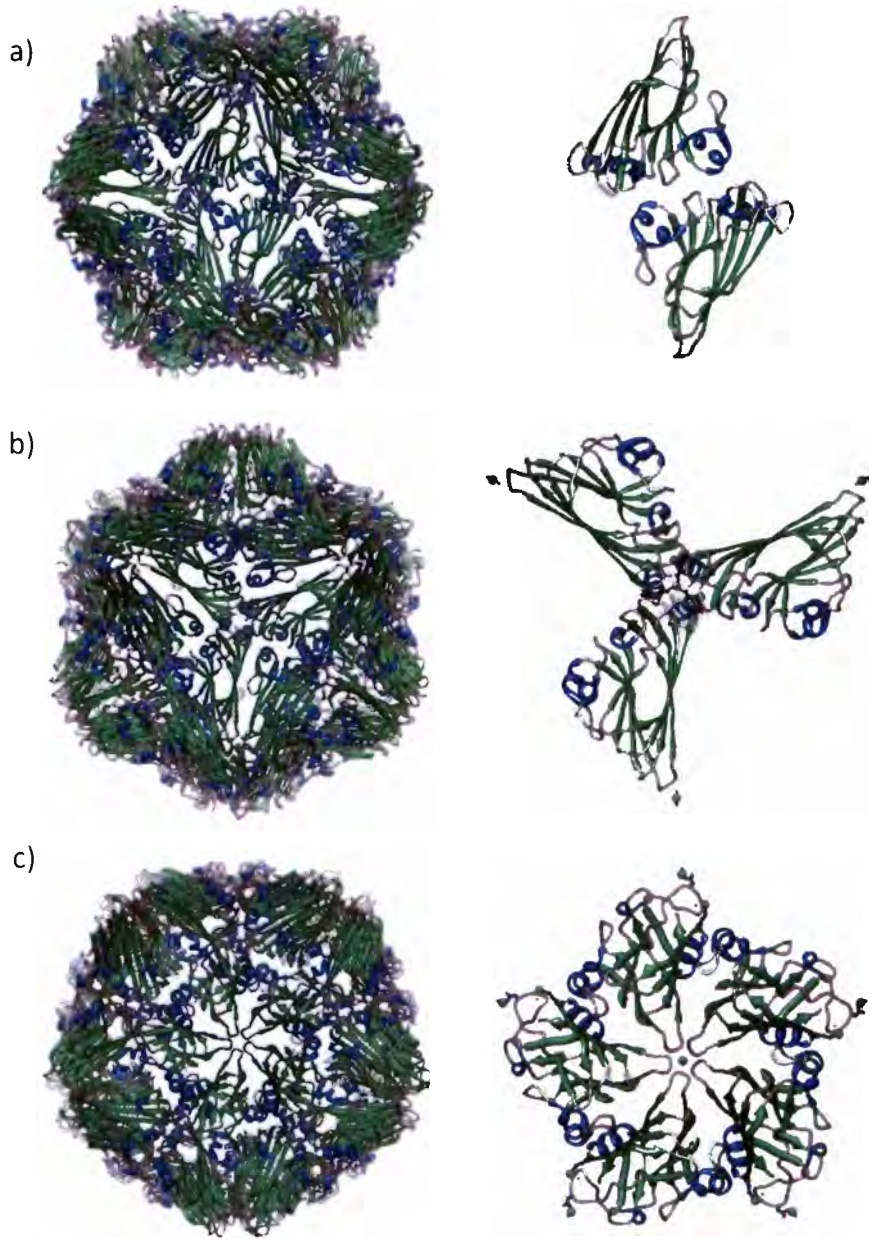

Figure 1-2 Structure and architecture of *Satellite tobacco necrosis virus* (STNV). The STNV capsid is assembled from 60 copies of a 195 amino acid sequence which folds to adopt the B-jellyroll conformation discussed in **Figure 1-1**. The assembly possesses T = 1 icosahedral symmetry, and characteristic views along the 2-fold, 3-fold, and 5-fold axes are shown in a), b) and c) respectively. Association between subunits is made via 3 types of interaction either across (i.e. the 2-fold) or around (i.e. 3- and 5-fold) the symmetry axes. The protein subunits are shown as ribbons with beta-strands coloured green, and α-helices coloured blue. Calcium ions are coloured green, and are coordinated by the assembly at 5-fold and 3-fold symmetry axes. Images rendered using UCSF Chimera (Pettersen *et al.*, 2004)

(Bakker *et al.*, 2012), *Cowpea chlorotic mosaic virus* (Speir *et al.*, 1995) as well as *canine parvovirus* (Chapman and Rossmann, 1995) as just a few examples. Depletion of these ions *in vitro*, may for certain viral species, result in the formation of an expanded capsid isoform (owing to "vibrational breathing"), which is thought to exposes previously buried regions of the coat proteins, inducing sensitivity to proteases, and reduced stabilities allowing subunits to dissociate in prelude for genome egress. This phenomenon has also been referred to as environmentally induced particle expansion (Speir *et al.*, 1995). Such behaviour has been demonstrated for African cassava mosaic virus, in which pentamers were seen to dissociate (Kittelmann and Jeske, 2008) at high pHs. A number of studies have shown that in addition to pH effects (all involving increased pH), chelation of the divalent ions is possible using chelating agents such as EDTA, and particle expansion is also observed under these conditions.

1.3.2.2 Architecture and assembly of ssDNA viruses

Four of the six ssDNA families (*Microviridae*, *Nanoviridae*, *Circoviridae* and *Paravoviridae*) assemble to form a $T = 1$ capsid (Bennett et al., 2008). The *Geminiviridae*, as discussed in detail in this thesis are also able to form $T = 1$ virus-like-particles but predominantly encapsidate full-length genomes into bipartite $T = 1$ capsids, while the Inoviridae, the sixth ssDNA family, are unique amongst viruses of this genome type as they are filamentous. The ssDNA viruses assembling to form $T = 1$ (or derivative architectures) also share the β-barrel CP tertiary structure, but do not show any significant homology at a sequence level. While the typical β-barrel fold architecture serves as a shared core amongst these viruses, the overall shape of each subunit is dramatically altered by very large insertions (added domains with specific functions) that are often found between β-strands; it is for this reason that at intermediate resolutions that the capsids of these viruses appear quite dissimilar. Despite this, important similarities with respect to architecture have been identified, but more investigation is required to identify and compare interactions that make up the CP interfaces (Bennett et al., 2008); hence motiving this study of MSV.

The *Microviridae*, type species ΦX174, possess genomes ~5.4 kb in size and encode 11 gene products. The capsid (270 Å diameter) consists of four different structural proteins, F, G, H and J. The H protein occurs 12 times per virion (McKenna et al., 1992), while the others occur in 60 copies each. Significantly, for these viruses assembly does not proceed via concerted mechanism, but requires involvement of two scaffolding proteins, B and D, that produce a procapsid (Dokland et al., 1997). The *Paravoviridae* (~5 kb genomes), assemble capsids (260 Å diameter) containing a total of 60 copies of three structural proteins, VP1, VP2, and VP3, at a ratio of 1:1:8 (Bleker et al., 2005). The assembly of the *Geminiviridae* is discussed in Section 1.6.3.5.

1.3.2.3 Viruses with T = 3 icosahedral symmetry

Plant and animal viruses with $T = 3$ symmetry are usually ~ 30 nm in diameter and have (+) ssRNA genomes. These viruses use a variety of different strategies to encode and construct their virions, and some of these are described here; these are relevant to this study as $T = 3$ symmetry could be said to represent an "evolutionary alternative" to the geminate architecture, on the basis that the volume available to package nucleic acid is similar. Most $T = 3$ viruses encode coat proteins with the β-barrel fold as the core structural unit of their capsids, although viruses of the *Leviviridae* encode their own unique fold (Valegård *et al.*, 1990).

The *Tombusviridae, Bromoviridae, Nodaviridae,* and *Leviviridae,* are examples of $T = 3$ viruses with capsids assembled from 90 dimers of a single *chemically unique* coat protein. 30 of these are structurally symmetric dimers (conformations C/C), and within the context of the global icosahedral symmetry of the capsid, are found to coincide with the global 2-fold axes of symmetry. The remaining 60 dimers are asymmetric (conformations A/B) and form locations of pseudo-2-fold symmetry between axes of 5- and 3-fold symmetry. The first virus found to adhere broadly, to the theory of quasi-equivalence was Tomato bushy stunt virus (TBSV, *Tombusviridae*). However, while core structural domains were placed within the assembly lattice as predicted by the QE theory, closer inspection of the atomic structure showed that unique sets of residues mediated interactions between the three unique CP conformers; thus not adhering to the fundamental tenets of quasi-equivalence completely. Subsequent study of *Cowpea chlorotic mosaic virus* (CCMV), revealed a geometric arrangement that also agreed closely with QE theory, but satisfyingly, was also the first structure to show that it was possible to form truly quasi-equivalent interactions between conformers. The *Nodaviridae* were to reveal yet another means of adhering to the geometric requires of QE theory. For these viruses the RNA genome has been found to influence the conformation of the assembling dimers, essentially regulating assembly of the $T = 3$ shell by determining where symmetric and asymmetric dimers are placed within the architecture (Cheng *et al.*, 1994). RNA duplexes are observed to associate with the C/C dimers, altering the conformation of and N-terminal arm which in turn effects the geometric arrangement of the homo-dimeric subunits. In the presence of RNA the dimer forms a 'low-curvature' form, while in the absence the N-terminal arm is displaced, allowing subunits to fold to form a high-curvature form. The curvature, or absence thereof, is the 'molecular switch' which determines whether the dimer contributes to a planer hexavalent site, or a curved pentavalent site. The *Picornaviridae* are particularly interesting in that they assemble what is termed a pseudo $T = 3$ capsid from three *chemically distinct* β-barrel domains. These domains are co-translated as a

fusion protein that is subsequently processed to produce 60 copies of VP1, VP2, VP3 and VP4 which assemble to form the capsid structure in a poor understood process that is somehow coupled to replication and translation of the viral genome (Nugent *et al.*, 1999).

The plant infecting *Bromovirdae* are perhaps most relevant to our consideration of the geminiviruses. By electron microscopy this viral capsid was observed to exhibit prominent morphological units with both hexavalent and pentavalent symmetry (Speir *et al.*, 1995). Both N- and C-terminal extensions play and important role in formation of the tertiary structure. C-termini appear to interact to stabilize the dimer association, while N-termini (in addition to interacting with the genome), also interact to stabilize formation of the hexavalent morphological unit (a trimer of dimers); this is similar behaviour to SV40 (refer Section 1.3.1.4). CCMV has 180 metal ions at interfaces, and depletion of these results in formation of an expanded form 10% larger in diameter (Speir *et al.*, 1995).

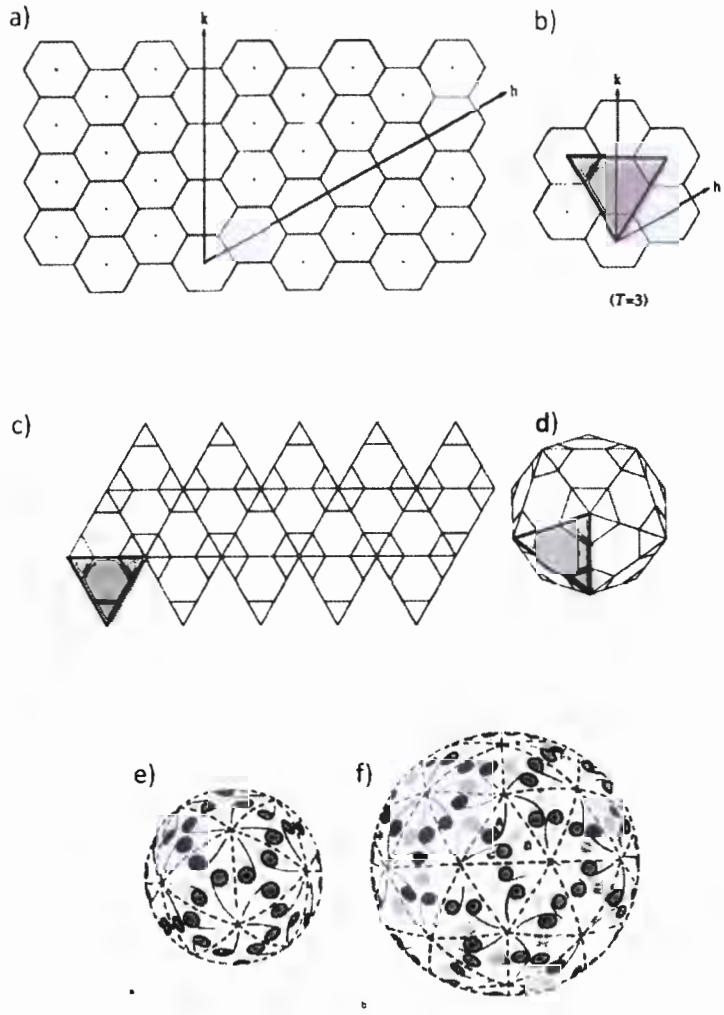

Figure 1-3 Construction of a *T = 3* icosahedron from a hexagonal lattice (p6 wallpaper group) according to principles of quasi-equivalence. The construction of icosahedral assemblies using more than 60 units (60 x T), is described by the theory of quasi-equivalence, which describes how a hexagonal lattice can be sub-triangulated and folded to produce the icosahedron. Sub-triangulation of the icosahedral triangular face, proceeds in a) according to the use of two vectors (h and k) which may originate from the center of an arbitrary hexagon in the lattice. For a T = 3 lattice, both h and k = 1, and hence a equilateral triangular face of the icosahedron may defined by assigning one of the edges of the triangular face to the origin of the vectors h and k, and the remaining two edges at the origins of the neighbouring hexagons specified by the integer values of h and k. In this case a displacement of 1 hexagon along both and h and k applies. As mentioned in the text, an icosahedron consists of 20 such faces in total, and hence the remaining 19 must be defined as shown in c). The outcome is that location of 5-fold vertices is defined at sites where hexagons are made incomplete, but which when folded (d) produce protruding vertices with 5-fold symmetry. Shown in e) and f) are the placement of repeating motifs within *T = 1* and *T = 3* icosahedral lattices, respectively. Whereas only one environment (motif) occurs in the *T = 1* architecture, three occur in the *T = 3* architecture and are shaded appropriately. That different, but similar, environments can exist in such a lattice arrangement is of relevance to this study because similar but distinct behavior is likely to occurred in the twinned architecture of the geminiviruses. Figures reproduced from (Baker et al., 1999) and (Harrison et al., 1996)

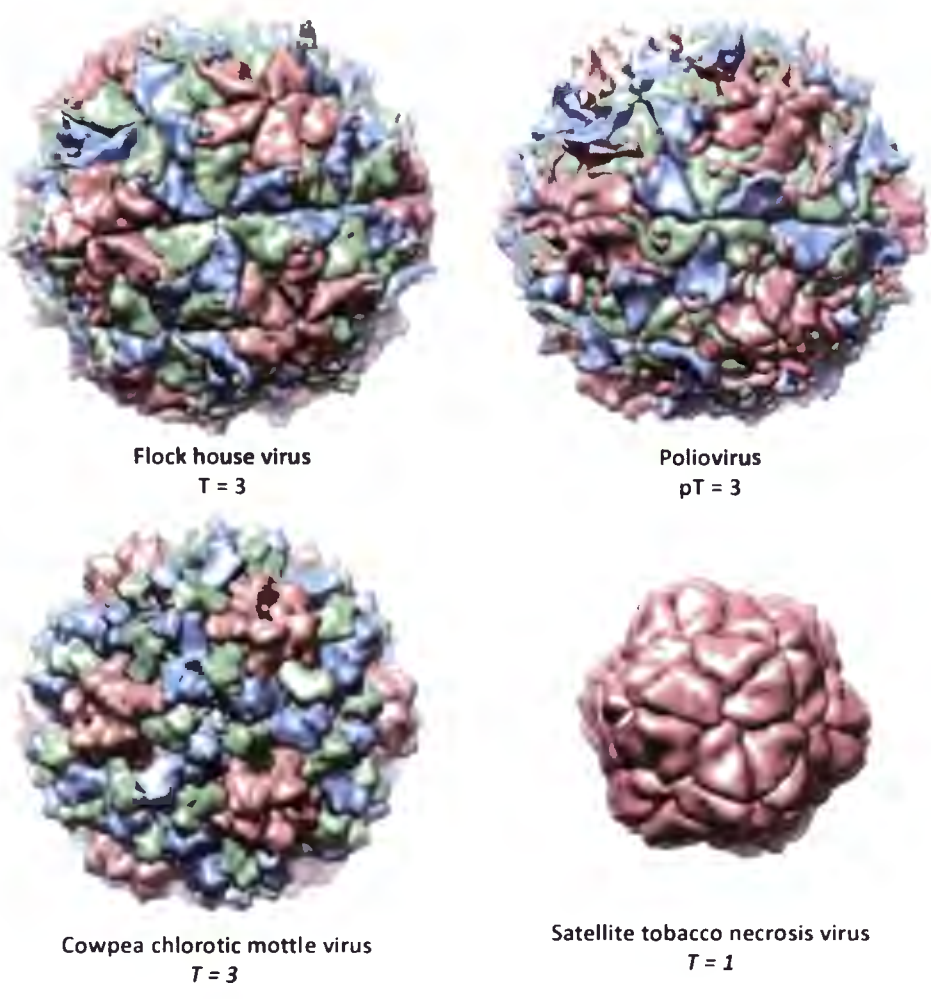

Figure 1-4 Icosahedral symmetry and subunit architectures of some simple viruses. A selection of three *T = 3* 'quasi-symmetric' viruses are shown alongside a *T = 1* symmetric virus. Subunit conformers of the T = 3 viruses are coloured as follows: A – red, B – green, C – blue. These structures are viewed along the 3-fold axis of symmetry. Figures rendered using UCSF Chimera (Pettersen *et al.*, 2004)

1.3.3 Overview of structural methods and physical principles

High-resolution imaging techniques require that the specimen is exposed to an incident beam of radiation (most often X-rays or electrons). The beam will interact with specimen according to physical laws causing changes to both the specimen (radiation damage), as well as the beam of radiation itself. Changes in the beam will encode information about the specimen, and consequently high-energy (or short wavelength) radiation must be used to encode information about the fine features that characterise macromolecules (e.g. inter-atomic bond lengths). A beam may have different properties such as coherence (all quanta travelling in same direction and phase) or monochromaticity (all quanta having the same energy) that will strongly influence the quality of information retrieved during this process. The potential consequences of the interaction of a high-energy particle with an atom are numerous, however, the most sought after is referred to as elastic scattering, in which the path of the incident particle is altered by the specimen without loss of energy. Inelastic scattering will cause transfer of energy into the specimen, causing radiation damage – an inevitable consequence when imaging with beams of radiation (Henderson, 1995).

From a mathematical perspective, scattering has the effect of physically calculating the Fourier transform of an object, and a benefit of this representation is that repeating patterns (as for symmetric macromolecular complexes or crystals of these) become evident. The reason for this is that the Fourier transform expresses a given signal as a sum of waves; waves are mathematical objects consisting of *frequency*, amplitude and phase. If the amplitudes and phases of a particular structure can be recorded over a useful range of spatial frequencies (on nanometre scales, over a resolution band corresponding to atomic or macromolecular resolutions), these can be used to calculate the structure of a macromolecule of interest at a defined resolution. Numerous challenges prevent this information from being recorded, and these include the nature of interaction of the radiation with the specimen, the quality of the specimen as well as the performance of the detector.

Even prior to the imaging experiment, the specimen of interest must be prepared for imaging. This most often involves isolating the macromolecular complex (or virus) from the other components of the biological cell, and these considerations are discussed in the context of virological studies in the following section.

1.3.3.1 Virus purification and sample preparation

Prior to structural characterization, virus particles must be harvested and enriched from a host organism. Bacteriophages can be grown using prokaryotic hosts, usually to very high yields. For animal and insect viruses these are typically grown in tissue culture; or for some viruses chicken

eggs can be used. However, in the case of plant viruses these must usually be isolated from infected plant tissue (often leaves), or in some cases expressed in insect tissue culture systems.

Viruses are most often isolated using differential gradient centrifugation. Density gradients using either sucrose or caesium salts (e.g. caesium chloride), result in banding of viruses (and contaminants) at their buoyant densities, i.e. where their density matches that of the surrounding buffered medium (also known as isopycnic density gradient centrifugation). The virus particles will usually have a density different to that of the contaminating materials, and hence the density band containing the virus can be taken further, and usually subjected to another purification step. Alternatively the virus particles may be large enough that they can be pelleted out of solution either onto a "cushion" of sucrose, or if they are stable enough they can be spun directly onto the bottom surface of centrifugation tube. Following this, viruses are suspended and dialysed into a pH buffered solution that is suitable for structural experiments. Such buffers typically have low concentrations of salt as this can interfere with visualization using methods such as cryo-EM, as the buffer will produce a background that reduces the signal-to-noise ratio of the images, especially at higher resolutions.

1.3.3.2 X-ray crystallography and fibre diffraction

Structural virology came into its own with the structure determination of the capsid proteins of two plant viruses: TBSV, and TMV by research groups of Stephen Harrison, and Aaron Klug (Bloomer *et al.*, 1978; Harrison *et al.*, 1978). These were high-resolution structures (2.5 – 3.0 Å) that allowed the amino-acid chain of the viral nucleocapsid proteins to be modelled in detail that included side-chain and polypeptide backbone interactions.

X-ray diffraction relies on the order of a macromolecular crystal to produce a diffraction pattern when it is irradiated with a monochromatic beam of high-energy photons (X-rays, e.g. of a wavelength of 0.9 Å). A crystal consists of a series unit cells, all identical in size and composition. It is this order that causes diffraction to occur, as many scattering events interfere either constructively or destructively to produce the diffraction pattern. The size and order of the crystal relates to the intensity of the signal across a range of scattering angles, generally necessitating that the largest possible crystals are used for this process. Diffraction intensities (which are the recorded experimental data) can be combined to create an *in silico* representation of the electron density, by inverse Fourier transform. However, the X-ray crystallography experiment does not record phase information, and hence separate methods have been developed to overcome what is known as the "phase problem". Different experimental methods have been developed to acquire this information such as the original method of multiple isomorphous replacement, in which heavy atoms are soaked into crystals of identical space group (Kendrew *et al.*, 1958; Rossmann,

1960), or multiple-wavelength anomalous dispersion (MAD) which relies on recording a diffraction pattern of a crystal using a wavelength near the absorption edge of an atom such as selenium (introduced into the proteins as selenomethionine) to derive initial phase information which can then be applied to the remainder of the data (Hendrickson *et al.*, 1990). Computational methods for phase determination have also been developed, and these are referred to as "molecular placement" (Lattman, 1985; Rossmann and Blow, 1962). Structurally homologous proteins (of close tertiary structure) can be used to calculate models of the crystallographic unit cell, and the phase information of the model applied to the experimental data of the macromolecular complex under study. In such cases, great care must be taken to not bias the result with the original model. In the case of viruses, the icosahedral symmetry of the specimen (e.g. the crystallized virions) has aided "phase extension" to higher resolution by a process such as iterative non-crystallographic symmetry averaging (Stuart *et al.*, 1999). The orientation of virus particles within each unit cell can be assessed using self-rotation functions that reveal the location and orientation of 5-fold symmetry axes (repeating every 72°) of each virion, and icosahedral symmetry is now routinely exploited by X-ray crystallographers to improve the resolution and quality of electron density maps. In the case of STNV, non-crystallographic symmetry allowed the resolution to be extended to 2.5 Å (Jones and Liljas, 1984b), allowing the side-chains of the almost all of the amino acids to be modelled with greatly improved accuracy in comparison to the original determination to 3.8 Å.

Modern day X-ray crystallography has benefited from the development of technologies such as synchrotron X-ray sources and cryo-preservation methods that overcome radiation damage allowing more data to be collected from fewer crystals. Improvements in detector design, have improved the quality of data recorded, as well as reducing the time required for data collection immensely. Improved software and computing power has also improved the speed that structures can be solved and modelled.

1.3.3.3 Transmission electron microscopy and 3D reconstruction

The transmission electron microscope (TEM) consists of an electron gun (built from an electron emitting source, as well as a high-voltage extractor) that fires electrons down an evacuated column of electromagnetic lenses. Electrons pass through a specimen, and are focused and magnified onto a detector (usually either film, CCD camera). The electromagnetic lenses are numerous, but two primary systems of lenses exist, the illumination system consisting of condenser lenses, and the image formation system consisting of an objective lens, as well as projection lenses that will magnify the image produced by the objective lens. Electron sources may be either thermionic, such as tungsten filament or lanthanum hexaboride crystal (LaB_6), or field emission gun (FEG)

sources. These vary in the quality of the electron beam produced as well as brightness. LaB_6 is an order of magnitude brighter than the tungsten filament, and because the source is physically smaller, also allows higher resolution. FEG sources offer a very bright, but also coherent source of electrons (also approaching monochromacity), which is most suitable for high-resolution phase contrast imaging.

Transmission electron microscopy has allowed many viruses, particularly larger viruses, to be studied without the requirement of crystallization, but original studies were often of considerably lower resolution (15-30 Å), this owing to the specimen preparation techniques used, as well as the capabilities of the instrumentation at the time. Indeed, in cases where sufficient quantities of the virus are difficult to produce (e.g. grams) TEM is able to provide valuable information, because the sample concentration need only be a fraction of that typically required for crystallization experiments (milligram vs gram concentrations). For viruses that are difficult to grow in culture, these benefits are attractive, as was the case for work described in this thesis. More recently, with the advent of improve microscopes and detectors, routine structure determination of viruses and other macromolecular machines such as the ribosome and proteosome has progressed from sub-nanometre resolutions (6-8 Å) to near-atomic resolutions (3-4 Å) (Bai *et al.*, 2013; Li *et al.*, 2013).

Fundamentally, transmission of electrons through a specimen, allows nanoscale structures to be viewed in projection, i.e. in a manner similar to how a "shadow" is cast, but in which all structural information is superimposed onto the plane of the image, and in which the image is quantitatively representative of the structures entire mass. Figure 1-5 illustrates this concept, and reveals that from a mathematical perspective, the Fourier transform of a 2D projection of 3D structure will represent a central section (or plane passing through the origin) of the 3D Fourier transform of the object; a fact that has important implications for 3D reconstruction. The relatively high depth-of-field of the TEM (~ 1 μm), in comparison to the typical light microscope, means that the entire specimen is imaged at the same focus; which is important for image processing and 3D reconstruction described below and in Section 1.4.2. The primary limitation of TEM is specimen thickness, and a variety of preparation methods have been developed for this reason. Electrons tend to interact with matter more strongly than X-rays do, and consequently multiple scattering events are likely to occur for thick specimens (> 300 nm deep). The probability of an electron interacting with an atom is influenced strongly by the kinetic energy of the electron as well as the charge density (determined by the atomic number) of the specimen atom itself. An incident electron may interact either elastically or inelastically with a specimen. Elastic interactions preserve the kinetic energy of the electron, but change its direction (phase). These interactions are

desirable for structural studies because they result in contribution of high-resolution information during image formation. Inelastic interactions involve a transfer of energy from an incident electron to an atom of the specimen – damaging it. These electrons are scattered at low angles, and owing to their different energies are poorly focused during image formation resulting in additional background noise. Energy filters have been developed, which work essentially as electromagnetic prisms, and allow the microscopist to exclude electrons outside of a defined energy range.

High-energy electrons such as those used for imaging in TEMs are a form of ionizing radiation, which can disrupt molecular bonds, and induce formation of free radicals which have the potential to further damage the atomic structure of a biological specimen. For this reason, strict dose limits are applied for data-collection of high-resolution information, as it is these features that are disrupted first.

Electron microscopy does not suffer from the phase problem (as does X-ray diffraction), because the charged nature of electrons allows them to be focused or recombined to form an image. While projection to form a 2D image is useful for structure investigation, projection also introduces significant structural ambiguity in images that complicates analysis and interpretation of the image by the investigator. One drawback of projection is that information encoding the handedness of the object in question is lost. However, comparison with other information such as crystallized structural components, or even more recently scanning electron microscopy (SEM) studies of unidirectional shadowed virus particles can provide the information required to assign the correct hand to a 3D reconstruction (Miller *et al.*, 2011; Woodward *et al.*, 2009). Richard Henderson (MRC-LMB, UK) and co-workers have suggested a procedure "tilt-pair analysis", that ensures accurate hand is applied to single-particle EM structure by comparison of the reconstruction with data recorded with tilt, and is a procedure that can be routinely applied as part of the EM data-collection strategy (Rosenthal and Henderson, 2003).

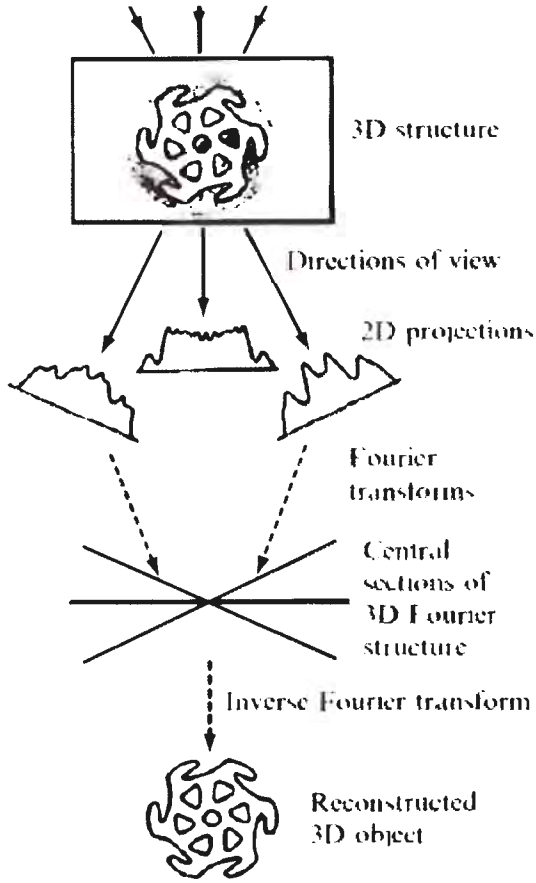

Figure 1-5 General principles of 3D reconstruction of an object from 2D projection images. The 3D structure under study, in this case an object with C6 rotational symmetry is observed under experimental conditions over a certain angular range to produce a set of 2D projections (the experimental data). The object can be reconstructed by calculating the Fourier transform of each image, and orienting the projections in 3D reciprocal space. Each experimental image corresponds to a central section of the objects 3D Fourier transform. Once the data have been combined in reciprocal space with adequate angular sampling, the reciprocal space volume can be inverse Fourier transformed to generate a real space representation of the object that can be analyzed further *in silico* to gain biological insight. Figure reproduced from DeRosier and Klug (1968)

1.3.4 TEM image formation, correction and resolution

Owing to the high kinetic energy of the electron beam (120-300 kV), very little adsorption of electrons by the specimen is expected, and this results in low image contrast. Hence methods of introducing contrast into electron micrographs must be used, especially for unstained specimens. As discussed above, interactions of two primary types (elastic and inelastic) occur between the electron beam and the specimen, and these result in changes in the energy and direction of electrons. It is important to note that how these changes occur depends on the atomic composition of the specimen, and for biological specimens the precise nature of how they are prepared. For thin specimens (10-30 nm) prepared by negative-staining, the heavy metal atoms (e.g. tungsten or uranyl) form an envelope which surrounds the structure of interest. The heavy atoms are strongly scattering, and deflect electrons at high angles with relatively high probability. In such cases amplitude contrast can be introduced into the images by inserting an aperture into the back-focal plane of the objective lens (referred to as the objective aperture), that then excludes the high-angle electrons from image formation.

At the electron energies (or wavelengths) used for TEM, these instruments can in principal yield sub-Angstrom resolution images of materials. Unfortunately it is the aberrations of the objective lens of the microscope, as well as radiation sensitivity of biological specimens that will limit the resolution of images in the context of biological microscopy. For unstained and frozen specimens, high-resolution imaging is considerably more challenging. For these specimens that consist primarily of carbon, nitrogen, phosphate and oxygen, the strength with which electrons are scattered is relatively low (modelled mathematically as the weak-phase approximation (Wade, 1992), relating that only a small fraction of the incident beam is actually scattered), and amplitude contrast provides no high-resolution information about the specimen at a sufficiently high signal-to-noise ratio. In such cases, the only alternative is to convert the weak phase shifts experienced by the electron beam into an amplitude varying signal that can be detected. The most commonly used approach is to defocus the images (i.e. to focus the objective lens such that the image-plane of the objective is shifted slightly below that of the detector). This procedure introduces phase-contrast by taking advantage of spherical aberration of the objective lens. This behaviour of the objective lens is well understood, and is described by the TEM microscope phase contrast transfer function or CTF (Erickson and Klug, 1970; Thon, 1971) and is depicted in Figure 1-6. The CTF can be considered the Fourier transform of the microscope point spread function (PSF), which models the blurring of a point source during imaging, and provides indication of the optical resolution of the microscope. By defocusing the microscope, the oscillations of the microscope CTF are increased, and these in turn introduce phase inversions and amplitude modulations into the signal of the specimen causing improved contrast to arise in the image. The reason for this is

that at either underfocus or overfocus, a phase contrast image will be produced by interference of scattered (phase "aberrant") electrons with the unscattered beam (Erickson and Klug, 1970). The further from focus the image is shifted, the more contrast is introduced, but at the cost of resolution as a separate "envelope" function, also arising from aberration, dampens the amplitudes of high-resolution information. Additionally, the sinusoidal oscillations of the CTF result in varying levels of contrast transfer with certain regions of no transfer (referred to herein as the "zeroes"). At these regions, and at frequencies adjacent, the signal-to-noise ratio is either zero or close to zero and consequently the full spectrum of information about the specimen is incompletely sampled.

These properties of the CTF influence how data for single-particle reconstructions is captured. First, data must be recorded over a defocus range (or spread, e.g. 1 – 3.5 μm underfocus) to ensure that the Fourier transform of the objects view is adequately sampled across the full spectrum of spatial frequencies. Second, and as a component of data processing, procedures for CTF correction must be applied. The phase inversions useful for "seeing" the specimen, have the effect of corrupting its true appearance. One approach would be to band-pass filter the images in reciprocal space so that spatial frequencies higher than the first CTF cross-over (zero) do not contribute to the images (i.e. all structure factors with corrupted phases removed). While valid, this has the effect of limiting the resolution of the reconstruction to resolutions typically > 1.5 nm (15 Å), and consequently methods for correcting these phase inversions have been developed. The first method is known as phase flipping which involves an inversion of phases only in the inverted spatial frequency bands. The second method is termed the Weiner filter, and involves phase inversion as well as amplitude adjustments at higher spatial frequencies. In addition to the procedure for correcting phases, a method for assessing the defocus parameters of the data is another challenging aspect to the resolution of this problem. Semi-automated methods which involve fitting model CTF functions to the data have been developed (Fernández *et al.*, 2006; Huang *et al.*, 2003; Zhou *et al.*, 1996), as well as more automated methods which have allowed magnification, defocus and astigmatism parameters to be modelled on noisy cryo-EM data without significant user intervention (Mindell and Grigorieff, 2003).

Resolution assessment is a complex (and highly debated) topic in the field of high-resolution EM of macromolecular complexes. For optical instruments, the term "resolution" relates the minimum distance between which two points can be distinctly identified (referred to as the Rayleigh criterion). However, this method is somewhat subjective, and when noisy images are being considered (as for dose-limited specimens), resolution becomes difficult to assess reliably. In the case of single-particle EM, in which macromolecular assembles are imaged individually and in the presence of very high-levels of noise, other means of estimating resolution have been

developed which specifically take the noise into account. These methods measure the "internal consistency" of the data by means of cross-validation. The most routinely applied of these is the Fourier shell correlation (FSC), which reveals the correlation between two images (2D, or 3D) as a function of spatial frequency (or resolution) in which a series of bands; rings for 2D or "shells" for 3D data are compared (Harauz and van Heel, 1986). A figure for resolution is assigned by inspecting the curve for the resolution at which the correlation drops below a certain value (i.e. a correlation "threshold"). The most popular of these is FSC 0.5 criterion (Bottcher *et al.*, 1997; Malhotra and Harvey, 1994; van Heel and Schatz, 2005), although some have argued that a correlation thresholds of 0.3 or 0.143 may have merit because by separating the data into half-sets, these procedures will inherently underestimate the resolution of the full dataset (Rosenthal and Henderson, 2003). It is generally agreed upon that any assigned value for the resolution (e.g. 8 Å) should be supported by observable features in the reconstruction, e.g. tubular densities corresponding with α-helical protein secondary structure.

1.3.5 Three-dimensional electron microscopy of viruses

A virus specimen prepared for EM can be reconstructed using a number of approaches. These are: single-particle image processing, electron tomography, and electron tomography combined with sub-tomographic averaging. The details of these are described in Figure 1-7. These approaches all rely on the central-section theorem but involve different data-collection strategies. The ultimate determinant of which is used is the size and morphology of the virus (whether regular and symmetric: single-particle, or pleiomorphic: tomography), and precisely which structural aspect of the virus is under study; for instance whether the whole of virus, or just a single structural component of this is under study. Tomographic approaches are most useful for specimens that are structurally unique (pleiomoprhic viruses, or when studying viruses inside cells) or when a macromolecular complex presents only a limit angular distribution of views when visualized by cryo-EM. Tomography is limited by the overall dose that can be applied to the specimen, as this must be fractionated across a wide angular range (e.g. ± 65°) and tomograms will suffer from radiation damage as well as low contrast. By directly visualizing cell entry of herpesvirus in 3D, cryo-tomographic methods have informed us of important aspects of the life-cycle of this virus at molecular resolution (Fu and Johnson, 2011; Grunewald and Cyrklaff, 2006). Sub-tomographic averaging has improved our understanding of viral glycoprotein complexes (Huiskonen *et al.*, 2009; White *et al.*, 2010), as well as the entry mechanisms and structural intermediates of a number of viruses (Bostina *et al.*, 2011; Bubeck *et al.*, 2005; Chang *et al.*, 2011).

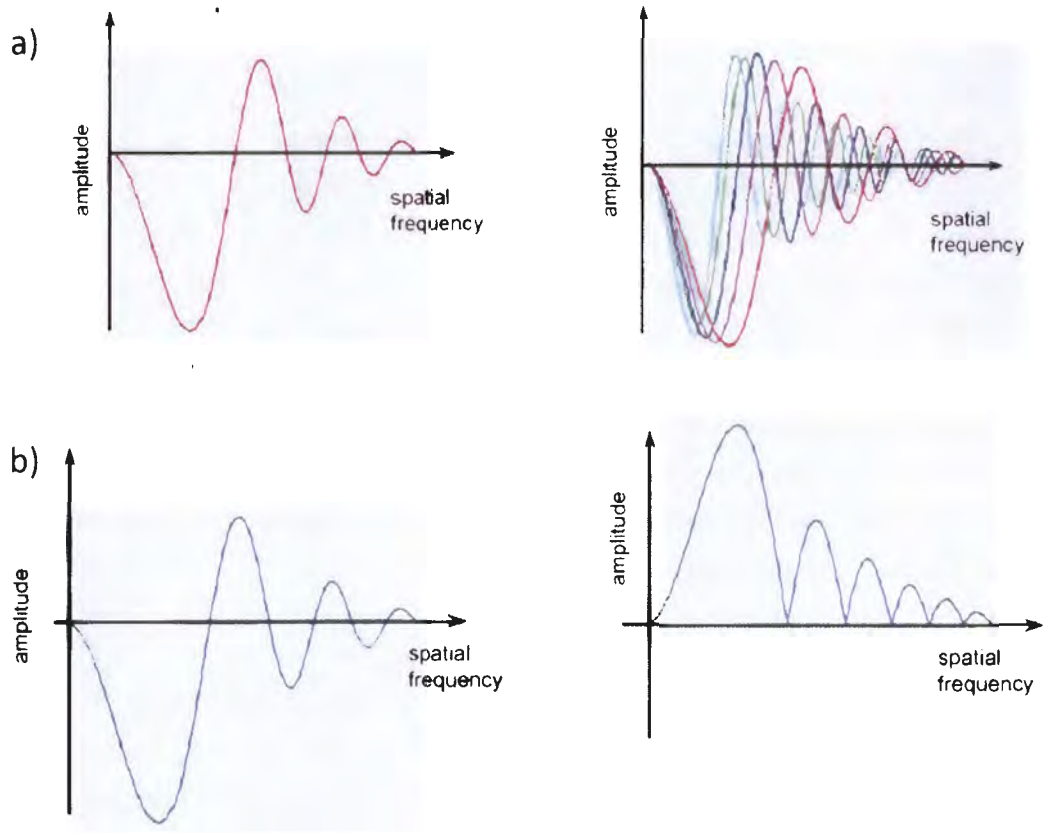

Figure 1-6 The microscope contrast transfer function (CTF). a, left) A typical microscope CTF when an image is recorded at a particular defocus value. The oscillating function implies phase inversions in the recorded image and variations in the amount of contrast transferred as a function of spatial frequency. The phase inversions result in useful contrast in the images, however these inversions essentially represent a corruption of the signal. Where the function crosses the horizontal axes, no information is recorded in the images, and the SNR ratio of information in regions immediately adjacent to these regions (the "zeroes" of the CTF) are characterized are very low. a, right) These properties of the CTF necessitate that images are recorded using multiple defocuses, such that missing information in Fourier space is filled in. b) Phase inversions can be corrected (and true contrast restored) if the CTF parameters can be determined. b, right) The appearance of the CTF after 'phase-flipping' has been applied; a procedure for CTF correction which restores the phases in corrupted areas, but does not attempt to scale the amplitudes to compensate for the microscopes attenuating 'envelope' function (Orlova and Saibil, 2011)

Single-particle methods, as used for the work described in this thesis, are best applied to symmetric specimens and achieve high resolutions because dose is not limited as many instances of an identical structure are merged to produce a reconstruction.

1.4 Single-particle electron microscopy

Single-particle electron microscopy uses one of two primary specimen preparation methodologies to suspend a sample containing a macromolecular complex of interest over a plane, and these can then be imaged conveniently by TEM; ultimately for computational image processing. The methods for specimen preparation are known as negative-stain (NS) EM (Brenner and Horne, 1959), and electron cryo-microscopy (cryo-EM) of thin aqueous layers (Adrian *et al.*, 1984; Dubochet *et al.*, 1988). NS requires that macromolecular complexes are adsorbed to a thin carbon film, enveloped in a heavy metal stain and then dehydrated, while in cryo-EM the complexes are imaged unstained, preserved in a thin-layer of vitreous ice. The result of both methods is that multiple instances of the complex can be viewed in different orientations at approximately the same magnification and focus, and ensures that the fragile biological material is rendered resistant to the high vacuum environment of the EM column, as well as the ionizing effects of the electron beam used for imaging.

Single-particle electron cryo-microscopy (cryo-EM) allows the solution structure of a macromolecular complex to be visualized at sub-nanometer resolution (< 1 nm). A significant benefit of cryo-EM is that crystallization conditions do not need to be found for the investigation to proceed, and the complex is visualized outside a crystal lattice. However, high resolution cryo-EM studies will often require that sample preparation conditions such as buffer pH, ionic strength, or the addition of detergents are optimized to achieve an adequate spread of particles within the ice layer, and at the highest sample concentration reasonably possible. NS EM is known to introduce structural heterogeneity into the dataset due to distortions that are undergone by the structure during sample preparation. Larger macromolecular assemblies (such as viruses) are usually more strongly affected by these than are smaller assemblies. In some cases actual conformational changes may also be introduced in addition to distortion effects caused by compression (or flattening) under dehydration of the sample. Cryo-EM allows biological samples to be visualized in a hydrated state. The benefits of this are the preservation of the sample in a close to native state, whilst also providing the potential to record near atomic resolution information of the sample itself, and not a stain envelope.

During cryo-EM of vitreous aqueous films, the film thickness is typically 100 nm thick, and it is generally considered undesirable for the thickness to increase unless the dimensions of the object warrant thicker ice. Sample preparation will impose a limit on the resolution of the reconstruction

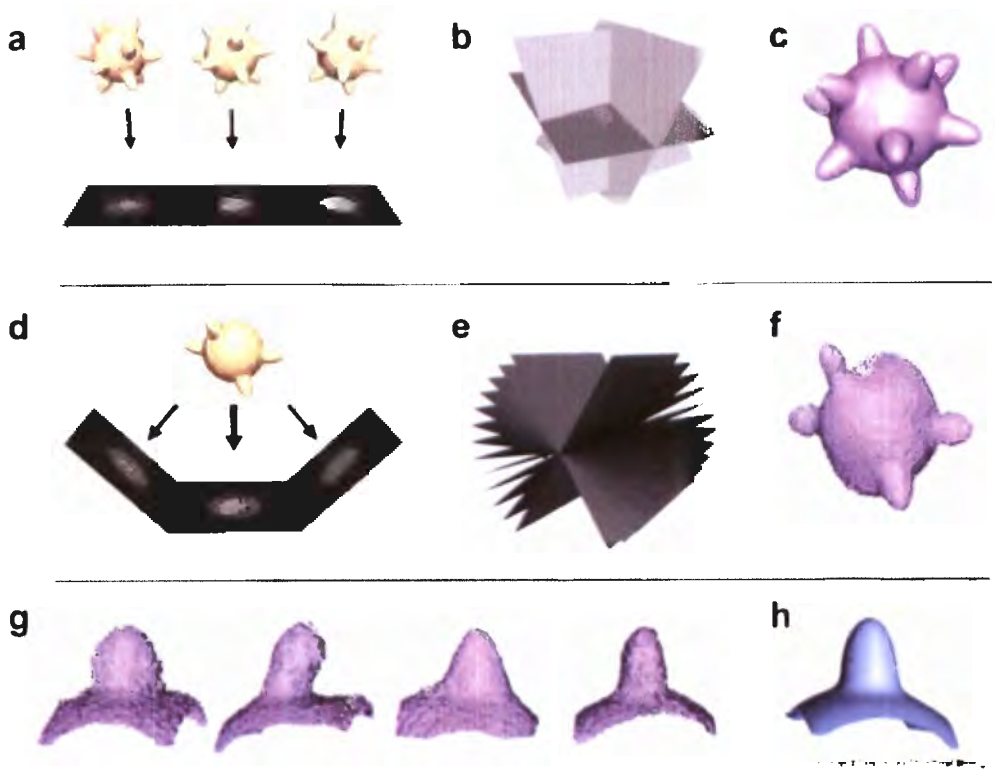

Figure 1-7 3D reconstruction of viruses and their components from electron microscopy data. Regular viruses (possessing icosahedral symmetry) are most often reconstructed using single-particle approaches with imposition of 532 point group symmetry. In this approach the viruses are a) visualized and low signal-to-noise images recorded on either film or CCD camera. The images correspond to projection images, which according to the projection theorem, correspond to central sections through the objects Fourier transform as shown in b). By determining the orientation parameters of each image, the central sections can be accurately oriented in 3D reciprocal space and a 3D reconstruction calculated with imposition of 532 symmetry. The 532 symmetry of an icosahedral particle implies a 60-fold redundancy of information in the data at hand, and averaging which takes this symmetry into account can result in large improvements in SNR and resolution (sub-nanometer resolutions are now often routinely achieved). Some viruses do not have icosahedral capsids (e.g. pleiomorphic viruses) or contain structural features that do not coincide with the symmetry of the protein shell. In such cases electron cryo-tomography (cryo-ET) is used. Cryo-ET involves d) recording a tilt-series around a single tilt-axis. The images of known angular relationship are then oriented in reciprocal space (e) as for single-particle reconstruction, however, a missing wedge of information is present owing to physical limitations associated with imaging a planar specimen at high tilt angles (i.e. a doubling of specimen thickness at 60°). 3D reconstruction to produce a tomogram (f), will yield a low SNR reconstruction of the object which will suffer from missing wedge artefacts (considered most simply as a smearing along the direction of the electron beam, or more technically, as a result of anisotropic resolution as a consequence of a large amount of missing information). g) Structural features that repeat along the surface of the virus, can be excised from the tomogram and averaged to produce a sub-tomographic average (h) which has improved resolution and SNR. Because the repeating features are visualized in multiple different orientations, sub-tomographic averaging can overcome the missing wedge allowing molecular resolutions (2-4 nm) to be achieved. Figure reproduced from (Subramaniam *et al.*, 2007).

in other ways. The resolution limit of a negative stain reconstruction is the grain size of the heavy metal stain, while considerations pertinent to cryo-EM are the thickness of the vitreous ice as well as the size of the complex under study. Thicker ice will result in reduced contrast (especially at higher resolutions), and a greater distribution of particles along the Z-axis of the microscope resulting in particles imaged at different focii and magnifications. The structure of complexes ≥ 250 kDa can be confidently studied using single-particle methods, however, the larger a complex is the more contrast it will produce. Contrast is particularly relevant, as the performance of computational alignment algorithms is heavily influenced by this property of the data, and large symmetric viruses typically produce the highest resolution 3D structures for this reason (Grigorieff and Harrison, 2011).

1.4.1 Vitrification of thin aqueous layers

Vitrification is a process in which an aqueous specimen (e.g. viruses prepared in a buffered solution) are cooled so rapidly that the water molecules freeze prior to adopting a crystalline organization (e.g. either cubic or hexagonal ice). Viruses suspended in such a solution will be preserved in near-native state, immobilized in a layer of vitreous ice with their hydration shell, and macromolecular conformations unaltered. Freezing is carried out by plunging the specimen into liquid ethane cooled by liquid nitrogen (~ -180°C). For single-particle studies, a droplet of the specimen will be applied to a perforated carbon film, and blotted so that a small fraction of the suspension remains on the film as a thin-layer held across the perforations by virtue of surface tension. Such layers may be 50-200 nm thick, and are plunged immediately, and in a high humidity environment, in an attempt to minimize evaporation that might result in concentration of buffer components that might reduce contrast in the EM. Upon rapid entry of the thin-specimen into the liquid ethane, the temperature of the specimen is thought to decrease at ~10,000-1,000,000°C/s; ensuring vitrification. Because the macromolecular complexes (virions) are freely dispersed they should be randomly oriented within the thin- layer. This is required for 3D reconstruction because it ensures good angular coverage within the dataset. However, this may not always be the case owing to interactions with the air-water interface, or a carbon support, which if present may adsorb the particles, restricting their orientations in a manner similar to that of negative-stain EM. This is the case for the work described in this study.

1.4.2 Three-dimensional electron microscopy

Electron microscopy of biological samples, particularly at the high magnifications afforded by EM, is characterized by a noisy visualization of the specimen and much of image processing is intended to overcome this problem. Low-dose techniques attempt to reduce radiation damage of the specimen, however the signal-to-noise ratio of such images is extremely low and further

worsened by noise arising during detection (especially in the case of CCD cameras). Other unintended processes such as specimen charging and beam induced movement of the specimen, further worsen the quality of recorded images. 3D EM techniques have developed complex procedures that attempt to overcome these limitations in data-collection.

For the single-particle EM experiment, the recorded electron micrograph will typically contain images of the structure of interest in a number of random orientations, and as result, reconstruction from a large set of images is essentially a problem of 'grouping' similar views of the structure by classification, and then determining the three Euler angles - phi (ϕ), theta (θ) and psi (ψ) - that describe how each image is oriented in relation to a common 3D coordinate system.

1.4.2.1 2D alignment and classification

The signal-to-noise ratio (SNR) of the data is usually improved by averaging images containing the same information (identical views), and 2D image alignment algorithms based on statistical comparison of the images are used to determine the in-plane angle of rotation, as well as translation parameters that allow 2D images representing identical views to be brought into register before averaging to create "class average views". Methods to achieve grouping and alignment of images are inherently based in statistical evaluation of the data and include those based on multivariate statistical analysis (MSA) (e.g. principal component analysis combined with hierarchical clustering methods (van Heel and Schatz, 2009)), or alternatively methods based on K-means clustering (Penczek *et al.*, 1996) which involve non-hierarchical clustering by dividing the dataset into a number of K groups.

MSA methods attempt to reduce the complexity of the task by finding the largest variations within a dataset, and after this, arrange the images as vectors in a multidimensional space using these principal components of variation as orthogonal axes. With the image dataset expressed in this way, clustering methods then allow the images to be grouped according to their distances from one another within the multidimensional space. The precise methods of calculating these distances, and mechanisms of clustering vary from method to method, however, the intended result is the same. Both methods are implemented iteratively, until the procedures converge on a solution of improved quality.

For the rotationally invariant K-means algorithm used in this study, the fact that images can usually be centered according to their centers of gravity relatively easily is exploited. For a dataset containing many views, the next stage is to determine the angles of in-plane rotation that would allow the images to be brought into register (while grouping these into like views that can be average together). For this to be achieved a number (K) of seed images are selected and the

"distance" between each image in the dataset and the seed images are assessed using a rotation squared-discrepancy function which imposes the property of "rotational invariance" on the similarity assessment. Following this, class averages are calculated using the rotational parameters determined by the rotation function, and applied to the images prior to averaging to create new set of class images. This process is then iterated until images no longer move from class to class; i.e. they reach convergence.

A fundamental assumption associated with the single-particle method is structural equivalence of all instances of the macromolecular complex under study. Structural or conformational flexibility will result in class average views, and 3D reconstructions that are not faithful representations of the object or that are of decreased resolution (detail). While these methods of classification can be used to evaluate the characteristic views of a dataset, they can also allow classes representing alternative conformations of the macromolecular complex to be discerned. The ability to discern multiple conformational states is highly desirable, as this would offer the potential of elucidating the mechanistic states of macromolecular machines such as the ribosome, proteasome or Gro-EL/ES complexes by simply imaging the complex while functioning, and using image reconstruction to yield an array of conformational states; the so-called "story in a sample" approach (Frank, 2010). More recent methods of image classification rely on a maximum-likelihood approach to multi-reference refinement and classification and are showing great promise towards these goals (Scheres *et al.*, 2005).

1.4.2.2 Orientation determination

Before a 3D structure can be calculated from the class average views (described in the previous section), the relative orientations (or projection direction) of each view must be determined. Both experimental methods (relying on images of tilted particles), as well as computational methods (using un-tilted images) can be used. Experimental methods are important when a macromolecular complex of unknown structure is studied for the first time. This is especially the case for an asymmetric structure such as the ribosome, as the shear number of views required to reconstruct the structure in 3D, the low signal-to-noise ratio of images, as well as other ambiguities inherent to projection (such as those pertaining to handedness) make finding the correct solution immensely challenging. A method known as random conical tilt (Radermacher *et al.*, 1987) was developed for situations such as this, and is most often used when a complex adopts a preferred orientation when placed on a support film (as is relatively common for negative-staining). By imaging the specimen in both un-tilted and tilted orientations, the preferred orientation (which will usually have different in-plane tilts or rotational orientations on the grid) will present a range of views of the complex when tilted. The data-collection geometry in such

cases is conical, and allows generally good coverage of angular space to be attained, albeit harbouring a "missing cone" which will introduce some artefacts. A 3D reconstruction calculated with these experimentally determined orientations will also be of reduced resolution (owing to defocus gradient across images recorded with tilt) but should serve as a suitable starting model for further studies which might use cryo-EM for example to overcome the preferred orientation at 0° tilt. The benefit of this technique is that the true handedness of the structure can be assigned from the outset – a feature essential for accurate interpretation of a biological molecule. The methods described in the following paragraphs are unable to determine handedness from the projection data alone, and rely on additional experimentally derived information (Cheng *et al.*, 1995).

A purely computational method for orientation determination is known as angular reconstitution or sinogram correlation (van Heel, 1987). This method takes advantage of the central-section theorem directly, and states that the relative orientation between two projections can be specified according to a common tilt-axis, and determined with respect to this by identifying the "common-line" which any two projections of the a structure will share. In real space, any two 2D projections will share at least a single 1D line projection in common, and furthermore the line projection will pass through the origin of the specimen (Penczek *et al.*, 1996). For viruses there may be numerous common-lines, and this approach has seen extensive use for icosahedral reconstruction (Crowther *et al.*, 1970; Fuller *et al.*, 1996). The common-lines are found using a procedure called sinogram correlation after a sinogram is calculated by calculating a series of 1D line projections of the 2D image (usually a class average for improved SNR), and compared by sinogram correlation function. The Euler angles relating the orientations can then be determined from the tilt-axis shared by the two projections. If three projections are used, the orientations of each image can be determined with respect to three different planes, and 3D model reconstructed to an intermediate resolution using these angular approximations (van Heel, 1987).

For complexes where another structure is available to serve as starting model a technique known as iterative projection matching (or 3D projection matching) may be used (Penczek *et al.*, 1994). This method was originally applied to a cryo-EM investigation of Flock house virus (family *Nodaviridae*) RNA packaging using a X-ray structure of the capsid as a starting model (Cheng *et al.*, 1994). In this approach the starting model (or 3D reconstruction calculated in the previous step) is re-projected at known angles (covering the asymmetric unit of the structure) inside the computer (*in silico*) to create a series of reference projections. These references then serve as the basis for multi-reference alignment of the dataset (experimental images) by cross-correlation. Following cross-correlation, each experimental image will be assigned an orientation according

to the reference image with which it correlated with the highest score. Because the reference images are of substantially higher SNR, the approach is able to differentiate orientations of very small angular separation, and hence improve resolution over a number of iterations in which the 3D reconstruction is re-projected with finer angular spacing. Ultimately, most reconstructions (whether starting from a RCT reconstruction, or angular reconstitution) are refined using this approach, however it has been shown that in some cases accurate reconstructions can be determined using this approach using only a geometric starting models that matches the overall dimensions of the object. This is especially the case for objects with high symmetry such as helices or icosahedral assemblies, although potential pitfalls exist, and efforts must be made to validate the final result (Henderson *et al.*, 2012).

1.4.2.3 Three-dimensional reconstruction

A 3D density map can be calculated from 2D EM images or class averages using a procedure known as "back-projection". 3D reconstruction can be thought of as the inverse of projection and may be implemented in real or Fourier space. True back-projection algorithms are real space methods, which smear the 2D density backwards along lines determined by their orientation with respect to the 3D object (Harauz and van Heel, 1986). However, this procedure if implemented directly would result in a 3D reconstruction that would not accurately reflect the specimen under study. Weighted (or filtered), back-projection algorithms overcome some of these problems by weighting terms as a function of spatial frequency so that low frequency terms do not dominate the reconstruction, and therefore high-resolution information is more readily visible in the maps and iterative refinement by projection matching successful. Other procedures for back-projection have also been developed and these are the algebraic and simultaneous iterative methods. Fourier inversion (shown in Figure 1-5) represents a literal application of the central-section theorem to the problem of 3D reconstruction and was the first reconstruction approach used for this class of objects (De Rosier and Klug, 1968). In this scheme, 2D images are Fourier transformed and their transform, which is a plane or section, oriented as a central section in the 3D transform of the reconstruction. Inverse Fourier transform then results in a real space density map of the object in question.

For meaningful 3D reconstruction to be performed views covering a wide angular distribution of a single structure are typically required, and furthermore the angular space should be finely sampled to adequately sample of the Fourier transform of the object to high resolution. For an asymmetric object the angular range could be represented as hemisphere, however, symmetry inherent to the macromolecular complex, will greatly aid 3D reconstruction by reducing the angular distribution that needs to be visualized to just a fraction of this. Symmetry averaging

during 3D reconstruction will also increase the overall signal-to-noise ratio of the reconstruction as a consequence of a redundancy in the information required to describe the object. Applying this thinking to the geminiviruses (MSV and ACMV), hundreds of class averages were calculated using the IMAGIC software package, and orientations determined according to the common-lines method. 3D projection matching was then carried out using re-projections of the structure covering the angular space of just $1/10^{th}$ of a sphere. In this case, the global D52 symmetry of the particle was compared to a C1 and C5 reconstruction, and this confirmed that assuming the 10-fold symmetry was valid (Bottcher *et al.*, 2004).

1.4.2.4 Relevance of biological symmetry

Biological symmetry arises from the repeated association of equivalent structural units according to a limited number of specific binding rules (refer Section 1.3.1). While not only visually striking, the symmetry possessed by some macromolecular assemblies also provides a redundancy of information that is of great use to the structural biologist, as this redundancy can be used to further improve the signal-to-noise ratio of a 3D reconstruction by averaging. For an object possessing icosahedral symmetry, the asymmetric unit of the structure is always viewed in sixty different orientations in each projection, and once the orientation of a particle has been determined, and the central-section (single-particle) oriented in reciprocal space, averaging may be performed with all other equivalent orientations (i.e. corresponding central sections). This procedure not only allows noise to be reduced greatly, especially at higher spatial frequencies, but also serves to improve the angular coverage of the reconstruction; a feature of particular value when 3D reconstruction is attempted from a limited number of views.

Symmetric averaging can also be problematic as not all features of a particular structure may agree with the symmetry being imposed. For virus structures, the genome will without doubt deviate from icosahedral symmetry. Additionally, flexible regions of the CPs responsible for interacting with the genome (or other viral or host factors) may also deviate causing them to be "smeared out", preventing us from understanding the biological role of these aspects of the complex. It may be asked, why we simply don't attempt the asymmetric reconstruction of viruses more often in an attempt to overcome these limitations? Asymmetric studies have been carried out in some cases (Chang *et al.*, 2006; Duda and Conway, 2008), but these are limited to viruses with sufficiently large asymmetric features (for instance, portal complexes) that provide visible features allowing asymmetric alignment of the images to take place despite the extremely poor SNR of the experimental images. For simple viruses, which do not have such obvious features, the signal-to-noise, as well as the strong signal of the spherical and symmetric capsid makes accurate alignment in the presence of the 60-fold symmetry nearly impossible with present day

technology. Future studies, benefiting from improvements in image quality owing to use of improved detectors (Faruqi and Henderson, 2007) as well as phase-plate technology (Chang *et al.*, 2010; Murata *et al.*, 2010; Nagayama and Danev, 2009, 2008) will significantly improve contrast in EM images allowing features that deviate from symmetry to be discerned in a statistically meaningful way – essentially paving our way for the next step in our understanding of the structure and function of many viruses.

1.4.3 The structure of viral genomes

X-ray diffraction, neutron scattering, as well as electron microscopy studies have each been used to observe viral genomes. Techniques that use signal averaging (i.e. all of structural biology methods) requires that the structure concerned be highly ordered for it to be accurately represented in the averaged map. In the case of protein capsids that are all considered structurally identical, many of these can be averaged together to overcome signal-to-noise limitations imposed by dose applied and detector performance. Furthermore the icosahedral symmetry can be exploited to improve the signal-to-noise ratio 60-fold. However, nucleic acid genomes are asymmetric and most likely disordered, making averaging with symmetry averaging inappropriate. Hence visualization of viral genomes inside the viral capsids using these same techniques is uncommon. In cases where the genome associates with the capsid in a repetitive way, and by virtue of these bonding constraints, icosahedral symmetry may be adopted by the genome in some regions, and hence it is visible in the resolved electron density. Excellent examples of this include *canine parvovirus* (Chapman and Rossmann, 1995), *bacteriophage MS2* (Valegård *et al.*, 1994), *bean pod mottle virus* (Lin *et al.*, 2003), as well as *Pariacoto virus* (Tang *et al.*, 2001).

Using X-ray crystallography, different motifs have been observed to associate with the inner surface of the capsid, these may be stem-loop structures (Valegård *et al.*, 1994), duplexes (Tang *et al.*, 2001) or regions of single-stranded RNA (Lin *et al.*, 2003). In such cases where a precise conformation for the RNA can be described, investigators have described the percentage (e.g. 20-40%) of the genome that is clearly visible within the density. It is important that this does not mislead, as the genome being an inherently asymmetric structure can never be visualized "as it is" by any method which involves symmetric averaging, and hence the precise sequence visualized is only ever a "consensus sequence" being derived from the averaged asymmetric unit of the map. However, where interactions show some degree of sequence dependence, investigators have noted that the consensus sequence observed in the structural data can often be matched with many regions of the genome (Chapman and Rossmann, 1995; Lin *et al.*, 2003).

By providing phase information over a wide range of spatial frequencies, EM techniques have most often provided low-resolution information that is often complementary to X-ray studies

(Tang and Johnson, 2002). EM tends to reveal more disordered regions of RNA, and it is this insight which has revealed that may RNA viruses package their genomes in a series of concentric shells (Bakker *et al.*, 2012; Toropova *et al.*, 2008; van den Worm *et al.*, 2006). The outermost, or protein associated, shell is the layer most easily visualized as a consequence of the stabilizing influence of protein-nucleic interactions. The inner layers are significantly more disordered but are thought to adopt secondary structure, and organize into different layers under forces of charge repulsion or interactions with terminal regions of the coat protein subunits that may extend into the core of the particle. A striking example of the viral CPs ordering the genome are members of the *Nodaviridae*: *Flock house virus*, and *Pariacoto virus* in which the protein coat appears to induce a duplex conformation on RNA irrespective of sequence, and orders these segments in a dodecahedral arrangement (Tang *et al.*, 2001). In these cases, association of the genome with the capsid in these regions is thought to have a precise role in controlling formation of the capsid.

1.5 Structural bioinformatics and protein modelling

Many biological protein sequences will fold to adopt a conformation termed the proteins tertiary structure. Because protein folding is governed by its amino acid sequence (Anfinsen, 1973), it is reasonable to expect that knowledge of this sequence will provide information about its tertiary structure, possibly allowing a useful model of the proteins actual structure to be created. Methods for protein structure prediction may be either knowledge based, such as homology (or comparative) modelling (Martí-Renom *et al.*, 2000), or other so called *ab initio* methods (Bonneau and Baker, 2001). However, the "pseudo-atomic" model constructed from constraints not arising from direct observation of the molecule concerned, should only be taken seriously as far as their usefulness as a *guide* for assisting such an interpretation, and should ideally be validated by further experimental investigation of a preferably biochemical or molecular nature - e.g. site-directed mutagenesis experiments (Dill *et al.*, 2007).

Homology - or comparative - modelling procedures are a popular, and computationally inexpensive way of going about protein structure modelling, and rely on the assumption that protein evolution favours the conservation of structure rather than merely that of amino acid sequence. In addition, it has been observed that protein structure is resistant to site mutation, and furthermore, sequences of even marginal similarity can nevertheless result in an identical protein fold. Based on these premises, which are reviewed in (Xia and Levitt, 2004), homology modelling software packages are able to assign a probable backbone conformation to a query sequence based on information derived from a sequence alignment of the query sequence to that of an evolutionarily related (or homologous) "template" structure. A template is an experimentally determined structure whose sequence may be aligned to a query sequence with sufficiently high

confidence (Sali and Blundell, 1993). Homology modelling is considered reliable if performed using templates of at least 20% identity to that of the query sequence, and it has been asserted that automatic alignment and model building of sequence-template pairs of >= 40% identity may allow much of the experimentally unresolved conformational space to be filled by solely by computational means (Šali *et al.*, 1995). For sequences <= 40 % identity, the models produced are likely to contain errors, but are likely to be useful when applied in different contexts such as interrogation of intermediate to low resolution EM density maps; as in this study.

Viral coat proteins are known to possess one of five prototypical protein folds (Liljas, 2004), whilst at the same time showing considerable sequence divergence amongst members that share the same fold. Template identification, in the context of structural virology, therefore relies on methods that are sensitive enough to identify distant structural homologs that may be used during homology modelling attempts. The 'fold recognition' problem defines this requirement, and has resulted in a number of computational methods showing great promise in achieving this end (Jones, 1999a; Källberg *et al.*, 2012; Kelley and Sternberg, 2009; Remmert *et al.*, 2012). Specifically, computational algorithms for sequence threading are particularly important as they allow the compatibility (or chemical properties) of a particular sequence to be evaluated in a particular structural context. Amino acid residues will exhibit different compatibilities with particular chemical environments - as defined by a particular folds topology and architecture – these terms describing the sequence and nature of secondary structure elements, and how these are arranged in 3D space, respectively. A fundamental example is the observation that hydrophobic residues generally favour solvent inaccessible (buried) regions, whereas hydrophilic residues favour surface exposed (solve accessible regions). Based on such criteria an alignment of the sequence can be scored against a particular experimental structural atomic model. An algorithm called GenTHREADER (Jones, 1999a) uses a profile based alignment strategy to produce an initial alignment of the sequence against the structure. The alignment is used to generate pairwise scores of 'mean force' and solvation energy, and series of scores such as these are then evaluated using a neural network scoring routine constructed based on sequence comparisons to a database of high quality (and resolution) X-ray crystal structures. The neural network can be taught to carry out complex decisions using a large number of input variables, and in principal should work sufficiently well that the human supervision is not required.

Proteins fold (self-interact) to adopt tertiary structure, but their function is often dependent on the formation of stable interactions with other molecules (multimerization or oligmerization) which may be instances of themselves (homomeric assemblies) or other proteins (heteromeric

assemblies). Attempts to predict the formation of these complexes *in silico* have been made, however 3D reconstruction from EM data has the potential to empirically determine these, if the interacting partners can be fitted into the 3D density map at an appropriate resolution. The process of atomic structure 'fitting' is referred to as EM map 'docking' when carried out quantitatively using cross-correlation based six-dimensional (6D) searches implemented in computer programs (Roseman, 2000; Wriggers, 2010). More recently, methods for comparative modelling have been developed to include EM maps as conformational constraints during model refinement. In cases that the EM density is of sufficient resolution (generally $\leq$ 8 Å), this can serve as a constraint during the modelling process which governs the placement of secondary structures (Chan *et al.*, 2012; Topf *et al.*, 2008). In this study, standard 'rigid body' docking of the β-barrel domains of the viral pentameric capsomer (CP_5) was carried out using standard cross-correlation based approaches.

1.6 Geminiviridae

Geminiviruses (viruses of the family *Geminiviridae)* are agronomically relevant plant pathogens that significantly reduce yields of crops such as maize, cassava, beans, squash, cotton and tomatoes every year. The family consists of seven genera: *Mastrevirus*, *Begomovirus*, *Topocuvirus*, *Curtovirus, Becurtovirus, Eragrovirus, and Turncurtovirus* which differ in host and vector species specificities, as well as genome organization. Geminiviruses are heavily prone to genetic recombination, both within, and across genera, and this has been proposed to be a significant cause of emergence of new viral species (Padidam *et al.*, 1999). In support of this hypothesis, the establishment of efficient global transportation networks over the past few decades has led to increased spread of these viruses as well as their vector species, causing significant outbreaks in 39 countries; events that have further promoted the importance of studying these pathogens (Moffat, 1999).

1.6.1 Host and vector specificity

The host range and vector specificity of a geminivirus are important factors used for their classification, and will also influence their geographic distribution. Mastreviruses (e.g. *Maize streak virus*) will usually infect monocotyledonous grass species (*Gramineae*), but in some cases dicotyledonous plants are also infected (Briddon and Markham, 1995). MSV is transmitted by the insect vector *Cicadulina mbila* Naudé (Storey 1925); a leaf hopper. Begomoviruses, distinct in that they typically consist of bipartite genomes, usually only infect dicotyledonous hosts, and are transmitted by the whitefly (*Bemisia tabaci*). *African cassava mosaic virus* (ACMV), a Begomovirus, significantly reduces crop yields of the cassava plant (*Manihot esculenta*), also a dietary staple in Africa. *Beet curly top virus* (BCTV) is a member of the Curtovirus genus, is

transmitted by the leafhopper (*Circulifer tenellus*), and is known for its cause of commercial loss of sugar beets (*Beta vulgaris*) in western USA, but is also known to infect 300 other dicotyledonous plant species. *Tomato pseudo curly top virus* (TpCTV) a Topocuvirus, is responsible for agricultural loss of tomatoes in Florida, but will also infect some legumes and other dicotyledonous plants. It shows many symptoms in common with BCTV, but is transmitted by the treehopper *Micrutalis malleifera;* the only geminivirus known to be transmitted by such a vector.

1.6.1.1 Transmission and relationship with vector

Vectors are organisms required to carry out viral transmission from one host organism to another, and are especially required by plant viruses because plants are themselves immobile. Once an infectious virion arrives at a new host it must find a means of entering the cytoplasm of a permissive cell allowing viral replication. Mechanical disruption of the cell wall by the vector may allow this, however, geminiviruses are not generally considered mechanically transmissible, and must be introduced into the vascular system of the plant (i.e. phloem sieve tubes) by an insect that feeds on the sugar containing fluids of these tissues. The virus particles will be ingested by the vector (a leaf- or tree-hopper) at the time that it feeds on an infected plant, and will move from the vectors gut into the salivary glands where it will be available for re-introduction into an un-infected host during another feed. The virus is unable to replicate in the vector, i.e. it is said to display a 'non-propagative' relationship.

1.6.2 Genome organization and replication

Important differences at the level of genome organization as revealed by viral genome sequencing and are shown in Figure 1-8. Geminivirus genomes are relatively small (2.5 - 3 kb, i.e M_w ~760 – 915 kDa) circular single-stranded DNA (ssDNA) molecules that are replicated in the nucleus of the host cell (Boulton, 2002; Davies *et al.*, 1987; Nagar *et al.*, 1995). Consequently, geminiviruses are class II viruses according to the Baltimore classification scheme (Baltimore, 1971). Other examples of a single-stranded viruses include the DNA bacteriophage ϕX174 (the type species *Microviridae*) as well as well as members of the *Nanoviridae*. The *Parvoviridae* are ssDNA viruses with linear genomes, and infect mammals and insects (Chapman and Rossmann, 1993). Notably, these virus families have a $T = 1$ capsid architecture, and replicate by mechanisms known as rolling-circle or rolling-hairpin replication, respectively.

Upon entry into the nucleus, the ssDNA geminivirus genome is converted to a double-stranded DNA (dsDNA) intermediate by host DNA replication machinery. The dsDNA intermediates associates with histones (structural components of the nucleosome) to form minichromosomes (Pilartz and Jeske, 1992). This dsDNA intermediate serves as the template for bidirectional

transcription of viral genes as well as genome replication by rolling circle replication (Heyraud *et al.* 1993a; Heyraud *et al.* 1993b; Saunders *et al.* 1991, 1992). Replication is initiated by the viral replication associated protein (Rep) from a conserved nonanucleotide site (TAATATT/AC) in the long intergenic region (LIR) of the geminivirus genomes (Gutierrez 2000a, b; Hanley-Bowdoin *et al.* 2000) and is illustrated in Figure 1-8. This sequence motif adopts a stem-loop structure, and presents a site for nucleolytic cleavage, the substrate of the Rep protein, allowing rolling-circle replication to proceed.

1.6.3 Maize streak virus

Maize streak virus (MSV), the subject of this study, is the prototypical member of the *Mastrevirus* genus, and is considered the most severe maize (*Zea mays*) disease in sub-Saharan Africa, and has as a result, been the subject of extensive research (Bosque-Perez 2000; Boulton 2002). MSV has the most dramatic consequences for subsistence farmers who rely on maize as a dietary staple, and consequently can place families and communities at risk. Breeding efforts have resulted in MSV resistant maize stains, as well as the development of recombinant maize species that have induced resistance as a result of stable expression of a mutated viral replicase protein gene from a construct which was integrated into genome of the plant strain (Shepherd *et al.*, 2007).

Named after its characteristic symptoms; originally described as 'mealie variegation' (Fuller 1901). It was later renamed 'maize streak' by the researcher who identified the insect vector responsible for virus transmission to be *Cicadulina mbila* Naudé (Storey 1925). MSV encodes four gene products on a single (monopartite) single-stranded DNA (ssDNA) molecule that is 2687 nt in size (Francki *et al.*, 1980; Lazarowitz, 1988; Mullineaux *et al.*, 1984). Two of the gene products (Rep and RepA) are associated with replication of the genome and are expressed early in the virus life-cycle from a complimentary sense (C-sense) transcript. The remaining two genes are virion-sense (V-sense), and encode a movement (V2) and coat protein (V1), respectively.

1.6.3.1 Life-cycle, tissue specificity and role of viral genes

The life-cycle of MSV (Figure 1-10) begins with entry of the circular ssDNA genome into an uninfected cell by passage from the phloem sieve tubes of vascular bundle into an uninfected cell. The ssDNA genome must then be trafficked into the nucleus where the host cell DNA repair and replication machinery resides. Nuclear trafficking is mediated by the karyophillic coat proteins (CPs) that each possess a nuclear localization signal (NLS) (Liu et al., 1999). By an unknown mechanism, the virion is then disassembled to render the genome accessible to host factors, and the ssDNA genome converted to a double-stranded form by host replicases within the nucleus (Palmer and Rybicki, 1998). This dsDNA form is then transcribed

in other ways. The resolution limit of a negative stain reconstruction is the grain size of the heavy metal stain, while considerations pertinent to cryo-EM are the thickness of the vitreous ice as well as the size of the complex under study. Thicker ice will result in reduced contrast (especially at higher resolutions), and a greater distribution of particles along the Z-axis of the microscope resulting in particles imaged at different focii and magnifications. The structure of complexes ≥ 250 kDa can be confidently studied using single-particle methods, however, the larger a complex is the more contrast it will produce. Contrast is particularly relevant, as the performance of computational alignment algorithms is heavily influenced by this property of the data, and large symmetric viruses typically produce the highest resolution 3D structures for this reason (Grigorieff and Harrison, 2011).

1.4.1 Vitrification of thin aqueous layers

Vitrification is a process in which an aqueous specimen (e.g. viruses prepared in a buffered solution) are cooled so rapidly that the water molecules freeze prior to adopting a crystalline organization (e.g. either cubic or hexagonal ice). Viruses suspended in such a solution will be preserved in near-native state, immobilized in a layer of vitreous ice with their hydration shell, and macromolecular conformations unaltered. Freezing is carried out by plunging the specimen into liquid ethane cooled by liquid nitrogen (~ -180°C). For single-particle studies, a droplet of the specimen will be applied to a perforated carbon film, and blotted so that a small fraction of the suspension remains on the film as a thin-layer held across the perforations by virtue of surface tension. Such layers may be 50-200 nm thick, and are plunged immediately, and in a high humidity environment, in an attempt to minimize evaporation that might result in concentration of buffer components that might reduce contrast in the EM. Upon rapid entry of the thin-specimen into the liquid ethane, the temperature of the specimen is thought to decrease at ~10,000-1,000,000°C/s; ensuring vitrification. Because the macromolecular complexes (virions) are freely dispersed they should be randomly oriented within the thin- layer. This is required for 3D reconstruction because it ensures good angular coverage within the dataset. However, this may not always be the case owing to interactions with the air-water interface, or a carbon support, which if present may adsorb the particles, restricting their orientations in a manner similar to that of negative-stain EM. This is the case for the work described in this study.

1.4.2 Three-dimensional electron microscopy

Electron microscopy of biological samples, particularly at the high magnifications afforded by EM, is characterized by a noisy visualization of the specimen and much of image processing is intended to overcome this problem. Low-dose techniques attempt to reduce radiation damage of the specimen, however the signal-to-noise ratio of such images is extremely low and further

worsened by noise arising during detection (especially in the case of CCD cameras). Other unintended processes such as specimen charging and beam induced movement of the specimen, further worsen the quality of recorded images. 3D EM techniques have developed complex procedures that attempt to overcome these limitations in data-collection.

For the single-particle EM experiment, the recorded electron micrograph will typically contain images of the structure of interest in a number of random orientations, and as result, reconstruction from a large set of images is essentially a problem of 'grouping' similar views of the structure by classification, and then determining the three Euler angles - phi (ϕ), theta (θ) and psi (ψ) - that describe how each image is oriented in relation to a common 3D coordinate system.

1.4.2.1 2D alignment and classification

The signal-to-noise ratio (SNR) of the data is usually improved by averaging images containing the same information (identical views), and 2D image alignment algorithms based on statistical comparison of the images are used to determine the in-plane angle of rotation, as well as translation parameters that allow 2D images representing identical views to be brought into register before averaging to create "class average views". Methods to achieve grouping and alignment of images are inherently based in statistical evaluation of the data and include those based on multivariate statistical analysis (MSA) (e.g. principal component analysis combined with hierarchical clustering methods (van Heel and Schatz, 2009)), or alternatively methods based on K-means clustering (Penczek *et al.*, 1996) which involve non-hierarchical clustering by dividing the dataset into a number of K groups.

MSA methods attempt to reduce the complexity of the task by finding the largest variations within a dataset, and after this, arrange the images as vectors in a multidimensional space using these principal components of variation as orthogonal axes. With the image dataset expressed in this way, clustering methods then allow the images to be grouped according to their distances from one another within the multidimensional space. The precise methods of calculating these distances, and mechanisms of clustering vary from method to method, however, the intended result is the same. Both methods are implemented iteratively, until the procedures converge on a solution of improved quality.

For the rotationally invariant K-means algorithm used in this study, the fact that images can usually be centered according to their centers of gravity relatively easily is exploited. For a dataset containing many views, the next stage is to determine the angles of in-plane rotation that would allow the images to be brought into register (while grouping these into like views that can be average together). For this to be achieved a number (K) of seed images are selected and the

"distance" between each image in the dataset and the seed images are assessed using a rotation squared-discrepancy function which imposes the property of "rotational invariance" on the similarity assessment. Following this, class averages are calculated using the rotational parameters determined by the rotation function, and applied to the images prior to averaging to create new set of class images. This process is then iterated until images no longer move from class to class; i.e. they reach convergence.

A fundamental assumption associated with the single-particle method is structural equivalence of all instances of the macromolecular complex under study. Structural or conformational flexibility will result in class average views, and 3D reconstructions that are not faithful representations of the object or that are of decreased resolution (detail). While these methods of classification can be used to evaluate the characteristic views of a dataset, they can also allow classes representing alternative conformations of the macromolecular complex to be discerned. The ability to discern multiple conformational states is highly desirable, as this would offer the potential of elucidating the mechanistic states of macromolecular machines such as the ribosome, proteasome or GroEL/ES complexes by simply imaging the complex while functioning, and using image reconstruction to yield an array of conformational states; the so-called "story in a sample" approach (Frank, 2010). More recent methods of image classification rely on a maximum-likelihood approach to multi-reference refinement and classification and are showing great promise towards these goals (Scheres *et al.*, 2005).

1.4.2.2 Orientation determination

Before a 3D structure can be calculated from the class average views (described in the previous section), the relative orientations (or projection direction) of each view must be determined. Both experimental methods (relying on images of tilted particles), as well as computational methods (using un-tilted images) can be used. Experimental methods are important when a macromolecular complex of unknown structure is studied for the first time. This is especially the case for an asymmetric structure such as the ribosome, as the shear number of views required to reconstruct the structure in 3D, the low signal-to-noise ratio of images, as well as other ambiguities inherent to projection (such as those pertaining to handedness) make finding the correct solution immensely challenging. A method known as random conical tilt (Radermacher *et al.*, 1987) was developed for situations such as this, and is most often used when a complex adopts a preferred orientation when placed on a support film (as is relatively common for negative-staining). By imaging the specimen in both un-tilted and tilted orientations, the preferred orientation (which will usually have different in-plane tilts or rotational orientations on the grid) will present a range of views of the complex when tilted. The data-collection geometry in such

cases is conical, and allows generally good coverage of angular space to be attained, albeit harbouring a "missing cone" which will introduce some artefacts. A 3D reconstruction calculated with these experimentally determined orientations will also be of reduced resolution (owing to defocus gradient across images recorded with tilt) but should serve as a suitable starting model for further studies which might use cryo-EM for example to overcome the preferred orientation at 0° tilt. The benefit of this technique is that the true handedness of the structure can be assigned from the outset – a feature essential for accurate interpretation of a biological molecule. The methods described in the following paragraphs are unable to determine handedness from the projection data alone, and rely on additional experimentally derived information (Cheng *et al.*, 1995).

A purely computational method for orientation determination is known as angular reconstitution or sinogram correlation (van Heel, 1987). This method takes advantage of the central-section theorem directly, and states that the relative orientation between two projections can be specified according to a common tilt-axis, and determined with respect to this by identifying the "common-line" which any two projections of the a structure will share. In real space, any two 2D projections will share at least a single 1D line projection in common, and furthermore the line projection will pass through the origin of the specimen (Penczek *et al.*, 1996). For viruses there may be numerous common-lines, and this approach has seen extensive use for icosahedral reconstruction (Crowther *et al.*, 1970; Fuller *et al.*, 1996). The common-lines are found using a procedure called sinogram correlation after a sinogram is calculated by calculating a series of 1D line projections of the 2D image (usually a class average for improved SNR), and compared by sinogram correlation function. The Euler angles relating the orientations can then be determined from the tilt-axis shared by the two projections. If three projections are used, the orientations of each image can be determined with respect to three different planes, and 3D model reconstructed to an intermediate resolution using these angular approximations (van Heel, 1987).

For complexes where another structure is available to serve as starting model a technique known as iterative projection matching (or 3D projection matching) may be used (Penczek *et al.*, 1994). This method was originally applied to a cryo-EM investigation of Flock house virus (family *Nodaviridae*) RNA packaging using a X-ray structure of the capsid as a starting model (Cheng *et al.*, 1994). In this approach the starting model (or 3D reconstruction calculated in the previous step) is re-projected at known angles (covering the asymmetric unit of the structure) inside the computer (*in silico*) to create a series of reference projections. These references then serve as the basis for multi-reference alignment of the dataset (experimental images) by cross-correlation. Following cross-correlation, each experimental image will be assigned an orientation according

to the reference image with which it correlated with the highest score. Because the reference images are of substantially higher SNR, the approach is able to differentiate orientations of very small angular separation, and hence improve resolution over a number of iterations in which the 3D reconstruction is re-projected with finer angular spacing. Ultimately, most reconstructions (whether starting from a RCT reconstruction, or angular reconstitution) are refined using this approach, however it has been shown that in some cases accurate reconstructions can be determined using this approach using only a geometric starting models that matches the overall dimensions of the object. This is especially the case for objects with high symmetry such as helices or icosahedral assemblies, although potential pitfalls exist, and efforts must be made to validate the final result (Henderson *et al.*, 2012).

1.4.2.3 Three-dimensional reconstruction

A 3D density map can be calculated from 2D EM images or class averages using a procedure known as "back-projection". 3D reconstruction can be thought of as the inverse of projection and may be implemented in real or Fourier space. True back-projection algorithms are real space methods, which smear the 2D density backwards along lines determined by their orientation with respect to the 3D object (Harauz and van Heel, 1986). However, this procedure if implemented directly would result in a 3D reconstruction that would not accurately reflect the specimen under study. Weighted (or filtered), back-projection algorithms overcome some of these problems by weighting terms as a function of spatial frequency so that low frequency terms do not dominate the reconstruction, and therefore high-resolution information is more readily visible in the maps and iterative refinement by projection matching successful. Other procedures for back-projection have also been developed and these are the algebraic and simultaneous iterative methods. Fourier inversion (shown in Figure 1-5) represents a literal application of the central-section theorem to the problem of 3D reconstruction and was the first reconstruction approach used for this class of objects (De Rosier and Klug, 1968). In this scheme, 2D images are Fourier transformed and their transform, which is a plane or section, oriented as a central section in the 3D transform of the reconstruction. Inverse Fourier transform then results in a real space density map of the object in question.

For meaningful 3D reconstruction to be performed views covering a wide angular distribution of a single structure are typically required, and furthermore the angular space should be finely sampled to adequately sample of the Fourier transform of the object to high resolution. For an asymmetric object the angular range could be represented as hemisphere, however, symmetry inherent to the macromolecular complex, will greatly aid 3D reconstruction by reducing the angular distribution that needs to be visualized to just a fraction of this. Symmetry averaging

during 3D reconstruction will also increase the overall signal-to-noise ratio of the reconstruction as a consequence of a redundancy in the information required to describe the object. Applying this thinking to the geminiviruses (MSV and ACMV), hundreds of class averages were calculated using the IMAGIC software package, and orientations determined according to the common-lines method. 3D projection matching was then carried out using re-projections of the structure covering the angular space of just $1/10^{th}$ of a sphere. In this case, the global D52 symmetry of the particle was compared to a C1 and C5 reconstruction, and this confirmed that assuming the 10-fold symmetry was valid (Bottcher *et al.*, 2004).

1.4.2.4 Relevance of biological symmetry

Biological symmetry arises from the repeated association of equivalent structural units according to a limited number of specific binding rules (refer Section 1.3.1). While not only visually striking, the symmetry possessed by some macromolecular assemblies also provides a redundancy of information that is of great use to the structural biologist, as this redundancy can be used to further improve the signal-to-noise ratio of a 3D reconstruction by averaging. For an object possessing icosahedral symmetry, the asymmetric unit of the structure is always viewed in sixty different orientations in each projection, and once the orientation of a particle has been determined, and the central-section (single-particle) oriented in reciprocal space, averaging may be performed with all other equivalent orientations (i.e. corresponding central sections). This procedure not only allows noise to be reduced greatly, especially at higher spatial frequencies, but also serves to improve the angular coverage of the reconstruction; a feature of particular value when 3D reconstruction is attempted from a limited number of views.

Symmetric averaging can also be problematic as not all features of a particular structure may agree with the symmetry being imposed. For virus structures, the genome will without doubt deviate from icosahedral symmetry. Additionally, flexible regions of the CPs responsible for interacting with the genome (or other viral or host factors) may also deviate causing them to be "smeared out", preventing us from understanding the biological role of these aspects of the complex. It may be asked, why we simply don't attempt the asymmetric reconstruction of viruses more often in an attempt to overcome these limitations? Asymmetric studies have been carried out in some cases (Chang *et al.*, 2006; Duda and Conway, 2008), but these are limited to viruses with sufficiently large asymmetric features (for instance, portal complexes) that provide visible features allowing asymmetric alignment of the images to take place despite the extremely poor SNR of the experimental images. For simple viruses, which do not have such obvious features, the signal-to-noise, as well as the strong signal of the spherical and symmetric capsid makes accurate alignment in the presence of the 60-fold symmetry nearly impossible with present day

technology. Future studies, benefiting from improvements in image quality owing to use of improved detectors (Faruqi and Henderson, 2007) as well as phase-plate technology (Chang *et al.*, 2010; Murata *et al.*, 2010; Nagayama and Danev, 2009, 2008) will significantly improve contrast in EM images allowing features that deviate from symmetry to be discerned in a statistically meaningful way – essentially paving our way for the next step in our understanding of the structure and function of many viruses.

1.4.3 The structure of viral genomes

X-ray diffraction, neutron scattering, as well as electron microscopy studies have each been used to observe viral genomes. Techniques that use signal averaging (i.e. all of structural biology methods) requires that the structure concerned be highly ordered for it to be accurately represented in the averaged map. In the case of protein capsids that are all considered structurally identical, many of these can be averaged together to overcome signal-to-noise limitations imposed by dose applied and detector performance. Furthermore the icosahedral symmetry can be exploited to improve the signal-to-noise ratio 60-fold. However, nucleic acid genomes are asymmetric and most likely disordered, making averaging with symmetry averaging inappropriate. Hence visualization of viral genomes inside the viral capsids using these same techniques is uncommon. In cases where the genome associates with the capsid in a repetitive way, and by virtue of these bonding constraints, icosahedral symmetry may be adopted by the genome in some regions, and hence it is visible in the resolved electron density. Excellent examples of this include *canine parvovirus* (Chapman and Rossmann, 1995), *bacteriophage MS2* (Valegård *et al.*, 1994), *bean pod mottle virus* (Lin *et al.*, 2003), as well as *Pariacoto virus* (Tang *et al.*, 2001).

Using X-ray crystallography, different motifs have been observed to associate with the inner surface of the capsid, these may be stem-loop structures (Valegård *et al.*, 1994), duplexes (Tang *et al.*, 2001) or regions of single-stranded RNA (Lin *et al.*, 2003). In such cases where a precise conformation for the RNA can be described, investigators have described the percentage (e.g. 20-40%) of the genome that is clearly visible within the density. It is important that this does not mislead, as the genome being an inherently asymmetric structure can never be visualized "as it is" by any method which involves symmetric averaging, and hence the precise sequence visualized is only ever a "consensus sequence" being derived from the averaged asymmetric unit of the map. However, where interactions show some degree of sequence dependence, investigators have noted that the consensus sequence observed in the structural data can often be matched with many regions of the genome (Chapman and Rossmann, 1995; Lin *et al.*, 2003).

By providing phase information over a wide range of spatial frequencies, EM techniques have most often provided low-resolution information that is often complementary to X-ray studies

(Tang and Johnson, 2002). EM tends to reveal more disordered regions of RNA, and it is this insight which has revealed that may RNA viruses package their genomes in a series of concentric shells (Bakker *et al.*, 2012; Toropova *et al.*, 2008; van den Worm *et al.*, 2006). The outermost, or protein associated, shell is the layer most easily visualized as a consequence of the stabilizing influence of protein-nucleic interactions. The inner layers are significantly more disordered but are thought to adopt secondary structure, and organize into different layers under forces of charge repulsion or interactions with terminal regions of the coat protein subunits that may extend into the core of the particle. A striking example of the viral CPs ordering the genome are members of the *Nodaviridae*: *Flock house virus*, and *Pariacoto virus* in which the protein coat appears to induce a duplex conformation on RNA irrespective of sequence, and orders these segments in a dodecahedral arrangement (Tang *et al.*, 2001). In these cases, association of the genome with the capsid in these regions is thought to have a precise role in controlling formation of the capsid.

1.5 Structural bioinformatics and protein modelling

Many biological protein sequences will fold to adopt a conformation termed the proteins tertiary structure. Because protein folding is governed by its amino acid sequence (Anfinsen, 1973), it is reasonable to expect that knowledge of this sequence will provide information about its tertiary structure, possibly allowing a useful model of the proteins actual structure to be created. Methods for protein structure prediction may be either knowledge based, such as homology (or comparative) modelling (Martí-Renom *et al.*, 2000), or other so called *ab initio* methods (Bonneau and Baker, 2001). However, the "pseudo-atomic" model constructed from constraints not arising from direct observation of the molecule concerned, should only be taken seriously as far as their usefulness as a *guide* for assisting such an interpretation, and should ideally be validated by further experimental investigation of a preferably biochemical or molecular nature - e.g. site-directed mutagenesis experiments (Dill *et al.*, 2007).

Homology - or comparative - modelling procedures are a popular, and computationally inexpensive way of going about protein structure modelling, and rely on the assumption that protein evolution favours the conservation of structure rather than merely that of amino acid sequence. In addition, it has been observed that protein structure is resistant to site mutation, and furthermore, sequences of even marginal similarity can nevertheless result in an identical protein fold. Based on these premises, which are reviewed in (Xia and Levitt, 2004), homology modelling software packages are able to assign a probable backbone conformation to a query sequence based on information derived from a sequence alignment of the query sequence to that of an evolutionarily related (or homologous) "template" structure. A template is an experimentally determined structure whose sequence may be aligned to a query sequence with sufficiently high

confidence (Sali and Blundell, 1993). Homology modelling is considered reliable if performed using templates of at least 20% identity to that of the query sequence, and it has been asserted that automatic alignment and model building of sequence-template pairs of >= 40% identity may allow much of the experimentally unresolved conformational space to be filled by solely by computational means (Šali *et al.*, 1995). For sequences <= 40 % identity, the models produced are likely to contain errors, but are likely to be useful when applied in different contexts such as interrogation of intermediate to low resolution EM density maps; as in this study.

Viral coat proteins are known to possess one of five prototypical protein folds (Liljas, 2004), whilst at the same time showing considerable sequence divergence amongst members that share the same fold. Template identification, in the context of structural virology, therefore relies on methods that are sensitive enough to identify distant structural homologs that may be used during homology modelling attempts. The 'fold recognition' problem defines this requirement, and has resulted in a number of computational methods showing great promise in achieving this end (Jones, 1999a; Källberg *et al.*, 2012; Kelley and Sternberg, 2009; Remmert *et al.*, 2012). Specifically, computational algorithms for sequence threading are particularly important as they allow the compatibility (or chemical properties) of a particular sequence to be evaluated in a particular structural context. Amino acid residues will exhibit different compatibilities with particular chemical environments - as defined by a particular folds topology and architecture – these terms describing the sequence and nature of secondary structure elements, and how these are arranged in 3D space, respectively. A fundamental example is the observation that hydrophobic residues generally favour solvent inaccessible (buried) regions, whereas hydrophilic residues favour surface exposed (solve accessible regions). Based on such criteria an alignment of the sequence can be scored against a particular experimental structural atomic model. An algorithm called GenTHREADER (Jones, 1999a) uses a profile based alignment strategy to produce an initial alignment of the sequence against the structure. The alignment is used to generate pairwise scores of 'mean force' and solvation energy, and series of scores such as these are then evaluated using a neural network scoring routine constructed based on sequence comparisons to a database of high quality (and resolution) X-ray crystal structures. The neural network can be taught to carry out complex decisions using a large number of input variables, and in principal should work sufficiently well that the human supervision is not required.

Proteins fold (self-interact) to adopt tertiary structure, but their function is often dependent on the formation of stable interactions with other molecules (multimerization or oligmerization) which may be instances of themselves (homomeric assemblies) or other proteins (heteromeric

assemblies). Attempts to predict the formation of these complexes *in silico* have been made, however 3D reconstruction from EM data has the potential to empirically determine these, if the interacting partners can be fitted into the 3D density map at an appropriate resolution. The process of atomic structure 'fitting' is referred to as EM map 'docking' when carried out quantitatively using cross-correlation based six-dimensional (6D) searches implemented in computer programs (Roseman, 2000; Wriggers, 2010). More recently, methods for comparative modelling have been developed to include EM maps as conformational constraints during model refinement. In cases that the EM density is of sufficient resolution (generally $\leq$ 8 Å), this can serve as a constraint during the modelling process which governs the placement of secondary structures (Chan *et al.*, 2012; Topf *et al.*, 2008). In this study, standard 'rigid body' docking of the β-barrel domains of the viral pentameric capsomer (CP_5) was carried out using standard cross-correlation based approaches.

1.6 Geminiviridae

Geminiviruses (viruses of the family *Geminiviridae)* are agronomically relevant plant pathogens that significantly reduce yields of crops such as maize, cassava, beans, squash, cotton and tomatoes every year. The family consists of seven genera: *Mastrevirus*, *Begomovirus*, *Topocuvirus*, *Curtovirus, Becurtovirus, Eragrovirus, and Turncurtovirus* which differ in host and vector species specificities, as well as genome organization. Geminiviruses are heavily prone to genetic recombination, both within, and across genera, and this has been proposed to be a significant cause of emergence of new viral species (Padidam *et al.*, 1999). In support of this hypothesis, the establishment of efficient global transportation networks over the past few decades has led to increased spread of these viruses as well as their vector species, causing significant outbreaks in 39 countries; events that have further promoted the importance of studying these pathogens (Moffat, 1999).

1.6.1 Host and vector specificity

The host range and vector specificity of a geminivirus are important factors used for their classification, and will also influence their geographic distribution. Mastreviruses (e.g. *Maize streak virus*) will usually infect monocotyledonous grass species (*Gramineae*), but in some cases dicotyledonous plants are also infected (Briddon and Markham, 1995). MSV is transmitted by the insect vector *Cicadulina mbila* Naudé (Storey 1925); a leaf hopper. Begomoviruses, distinct in that they typically consist of bipartite genomes, usually only infect dicotyledonous hosts, and are transmitted by the whitefly (*Bemisia tabaci*). *African cassava mosaic virus* (ACMV), a Begomovirus, significantly reduces crop yields of the cassava plant (*Manihot esculenta*), also a dietary staple in Africa. *Beet curly top virus* (BCTV) is a member of the Curtovirus genus, is

transmitted by the leafhopper (*Circulifer tenellus*), and is known for its cause of commercial loss of sugar beets (*Beta vulgaris*) in western USA, but is also known to infect 300 other dicotyledonous plant species. *Tomato pseudo curly top virus* (TpCTV) a Topocuvirus, is responsible for agricultural loss of tomatoes in Florida, but will also infect some legumes and other dicotyledonous plants. It shows many symptoms in common with BCTV, but is transmitted by the treehopper *Micrutalis malleifera;* the only geminivirus known to be transmitted by such a vector.

1.6.1.1 Transmission and relationship with vector

Vectors are organisms required to carry out viral transmission from one host organism to another, and are especially required by plant viruses because plants are themselves immobile. Once an infectious virion arrives at a new host it must find a means of entering the cytoplasm of a permissive cell allowing viral replication. Mechanical disruption of the cell wall by the vector may allow this, however, geminiviruses are not generally considered mechanically transmissible, and must be introduced into the vascular system of the plant (i.e. phloem sieve tubes) by an insect that feeds on the sugar containing fluids of these tissues. The virus particles will be ingested by the vector (a leaf- or tree-hopper) at the time that it feeds on an infected plant, and will move from the vectors gut into the salivary glands where it will be available for re-introduction into an un-infected host during another feed. The virus is unable to replicate in the vector, i.e. it is said to display a 'non-propagative' relationship.

1.6.2 Genome organization and replication

Important differences at the level of genome organization as revealed by viral genome sequencing and are shown in Figure 1-8. Geminivirus genomes are relatively small (2.5 - 3 kb, i.e M_w ~760 – 915 kDa) circular single-stranded DNA (ssDNA) molecules that are replicated in the nucleus of the host cell (Boulton, 2002; Davies *et al.*, 1987; Nagar *et al.*, 1995). Consequently, geminiviruses are class II viruses according to the Baltimore classification scheme (Baltimore, 1971). Other examples of a single-stranded viruses include the DNA bacteriophage ϕX174 (the type species *Microviridae*) as well as well as members of the *Nanoviridae*. The *Parvoviridae* are ssDNA viruses with linear genomes, and infect mammals and insects (Chapman and Rossmann, 1993). Notably, these virus families have a $T = 1$ capsid architecture, and replicate by mechanisms known as rolling-circle or rolling-hairpin replication, respectively.

Upon entry into the nucleus, the ssDNA geminivirus genome is converted to a double-stranded DNA (dsDNA) intermediate by host DNA replication machinery. The dsDNA intermediates associates with histones (structural components of the nucleosome) to form minichromosomes (Pilartz and Jeske, 1992). This dsDNA intermediate serves as the template for bidirectional

transcription of viral genes as well as genome replication by rolling circle replication (Heyraud *et al.* 1993a; Heyraud *et al.* 1993b; Saunders *et al.* 1991, 1992). Replication is initiated by the viral replication associated protein (Rep) from a conserved nonanucleotide site (TAATATT/AC) in the long intergenic region (LIR) of the geminivirus genomes (Gutierrez 2000a, b; Hanley-Bowdoin *et al.* 2000) and is illustrated in Figure 1-8. This sequence motif adopts a stem-loop structure, and presents a site for nucleolytic cleavage, the substrate of the Rep protein, allowing rolling-circle replication to proceed.

1.6.3 Maize streak virus

Maize streak virus (MSV), the subject of this study, is the prototypical member of the *Mastrevirus* genus, and is considered the most severe maize (*Zea mays*) disease in sub-Saharan Africa, and has as a result, been the subject of extensive research (Bosque-Perez 2000; Boulton 2002). MSV has the most dramatic consequences for subsistence farmers who rely on maize as a dietary staple, and consequently can place families and communities at risk. Breeding efforts have resulted in MSV resistant maize stains, as well as the development of recombinant maize species that have induced resistance as a result of stable expression of a mutated viral replicase protein gene from a construct which was integrated into genome of the plant strain (Shepherd *et al.*, 2007).

Named after its characteristic symptoms; originally described as 'mealie variegation' (Fuller 1901). It was later renamed 'maize streak' by the researcher who identified the insect vector responsible for virus transmission to be *Cicadulina mbila* Naudé (Storey 1925). MSV encodes four gene products on a single (monopartite) single-stranded DNA (ssDNA) molecule that is 2687 nt in size (Francki *et al.*, 1980; Lazarowitz, 1988; Mullineaux *et al.*, 1984). Two of the gene products (Rep and RepA) are associated with replication of the genome and are expressed early in the virus life-cycle from a complimentary sense (C-sense) transcript. The remaining two genes are virion-sense (V-sense), and encode a movement (V2) and coat protein (V1), respectively.

1.6.3.1 Life-cycle, tissue specificity and role of viral genes

The life-cycle of MSV (Figure 1-10) begins with entry of the circular ssDNA genome into an uninfected cell by passage from the phloem sieve tubes of vascular bundle into an uninfected cell. The ssDNA genome must then be trafficked into the nucleus where the host cell DNA repair and replication machinery resides. Nuclear trafficking is mediated by the karyophillic coat proteins (CPs) that each possess a nuclear localization signal (NLS) (Liu et al., 1999). By an unknown mechanism, the virion is then disassembled to render the genome accessible to host factors, and the ssDNA genome converted to a double-stranded form by host replicases within the nucleus (Palmer and Rybicki, 1998). This dsDNA form is then transcribed

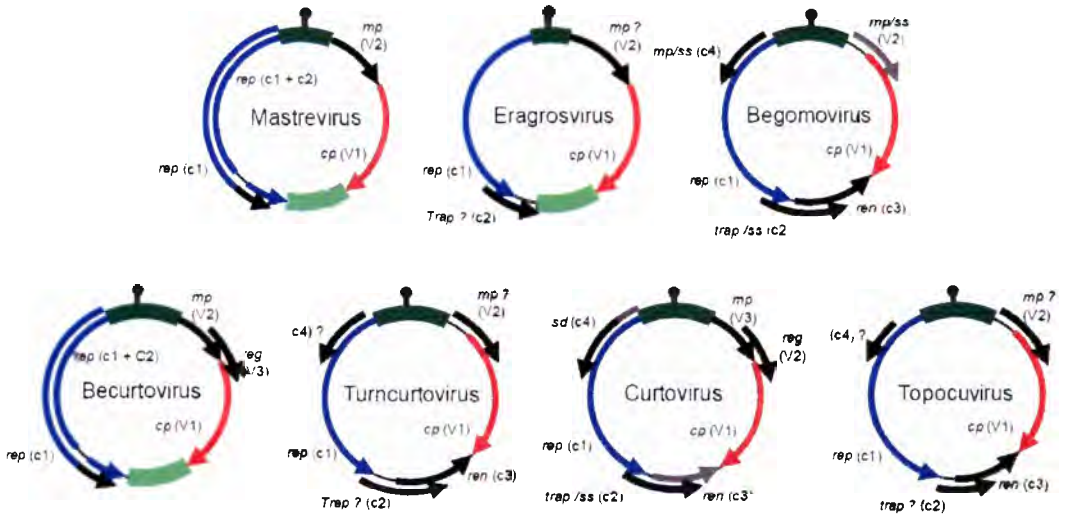

Figure 1-8 Genome organisation of geminiviruses. The organization of circular single-stranded DNA genomes of the geminivirus genera are shown by schematic and described here with specific reference to MSV. Geminiviruses of the genera *Mastrevirus*, *Becurtovirus*, *Turncurtovirus*, *Curtovirus* and *Topocuvirus* possess monopartite genomes (i.e. a single component or segment of circular ssDNA), whilst members of the genus *Begomovirus* most often possess bipartite genomes (two segments referred to as A and B, but only the A segment is shown); however a monopartite begomovirus has been identified (ToLCV). The site of replication initiation (a conserved nonanucleotide motif, TAATATTAC) is illustrated schematically for each genome as grey stem-loop. c1 (blue ORF), encodes the replication associated protein (rep) required for virus replication. Mastreviruses and becurtoviruses are the only geminiviruses to encode Rep as a spliced mRNA produced by processing of a transcript of c1 and c2. In Mastreviruses expression of c1 alone produces a product known as RepA. The virion sense strand encodes movement (mp, grey ORF) and coat (cp, red ORF) proteins. The MSV genome is characterized by a long-intergenic region (LIR, dark green) and a short-intergenic region (SIR, light green) that separate genes encoded in virion (V) and complementary (C) senses. Figure provided by Dr Arvand Varsani.

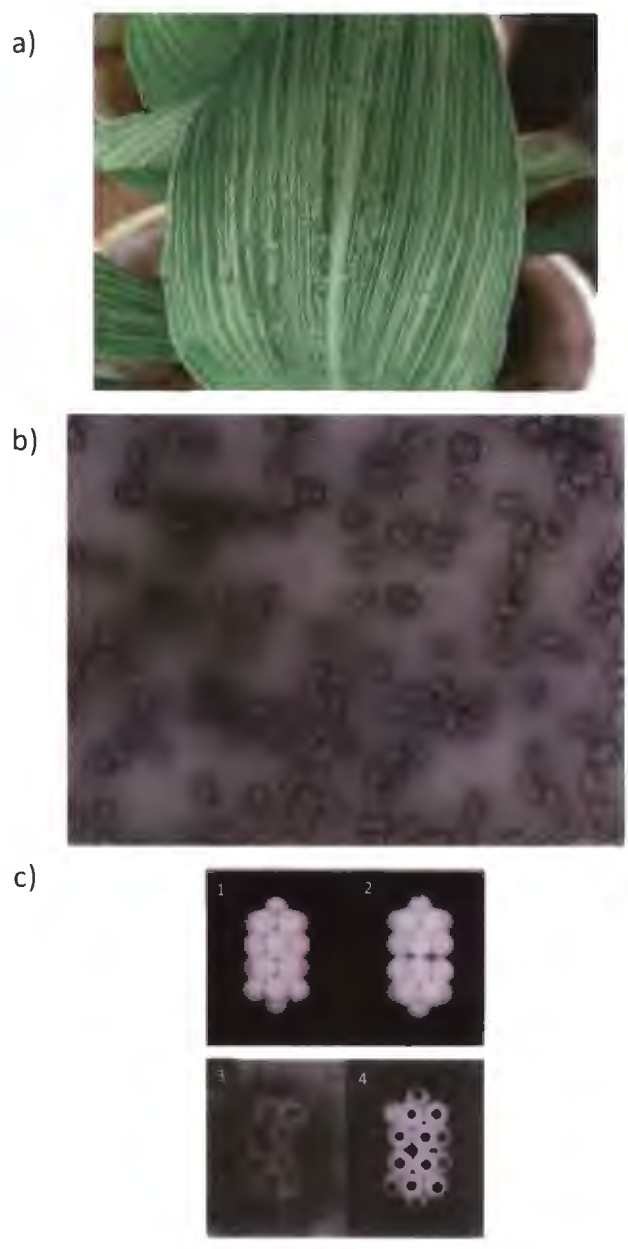

Figure 1-9 Symptoms, negative-EM and early architectural modelling of MSV. a) A leaf of a maize plant infected with Maize streak virus. The characteristic "streaked variegation" pattern is apparent. Figure reproduced from (Shepherd et al., 2010). b) Negative-stain electron microscopy of purified capsids of the geminivirus species Chloris striate mosaic virus (CSMV). c) "Ping pong" models of MSV architecture originally proposed by Hatta and Francki (1979). The ping-pong spheres are intended to represent the pentameric capsomers (CP_5) which self-assemble to produce the capsid. Two models in c,1) and c,2) depict different rotational offsets between the two incomplete icosahedral heads. In c,3) and c,4) locations of pentameric units is mapped onto the experimentally determined negative-stained image of the particle and compared to the ping pong models. The model produced during this study did not identify an offset between the heads leaving a 'gap' in our understanding of the bipartite architecture. Figure a) adapted from Figures b) and c) adapted from Hatta and Francki 1979.

to produce viral replication-associated proteins (Rep). After translation in the cytoplasm, Rep enters the cell nucleus and converts the dsDNA to a replicative form, which is then continuously replicated to produce many progeny viral genomes by a process known as rolling-circle replication (RCR) (Palmer and Rybicki, 1998). As ssDNA virion sense strands are displaced during this process, they are ligated to produce the closed-circular ssDNA genomes (Erdmann et al., 2010).

Systemic infection and spread of the virus through the host to eventually produce disease symptoms is mediated by the virion sense gene products: movement protein (MP or V1) and coat protein (CP or V2), and these are only expressed late in the viral life cycle (Liu, 1997). The CP is expressed from the same transcript as the MP and serves multiple functions including accumulation and subsequent encapsidation of genomic ssDNA (Liu et al., 1997a). MSV virions accumulate in infected host cell in quasi-crystalline arrays (Pinner et al., 1993), but these will be moved to the cell-periphery by the MP where they will either pass to an adjacent cell or to the phloem where they will be ingested by vector insects when they feed on the plant. Vectors acquire the virus as a result of specific interaction with the CP, which has been identified as the primary determinant of vector specificity (Boulton *et al.*, 1989b; Briddon *et al.*, 1990). The MP facilitates egress of the newly produced virions from the nucleus of the host cell by a specific interaction with the coat protein-genome complex (Liu *et al.*, 2001a). Two types of movement have been identified and are cell-to-cell (via plasmodesmata) and systemic movement (via phloem vessels); both are required for systemic infection. By mechanisms currently unknown, MP is thought to alter the selectivity properties of both the nucleus and plasmodesmata to allow this movement to take place (Liu et al., 2001a).

The host cell types that are permissive for MSV replication are important considerations for understanding how this organism spreads. Because MSV is delivered into the phloem by its vector it may be expected that the virus is limited to phloem-associated cells. It has been demonstrated that in mature leaves, MSV is able to move past cells forming the vascular bundle and is also found is mesophyll cells (Pinner *et al.*, 1993). Investigation relying on immunocytochemistry and *in situ* hybridization of prepared tissue samples has shown that MSV can infect a range of cell types in the context of leaf tissue (e.g. bundle sheath, parenchyma, and mesophyll cells), but in stem regions it limited to cells of vascular bundle, and not able to enter the apical meristem (Lucy *et al.*, 1996). It has been suggested that cell division is required for geminiviral replication (Townsend *et al.*, 1986), a fact supported by work on cell protoplasts as accumulation of viral DNA correlated with the onset of cell division (Boulton *et al.*, 1993), with further work suggesting that the S phase in particular was important. Evidence also exists to suggest that host cell mitosis

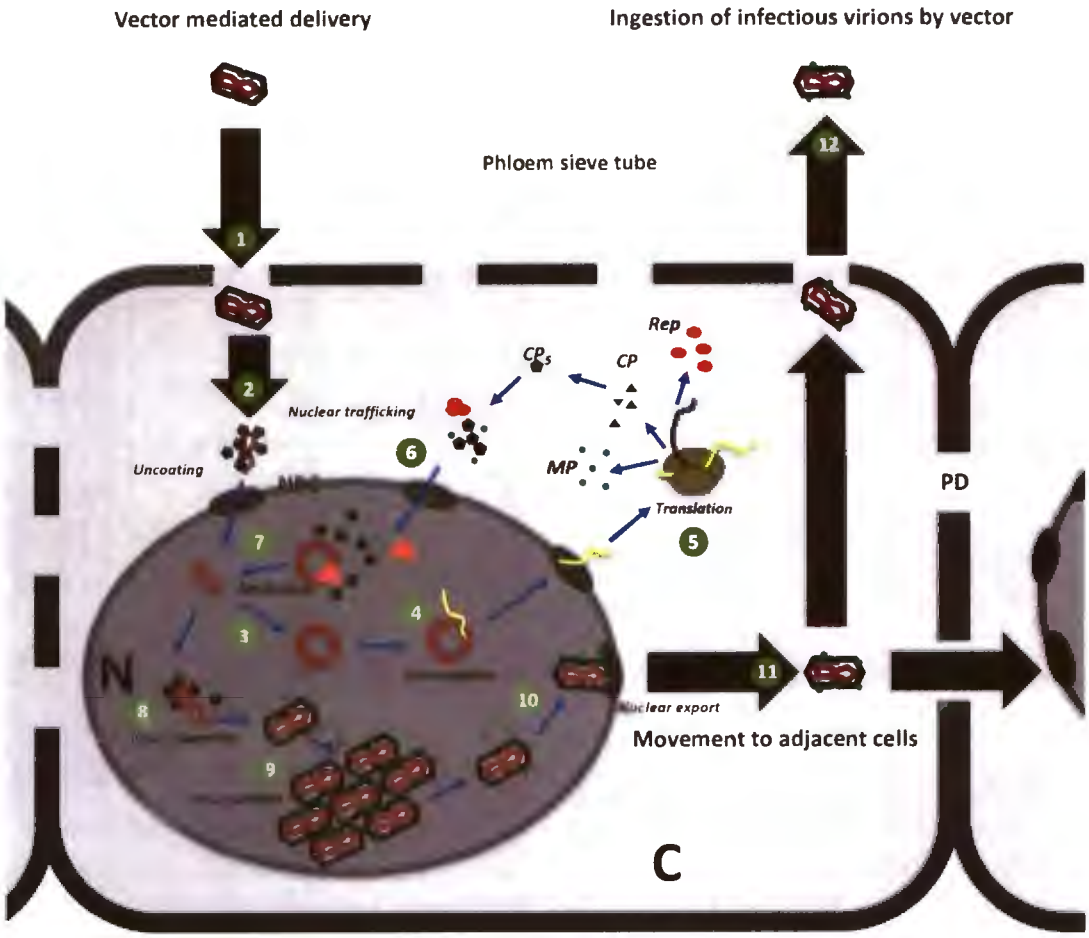

Figure 1-10 Infectious cycle of Maize streak virus. The life-cycle of MSV is summarized into 12 steps. Initially the virus must enter the host cell cytoplasm (C) (1), following this the viral capsid is destabilized or disassembled to expose nuclear localization signals on the N-terminal arms of the viral coat protein subunits. Following nuclear trafficking into the nucleus (N) (2), the single-stranded viral genome will be converted into a double-stranded intermediate (3), which will be bi-directionally transcribed by the host RNA polymerase. Viral mRNAs will be translated by ribosomes in the host cytoplasm, and gene products such as the MP (small green spheres) viral Rep (red ellipsoids) and CP_5 units (black pentagons), will also be trafficked into the nucleus (6). The viral Rep carries out viral replication from the double-stranded template to produce progeny ssDNA genomes (7). Binding of viral CPs late in the infectious cycle will prevent conversion to dsDNA (step 3), and result in assembly of infectious virions (8). Virions will accumulate inside the nuclei of infected cells as quasi-crystalline aggregates (9). By poorly understood mechanisms, the viral MP will interact with the viral CP to effect nuclear egress of virions such that cell-to-cell movement can occur via plasmodesmata (PD, 11). Systemic movement will involve cell egress into the vascular bundle (phloem vessels), where host-to-host transmission will be facilitated by vector insects by feeding on the host plant (12). Figure drawn to summarize information available in the literature, refer to text for further information.

is not required for viral replication to take place (Lucy *et al.*, 1996; Timmermans *et al.*, 1992). More detail of the roles of each viral component are discussed in the following sections.

1.6.3.1.1 Replication-associated proteins (Rep and RepA)

All geminiviruses express a replication-associated protein (Rep) and replicate their ssDNA genomes using rolling circle replication (RCR) (Morris-Krsinich *et al.*, 1985). Other modes of replication have also been described such as complementary-strand replication (CSR) and recombination-dependent replication (RDR) (Jeske *et al.*, 2001; Jovel *et al.*, 2007; Preiss and Jeske, 2003; Saunders *et al.*, 1991). Replication of MSV occurs via a series of intermediates that have been characterized by 2D gel electrophoresis (Erdmann *et al.*, 2010), and have confirmed that MSV genomes may replicate via all three of the above mentioned mechanisms.

Rep will bind to the dsDNA intermediate at the replication origin and create a nick in the positive-strand of the stem-loop to form an open-circular dsDNA intermediate that serves as the template for RCR. The freed 3'-OH group of the positive-strand then serves as a template for rolling circle replication by the host cell polymerase while the 5'-OH remains covalently associated with Rep. The newly formed circular ssDNA molecule will either be packaged by the CP for transmission, or be converted to another dsDNA intermediate for further transcription or replication. The complementary (negative) sense open reading frames (ORFs) C1 and C2 are responsible for encoding Rep and RepA. Rep is produced by splicing a C1-C2 transcript to remove a single intron; in this manner C2 encodes for the C-terminal domain of the full-length Rep. The *Mastreviruses* are unique amongst geminiviruses in that they also express a truncated form of Rep known as RepA. RepA is the translation product of the unspliced transcript with translation terminating in the C2 ORF (refer Figure 1-8). Genome replication can only be initiated after the Rep gene has been expressed. While RepA, unlike Rep, is not absolutely required for genome replication (Liu *et al.*, 1998; Schalk *et al.*, 1989) is it believed to have important roles which are mediated by interaction with both virus and host factors (Boulton, 2002; Castellano *et al.*, 1999; Shepherd *et al.*, 2005).

1.6.3.1.2 Coat protein (CP)

The CP is thought to bind both dsDNA and ssDNA in a non-sequence dependent manner (Liu *et al.*, 1997b). Its primary function is the encapsidation of viral genome, however, CP is also heavily involved in the movement of the genome through the host. CP is required not only for translocation of the ssDNA genome into the nucleus immediately post-infection (Liu *et al.*, 1999), but also cell-to-cell, and systemic movement through the host via the vascular bundle (Boulton *et al.*, 1993, 1989b; Lazarowitz *et al.*, 1989; Liu *et al.*, 2001a). Whether the capsid is assembled and intact during nuclear trafficking (nuclear entry) is presently unknown, but a single amino-acid

mutation in CP that abrogates capsid self-assembly has been shown to abolish viral movement (Liu *et al.*, 2001b), suggesting that assembly to form the virion is essential. Nuclear trafficking of the genome is mediated by a bipartite nuclear localization signal (Robbins *et al.*, 1991), present in the N-terminal residues of the coat protein (Liu *et al.*, 1997a).

In addition to multiple host interactions, the CP of many geminiviruses also determines vector specificity (Briddon *et al.*, 1990). The nucleic-acid binding domain has been shown by truncation experiment to reside in the N-terminal domain of the protein (Liu, 1997). The first 20 N-terminal amino acids are rich in basic residues that are presumed to bind the genome during assembly. At the time of formation of the virion, the CP must sequester ssDNA for packaging, and prevent it from being converted to a double-stranded (ds) intermediate.

1.6.3.1.3 Movement protein (MP)

The MP (101 amino acids, 10.9 kDa) is expressed late in the viral life cycle and is responsible for translocation of progeny virions from the nucleus to the cytoplasm, as well as from the cytoplasm of one cell to another adjacent cell via plasmodesmata. MP is thought to facilitate cell-to-cell spread by increasing size exclusion limit of plasmodesmata (Boulton, 2002; Liu *et al.*, 2001a), a fundamental constraint on the size of the viral genome (Gilbertson *et al.*, 2003), and immune-electron microscopy has revealed its association with plasmodesmata at the same time that symptoms (viral lesions) are observed (Dickinson *et al.*, 1996). Notably, aside from the non-coding LIR of the genome, the MP is the most sequence variable component of the mastreviruses (Rybicki, 1994).

1.6.3.2 The structure and symmetry of geminiviruses

The coat protein is the only protein present in the MSV virus particle (Bock *et al.*, 1977). Previous studies have reported that the coat protein is able to undergo self-assembly to form a architecturally unique bipartite structure that possesses global D5 symmetry and was originally determined to be approximately 18 by 30 nm in size. The first architectural model was reported by Hatta and Francki (Francki et al., 1980; Hatta and Francki, 1979) after inspection of a negatively stained preparation of virions and describes an architecture in which the global fivefold axis coincides with the 5-fold point group symmetry of the apical capsomers. In addition to this C5 axis, the structure has 5 global 2-fold symmetry axes perpendicular to the long axis, at the equator of the particle. This arrangement is referred to as dihedral symmetry, abbreviated in this case as 'D5'. The model describes a capsid that is formed by the association of two incomplete icosahedra, which are each formed by 11 pentameric capsomers (55 coat protein molecules per incomplete icosahedral unit). Hence the full virion consists of 22 pentameric capsomers and 110 subunits (Hatta and Francki, 1979). By analysis of thin sections of plant tissue as well as pelleted

nuclei of cells from *Chloris guyana*, Hatta and Francki went on to show that the geminate morphology was also the predominant species present *in situ*.

There have been reports that the capsid is able to form monopartite and tripartite assemblies as well, and that the formation of these is genome size dependent. Monopartite capsids have been shown to package sub-genomic fragments in the case of the monopartite capsids (Casado *et al.*, 2004), and multi-partite (> 2 heads) are likely to encapsidate concatameric species (nucleic acids representing multiple sequential copies of the full-length genome) that may arise during rolling-circle replication, respectively (Dr Arvind Varsani, personal communication).

1.6.3.3 Electron microscopy and 3D reconstruction of geminiviruses

While negative-stain EM has allowed a general model for the capsid architecture of geminiviruses to be proposed, it was only *3D reconstruction* of virus frozen in vitreous ice (i.e. cryo-EM) that allowed the architecture of this family of viruses to be studied further in a physiologically relevant state; *MSV-N* (Zhang *et al.*, 2001); *ACMV* (Bottcher *et al.*, 2004). The first 3D cryo-EM reconstruction of a geminivirus (MSV) was published 100 years after the first description of this disease. Cryo-EM structures of African cassava mosaic virus (Begomovirus, ACMV) followed. In the case of ACMV, the density gradients used to purify the particles showed a "top" (T) and "bottom" (B) fraction. 3D reconstruction supported the experimental observation that the bottom virions are denser than the top (presumably containing more nucleic acid for an unknown reason). Both studies reported that the virion is larger than previously estimated (~ 22 x 38 nm). Additionally, both studies show that the architecture exhibits an offset between the two heads, and showed that the geminate architecture is likely to be conserved across the four genera. In the case of ACMV, the diameter of the $T = 1$ heads appeared to be different by 10 Å (1 nm), with the bottom being estimated as having a diameter of 22.5 nm. The findings of these studies are summarized in Table 1.0.

If the geminivirus CPs were to assemble to form a $T = 1$ icosahedron, the capsid subunits would exist in identical conformations, as observed STNV (Section 1.3.2.1). Even at the relatively low resolutions achieved by the cryo-EM studies, it was obvious that this was not the case for the bipartite $T = 1$ structure of geminiviruses. Conformational differences are observed locally when comparing pentamers in different regions of the reconstructed capsid. Equatorial capsomers (i.e. those mediating interactions between geminate heads) show a deviation from the local 5-fold symmetry at all other capsomers, however, the precise nature of these conformational differences was difficult to assess owing to the resolution achieved.

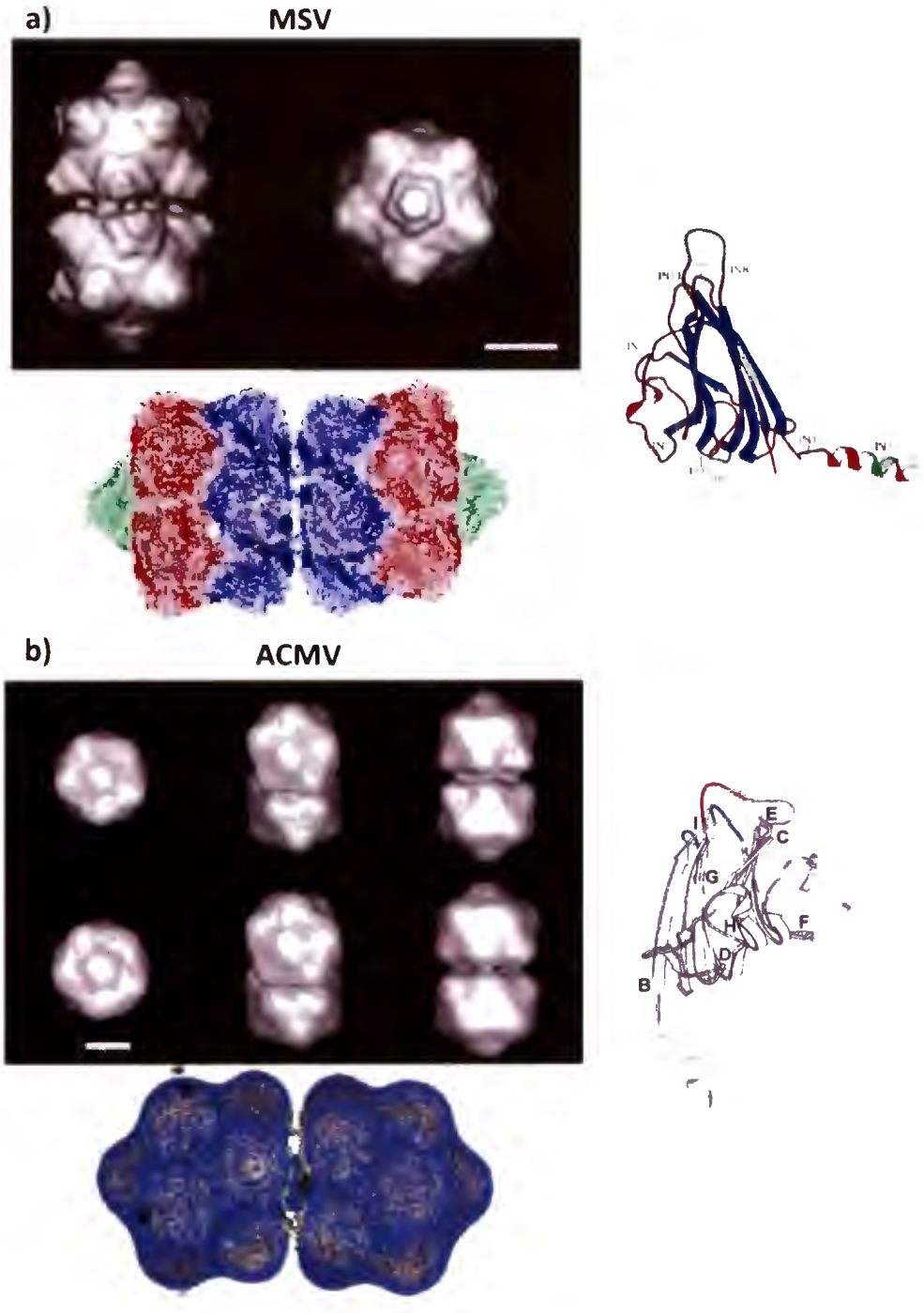

Figure 1-11 Structural studies of geminiviruses by single-particle electron microscopy. 3D reconstructions, and modeling of a) *Maize streak virus* (MSV), and b) *African cassava mosaic virus* (ACMV) are shown. a/b top, left) Surface representations of the 3D reconstructions are showed viewed along principal axes of global symmetry (5-fold, and 2-fold point-group rotational symmetry). (a/b, right) Atomic models of the geminivirus coat proteins are shown as ribbon diagrams in which the β-strands (B-I) are annotated. a/b, bottom, left) The fit of the comparative models is shown with respect to the experimental density maps. Figures adapted from a) (Zhang *et al.*, 2001) and b) (Bottcher *et al.*, 2004).

Table 1 Properties of 3D cryo-EM reconstructions of geminiviruses

Virus	Size (nm)	No. of images	Specimen prep.	Resolution (FSC 0.5)	Reference	Host
MSV-N	22.0x 38.0	1516	Cryo-EM In solution	25 Å	Zhang *et al.* 2001	*Zea mays*
ACMV-B	22.5x 38.0	5530	Cryo-EM (carbon film)	19 Å	Bottcher *et al.* 2004	*Nicotinia benthemania*
ACMV-T	21.5x 38.0	Unknown	Cryo-EM (carbon film)	19 Å	Bottcher *et al.* 2004	*Nicotinia benthemania*

1.6.3.4 Topology and architecture of the coat protein of geminiviruses

Owing to low resolution, the EM studies carried out so far have not been able to determine the precise geometry of the geminivirus CP polypeptide backbone. Hence, assignment of the geminivirus CP fold has relied on structural bioinformatics and protein modelling. Both previous cryo-EM studies relied on protein sequence similarity search and alignment of hundreds of related sequences to confirm that the closest structural homologue - presently known - is the CP of STNV. The authors (Bottcher *et al.*, 2004; Zhang *et al.*, 2001) noted that with significant divergence between geminivirus CPs (at a sequence level), direct alignment of any two sequences was impossible. With such multiple sequence alignments in place, structural prediction of the CP by computational homology modelling (described in Section 1.5) could be carried out to produce a "pseudo-atomic" model that could be used to probe the EM reconstructions.

1.6.3.4.1 Structural interfaces of geminivirus capsids

As described in the previous section, the previous cryo-EM studies used homology modelling to probe the manner in which pentameric capsomers (abbreviated to "CP_5" herein) interact to form the assembly. Both studies found that in comparison to STNV the radius at which the capsomers had to be placed in order to fit in the maps was greater. Both studies observed a regions of density between CP_5 units that was not accounted for by their atomic models, and both studies suggested that the N-terminal arm may exist in an atypical conformation that would mediate interactions between CP_5s by filling this density; although the resolution of the reconstructions was insufficient to confirm that this was the case without some ambiguity.

With regard to regions of the CP responsible for interactions across the equator of the geminate particle (between heads), the two studies differed. Zhang *et al.* (2001) observed a total of 10 interactions across the equator, and were able to adjust the conformation of the N-terminal arms such that 5 of these interactions were accounted for by the structure. The conclusion of both studies was that specific and rigid interactions link the two quasi-icosahedral heads, but no description of the conformational differences allowing the formation of these interactions was reported.

1.6.3.5 Geminate assembly and disassembly

Assembly of MSV virions takes place in the nucleus of the cell as circular virion-sense ssDNA species accumulate during viral replication. An interesting aspect of virion assembly is the ability of the CP subunits to discern virus nucleic acids from host, although how this is achieved in the case of the geminiviruses still needs remains to be elucidated. It has been proposed that this packaging selectivity or specificity might be signalled by the nonanucleotide stem-loop that acts as a site of nucleation for the viral coat protein (Qin *et al.*, 1998). In this way such an association would also prevent conversion of the single-stranded genome into a dsDNA intermediate required for replication. The structural basis for this stem-loop DNA-protein interaction has not been determined, however, it has been demonstrated that stem-loop interactions are responsible for packing and assembly of STNV (Bunka *et al.*, 2011; Ford *et al.*, 2013), and this behaviour was originally observed for the helical virus TMV. It is possible that other "DNA sequence requirements", besides a single specific encapsidation signal also exist, and the ssDNA genome has also been implicated as being involved in geminate particle formation by the basic observation that no empty particles are observed (Francki *et al.*, 1980). It would seem that unlike some viral families (particularly of the Picornavirus superfamily), geminiviruses do not produce empty capsids. This observation heavily supports an important role of the genome in particle formation.

The MSV CP has been observed to self-assemble to form two distinct closed assemblies. These are a icosahedral assemblies that consists of 60 copies of the CP, and a geminate particle that consists of 110 copies of the CP, however, the processes that affect this multimerization are poorly understood. Defective interfering particles of ACMV have been described, which formed icosahedral particles that appeared to be missing one CP_5 (Frischmuth *et al.*, 2001). However, in the case of MSV, density gradient centrifugation allowed apparently 'complete' icosahedral capsids to be separated from geminate capsids. Analysis of the nucleic acid species present in the fraction corresponding to single capsids showed that up to 1.4 kb of ssDNA was packaged in these particles, as either a single segment or possibly two segments of smaller sizes. These sub-genomic DNAs were typically of included regions of the C-sense genes, and more often than not

included the stem-loop structure (origin of replication) required to initiate rolling-circle replication (Casado *et al.*, 2004).

Conditions favouring disassembly of geminiviruses have been investigated for ACMV using negative-stain and rotatory shadowing EM (Kittelmann and Jeske, 2008). These particles demonstrate reduced stability at alkaline pH conditions and investigators were able to observe disassociation of capsomers from certain locations in the virus particle. At these sites egress of the genome was also observed suggesting that within the cell, conditions must arise that result in similar behaviour. The study of ACMV disassembly revealed that either apical or peri-apical capsomers dissociate in the majority of cases. The investigators proposed that even though the genome emerges from the interior of the virion it may still be bound to the dissociated capsomers.

A review of the literature showed that no studies have been carried out to reconstitute geminate virions from disassembled and subsequently isolated viral components. Factors influencing viral assembly have been inferred based on mutagenesis, and inspection of mutant particles isolated from plant material.

1.6.3.6 Emergence and evolution of geminiviruses

All viruses can be considered modular entities, consisting of molecular components that each have independent functions and also independent evolutionary lineages (Abrescia *et al.*, 2012). It is for this reason that recombination is an important aspect of DNA viral evolution, because it allows modules to be swapped between two parent viral strains, helping to create variants with potentially improved fitness within a particular environmental context. Experiments has shown that recombination in geminiviruses will not always produce viable variants, and it has been demonstrated that recombinant variants are selected for those which preserve "intra-genomic interaction networks", a property which favours recombination between closely related viruses (Martin *et al.*, 2005). Further experiments using divergent geminiviruses have been used to show that this approach can actually be used to predict such interaction networks, providing fundamental insight into the molecular interactions underpinning the life-cycles of these organisms (Martin *et al.*, 2011).

The replication-associated protein (Rep) is one such modular component that has fuelled discussion regarding geminivirus emergence. The geminivirus Rep proteins show three important similarities in key motifs required for replication initiation that were first identified in the Rep proteins of Gram-positive bacterial plasmids, suggesting that geminiviruses may have descended from these plasmids (Koonin and Ilyina, 1992). Previous studies have suggested that geminiviral genomes may have descended from an extrachromosomal plasmid of red algae, some 450 million

years ago (Nawaz-ul-Rehman and Fauquet, 2009). More recently, phylogenetic analysis has revealed that the geminivirus Rep is also closely related to those of phycoplasmal plasmids (Saccardo *et al.*, 2011); phycoplasma being pathogenic phloem limited bacteria that lack a cell wall. Like geminiviruses, phycoplasma also rely on insect vectors for movement. How such replicons acquired a coat protein to allow them to move from host to host becomes an important question, although it is presumed that a rare recombination event is at the foundation of the emergence of these entities.

There is some debate regarding which component determines the "virus self" – genome or capsid. Considering that it is the encapsidation behaviour of viruses that distinguishes them from plasmids (Forterre, 2010; Raoult and Forterre, 2008), it has been proposed that it is the viral coat protein that gives a virus its "identity" (Abrescia *et al.*, 2012, 2010). It has been shown by empirical observation of structural homology at the protein tertiary level (i.e. structural similarity of viral coat proteins), is able to inform us about homology which cannot be observed at either nucleic acid or protein sequence level, and has allowed evolutionary relatedness to be inferred between viruses infecting all kingdoms of life; bacteria, archaea, eukarya (Bamford, 2002; Bamford *et al.*, 2005). Applying this approach of "structure-based viral lineages" to the geminiviruses, it has been proposed that these arose as a consequence of a recombination event between a phycoplasmal ssDNA plasmid – encoding the Rep, and a (+) ssRNA plant virus – encoding the coat protein CP (Krupovic *et al.*, 2009). Hence the coat protein (CP) of the plant virus was 'recruited' to become the "packaging machine" of a self-replicating nucleic acid that encoded a specific replication associated protein (Rep), while also possessing the nucleic acid secondary structure which allowed it to be replicated by RCR. This hypothesis has since been refuted by the argument that such a scenario is unlikely, and that, based on sequence analysis, the phycoplasmal plasmid itself is instead likely to have arisen from geminiviruses - and not the other way round (Saccardo *et al.*, 2011). Saccardo *et al.* went further to show that the acquisition of a coat protein is more likely to have occurred within a marine environment from another donor, rather than a plant virus replicating in a terrestrial plant host.

In addition to high rates of recombination, geminiviruses show high rates of mutation (10^{-4} substitutions per site per year); rates that are comparable to that of RNA viruses (Duffy *et al.*, 2008; Harkins *et al.*, 2009; Monjane *et al.*, 2011). This is intriguing because these viruses are replicated by high fidelity host DNA polymerases, unlike RNA viruses. The mutational processes that account for this behaviour have been observed to be single-strand specific, leading to various hypotheses that have attempted to explain the behaviour. A recent study has led to the suggestion

that it may be high-levels of oxidative damage to the single-stranded form that accounts for these relatively high rates of mutation (Monjane *et al.*, 2012).

1.7 Aims of this work

Investigating the structure and assembly of infectious virions or virus-like-particles (VLPs) serves to inform our understanding of the underlying principles governing macromolecular interaction. Molecular recognition, and subsequent association and assembly are fundamental processes in the "molecular life" of the cell (Robinson *et al.*, 2007), and viruses have long proved important model systems for studying the structural basis of these mechanisms. As discussed in the previous sections, the geminate capsid morphology is unique, and certainly a curiosity that demands study from the perspective of self-assembly. Is this assembly built on principles of conformational switching and interaction used by typical $T = 3$ viruses, or are different factors at play in the assembly and evolution of this particle? The morphology certainly implies unique behaviour during virus assembly and in spite of previous study, the quasi-equivalent coat protein (CP) conformations required to mediate assembly of the geminate particle have not been characterized. Consequently, in an attempt to shed light on these properties the geminate capsid architecture, as well as its implications for assembly of the geminate virion, the capsid of MSV is the subject of study described in this thesis, as previous work (Bottcher *et al.*, 2004; Zhang *et al.*, 2001) has shown that MSV virions can be isolated at sufficient concentrations for such work, and there is a long history of work carried out on this virus at the University of Cape Town.

The principal aim was to use already established structural biological methods (cryo-EM and homology modelling) to infer the CP-CP interactions, and architecture of MSV. If a cryo-EM map of sufficiently high resolution can be determined (e.g. 7.5 – 15 Å), this can be used in conjunction with protein structure prediction and flexible fitting to derive a pseudo-atomic model of the capsid without the requirement of crystallization. Using these methods, a plausible and hence empirically relevant model of residue-residue interactions can be constructed *in silico,* which may further inform antiviral strategies undertaken in the future, as well as elucidating the functional requirements which underpin the evolutionary divergence that the gemini-architecture represents. The number of biochemical and molecular studies already carried out on geminivirus coat proteins would benefit greatly from such a model if it were to be made publically available (e.g. in the Electron Microscopy Data Bank (EMDB;http://www.ebi.ac.uk/msd), and further experiments could be designed in a rational manner using the experimental model as a reference point.

In this thesis, cryo-EM, 3D reconstruction, and protein modelling were applied to achieve the following ends:

- Determine the structure of MSV geminate capsid to an improved resolution (< 25 Å).

- Determine the structure of the MSV icosahedral capsid to ~ 20 Å resolution.

- Determine the geometric parameters associated with each architecture; i.e. understand how the capsid building block (CP_5) is arranged within each capsid.

- Construct pseudo-atomic models using information from the EM study, as well as other sources (e.g. atomic structures of structural homologues).

- Use the above mentioned, to determine interacting regions/residues between MSV CP subunits/pentameric building blocks.

With these objectives in place, it was hoped that analysis would shed some light on the following questions:

-What is the structural/geometric basis of geminate morphology?

-How is the geminate morphology related to $T = 3$ viral capsid architectures? As both are essentially effective solutions to the problem of packaging a genome larger than what can fit inside a $T = 1$ architecture.

- How is the ssDNA genome organized within the MSV capsid, and how does it interact with the MSV capsid?

- What possible roles does the genome fulfill during assembly?

- Based on sequence alignment with other geminivirus CPs, can structural information about MSV be used as a basis for statements regarding the architecture of other geminiviruses? Specifically, is the precise geometry of the architecture conserved amongst related geminiviruses?

- Are interactions between pentameric capsomers conserved in the architectures of MSV and Satellite tobacco necrosis virus (STNV), or have these interactions diverged along with the coat protein sequences?

- What changes in conformation take place in equatorial capsomers?

- Can structural information inform us as to how this architecture might constrain viral evolution?

- Given that the genomes of other simple viruses have been implicated in capsid assembly, is it possible that some structural property of the geminivirus genome favours assembly of the geminate architecture?

- It is important that these requirements for particle formation, and role of the genome in this process, are understood, as any attempts to exploit the geminate coat proteins in biotechnology will require that their assembly behaviour and requirements are well understood.

The original reconstruction was performed without correction of the microscope contrast transfer function (CTF), and a result was limited to a resolution of 25 Å. Many questions remain unanswered in the literature regarding the nature of interactions between the MSV CP, its genome as well as the movement protein, it was felt necessary to repeat the previous study using more recent tools of structural modelling and analysis.

1.7.1 A note on terminology

Terms used to describe structural features of viruses with respect to those observed in raw electron micrographs, and also 3D reconstructions, are distinct from those used to describe structural and chemical features of the actual virions themselves. The terms 'capsomere' and 'morphological unit' refer to pronounced observable features in EM data; and are particularly revealed in low-resolution studies because they reflect clusters of subunits; these being the most visible features at 'macromolecular resolutions' of 3-5 nm (30-50 Å). In reference to the capsid as an assembly of coat protein subunits, the term 'capsomer' refers to a physical building block of the capsid, e.g. in this study the capsid of MSV is thought to be assembled from a series of pentameric capsomers (CP_5), each consisting of five coat protein subunits. Owing to the arrangement of these in the MSV capsid shell, the CP_5 also gives rise to the morphological units (or capsomeres) apparent in the EM reconstructions. Depending on the virus under investigation, these two terms may or may not refer to the same biological/macromolecular arrangement. With respect to quasi-equivalence, coat protein subunits identical in sequence, may be described to exhibit 'structural polymorphism' in which cases regions of the coat protein adopt different structures or conformations. These different but closely related conformations are referred to as 'conformers'. Herein, the monopartite (icosahedral) capsid reconstruction carried out by application of 60-fold symmetry is sometimes referred to as the '532 reconstruction'. Bipartite (geminate) reconstructions, carried

out by applying 10-fold (D5) symmetry, possess D5 symmetry and may also be referred to as such.

2 Materials and Methods

2.1 Sample preparation

2.1.1 Agro-inoculation of maize

Maize seedlings were infected by Agrobacterium-mediated DNA transfer (Boulton *et al.*, 1989a; Grimsley *et al.*, 1989); a procedure termed "agro-inoculation". An agro-infectious construct with MSV-MatA was provided by Darren Martin (Martin *et al.*, 2001). Briefly, sweetcorn kernels (Golden Bantam) were germinated and grown for three days at 30°C. Seedlings were cleaned and injected with a solution of agrobacterium cells at a site near the plant meristem. Plants were allowed to grow for 4 weeks, and showed obvious signs of successful infection at 2 weeks.

2.1.2 Virus isolation

Virions were isolated from symptomatic leaves according to a previously established protocol (Palmer et al 1998). Symptomatic maize leaves were harvested from infected plants and homogenized in a blender in an equal weight/volume of acetate buffer at pH 4.8. The homogenate was then passed through two layers of muslin cloth to remove crude plant material, and the pH adjusted to 4.8 using glacial acetic acid. Finer cellular materials that had aggregated owing to the acidic pH, were then spun out of solution by low-speed centrifugation at 12,000 g for 10 minutes using a Sorval GSA rotor. The virions themselves were then pelleted by ultracentrifugation at 130,000 g for 150 minutes using a Beckman Type 35 rotor. The virions, which formed a white translucent pellet were then re-suspended using 0.5 ml sodium acetate buffer at pH 4.8. Membranous contaminants evident only by EM (and not SDS-PAGE) were subsequently removed by centrifugation at 13,000 g using a desktop microfuge.

2.1.3 Analysis of preparations by SDS-PAGE

Virus preparations were routinely analysed by sodium dodecyl sulphate - polyacrylamide gel electrophoresis (SDS-PAGE) to assess the quality and purity of the preparations. Standard methods (Laemmli, 1970) were followed to produce 12.5% acrylamide gels that were subsequently stained using Coomassie blue.

2.2 Specimen preparation

Specimen preparation (negative-staining or plunge-freezing) involved immobilization of MSV virions onto the carbon support of standard 3 mm diameter EM copper grids. Carbon coating of 400 mesh copper grids (for negative-staining), and thin-carbon coating of Quantifoil grids (for cryo-EM) was performed by Mohammed Jaffer (Electron Microscope Unit, University of Cape

Town) and is gratefully acknowledged. Precise details of procedures followed is described in the following sections.

2.2.1 Negative-staining

Negative-staining was carried out using 2 % (w/v) uranyl acetate (pH 4.8) according to the droplet technique (Brenner and Horne, 1959; Ohi *et al.*, 2004). Three 10µl droplets of distilled water were applied to Parafilm paper alongside two drops of uranyl acetate. Initially, 3 µl of the sample was allowed to adsorb to the EM grid for 15-30 seconds. The grid was subsequently blotted to remove excess sample (ensuring that the grid did not dry out), and then droplets of distilled water were sequentially lifted onto the grid to wash the specimen of buffer salts, and then immediately blotted away. The same procedure was followed for two drops of stain, however the final drop was left on the grid for 20 seconds, after which blotting was carefully controlled to ensure that stain of a defined thickness could be achieved. 400 mesh copper standard 3 mm EM grids had been coated with a layer of evaporated carbon (estimated thickness 20-30 nm).

2.2.2 Vitrification of a thin-aqueous layers

Plunge-freezing (Adrian *et al.*, 1984; Dubochet *et al.*, 1988) of MSV into liquid nitrogen cooled liquid ethane allowed a vitrified layer of ice of an estimated thickness of 100 nm to be produced regularly. This was carried out using a manual plunging apparatus powered by gravitational forced originally designed at the EMBL Heidelberg but constructed at the Electron Microscope Unit, University of Cape Town. Quantifoil grids had been cleaned of plastic residues by immersion in acetone for 30 minutes and allowed to dry. Either plain Quantifoil (R1/2), or C-flat™ grids which had been overlayed with a thin layer of continuous carbon (estimated thickness 5-10 nm), were glow-discharged for 20 seconds in nitrogen gas at 20 mA using a EMS100X (Electron Microscopy Sciences, Hatfield, Pennsylvania). Glow-discharge renders the carbon layer hydrophillic, and allows a droplet of the sample to be spread more easily.

The procedure for plunge freezing is as follows: All liquid nitrogen used was filtered to remove contaminating ice crystals, and all containers were blow-dried with hot air prior addition of liquid nitrogen to avoid contaminating ice. Liquid nitrogen was added to the base dewar of the plunging device, and ethane gas was subsequently liquefied into a small pot. A 3 µl droplet of the MSV preparation was then applied to each grids, and each was then blotted by hand from one side for ~5 seconds using Whatman No. 1 filter paper. The grid was then immediately plunged into liquid ethane under gravitational force. The liquid ethane had been allowed to cool to close to freezing temperature (at which point it became murky white) to ensure most rapid freezing of the specimen. The perforated carbon of the C-flat grids was found to be fragile; consequently the

included the stem-loop structure (origin of replication) required to initiate rolling-circle replication (Casado *et al.*, 2004).

Conditions favouring disassembly of geminiviruses have been investigated for ACMV using negative-stain and rotatory shadowing EM (Kittelmann and Jeske, 2008). These particles demonstrate reduced stability at alkaline pH conditions and investigators were able to observe disassociation of capsomers from certain locations in the virus particle. At these sites egress of the genome was also observed suggesting that within the cell, conditions must arise that result in similar behaviour. The study of ACMV disassembly revealed that either apical or peri-apical capsomers dissociate in the majority of cases. The investigators proposed that even though the genome emerges from the interior of the virion it may still be bound to the dissociated capsomers.

A review of the literature showed that no studies have been carried out to reconstitute geminate virions from disassembled and subsequently isolated viral components. Factors influencing viral assembly have been inferred based on mutagenesis, and inspection of mutant particles isolated from plant material.

1.6.3.6 Emergence and evolution of geminiviruses

All viruses can be considered modular entities, consisting of molecular components that each have independent functions and also independent evolutionary lineages (Abrescia *et al.*, 2012). It is for this reason that recombination is an important aspect of DNA viral evolution, because it allows modules to be swapped between two parent viral strains, helping to create variants with potentially improved fitness within a particular environmental context. Experiments has shown that recombination in geminiviruses will not always produce viable variants, and it has been demonstrated that recombinant variants are selected for those which preserve "intra-genomic interaction networks", a property which favours recombination between closely related viruses (Martin *et al.*, 2005). Further experiments using divergent geminiviruses have been used to show that this approach can actually be used to predict such interaction networks, providing fundamental insight into the molecular interactions underpinning the life-cycles of these organisms (Martin *et al.*, 2011).

The replication-associated protein (Rep) is one such modular component that has fuelled discussion regarding geminivirus emergence. The geminivirus Rep proteins show three important similarities in key motifs required for replication initiation that were first identified in the Rep proteins of Gram-positive bacterial plasmids, suggesting that geminiviruses may have descended from these plasmids (Koonin and Ilyina, 1992). Previous studies have suggested that geminiviral genomes may have descended from an extrachromosomal plasmid of red algae, some 450 million

years ago (Nawaz-ul-Rehman and Fauquet, 2009). More recently, phylogenetic analysis has revealed that the geminivirus Rep is also closely related to those of phycoplasmal plasmids (Saccardo *et al.*, 2011); phycoplasma being pathogenic phloem limited bacteria that lack a cell wall. Like geminiviruses, phycoplasma also rely on insect vectors for movement. How such replicons acquired a coat protein to allow them to move from host to host becomes an important question, although it is presumed that a rare recombination event is at the foundation of the emergence of these entities.

There is some debate regarding which component determines the "virus self" – genome or capsid. Considering that it is the encapsidation behaviour of viruses that distinguishes them from plasmids (Forterre, 2010; Raoult and Forterre, 2008), it has been proposed that it is the viral coat protein that gives a virus its "identity" (Abrescia *et al.*, 2012, 2010). It has been shown by empirical observation of structural homology at the protein tertiary level (i.e. structural similarity of viral coat proteins), is able to inform us about homology which cannot be observed at either nucleic acid or protein sequence level, and has allowed evolutionary relatedness to be inferred between viruses infecting all kingdoms of life; bacteria, archaea, eukarya (Bamford, 2002; Bamford *et al.*, 2005). Applying this approach of "structure-based viral lineages" to the geminiviruses, it has been proposed that these arose as a consequence of a recombination event between a phycoplasmal ssDNA plasmid – encoding the Rep, and a (+) ssRNA plant virus – encoding the coat protein CP (Krupovic *et al.*, 2009). Hence the coat protein (CP) of the plant virus was 'recruited' to become the "packaging machine" of a self-replicating nucleic acid that encoded a specific replication associated protein (Rep), while also possessing the nucleic acid secondary structure which allowed it to be replicated by RCR. This hypothesis has since been refuted by the argument that such a scenario is unlikely, and that, based on sequence analysis, the phycoplasmal plasmid itself is instead likely to have arisen from geminiviruses - and not the other way round (Saccardo *et al.*, 2011). Saccardo *et al.* went further to show that the acquisition of a coat protein is more likely to have occurred within a marine environment from another donor, rather than a plant virus replicating in a terrestrial plant host.

In addition to high rates of recombination, geminiviruses show high rates of mutation (10^{-4} substitutions per site per year); rates that are comparable to that of RNA viruses (Duffy *et al.*, 2008; Harkins *et al.*, 2009; Monjane *et al.*, 2011). This is intriguing because these viruses are replicated by high fidelity host DNA polymerases, unlike RNA viruses. The mutational processes that account for this behaviour have been observed to be single-strand specific, leading to various hypotheses that have attempted to explain the behaviour. A recent study has led to the suggestion

that it may be high-levels of oxidative damage to the single-stranded form that accounts for these relatively high rates of mutation (Monjane *et al.*, 2012).

1.7 Aims of this work

Investigating the structure and assembly of infectious virions or virus-like-particles (VLPs) serves to inform our understanding of the underlying principles governing macromolecular interaction. Molecular recognition, and subsequent association and assembly are fundamental processes in the "molecular life" of the cell (Robinson *et al.*, 2007), and viruses have long proved important model systems for studying the structural basis of these mechanisms. As discussed in the previous sections, the geminate capsid morphology is unique, and certainly a curiosity that demands study from the perspective of self-assembly. Is this assembly built on principles of conformational switching and interaction used by typical $T = 3$ viruses, or are different factors at play in the assembly and evolution of this particle? The morphology certainly implies unique behaviour during virus assembly and in spite of previous study, the quasi-equivalent coat protein (CP) conformations required to mediate assembly of the geminate particle have not been characterized. Consequently, in an attempt to shed light on these properties the geminate capsid architecture, as well as its implications for assembly of the geminate virion, the capsid of MSV is the subject of study described in this thesis, as previous work (Bottcher *et al.*, 2004; Zhang *et al.*, 2001) has shown that MSV virions can be isolated at sufficient concentrations for such work, and there is a long history of work carried out on this virus at the University of Cape Town.

The principal aim was to use already established structural biological methods (cryo-EM and homology modelling) to infer the CP-CP interactions, and architecture of MSV. If a cryo-EM map of sufficiently high resolution can be determined (e.g. 7.5 – 15 Å), this can be used in conjunction with protein structure prediction and flexible fitting to derive a pseudo-atomic model of the capsid without the requirement of crystallization. Using these methods, a plausible and hence empirically relevant model of residue-residue interactions can be constructed *in silico,* which may further inform antiviral strategies undertaken in the future, as well as elucidating the functional requirements which underpin the evolutionary divergence that the gemini-architecture represents. The number of biochemical and molecular studies already carried out on geminivirus coat proteins would benefit greatly from such a model if it were to be made publically available (e.g. in the Electron Microscopy Data Bank (EMDB;http://www.ebi.ac.uk/msd), and further experiments could be designed in a rational manner using the experimental model as a reference point.

In this thesis, cryo-EM, 3D reconstruction, and protein modelling were applied to achieve the following ends:

- Determine the structure of MSV geminate capsid to an improved resolution (< 25 Å).

- Determine the structure of the MSV icosahedral capsid to ~ 20 Å resolution.

- Determine the geometric parameters associated with each architecture; i.e. understand how the capsid building block (CP_S) is arranged within each capsid.

- Construct pseudo-atomic models using information from the EM study, as well as other sources (e.g. atomic structures of structural homologues).

- Use the above mentioned, to determine interacting regions/residues between MSV CP subunits/pentameric building blocks.

With these objectives in place, it was hoped that analysis would shed some light on the following questions:

-What is the structural/geometric basis of geminate morphology?

-How is the geminate morphology related to $T = 3$ viral capsid architectures? As both are essentially effective solutions to the problem of packaging a genome larger than what can fit inside a $T = 1$ architecture.

- How is the ssDNA genome organized within the MSV capsid, and how does it interact with the MSV capsid?

- What possible roles does the genome fulfill during assembly?

- Based on sequence alignment with other geminivirus CPs, can structural information about MSV be used as a basis for statements regarding the architecture of other geminiviruses? Specifically, is the precise geometry of the architecture conserved amongst related geminiviruses?

- Are interactions between pentameric capsomers conserved in the architectures of MSV and Satellite tobacco necrosis virus (STNV), or have these interactions diverged along with the coat protein sequences?

- What changes in conformation take place in equatorial capsomers?

- Can structural information inform us as to how this architecture might constrain viral evolution?

- Given that the genomes of other simple viruses have been implicated in capsid assembly, is it possible that some structural property of the geminivirus genome favours assembly of the geminate architecture?

- It is important that these requirements for particle formation, and role of the genome in this process, are understood, as any attempts to exploit the geminate coat proteins in biotechnology will require that their assembly behaviour and requirements are well understood.

The original reconstruction was performed without correction of the microscope contrast transfer function (CTF), and a result was limited to a resolution of 25 Å. Many questions remain unanswered in the literature regarding the nature of interactions between the MSV CP, its genome as well as the movement protein, it was felt necessary to repeat the previous study using more recent tools of structural modelling and analysis.

1.7.1 A note on terminology

Terms used to describe structural features of viruses with respect to those observed in raw electron micrographs, and also 3D reconstructions, are distinct from those used to describe structural and chemical features of the actual virions themselves. The terms 'capsomere' and 'morphological unit' refer to pronounced observable features in EM data; and are particularly revealed in low-resolution studies because they reflect clusters of subunits; these being the most visible features at 'macromolecular resolutions' of 3-5 nm (30-50 Å). In reference to the capsid as an assembly of coat protein subunits, the term 'capsomer' refers to a physical building block of the capsid, e.g. in this study the capsid of MSV is thought to be assembled from a series of pentameric capsomers (CP_5), each consisting of five coat protein subunits. Owing to the arrangement of these in the MSV capsid shell, the CP_5 also gives rise to the morphological units (or capsomeres) apparent in the EM reconstructions. Depending on the virus under investigation, these two terms may or may not refer to the same biological/macromolecular arrangement. With respect to quasi-equivalence, coat protein subunits identical in sequence, may be described to exhibit 'structural polymorphism' in which cases regions of the coat protein adopt different structures or conformations. These different but closely related conformations are referred to as 'conformers'. Herein, the monopartite (icosahedral) capsid reconstruction carried out by application of 60-fold symmetry is sometimes referred to as the '532 reconstruction'. Bipartite (geminate) reconstructions, carried

out by applying 10-fold (D5) symmetry, possess D5 symmetry and may also be referred to as such.

2 Materials and Methods

2.1 Sample preparation

2.1.1 Agro-inoculation of maize

Maize seedlings were infected by Agrobacterium-mediated DNA transfer (Boulton *et al.*, 1989a; Grimsley *et al.*, 1989); a procedure termed "agro-inoculation". An agro-infectious construct with MSV-MatA was provided by Darren Martin (Martin *et al.*, 2001). Briefly, sweetcorn kernels (Golden Bantam) were germinated and grown for three days at 30°C. Seedlings were cleaned and injected with a solution of agrobacterium cells at a site near the plant meristem. Plants were allowed to grow for 4 weeks, and showed obvious signs of successful infection at 2 weeks.

2.1.2 Virus isolation

Virions were isolated from symptomatic leaves according to a previously established protocol (Palmer et al 1998). Symptomatic maize leaves were harvested from infected plants and homogenized in a blender in an equal weight/volume of acetate buffer at pH 4.8. The homogenate was then passed through two layers of muslin cloth to remove crude plant material, and the pH adjusted to 4.8 using glacial acetic acid. Finer cellular materials that had aggregated owing to the acidic pH, were then spun out of solution by low-speed centrifugation at 12,000 g for 10 minutes using a Sorval GSA rotor. The virions themselves were then pelleted by ultracentrifugation at 130,000 g for 150 minutes using a Beckman Type 35 rotor. The virions, which formed a white translucent pellet were then re-suspended using 0.5 ml sodium acetate buffer at pH 4.8. Membranous contaminants evident only by EM (and not SDS-PAGE) were subsequently removed by centrifugation at 13,000 g using a desktop microfuge.

2.1.3 Analysis of preparations by SDS-PAGE

Virus preparations were routinely analysed by sodium dodecyl sulphate - polyacrylamide gel electrophoresis (SDS-PAGE) to assess the quality and purity of the preparations. Standard methods (Laemmli, 1970) were followed to produce 12.5% acrylamide gels that were subsequently stained using Coomassie blue.

2.2 Specimen preparation

Specimen preparation (negative-staining or plunge-freezing) involved immobilization of MSV virions onto the carbon support of standard 3 mm diameter EM copper grids. Carbon coating of 400 mesh copper grids (for negative-staining), and thin-carbon coating of Quantifoil grids (for cryo-EM) was performed by Mohammed Jaffer (Electron Microscope Unit, University of Cape

Town) and is gratefully acknowledged. Precise details of procedures followed is described in the following sections.

2.2.1 Negative-staining

Negative-staining was carried out using 2 % (w/v) uranyl acetate (pH 4.8) according to the droplet technique (Brenner and Horne, 1959; Ohi *et al.*, 2004). Three 10µl droplets of distilled water were applied to Parafilm paper alongside two drops of uranyl acetate. Initially, 3 µl of the sample was allowed to adsorb to the EM grid for 15-30 seconds. The grid was subsequently blotted to remove excess sample (ensuring that the grid did not dry out), and then droplets of distilled water were sequentially lifted onto the grid to wash the specimen of buffer salts, and then immediately blotted away. The same procedure was followed for two drops of stain, however the final drop was left on the grid for 20 seconds, after which blotting was carefully controlled to ensure that stain of a defined thickness could be achieved. 400 mesh copper standard 3 mm EM grids had been coated with a layer of evaporated carbon (estimated thickness 20-30 nm).

2.2.2 Vitrification of a thin-aqueous layers

Plunge-freezing (Adrian *et al.*, 1984; Dubochet *et al.*, 1988) of MSV into liquid nitrogen cooled liquid ethane allowed a vitrified layer of ice of an estimated thickness of 100 nm to be produced regularly. This was carried out using a manual plunging apparatus powered by gravitational forced originally designed at the EMBL Heidelberg but constructed at the Electron Microscope Unit, University of Cape Town. Quantifoil grids had been cleaned of plastic residues by immersion in acetone for 30 minutes and allowed to dry. Either plain Quantifoil (R1/2), or C-flat™ grids which had been overlayed with a thin layer of continuous carbon (estimated thickness 5-10 nm), were glow-discharged for 20 seconds in nitrogen gas at 20 mA using a EMS100X (Electron Microscopy Sciences, Hatfield, Pennsylvania). Glow-discharge renders the carbon layer hydrophillic, and allows a droplet of the sample to be spread more easily.

The procedure for plunge freezing is as follows: All liquid nitrogen used was filtered to remove contaminating ice crystals, and all containers were blow-dried with hot air prior addition of liquid nitrogen to avoid contaminating ice. Liquid nitrogen was added to the base dewar of the plunging device, and ethane gas was subsequently liquefied into a small pot. A 3 µl droplet of the MSV preparation was then applied to each grids, and each was then blotted by hand from one side for ~5 seconds using Whatman No. 1 filter paper. The grid was then immediately plunged into liquid ethane under gravitational force. The liquid ethane had been allowed to cool to close to freezing temperature (at which point it became murky white) to ensure most rapid freezing of the specimen. The perforated carbon of the C-flat grids was found to be fragile; consequently the

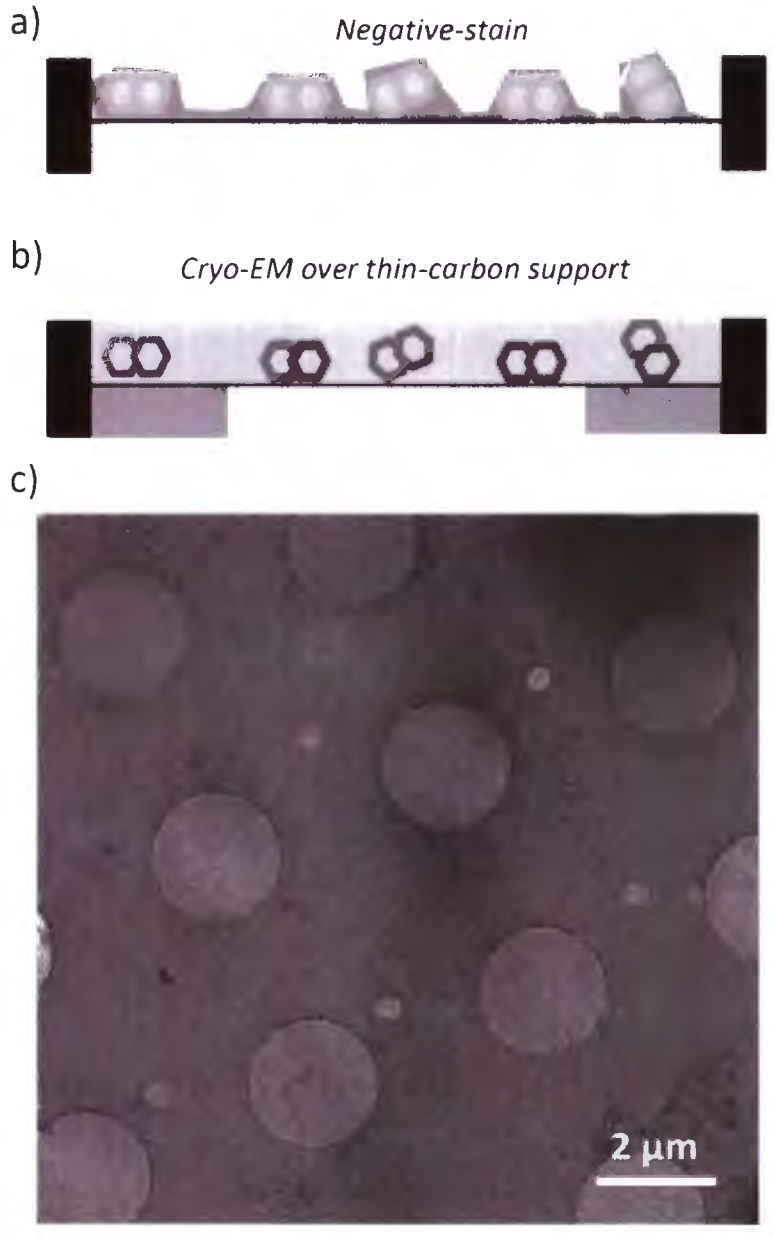

Figure 2-1 Sample preparation and data collection for low-dose electron microscopy. Negative-staining and rapid vitrification are illustrated by schematic in a) and b), respectively. c) Electron micrograph of a region of a grid (800 X). Holes are 2 µm in diameter. Figure created using schematic figures and experimental images created/acquired over the course of the investigation.

plunging apparatus was configured such that the grid was plunged to the minimum distance required to immerse the grid. Grids were transferred rapidly from the liquid ethane bath into the surrounding liquid nitrogen and deposited into a cryo-EM grid holder; this avoided contamination of the grid with humidity. Grids were stored under liquid nitrogen until transferred at cryogenic temperatures into the electron microscope.

2.3 Transmission electron microscopy and data-collection

2.3.1 Cryo-transfer

For electron cryo-microscopy, a Gatan 626 cryo-holder (Gatan Inc., Pleasanton, CA) was used to maintain the temperature of the specimen at cryogenic temperatures (~ -180 °C) during transfer and electron microscopy. The cryo-holder was prepared by evacuation of the jacket surround the LN dewar overnight. Frozen grids were retrieved from storage and transferred into the Gatan cryo-transfer workstation after it had been cooled to LN temperatures. All liquid nitrogen had been filtered prior to use. Grids were transferred into the cryo-holder under LN, and the cryo-holder transferred into the FEI Tecnai F20 (FEI Company).

2.3.2 Electron microscope alignment

Electron microscopy was carried out at either 120 kV or 200 kV using a JEOL EX II or FEI G^2 Tecnai F20, respectively. The microscope was aligned to minimize image aberrations during low-dose electron microscopy. First the gun alignment was performed to maximize brightness of the electron source. This was achieved by using the gun tilt and gun shift parameters to centre the filament "bright-spot" (FEI Tecnai F20, TF20). For the tungsten source (JEOL EX II), the gun intensity was de-saturated and gun tilt and shift parameters used to centre and symmetrize the image of the filament. The condenser aperture was centred, and condenser astigmatism corrected to ensure that the beam was circular. On the Tecnai F20, a negative-stain specimen was inserted and the Z-height of the holder brought to the eucentric height. The eucentric height is also known as the centre of rotation and provides a physical reference point used to calibrate the microscopes magnification. The objective lens was set to in-focus (i.e. minimum contrast), and the beam tilt pivot points carried out to ensure that the beam axis was parallel to the axis of the objective lens. The rotation centre alignments were carried out to ensure that the beam passed through the centre of the objective lens. This alignment was refined further by carrying out coma-free alignment at the magnification used for imaging using the microscopes video camera. Similar procedures were performed to ensure adequate alignment of the JEOL microscope.

plunging apparatus was configured such that the grid was plunged to the minimum distance required to immerse the grid. Grids were transferred rapidly from the liquid ethane bath into the surrounding liquid nitrogen and deposited into a cryo-EM grid holder; this avoided contamination of the grid with humidity. Grids were stored under liquid nitrogen until transferred at cryogenic temperatures into the electron microscope.

2.3 Transmission electron microscopy and data-collection

2.3.1 Cryo-transfer

For electron cryo-microscopy, a Gatan 626 cryo-holder (Gatan Inc., Pleasanton, CA) was used to maintain the temperature of the specimen at cryogenic temperatures (~ -180 °C) during transfer and electron microscopy. The cryo-holder was prepared by evacuation of the jacket surround the LN dewar overnight. Frozen grids were retrieved from storage and transferred into the Gatan cryo-transfer workstation after it had been cooled to LN temperatures. All liquid nitrogen had been filtered prior to use. Grids were transferred into the cryo-holder under LN, and the cryo-holder transferred into the FEI Tecnai F20 (FEI Company).

2.3.2 Electron microscope alignment

Electron microscopy was carried out at either 120 kV or 200 kV using a JEOL EX II or FEI G^2 Tecnai F20, respectively. The microscope was aligned to minimize image aberrations during low-dose electron microscopy. First the gun alignment was performed to maximize brightness of the electron source. This was achieved by using the gun tilt and gun shift parameters to centre the filament "bright-spot" (FEI Tecnai F20, TF20). For the tungsten source (JEOL EX II), the gun intensity was de-saturated and gun tilt and shift parameters used to centre and symmetrize the image of the filament. The condenser aperture was centred, and condenser astigmatism corrected to ensure that the beam was circular. On the Tecnai F20, a negative-stain specimen was inserted and the Z-height of the holder brought to the eucentric height. The eucentric height is also known as the centre of rotation and provides a physical reference point used to calibrate the microscopes magnification. The objective lens was set to in-focus (i.e. minimum contrast), and the beam tilt pivot points carried out to ensure that the beam axis was parallel to the axis of the objective lens. The rotation centre alignments were carried out to ensure that the beam passed through the centre of the objective lens. This alignment was refined further by carrying out coma-free alignment at the magnification used for imaging using the microscopes video camera. Similar procedures were performed to ensure adequate alignment of the JEOL microscope.

2.3.3 Data collection

Low-dose EM was carried out to prevent specimen damage arising from interaction of the high-energy electrons of the beam with the specimen. The basic principal of this approach is to focus on an area immediately adjacent to the recorded area (1-2 μm), and then apply the determined focusing settings to recorded area. In this way the total exposure of the recorded area to the electron beam was limited to a total of dose of 15-25 $e^-/Å^2$ in total. Image astigmatism was corrected during focusing using the fluorescent screen by inspection of the grain of the carbon support at 200,000X magnification. Low-dose cryo-EM images were recorded at 50,000 X on Kodak ISO 160 film, and were of significantly lower contrast in comparison to their negative-stain counterparts. Images were recorded over a range of defocus values (1.5 - 3 μm) in an attempt to account for oscillating nature of the microscope CTF, which exhibits reduced or no contrast transfer at specific spatial frequencies (referred to herein as "zeroes").

2.4 Single-particle 3D reconstruction

All image-processing steps were carried out on a Intel Core™ 2 Quad Q6600 (2.4 GHz) workstation running Ubuntu Linux (8.04, "Hardy Heron") with 4 GB RAM. Unless otherwise stated, all image processing operations were performed using SPIDER (Frank *et al.*, 1996). Micrographs, image stacks or 3D reconstructions were routinely inspected using the "*V2*" or "*V4*" programs of the EMAN software package version 1.9 (Ludtke *et al.*, 1999)

2.4.1 Image pre-processing

After micrographs had been recorded, film was retrieved from the microscope and developed according to standard procedures (Jiang and Chiu, 2007).

2.4.1.1 Digitization and format conversion

Micrographs were digitized using a HP Leafscan 45 densitometer using a pixel sampling of 2540 pixels/inch. The final sampling of the digitized electron micrographs was 2 Å/pixel at the specimen level.

2.4.1.2 Determination of Contrast Transfer Function (CTF) parameters

Micrograph CTF parameters were calculated using CTFFIND (Mindell and Grigorieff, 2003). CTFFIND determines the defocus (Δz; by convention positive for underfocus) and astigmatism parameters associated with a micrograph using a least-squares fitting algorithm. The algorithm is able to calculate the astigmatism by fitting a 2D contrast transfer function to the amplitude modulations visible in the power spectrum of the micrograph, and determines the fit in two perpendicular directions,. To boost the signal-to-noise ratio of the oscillations, the program was instructed to calculate an average power spectrum using a tile size of 128x128 pixels. The search

range for micrograph defocus values was constrained to Δz = 1 - 4 μm, using a step of 200 nm. CTF functions were fit over a resolution range of 80 to 5 Å. An amplitude contrast ratio of 0.07 and 0.18 were used for the vitreous water and negative-stain datasets, respectively. The accuracy of the parameters determined was confirmed by inspection of the diagnostic images outputted by the software which display the modelled CTF in juxtaposition with the average power spectrum calculated from micrograph.

2.4.1.3 Particle selection

Before image processing could proceed, individual particles (virion projections) were identified by an automated template matching algorithm within the program SIGNATURE (Chen and Grigorieff, 2007), and subsequently windowed (cut) out using the *"WI"* operation (SPIDER) from the digitized electron micrographs into square pixel intensity arrays (boxes) which are suitable for computational analysis.

The quality of the micrographs was assessed using the 'optical density histogram' and 'power spectrum' functions of SIGNATURE. It was considered desirable that micrographs showed an approximately Gaussian pixel intensity histogram and circular power spectrum. An elliptical power spectrum would indicate that the image was not focused equally in all directions (a sign of astigmatism). Automated picking proceeded using the template-matching algorithm of SIGNATURE using image classes (templates) generated from manual picking of a few micrographs. Regions of micrographs demonstrating unacceptable particle clumping, drift, astigmatism, or ice thickness were excluded by masking within SIGNATURE prior the automated particle searches. The co-ordinates of particles were inspected manually, and erroneous particles removed from the coordinate list of particle centres. The XY co-ordinates of each particle centre were then outputted as a SPIDER document file and windowed out using a SPIDER batch script. At the time of windowing, images were decimated by summation of neighbouring pixels by a factor of two or four in each dimension using "*IP*" operation, to produce two separate stacks, outputted at samplings of 4 and 8 Å/pixel, respectively. The 4 Å/pixel stack was outputted for 3D reconstruction (which required a higher sampling to achieve the intended resolution, and 8 Å/pixel for K-means classification. Additionally, for cryo-EM reconstruction carried out in SPIDER the pixel values were inverted such that density contributed by protein or nucleic acid was "white". At the time of windowing, particles were excluded if the difference of *DF1* and *DF2* was greater than 100 nm, as this was considered to be unacceptable astigmatism.

2.4.1.4 Particle centring

Windowed particles were first approximately centred according to their centres of gravity using the "*CG*" operation of SPIDER. A strategy for in-plane particle centring able to handle the

range for micrograph defocus values was constrained to Δz = 1 - 4 μm, using a step of 200 nm. CTF functions were fit over a resolution range of 80 to 5 Å. An amplitude contrast ratio of 0.07 and 0.18 were used for the vitreous water and negative-stain datasets, respectively. The accuracy of the parameters determined was confirmed by inspection of the diagnostic images outputted by the software which display the modelled CTF in juxtaposition with the average power spectrum calculated from micrograph.

2.4.1.3 Particle selection

Before image processing could proceed, individual particles (virion projections) were identified by an automated template matching algorithm within the program SIGNATURE (Chen and Grigorieff, 2007), and subsequently windowed (cut) out using the *"WI"* operation (SPIDER) from the digitized electron micrographs into square pixel intensity arrays (boxes) which are suitable for computational analysis.

The quality of the micrographs was assessed using the 'optical density histogram' and 'power spectrum' functions of SIGNATURE. It was considered desirable that micrographs showed an approximately Gaussian pixel intensity histogram and circular power spectrum. An elliptical power spectrum would indicate that the image was not focused equally in all directions (a sign of astigmatism). Automated picking proceeded using the template-matching algorithm of SIGNATURE using image classes (templates) generated from manual picking of a few micrographs. Regions of micrographs demonstrating unacceptable particle clumping, drift, astigmatism, or ice thickness were excluded by masking within SIGNATURE prior the automated particle searches. The co-ordinates of particles were inspected manually, and erroneous particles removed from the coordinate list of particle centres. The XY co-ordinates of each particle centre were then outputted as a SPIDER document file and windowed out using a SPIDER batch script. At the time of windowing, images were decimated by summation of neighbouring pixels by a factor of two or four in each dimension using "*IP*" operation, to produce two separate stacks, outputted at samplings of 4 and 8 Å/pixel, respectively. The 4 Å/pixel stack was outputted for 3D reconstruction (which required a higher sampling to achieve the intended resolution, and 8 Å/pixel for K-means classification. Additionally, for cryo-EM reconstruction carried out in SPIDER the pixel values were inverted such that density contributed by protein or nucleic acid was "white". At the time of windowing, particles were excluded if the difference of *DF1* and *DF2* was greater than 100 nm, as this was considered to be unacceptable astigmatism.

2.4.1.4 Particle centring

Windowed particles were first approximately centred according to their centres of gravity using the "*CG*" operation of SPIDER. A strategy for in-plane particle centring able to handle the

2.3.3 Data collection

Low-dose EM was carried out to prevent specimen damage arising from interaction of the high-energy electrons of the beam with the specimen. The basic principal of this approach is to focus on an area immediately adjacent to the recorded area (1-2 μm), and then apply the determined focusing settings to recorded area. In this way the total exposure of the recorded area to the electron beam was limited to a total of dose of 15-25 $e^-/Å^2$ in total. Image astigmatism was corrected during focusing using the fluorescent screen by inspection of the grain of the carbon support at 200,000X magnification. Low-dose cryo-EM images were recorded at 50,000 X on Kodak ISO 160 film, and were of significantly lower contrast in comparison to their negative-stain counterparts. Images were recorded over a range of defocus values (1.5 - 3 μm) in an attempt to account for oscillating nature of the microscope CTF, which exhibits reduced or no contrast transfer at specific spatial frequencies (referred to herein as "zeroes").

2.4 Single-particle 3D reconstruction

All image-processing steps were carried out on a Intel Core™ 2 Quad Q6600 (2.4 GHz) workstation running Ubuntu Linux (8.04, "Hardy Heron") with 4 GB RAM. Unless otherwise stated, all image processing operations were performed using SPIDER (Frank *et al.*, 1996). Micrographs, image stacks or 3D reconstructions were routinely inspected using the "*V2*" or "*V4*" programs of the EMAN software package version 1.9 (Ludtke *et al.*, 1999)

2.4.1 Image pre-processing

After micrographs had been recorded, film was retrieved from the microscope and developed according to standard procedures (Jiang and Chiu, 2007).

2.4.1.1 Digitization and format conversion

Micrographs were digitized using a HP Leafscan 45 densitometer using a pixel sampling of 2540 pixels/inch. The final sampling of the digitized electron micrographs was 2 Å/pixel at the specimen level.

2.4.1.2 Determination of Contrast Transfer Function (CTF) parameters

Micrograph CTF parameters were calculated using CTFFIND (Mindell and Grigorieff, 2003). CTFFIND determines the defocus (Δz; by convention positive for underfocus) and astigmatism parameters associated with a micrograph using a least-squares fitting algorithm. The algorithm is able to calculate the astigmatism by fitting a 2D contrast transfer function to the amplitude modulations visible in the power spectrum of the micrograph, and determines the fit in two perpendicular directions,. To boost the signal-to-noise ratio of the oscillations, the program was instructed to calculate an average power spectrum using a tile size of 128x128 pixels. The search

heterogeneity of capsid species (icosahedral and geminate capsids) needed to be developed, as it was found that the coincidence of these species within the dataset could interfere with the accuracy of centring (icosahedral capsids were seen to align to either of the two geminate heads apparent in the global average). The X and Y translational parameters could not be achieved reliably by aligning against a "global average" of the dataset, and the most robust method was found to be aligning each particle by cross-correlation, iteratively, against its own rotational average over three rounds, in which the rotational average was recalculated after each round of alignment.

2.4.2 K-means classification and partitioning of data

Once the particles had been adequately centred, an initial classification was carried out using the rotationally invariant K-means classification algorithm implemented in the "*AP C*" operation of SPIDER (Penczek *et al.*, 1996). This operation classifies images into groups, and also provides an in-plane rotational alignment parameter that allows images of the same class to be aligned prior to calculation of class averages. 30 seed images were selected randomly, and subsequent classes were refined over 50 iterations with re-normalization after each round of refinement. In order to improve the signal-to-noise ratio of the data in the case of the cryo-EM dataset, the images were down sampled to 8 Å/pixel (Nyquist limit of 16 Å/pixel), and searched over a ring radius corresponding to the full size of the geminate particle. To reduce the number of different views reported to only those possessing unique structural information, mirror checking was used during the classification. Class averages for K-means groups were calculated for classes consisting of two or more particles using the "*AS R*" operation to average images specified by image selection documents outputted by the "*AP C*" operation. Class averages were then masked with a circular mask and normalized using PROC2D of EMAN (Ludtke *et al.*, 1999) prior to inspection.

2.4.2.1 Segregation of images for reconstruction of capsid species

After K-means classification, images thought to be icosahedral capsids and those thought to be geminate capsids (possessing D5 symmetry) were copied into a separate image stacks for 3D reconstruction. This was achieved using a SPIDER script written to export images according to K-means class number. Additionally, particle parameters (i.e. estimates of magnification, defocus, micrograph number) were also outputted as separate document files that allowed this information to be tracked throughout further image processing steps.

2.4.3 3D reconstruction

The alignment parameters required to bring each single-particle image into register with respect to a shared 3D co-ordinate system can be determined using different strategies. Reconstructions

described in previous studies of MSV and ACMV were carried out using 3D projection matching using starting models constructed from the X-ray structure of STNV. Those investigators demonstrated convincingly that the starting model did not bias the final result, as the dimensions and features of the final EM maps were unlike that of the starting models. Consequently, we also decided to use 3D projection matching. This study differed in that two independent geometric starting models were used to "boot-strap" the alignment of the images so that resolution assessment could be carried out independent of any reference bias (arising from correlation of noise). Additionally, we hoped to record a high-resolution dataset to which CTF correction could be applied so that high-resolution structural features would be revealed in the 3D reconstruction.

As described previously, the MSV virion is a relatively large assembly that possesses 5-fold dihedral symmetry (D5). This implies a relatively high redundancy of information in each image (increasing the effective size of the dataset by an order of magnitude), and implies that the capsid has an "asymmetric unit" covering an angular range equivalent to $1/10^{th}$ of a sphere. These properties of the specimen influenced our choice in projection geometry with respect to both *in silco* calculation of 2D references, as well as back-projection. This also had the benefit of constraining the orientation search; greatly expediting the computation time for each iteration of the projection-matching algorithm. D5 symmetry was applied during 3D reconstruction, because it was expected that this would improve the signal-to-noise ratio of the final result substantially. The choice to take advantage of this symmetry limits the strict validity of the map to regions adhering to this symmetry, i.e. the core regions of the protein shell. This would allow the geometry of CP units to be investigated, but we did not expect to be able to see the precise structure of the genome, or the true nature of the capsids interactions with genome. However, the structure would nevertheless provide us with a means of assessing the radial distribution of genomic density, as well as to estimate the approximate locations of any interactions. No attempt to carry out asymmetric reconstruction of the capsids was made. Validation of our results relied on extensive information available from previous studies, structural bioinformatics and model fitting, as well as reasonable expectations of what structural features to expect in an assembly of pentameric capsomers.

2.4.3.1 Creation of starting models

Starting models are the first volumes used by the iterative-projection-matching algorithm, and allow initial estimates for alignment parameters (ϕ, θ, ψ, x and y) to be assigned to each image in a dataset. Because the choice of starting model may bias a reconstruction, it is important that each dataset be refined against two independent starting models. Additionally, it is important that the starting models are not too similar to one another, as these might introduce correlations into the

alignment of the dataset that may result in an overestimation of the resolution of the 3D reconstruction. For the negative-stain dataset, geometric models (a cylinder and bi-spherical model) were created and used as 3D approximations of the MSV virion. For the cryo-EM reconstruction, the negative-stain reconstruction low-pass filtered to 40 Å, was used as the starting model to refine the dataset.

A cylinder (radius of 200 Å, and a height of 350 Å) and "twin-spheres" (sphere diameter of 200 Å or 25 pixel radius) models were created using the "*MO 3*" SPIDER operation. The long axis of each model lay on the Z-axis. Models were low-pass filtered to 40 Å so that high-resolution features could not correlate with noise in the dataset. For the sphere model, a single sphere was created, the box windowed to remove a side of the sphere (thus simulating an incomplete icosahedron), displaced along the Z-axis, and then duplicated and rotated 180° to create a partner, which was added to the original volume. For the icosahedral (532) dataset, independent reconstructions were initiated using a sphere (200 Å diameter), which had been Gaussian low-pass filtered to 40 Å, and an icosahedral model of *Satellite tobacco necrosis virus* (PDB: 1BUK) that had been converted to density and Gaussian low-pass filtered to 30 Å.

2.4.3.2 Implementation of the iterative-projection-matching algorithm

For 3D reconstruction without CTF correction, the SPIDER script "*reconAPSH.kcd*" (refer to Appendix B) was written to execute the iterative-projection-matching algorithm that involved iterative alignment of a dataset to the 3D reconstruction calculated in the previous round. The script allowed parameters to be specified (e.g. the number of iterations to be carried out, as well as parameters pertaining to alignment search, and post reconstruction masking) and could be left to execute until completion. Statistics associated with refinement were simultaneously written out in a format that could be readily interpreted as the computation proceeded. Various details of the script are described in the paragraphs that follow.

2.4.3.2.1 *In silico* projection

First, a SPIDER document containing the projection angles used for *in silico* projection of the previous rounds model (or starting model) was generated using the "*VO EA*" operation. The operation produced a list of Euler angles over a range in phi (φ, rotation about the Z-axis) of 0 to 72°, and over a range of theta (θ) (rotation about the Y axis) of 0 to 90°. These limits span the asymmetric unit of an object with D5 symmetry ($1/10^{th}$ of a sphere). The step in theta (Δθ) used was 3°. For each iteration, reference projections were produced using the "*PJ3Q*" operation using the angular document file described above, and the volume generated during the previous iteration. A radius of 50 pixels was used and a total of 427 references were calculated.

2.4.3.2.2 Multi-reference projection matching

Multi-reference alignment was carried out using "*AP SH*" operation. A rotational search radius of 190 Å, and translational search range of 20 Å was used. 10878 images were included in the refinement of the negative-stain dataset, and 4300 images were included in the refinement of the cryo-EM dataset. Mirror checking was not used during alignment. Alignment parameters were written out into a separate SPIDER document file, and used by the *gen_classaverages.kcd* (Appendix B) script to carry out class averaging.

2.4.3.2.3 Weighted back-projection

3D reconstruction was carried using the "*BP 32F*" SPIDER operation, which produced two volumes using half-sets of the dataset. Symmetry was imposed during the reconstruction using a D5 symmetry document generated using the "*SY DOC*" operation. The half-sets reconstruction were used to carry out resolution assessment (refer Section 2.4.3.4)

2.4.3.2.4 3D masking

"Smooth-masking" ("*mask_vol.kcd*", Appendix B) was used to improve the signal-to-noise ratio (SNR) of each reconstructed volume, and consequently the following round of reference projections used to refine the dataset. Masking was used to remove densities both outside and inside the boundary of the MSV virion, and this meant that cross-correlation of volume projections on the subsequence iteration involved correlations of the capsid structure only.

At the end of each iteration a mask was created using the reconstructed volume itself. First the volume was thresholded ("*TH*" operation, 1.5 σ) to create a volume in which noise of weak densities had been removed. The thresholded volume was then low-pass filtered ("*FQ*" operation) heavily (50 Å) to increase the boundary of virion. A binary mask was then created ("*TH M*", 1.6 σ), which was subsequently filtered to create a continuous mask. Masking was then carried out using the "*MM C*" operation.

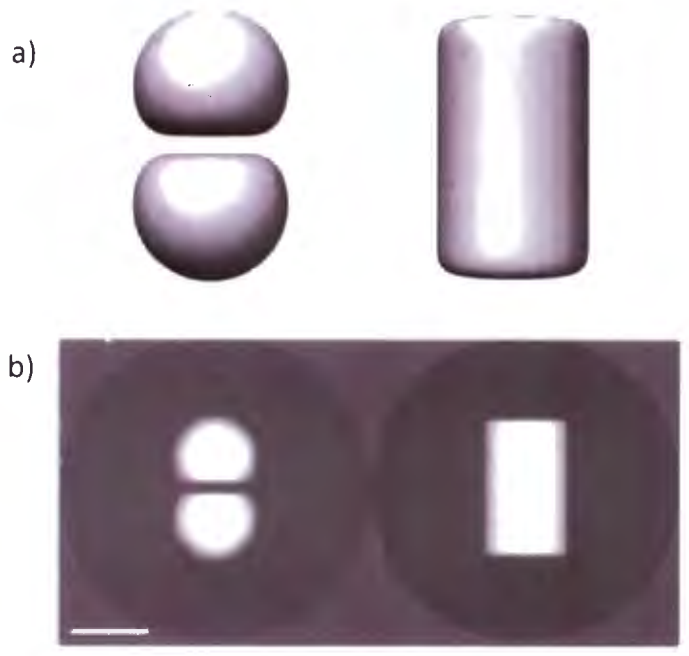

Figure 2-2 Starting models for 3D reconstruction of the negative-stain geminate dataset. a, left) A starting model consisting of two modified spheres, and a, right) a cylinder were used as "independent" starting models to produce two reconstructions from the negative-stain data. b) The models are shown as projections to assist their comparison to the experimental data. The models were created to have similar, but not precisely the same dimensions of the geminate capsid. This ensured that reference bias did not influence the outcome of the iterative projection matching algorithm. Scale bar represents 20 nm.

2.4.3.2.5 Convergence of alignment parameters

Convergence of the projection-matching algorithm, a state whereby an iteration failed to induce changes (or improvements) of particle alignment parameters, was monitored by calculating the average cross-correlation of the dataset with the re-projections of the most recent 3D reconstruction. Additionally, whether or not the images changed class (or 'view') was also monitored. These results, taken in consideration of resolution estimates produced after each iteration were used to assess the progress of a refinement and were implemented in the "*manage_stats.kcd*" SPIDER scripts (Appendix B). In general, average cross-correlation scores showed consistent improvement over the refinements carried out, and most images (>90%) were observed to fall into stable classes.

2.4.3.3 CTF correction and refinement of the cryo-EM reconstruction

The program FREALIGN version 0.86 (Grigorieff, 2007) was used to carry out 3D reconstruction of the cryo-EM dataset with CTF correction. FREALIGN allowed the magnification, defocus, and alignment parameters of each image in the dataset to be refined as part of an iterative projection matching reconstruction strategy (similar in conception to that implemented in SPIDER), and resulted in a reconstruction of improved quality.

The cryo-EM dataset had been reconstructed first using SPIDER, and subsequently SPIDER alignment parameters were converted to FREALIGN format using scripts kindly provided by Nikolaus Grigorieff (Brandeis University). The 3588 highest correlating images were copied from the original image stack and converted into a MRC formatted stack. Pixel densities were not altered by inversion (protein/nucleic acid appeared as black and not white as in the NS reconstruction) as is required by FREALIGN. An outer alignment search radius of 195 Å was used during all refinement iterations. Care was taken not to include information at spatial frequencies dominated by noise, and trial and error was used to establish that structure factors higher than $1/20Å^{-1}$ should not be included in the refinement. Attempts to extend the resolution beyond this limit resulted in erroneous alignments, and the appearance of 3D reconstruction suffered as a consequence. In total, 240 refinement iterations were carried out using FREALIGN to arrive at the final result.

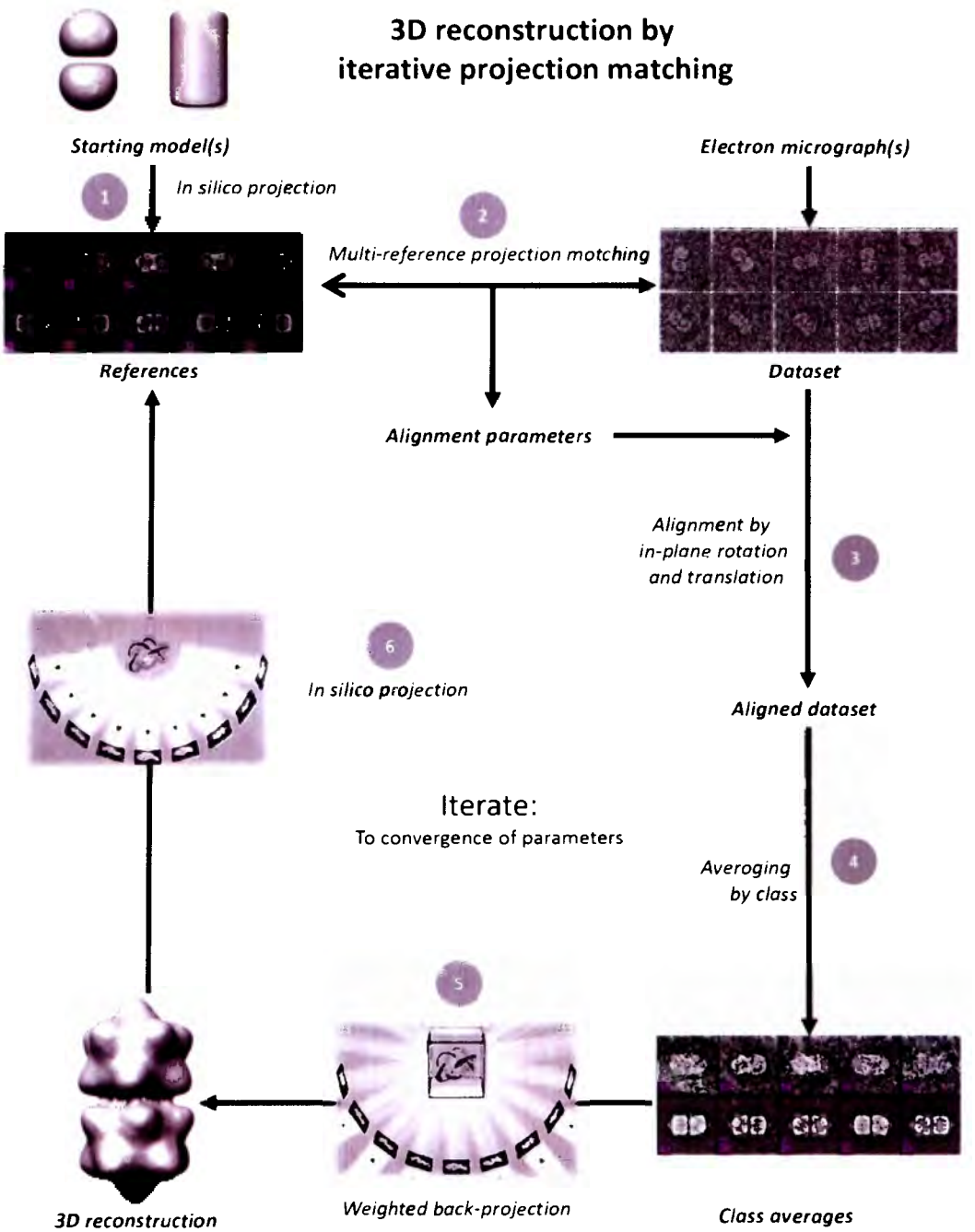

Figure 2-3 Procedure used for 3D reconstruction of the MSV virion from EM micrographs. Panels depicting projection and back-projection reproduced from (Baumeister and Grimm, 1999).

2.4.3.4 Resolution assessment

For all refinements and 3D reconstructions carried out using SPIDER, the "*BP 32F*" operation was used to back-project two volumes from each dataset (one for each odd/even half-set), and these were assessed using the "*RF 3*" operation. The FSC 0.5 criterion was used to assign the resolution in all cases.

2.4.4 *In silico* visualization

Electron density maps and atomic co-ordinates were routinely visualized using UCSF Chimera (Pettersen *et al.*, 2004). Casual fitting of atomic co-ordinates into volumes was performed using the Chimera "Fit in Map" function. Rigorous fitting studies are described in the sections that follow.

2.5 Structural bioinformatics

During this study, recently developed structural bioinformatics techniques were applied to identify appropriate templates for modelling the MSV CP sequence. Secondary structure prediction and well as sequence-structure threading was used to guide the creation of an accurate sequence alignment; a challenging task given the extremely low sequence identity between the template and query sequences (less than 15%), which was further used to assess the divergence of geminivirus CP sequences.

In overview, the sequence alignment and template PDB structure of STNV, Protein Data Bank (PDB) ID: 2BUK, were used to produce a comparative model of the MSV pentameric capsomer (CP_5). The comparative model was aligned in 3D space along its principal axes, and its centre of gravity was brought to the origin. The CP_5 model was then used to create a series of models of a single incomplete $T = 1$ icosahedral head which varied in radius and orientation of the capsomer around the 5-fold axes of symmetry. After parameters had been determined for production of an accurate estimate of the icosahedral architecture, a geminate model could be created by producing a series of geminate models that varied in their angular offset, as well as displacement between heads. The strategy and procedures for this component of the work are shown in Figure 2-4 and Figure 2-5. Following this, the homology model was also used to predict the electrostatic surface potential of the interior surfaces of the MSV and STNV CP_5. This was important, as any difference in in this property of the capsid would influence the arrangement of the packaged genome on the inside of the particle.

2.5.1 Comparative modelling

2.5.1.1 Fold prediction and template identification

Previous studies have already identified that the CP of STNV is a suitable template (Bottcher *et al.*, 2004; Zhang *et al.*, 2001). During this study we resubmitted the sequence to evaluate whether a better template may have become available over the past decade. As a result, the MSV CP sequence was submitted to mGenTHREADER (Jones, 1999a), a webserver for automated fold-recognition. mGenTHREADER was used because of its ability to identify distantly related structural homologues, and the output of this process provided indication of a suitable template PDB, as well as an initial alignment between the two sequences.

2.5.1.2 Query sequence-template alignment

The accuracy of the protein sequence alignment is the most important consideration when creating a homology model that accurately represents the actual protein. Model refinement, performed after backbone coordinates have been assigned according to the alignment with the structural template, will be unable to resolve errors introduced by misalignment (Xiang, 2007). Hence, the full-length sequence of MSV CP was initially aligned against the CP of STNV using CLUSTAL W (Thompson *et al.*, 1994). Secondary structure prediction using the PSIPRED server (Jones, 1999b; McGuffin *et al.*, 2000), as well as the alignment outputted by mGenTHREADER (Jones, 1999a) helped to resolve alignment uncertainties which arose due to virtually absent sequence similarity in many regions, particular those of loops. In particular the alignment of large loop insertions taking place between β-strands βF and βG were assisted by identifying the conserved calcium-binding site by the sequence-threading algorithms of mGenTHREADER. The sequences were further manually aligned using JalView multiple sequence alignment editor (Clamp *et al.*, 2004; Waterhouse *et al.*, 2009) so that the β-strands were brought into register. The first 26 N-terminal amino acids and last 4 C-terminal amino acids of the MSV sequence were not modelled owing to the absence of corresponding residues in the template, as well as the limited resolution of the EM reconstruction.

2.5.1.3 Creation of the template structure

The crystal structure of STNV was retrieved from the PDB (www.rcsb.org, PDB code: 2BUK). The atomic structure provided by the database contains a single subunit of the structure, representing the asymmetric unit of the biological complex. The entire capsid was recreated by application of non-crystallographic symmetry in UCSF Chimera (Goddard *et al.*, 2007). Water molecules and calcium ions were stripped from the model, and a single pentamer (CP_5) written out after the atomic numbers and chains designations had been reassigned (1-189, A-E). The centre of gravity of the CP_5 was determined using the PBD CG3 procedure of SPIDER. The CP_5

was then aligned to its principal axes (axis of C5 symmetry placed on the Z axis) using the "ORIENT" script made available by Paul Grayson for use with the VMD molecular graphics and simulation package (Humphrey *et al.*, 1996). The aligned and centered CP_5 would serve as the starting point for homology modelling as well as construction of quaternary models.

2.5.1.4 Comparative model building of the pentameric capsomer (CP_5)

Model building was performed using *automodel* class of MODELLER 9v6 (Eswar *et al*., 2007). Comparative modelling was carried out using the oriented and centered pentamer of STNV as the template, and C5 symmetry was imposed by specifying that distances and geometries of Cα atoms for chains B, C, D and E were equivalent to those in chain A. Secondary structures (as informed by PSIPRED, and subjective comparison with STNV) was also imposed for each of the five chains. The manually edited sequence alignment was inputted to MODELLER in PIR format.

2.5.2 Icosahedral and geminate capsid modelling

In the first stage of capsid modelling (Figure 2-5, a), a series of PDB models were created in which the axial displacement or radius (geminate parameter **a**) and orientation (geminate parameter α) of the MSV CP_5 were varied. First, the centred and oriented CP_5 homology model was displaced along the Z-axis by an amount representing the radius of a single head of the geminate architecture. Models were then generated over radial range of a = 76 - 98 Å (separated by a 2 Å step size). A series of rotations about the Z-axis were then applied (α = 0 - 72°, in 7.2° steps) to each of these models to produce 120 models in total. 12 symmetry operations (Euler transformations) were then applied to construct a separate icosahedral model for each CP_5 model.

These icosahedral capsid models were then fitted into the EM density maps (532-NS, or a single head of each of the D5-NS, and D5-cryo-EM reconstructions) using the COLORES program of the SITUS package (Wriggers, 2010). Density maps were searched over a limited angular range (ϕ = 0 - 72°, θ = 0 - 5 , ψ = 0 - 5°) at a resolution of 25 Å. Only the single highest correlation peak was then optimized by the Powell optimization routine in SITUS. Prior to fitting calculations, each EM density map had been inspected to determine an appropriate density cut-off to be applied during the orientation search and this was ~ 1.5 σ for each map. This density "threshold" was found to be essential for meaningful and accurate docking of the models. The first stage of capsid modelling was considered complete once the icosahedral model of highest correlation (representing the most accurate approximation of a, and α geminate approximations) had been determined.

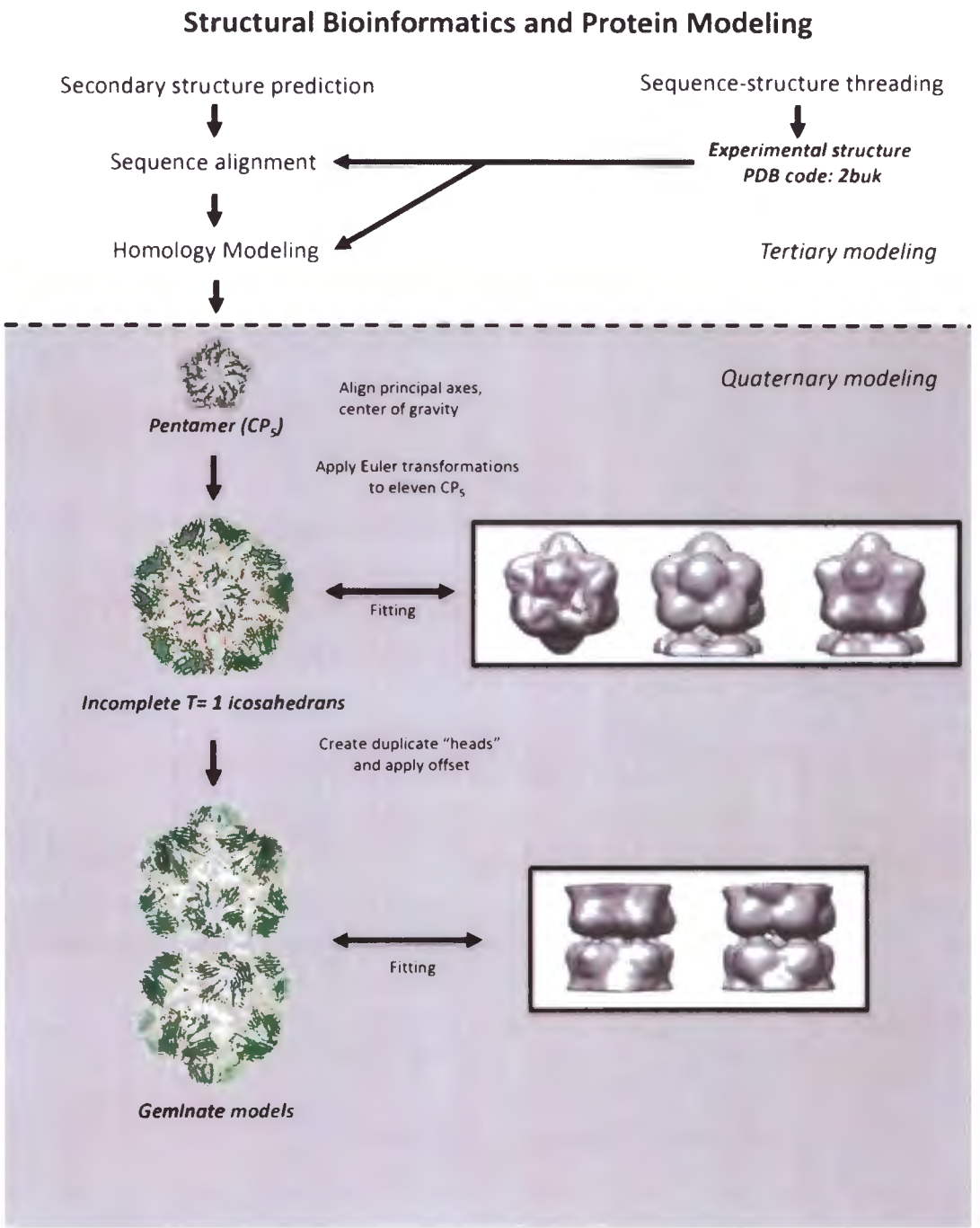

Figure 2-4 Overview of strategy for modelling of the MSV capsid architecture. The MSV sequence is modelled from secondary, tertiary and quaternary levels by homology modelling against the pentameric capsomer of STNV. Following this a series of capsid models are constructed and sequentially fitted into EM density maps in order to determine the most accurate approximations of geometric parameters for icosahedral and geminate capsid construction.

In the second stage of capsid modelling (Figure 2-5, b), a series of geminate models (22 CP_5 units) were created using the icosahedral parameters determined in the first stage. However, in this case a MSV "head" was produced by applying 11 icosahedral symmetry operations to produce an incomplete icosahedron (11 CP_5 capsomers). To produce the other half of the geminate assembly, the incomplete unit was duplicated, and rotated by 180° about the Y-axis. These "heads" were then displaced along the Z-axis by geminate parameter b = 72 – 84 Å, in 2 Å steps. A range of offsets (geminate parameter β = -36 - 36° in 3.6° steps) were then applied to the lower head. This produced 220 models that were then sequentially fitted into the D5-NS and D5-cryo-EM reconstructions. To avoid spurious correlations, as well as to speed the computational time of fitting, the EM reconstructions were windowed (or cropped) so that only most relevant features participated in the computational docking experiment (refer to Figure 2-4). For analysis, fitting (cross-correlation) scores for each map were independently normalized (or scaled) to a range of 0 – 1; 1 representing the highest score obtained. This allowed the scores for any two modelled structures to be more easily compared where fitted into the same EM density map (providing an indication of the geometric parameters associated with the map), but did not allow the scores between different density maps to be assessed. Docking results with highest correlation scores were visualized and analysed using UCSF Chimera (Goddard *et al.*, 2007).

2.5.3 Assessing the divergence of geminivirus coat proteins

A multiple sequence alignment of geminivirus CPs was created which included the type species of each geminivirus genus, as well as the STNV CP. To achieve this, the sequence alignment used for homology modelling, as well as unaligned geminivirus CP sequences were submitted to the MAFFT multiple sequence alignment webserver (Katoh *et al.*, 2002); http://mafft.cbrc.jp/alignment/software/. The MAFFT "*ADD*" function (Katoh and Frith, 2012) allowed homologous CPs to be aligned to the STNV CP using structural information present in the MSV-STNV CP sequence alignment. Sequence alignments were rendered and illustrated using UCSF Chimera (Goddard *et al.*, 2007)

2.5.4 Electrostatic surface potentials of atomic models

The electrostatic properties of viral pentameric capsomers were investigated to assess potential modes of nucleic acid association with these structural units. Electrostatic potential maps were calculated using the APBS (Baker *et al.*, 2001) plugin for the PyMol molecular graphics software package; PyMOLX11Hybrid version 0.99 (Schrodinger LLC, 2010) using default parameters. APBS applied the Poisson-Boltzmann equation to calculate the electrostatic potential of the MSV and STNV CP_5 atomic models, and these potential maps were rendered and inspected in UCSF Chimera (Goddard *et al.*, 2007).

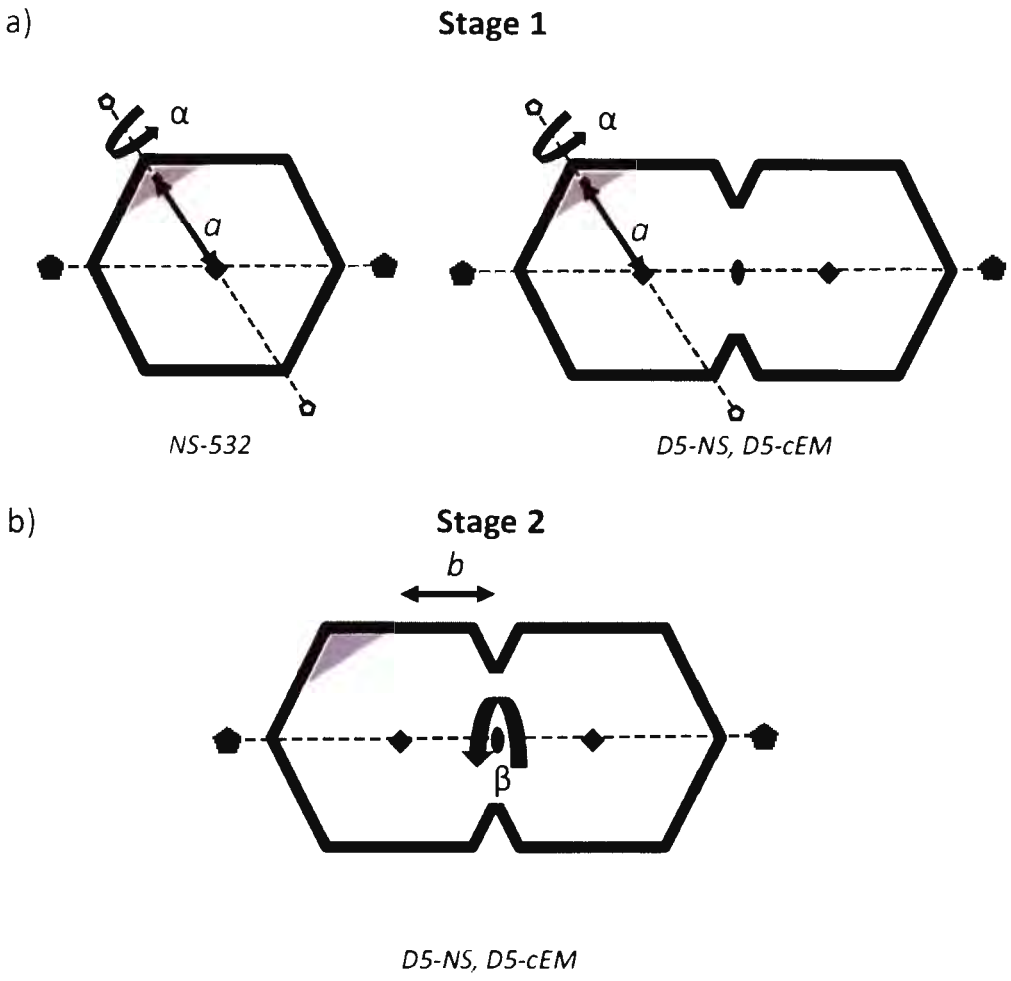

Figure 2-5 Capsid modelling and strategy for determining geometric parameters of the MSV capsid. The fitting of capsid models is carried out in two stages. In a) Stage 1, the orientation (parameter '*α*') and radial displacement (parameter '*a*') of the CP_5 unit (grey triangle) within the context of the quasi-icosahedral geminate "head" is explored. In b) Stage 2, the axial displacement (parameter '*b*') and axial orientation (parameter 'β') of the geminate head is explored. EM maps explored by computational docking are listed below each schematic.

3 Results and Discussion

3.1 Sample preparation and data collection

MSV virions were isolated from MSV-infected maize plants over 4 months. Each isolation successfully yielded sufficient virions for microscopy experiments (1-2 mg in total), but optimization was required to identify ways of removing plant cell contaminants which interfered with imaging, as well as to identify buffer conditions which reduced virion aggregation that was observed to occur over 3-7 days after the final purification step. Generally speaking, adsorption of virions to a carbon support during specimen preparation (either NS or cryo-plunging), appeared to minimized further aggregation which occurred during sample preparation itself, and no major changes to the purification protocol were required to achieve preparations of sufficient quality for the EM experiments.

Examples of infected and uninfected leaves from maize plants are shown in Figure 3-1, a. SDS-PAGE analysis routinely showed preparations consisting of a single protein species (Figure 3-1, b). Contaminants, that appeared to be membranous vesicles and sheets of a wide range of sizes, were assumed to be derived from lipid bilayers during cell disruption, as no band was observed for these on protein gels despite their prevalence in the preparations.

Micrographs of both negatively stained and vitrified MSV virions showed a heterogeneous distribution of particles consisting of multiples of the $T = 1$ quasi-icosahedral unit (~ 22 nm in diameter). The virus purification protocol played an important role in allowing this range of pH stable capsid species to be isolated and observed, as had density gradient centrifugation been used it is probable that rare species, such as those consisting of three or more units, would not have been observed. As expected, bipartite (geminate) assemblies were predominant, however, monopartite, tripartite and higher assemblies were also observed. An overview of species observed is shown in Figure 3-2. The following sections review the negative-stain and cryo-EM datasets recorded for single-particle image processing, as well as subsequent steps of preliminary image processing which further explored the range of particle views present in the data.

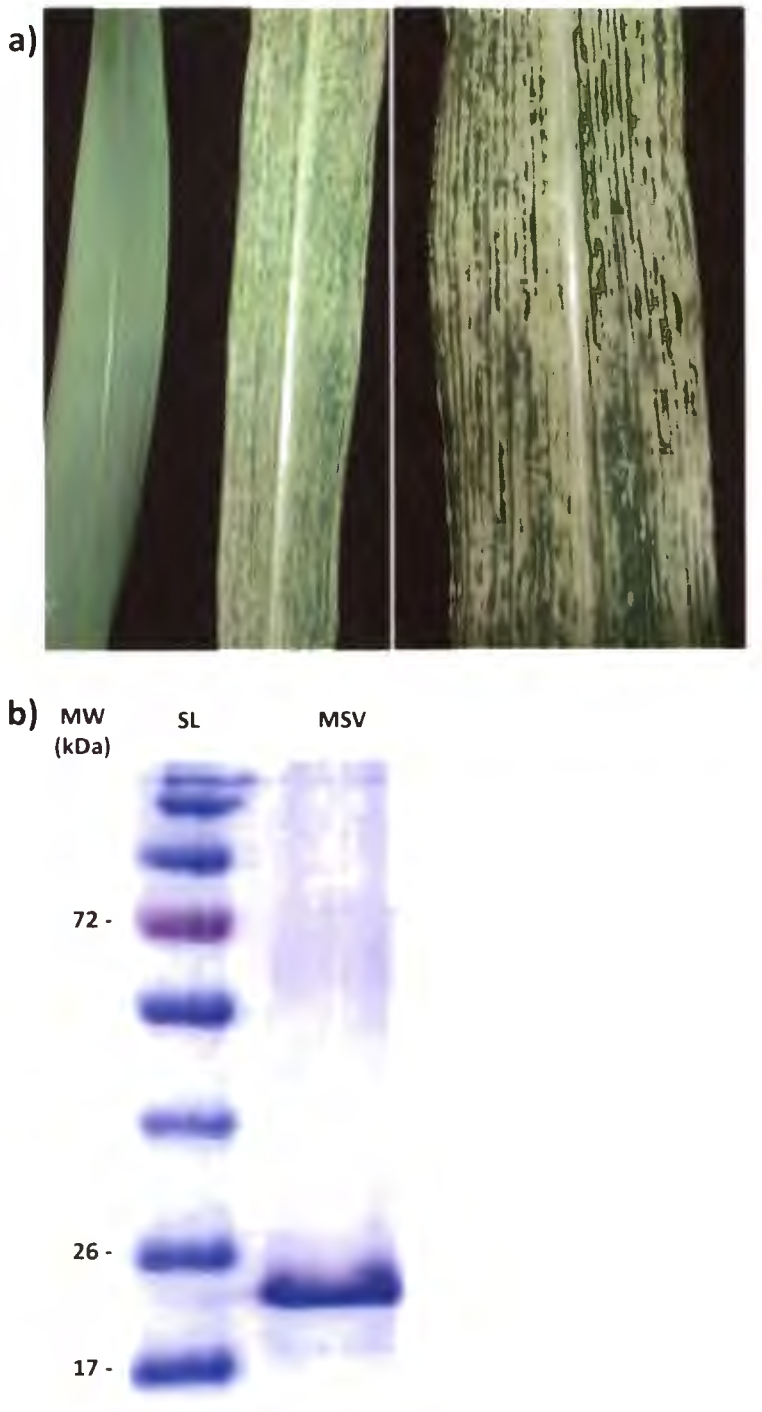

Figure 3-1 Propagation of *Maize streak virus* and analysis of purified fragments. Representative maize leaves from plants used in this study. a) An infected maize leaf (right) is shown photographed alongside an uninfected leaf (left). b) Close up view of an infected Maize leaf revealing the symptoms of diseased plants that were then stripped of their leaves for virus isolation. B) Polyacrylamide gel (SDS-PAGE) of a sample of purified MSV virions (MSV). A single band is observed in the size range of the MSV CP. The protein size ladder (SL) indicates that the protein is ~ 24 kDa in size, whereas the predicted size is 26.918 kDa.

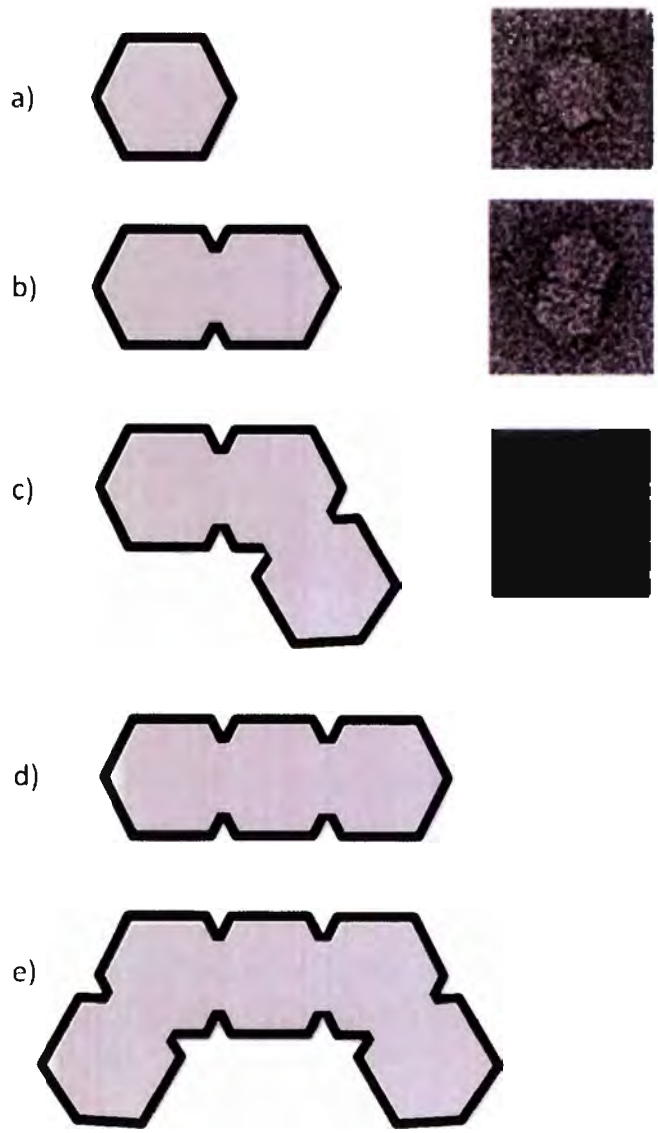

Figure 3-2 Capsid forms observed during the course of this study. Schematic representation of geminivirus capsids observed during negative-stain analysis of MSV virion preparations. The tripartite form (c) was observed relatively frequently, however the straight tri-partite (d) and 'penta-partite' (e) species were observed in a few instances each. Unfortunately, electron micrographs were not recorded of these species.

3.1.1 Negative-stain electron microscopy

3.1.1.1 Early datasets and the monopartite reconstruction

Early EM analyses of purified virions carried out using a JEOL EX II operating at 120 kV were complicated by the presence of membranous contaminants as already mentioned. However, these preliminary images provided important insights regarding the organization of pentameric capsomers (CP_5) within different capsid species. Additionally, only images of monopartite capsids recorded using this microscope could be used for 3D reconstruction and it is important to mention the reasons for this.

At 120 kV, the incident electrons demonstrate stronger interactions with the uranyl atoms of the negatively stained specimen, resulting in improved contrast in the images. This was found to be very useful for close-to-focus imaging, which increases the resolution or fineness of features directly discernable in the images. Micrographs recorded under these conditions were the only ones that revealed fine structures between capsomers, which appeared as serrated or punctuated interactions across the 2-fold axes of symmetry between CP_5 units. 3D reconstruction of these images (Section 3.2.2) was later to reveal these same identical features, and protein modelling allowed regions of the CP sequence to be placed in these areas. It is also worth noting, that 3D reconstruction of the monopartite (532 or icosahedral) capsid images only worked using this dataset. Because MSV has such pronounced capsomers (essentially conical in shape owing to the pronounced 'crowning' insertions with respect to STNV), its capsids show strong orientation bias when adsorbed to a carbon support. As Figure 3-5 reveals, classification performed on preparations free of contaminants showed that monopartite capsids presented predominantly 3-fold views, and this is not surprising because the particle in the shown orientation rest on the support on three separate capsomers (i.e. using these capsomers almost like the legs of a tripod – arguably the lowest energy orientation that the assembly can adopt when placed on a flat surface). The cryo-EM study was able to ameliorate these effects somewhat as the capsids do not interact as strongly with the support because dehydration forces are not experienced, but this only seemed to expand the number of views available to include those in the vicinity of the 2-fold axes. Hence, the studies of pure capsids were limited in angular coverage as well as limited in number of images, and attempts to carry out reconstructions failed for these reasons. In the 120 kV JEOL dataset the presence of contaminants appeared to create an uneven surface that resulted in a better presentation of views. This angular coverage was essential for the subsequently successful 3D reconstruction (described in Section 3.2.2).

3.1.1.2 Optimized dataset for the bipartite reconstruction

A preparation of MSV virions free of membranous contaminants was negatively stained, loaded into the TEM, and 110 micrographs recorded on film. These were then digitized at a sampling of 2 Å/pixel (densitometer step of 10 μm) and analysed for defects such as astigmatism and drift by inspection of local power spectra using functions provided by the SIGNATURE particle selection program (Chen and Grigorieff, 2007). Ninety four micrographs were deemed suitable for further image processing and virion projections (single-particles) were selected and windowed out into regular square arrays (boxes). An example of such a micrograph is shown in Figure 3-3. The inclusion of Triton X-100 into the sample during virus isolation successfully disrupted the membranous contaminant and resulted in a very clean virus preparation after another round of ultracentrifugation. However, an unwanted side-effect of this surfactant was the lowering of the surface tension of the droplet which meant that the sample droplet was prone to excessive spreading, resulting in both sides of the grid being covered when the droplet was applied. Washing with water meant that most of the Triton X-100 was then washed away, however, staining with the uranyl acetate was also problematic because both sides of the grid showed affinity for the stain droplet causing this to spread to both sides of the grid resulting in unequal staining on either side of the carbon support. Fortunately, virions showing appropriate degrees of staining could be identified and only these were used for 3D reconstruction and analysis.

3.1.2 Cryo-electron microscopy

Inclusion of an additive such as Triton-X100 is undesirable for imaging experiments for a number of reasons. Triton-X100 could possible induce the MSV virion to undergo a structural transition by disrupting hydrophobic interactions that might stabilize the virion (i.e. non-native conditions). In addition, it was found that Triton-X100 had a propensity to form bilayer-like layers over time which when adsorbed to the carbon support caused the virions to be "contrast matched" with the surrounding background. Consequently, another means of removing the membranous contaminant were sought for the purposes of cryo-electron microscopy in which contrast is inherently low, and would only be worsened by inclusion of a carbon-containing additive. It was subsequently discovered that the contaminant could be spun out of solution by low speed centrifugation. This procedure resulted in a preparation of suitable purity for cryo-EM, however when virions were plunge frozen (vitrified) to produce thin aqueous layers (estimated to be 100-150 nm thick), further complications arose. MSV virions demonstrated immensely high levels of aggregation when inspected by cryo-EM. At the time, it was wondered whether aggregation was occurring prior to plunge-freezing (in the test-tube), or whether the environment in the thin-aqueous layer immediately post-blotting and prior to plunging was somehow driving the virions out of solution. It has been observed that evaporation from the thin-layer during this brief window

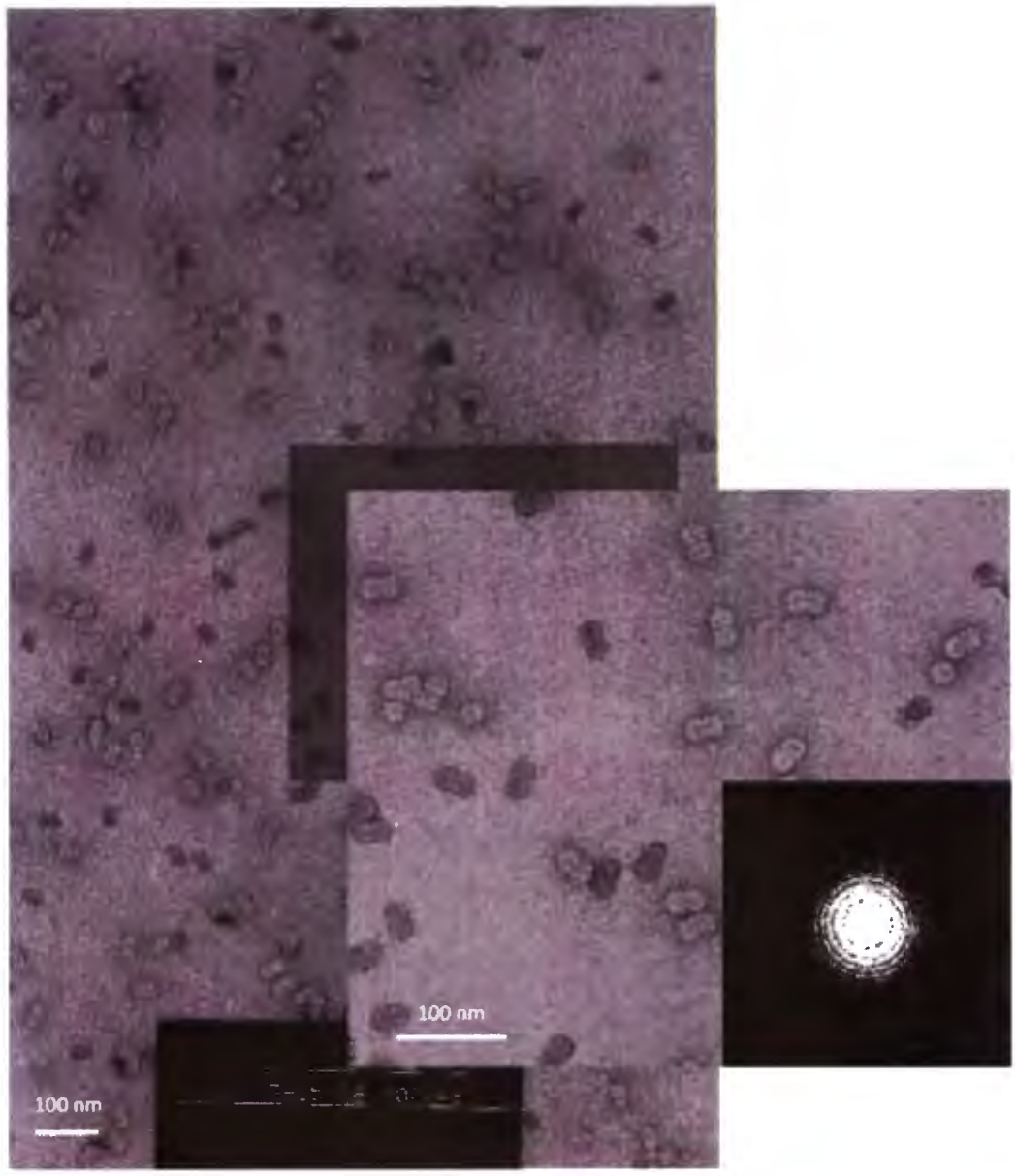

Figure 3-3 Negative-stain dataset used for calculation of the bipartite (geminate, D5) 3D reconstruction. A representative micrograph recorded on film using a FEI Tecnai F20 microscope, showing particles stained using 2% uranyl acetate. The inset shows a region of the micrograph at higher resolution, and provides a local power spectrum that reveals that the micrograph is free of astigmatism. The oscillations of the microscope CTF are also apparent in the power spectrum.

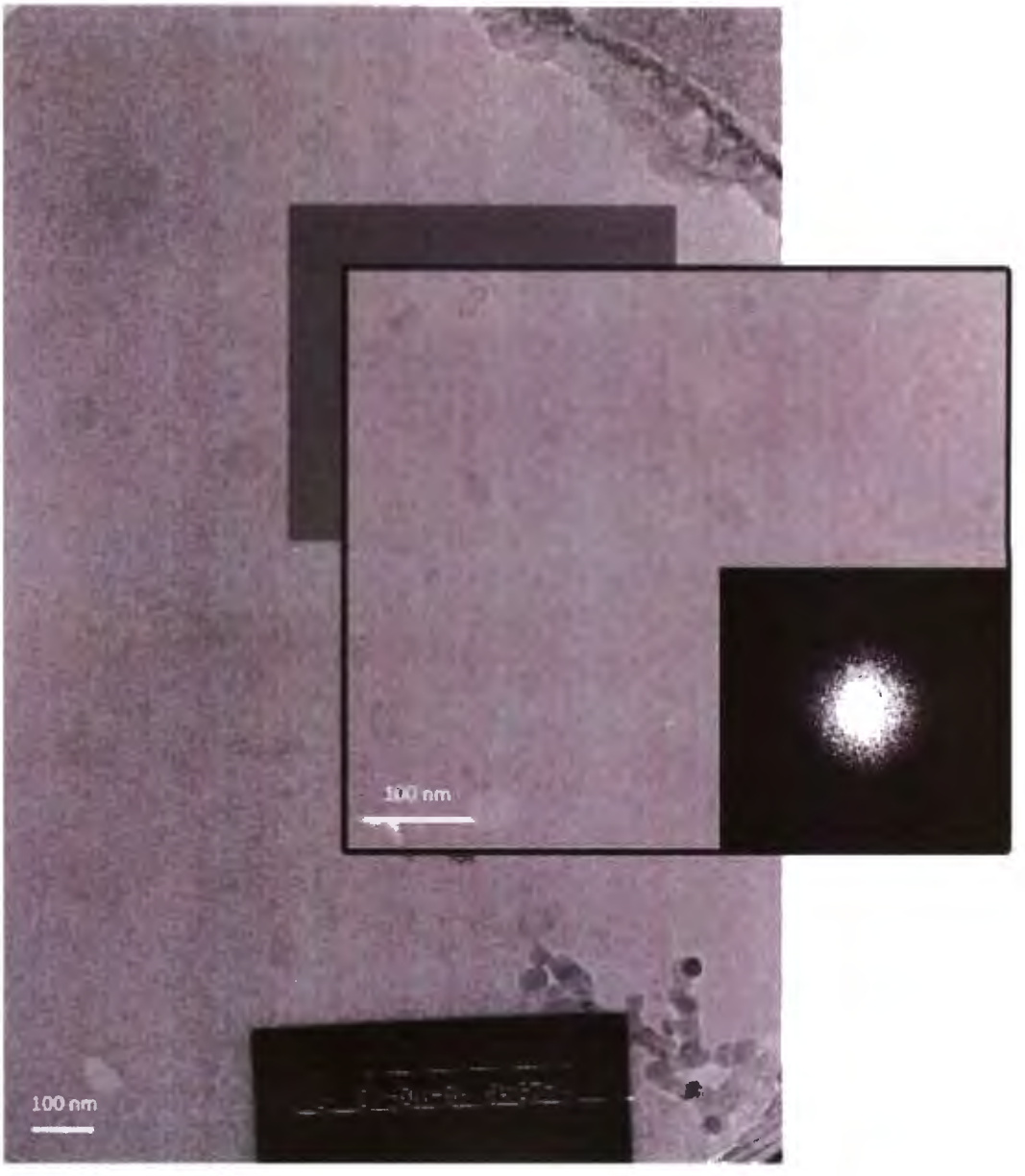

Figure 3-4 The cryo-EM dataset used for calculation of the bipartite (geminate, D5) 3D reconstruction. A representative micrograph recorded on film using a FEI Tecnai F20 showing particles that had been overlaid onto a thin carbon support and vitrified by rapid freezing. The inset shows a region of the micrograph at higher resolution, and provides a local power spectrum that reveals that the micrograph is largely free of astigmatism. The oscillations of the microscope CTF are also apparent in the power spectrum, but are of reduced strength owing to the low signal-to-noise ratio of the cryo-EM data.

of time can concentrate salts or solutes within the buffered medium, and that these can have variable effects on imaging – such as reduced contrast (De Carlo *et al.*, 1999). It is possible that a concentration 'artefact' such as this is responsible for this aggregative behaviour. The MSV virion is known to form pseudo-crystalline aggregates within infected cell nuclei, and it may be that this property has an evolutionary benefit. Hence, at the increased concentrations within the thin-aqueous-layer, the surface (electrostatic) properties of the virion come into play driving the observed aggregation. Nevertheless, the precise cause of this behaviour could not be determined owing to time constraints, but most importantly a means of overcoming this behaviour was found.

Attempts were made using perforated (or 'holey') carbon grids that had been prepared with a thin-carbon overlay over the pre-existing regular array of 2 μm holes (2-5 nm thick). The carbon overlay had been glow-discharged, and the resultant electrostatic properties of the carbon attracted the MSV virions, adsorbing them, and prevented them from aggregating with one another. An example electron micrograph is shown in Figure 3-4. While seemingly essential for cryo-EM imaging of MSV virions, the overlay is far from desirable as it has the effect of reducing contrast in the images. Images were recorded at relatively high defocus values in attempt to make the particles more easily identified in subsequent image processing steps. The carbon overlay has the benefit, however, of contributing a signal of weakly scattering material over a continuous and wide range of spatial frequencies. As such, this feature can improve the accuracy of CTF determination by providing a background against which the microscope CTF is 'imprinted', however the presence of the layer implies that larger than typical datasets would need to be recorded to achieve sufficiently high signal-to-noise ratios at high spatial frequencies during averaging and 3D reconstruction.

3.1.3 Preliminary image-processing and analysis

3.1.3.1 Alignment of particle centres

Alignment of particle centres is required before image clustering or classification can be attempted. During picking, the centres of the particles are only approximated, and hence the alignment of particle centres must be improved for meaningful image classification. The geminate architecture, possessing a long-axis, was found to be somewhat troublesome to centre using approaches which use a reference image owing to the presentation of views of greatly varying dimensions (ranging from circular top views: 220 Å, to rectangular side views: 380 Å) as well as the low signal-to-noise ratio of the images especially for the cryo-EM data. Reference-free alignment techniques were unable to overcome an apparent alignment ambiguity that arose due to most projections consisting of two adjacent spherical motifs (i.e. side views of the geminate architecture). The ambiguity resulted in images being aligned such that only a single head was

identified as the centre of each particle, and caused misalignment of the other half of the bipartite structure, as well as very poor particle centring.

Approximation of particle centres by iterative alignment of each image against its own rotational average was the only approach that improved particle centres consistently. This may be explained by the high signal-to-noise ratio of the rotational averages and the independence of each image during the alignment process.

3.1.3.2 Rotationally invariant K-means classification

K-means classification allowed the quality of the data and distribution of particle views to be assessed. Classification of available experimental images, both negative-stain and cryo-EM, each into 100 groups resulted in the following class averages (Figure 3-5, b and Figure 3-6, b). Importantly, the procedure was able to discriminate between icosahedral, geminate and tripartite capsid species. Because multiple experimental images are averaged, the signal-to-noise ratio of the class averages is significantly improved allowing a detailed analysis of each view, allowing the organization of morphological units to be more easily inspected, and revealed other interesting structural details of the capsids. However, these averages represent *projection* views of the structure, and are difficult to interpret because they represent a superimposition of all 3D structural information onto a 2D plane.

In comparison to most classes that were clearly projected side-views of geminate particles, some class averages were thought to be the averages of different views of an isometric (and not geminate) assembly. These were generally made up of a only a few hundred; refer to the last image for the negative-stain dataset Figure 3-5, b, as well as the last two images for the cryo-EM dataset Figure 3-6. These class averages showed approximately spherical particles with measured diameters of between 22 and 25 nm. These images were nevertheless removed from the datasets used for the geminate reconstruction.

3.1.3.3 Conclusions

The clustering was able to distinguish between icosahedral, geminate and tripartite species, allowing these to be separated and taken further as independent groups during subsequent image processing steps. The number of unique class average views showed (as suspected) that the datasets consisted of predominantly side views. In the case of the cryo-EM data of hydrated virions, the data suggested that an improved angular distribution of views around the long-axis of the particle was available, and that many views with small out-of-plane tilts were also present. Certainly, sufficient data was available to proceed with attempts to carry out 3D reconstruction

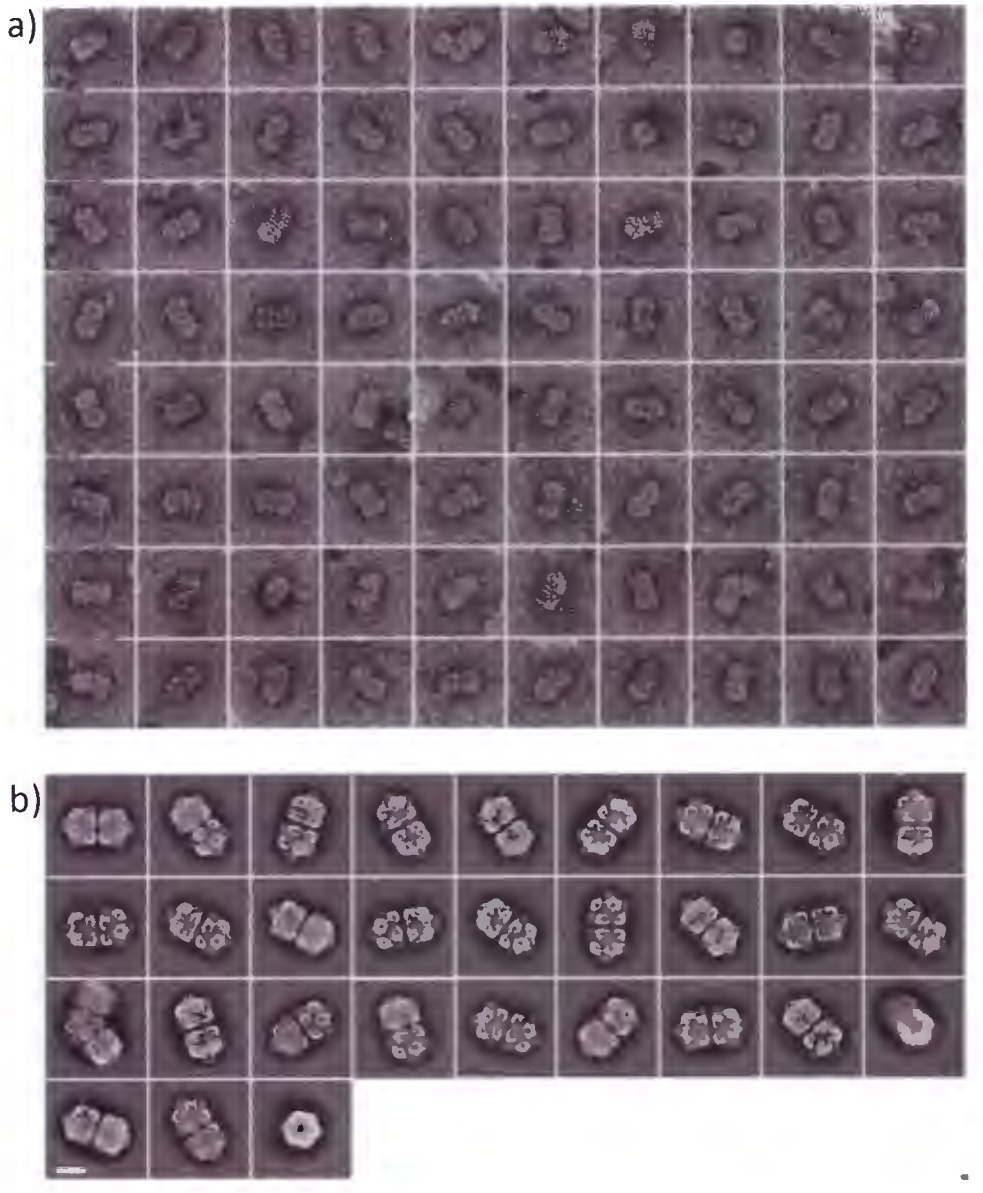

Figure 3-5 A selection of particles used for the negative-stain reconstruction of the geminate capsid. a) Single-particle images of MSV capsids visualized by negative stain EM. b) 30 class averages identified by a single round of rotationally invariant K-means classification performed on pre-centered images. Each image is 95 x 95 pixels, and represents the average of at least 40 experimental images and shown at a scale of 8 Å/pixel. All but one of the views of the geminate particle represent views around the long-axis of the particle (side views). The view in row 3, column 9 represents an oblique view which likely shows the particle adsorbed to the carbon layer on a local axis of 3-fold symmetry. A single class representing tripartite particles is shown in row 3, column 1.The last class shows a distinct view (3-fold) of the single headed (monopartite, icosahedral) capsid. The appearance of the particles reveals that the stain is unable to fill the capsids, as if this were the case, electron dense cores would be apparent. Rather it appears that stain penetration is limited owing to lack of accessibly of the stain to center (i.e. the capsid does not have sufficiently large holes for the stain to pass into the center), or that the cavity of the capsid is largely occupied by nucleic acid. The scale bar represents 20 nm.

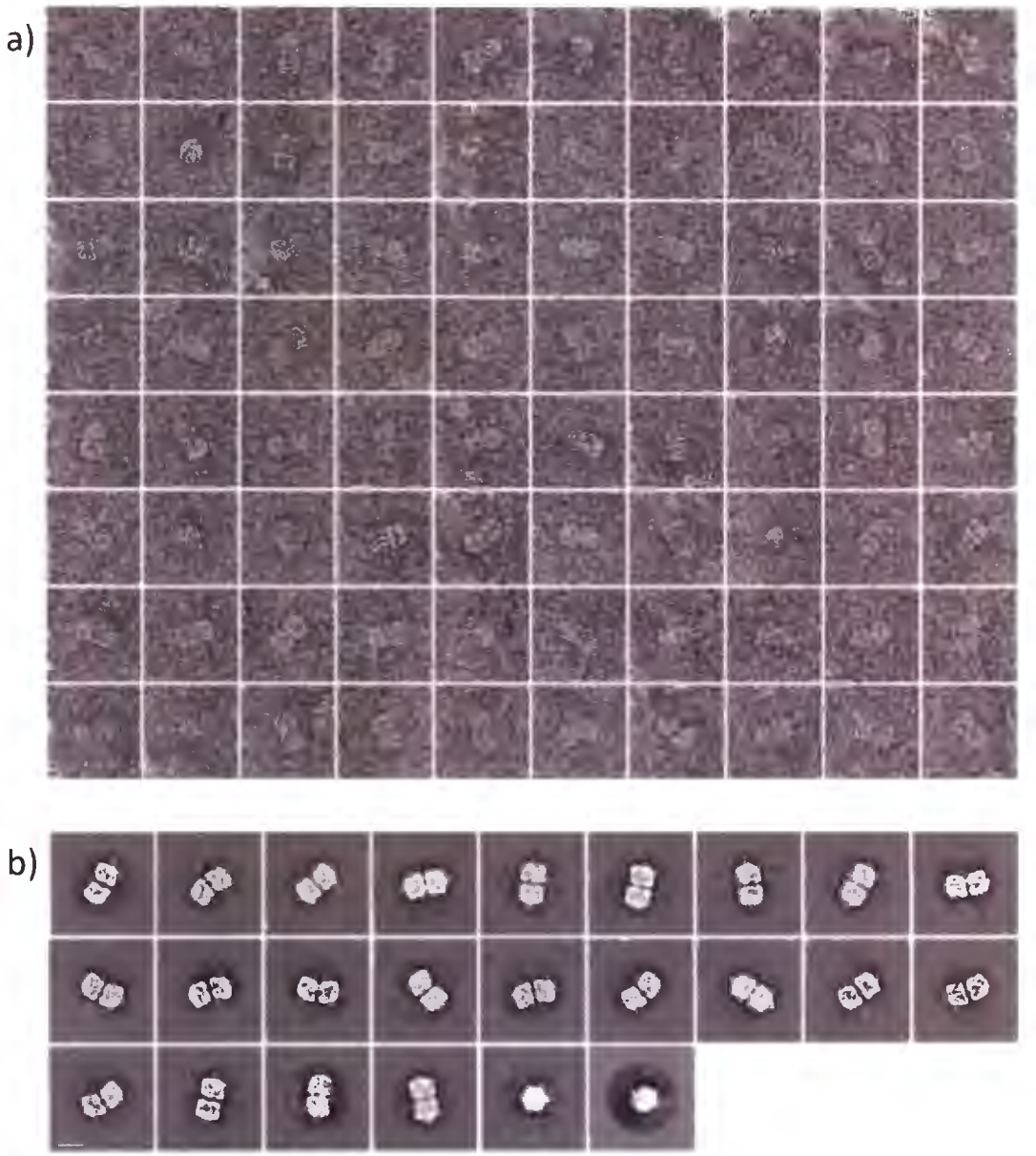

Figure 3-6 A selection of particles used for the cryo-EM reconstruction of the geminate capsid. a) Single-particle images of MSV capsids visualized by cryo-EM. The images show considerably less contrast in comparison to that of the negatively-stained particles shown in Figure x. b) 25 class averages identified by a single round of rotationally invariant K-means classification performed on pre-centered images. Each image is 95 x 95 pixels, and represents the average of at least 40 experimental images and shown at a scale of 8 Å/pixel. The last two classes show two distinct views (3-fold and 2-fold, respectively) of the single headed (monopartite) capsid form. The last and second last corresponding to 2-fold, and 3-fold views, respectively. The scale bar represents 20 nm.

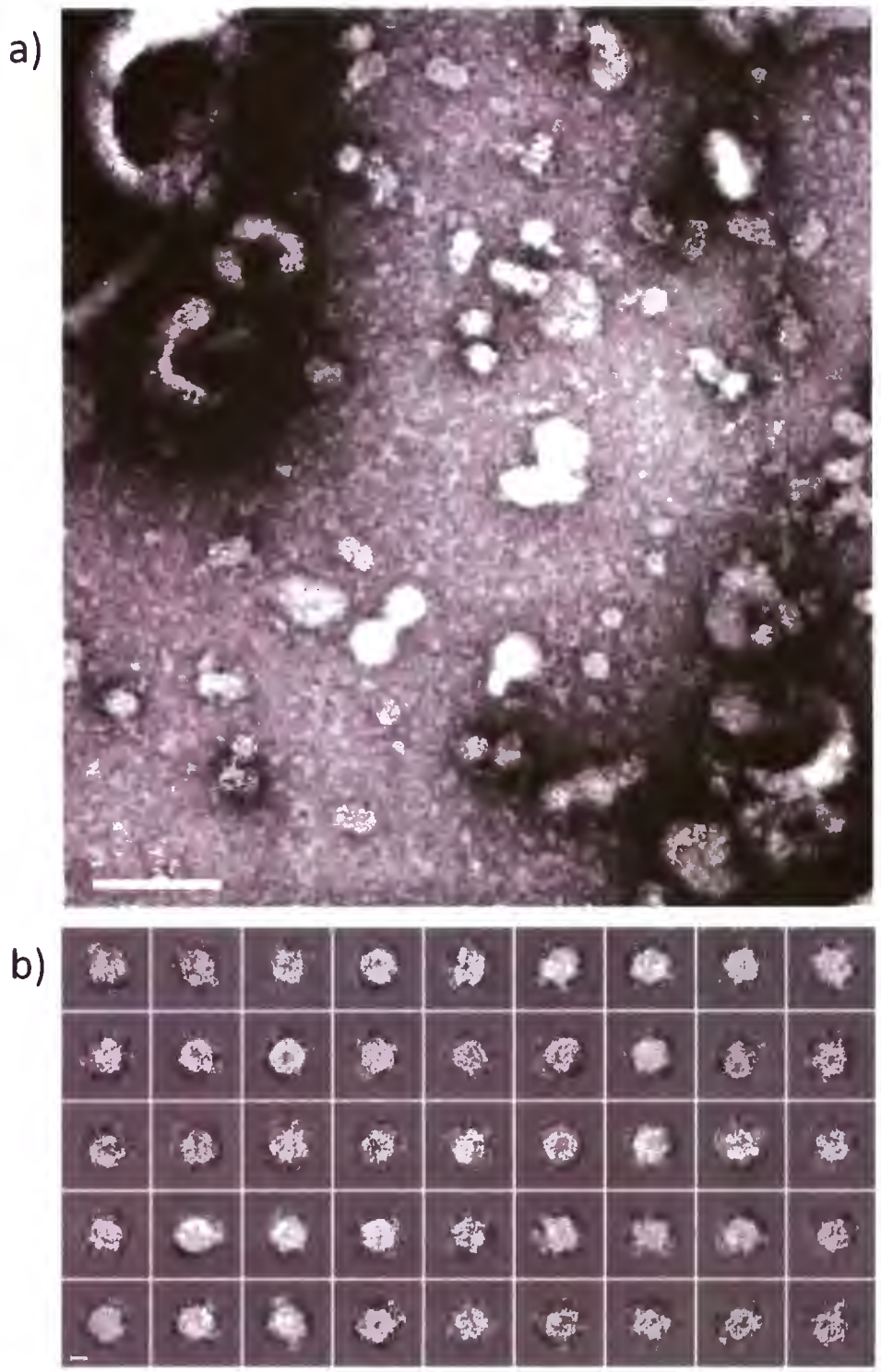

Figure 3-7 Negative-stain dataset used for calculation of the icosahedral (532) 3D reconstruction. a) Portion of a representative micrograph recorded on film using a JEOL EX II TEM. Scale bar 100 nm. b) A selection of images of the icosahedral MSV capsid, these were used for the icosahedral 3D reconstruction of the particle. Scale bar 10 nm.

of the geminate particle, however, whether or not 3D reconstruction of monopartite capsid species would be possible was not clear, as the amount of data available was unfortunately limited.

Notably, empty particles, i.e. particles devoid of ssDNA, were never observed, and virions appeared to always enclose a density that was assumed to be the viral ssDNA. This observation suggests that the ssDNA plays an important role in viral assembly, as well as stability of the assembled virion.

3.2 Single-particle 3D reconstruction

3.2.1 CTF Determination

Contrast transfer function (CTF) determination is a procedure required to identify the imaging parameters associated with an electron micrograph. Based on the knowledge regarding the microscope and imaging experiment (e.g. acceleration voltage, magnification and atomic composition of the specimen), the microscope defocus parameters can be estimated, and these can be used to model the CTF for each micrograph. The modelled CTF can then be manipulated in such a way that phase errors introduced during imaging can be corrected or reversed computationally. This procedure is typically only applied to cryo-EM data, however, the low-dose of cryo-EM micrographs makes this particularly challenging, as the only signal available is generally present in the protein or nucleic acid of the specimen itself. Fortunately, for the cryo-EM dataset described in this thesis, the addition of a thin carbon support strongly facilitates the accurate determination of micrograph CTF parameters.

The defocus range for each geminate dataset is shown in Figure 3-8, and reveal that a wide range of defocuses are represented between 1.2 – 3.3 μm, with very small amounts of astigmatism (data not shown). Defocuses were not evaluated for the monopartite 532 dataset recorded using the JEOL microscope.

3.2.2 Negatively-stained monopartite capsids

Single-particle images of monopartite (single-headed) viral capsids were selected from a dataset recorded using a JEOL EX II (described in Section 3.1.1.1), and subjected to icosahedral single-particle 3D reconstruction. The size of the dataset was very small (~300 images) and a selection of these images is shown in Figure 3-7, b. Reconstruction using a strategy that relied on 3D averaging alone (without calculating class averages by 2D averaging) was successful owing to the 60-fold symmetry (300 x 60 = 18000 asymmetric units) of the structure and produced an extremely featured density map after 10 iterations (Figure 3-9). Unfortunately, convergence to this result was only achieved once, and this refinement used the STNV crystal structure as a

starting model. This was attributed to the small size of the dataset, which meant that iteration from a sphere failed to accurately approximate the orientations of the particles to allow another independent reconstruction of particle. Resolution assessment by standard methods (FSC) was not possible because independent maps could not be calculated, however, the resolution was estimated to be between 15 and 20 Å by comparison to similar computationally calculated and filtered maps of STNV. Higher spatial frequency information may also be present at reduced weighting, and these are most easily inspected in the 2D slices of the density map shown in Figure 3-10.

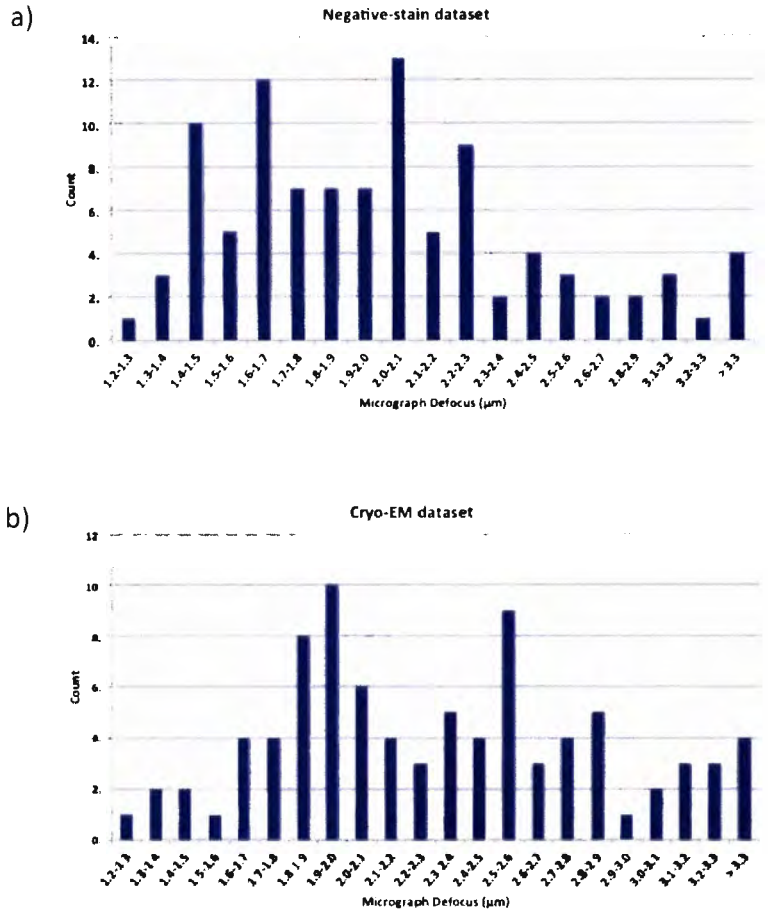

Figure 3-8 Defocus ranges for each of the geminate MSV single-particle datasets. Histograms reveal the distribution of imaging focus parameters used to record each respective dataset. Both datasets show a reasonable defocus spread, which implies that all spatial frequencies for the specimen should be well represented; assuming that CTF correction is applied to the dataset.

The map shows defined pentameric densities (referred to herein as 'capsomeres' or 'morphological units'), in agreement with the previous cryo-EM structure of MSV. Novel features may also be observed in the map, the most notable being a 'pore' on each of the axes of 5-fold symmetry. Additionally, at high isosurface thresholds, fine connections between pentamers are to be observed, and are discussed in a later section (Section 3.3.4.3.1). These agreed with features observed in the negative-stained particle projections. The high definition of the pentamers allowed the handedness of the reconstruction to be evaluated by comparison to the X-ray crystal structure of STNV.

Notably, the arrangement of the pentamers (i.e. capsid architecture) is obviously quite different to that of STNV. The buried surface area at the interfaces between capsomeres is reduced, suggesting vastly different inter-capsomer interactions, and owing to the increased diameter of the particle it clear that these differences result in assembly of a capsid with increased internal volume. It is this difference which explains why the MSV $T = 1$ "heads" are of an increased diameter in comparison to STNV, despite their structural units (pentamers) being of such similar sizes. The pentamers are clearly rotated about their 5-fold axes, and the principle of steric exclusion implies that the capsid must expand to accommodate this rotation.

As might be expected for a reconstruction of negatively stained particles, no clearly defined density for the genome is observed, although this does not indicate that none is present as it is apparent that an additional, somewhat lower density layer is present on the inside of the capsids (refer Figure 3-9 regarding interior of the EM map, and Figure 3-10 to inspect densities of a 2D central section through the reconstruction). 'Thresholding' the map to high levels, did not reveal any details about the possible structure of nucleic acid associating in this region, and it was hoped that the cryo-EM reconstruction might provide further details regarding such considerations. That a layer of nucleic acid is visible on the interior is a surprising result, and suggests that the particles are at least partially permeable to the uranyl acetate stain. It is possible that the stain entered the particles and accumulated to some extent within the core of the particle allowing the additional material to be represented in the final reconstruction. Alternatively these features appeared in the map as a result of a small component of phase contrast in the images, which with the high symmetry averaging used, is sufficiently reinforced to appear as a relatively strong density in the map. This result also implies that the nucleic acid in the monopartite species is in the majority associated with protein layer, and does not extend much beyond this layer into the central regions of the container (refer to Sections 3.3.6 and 3.4.2 for further discussion).

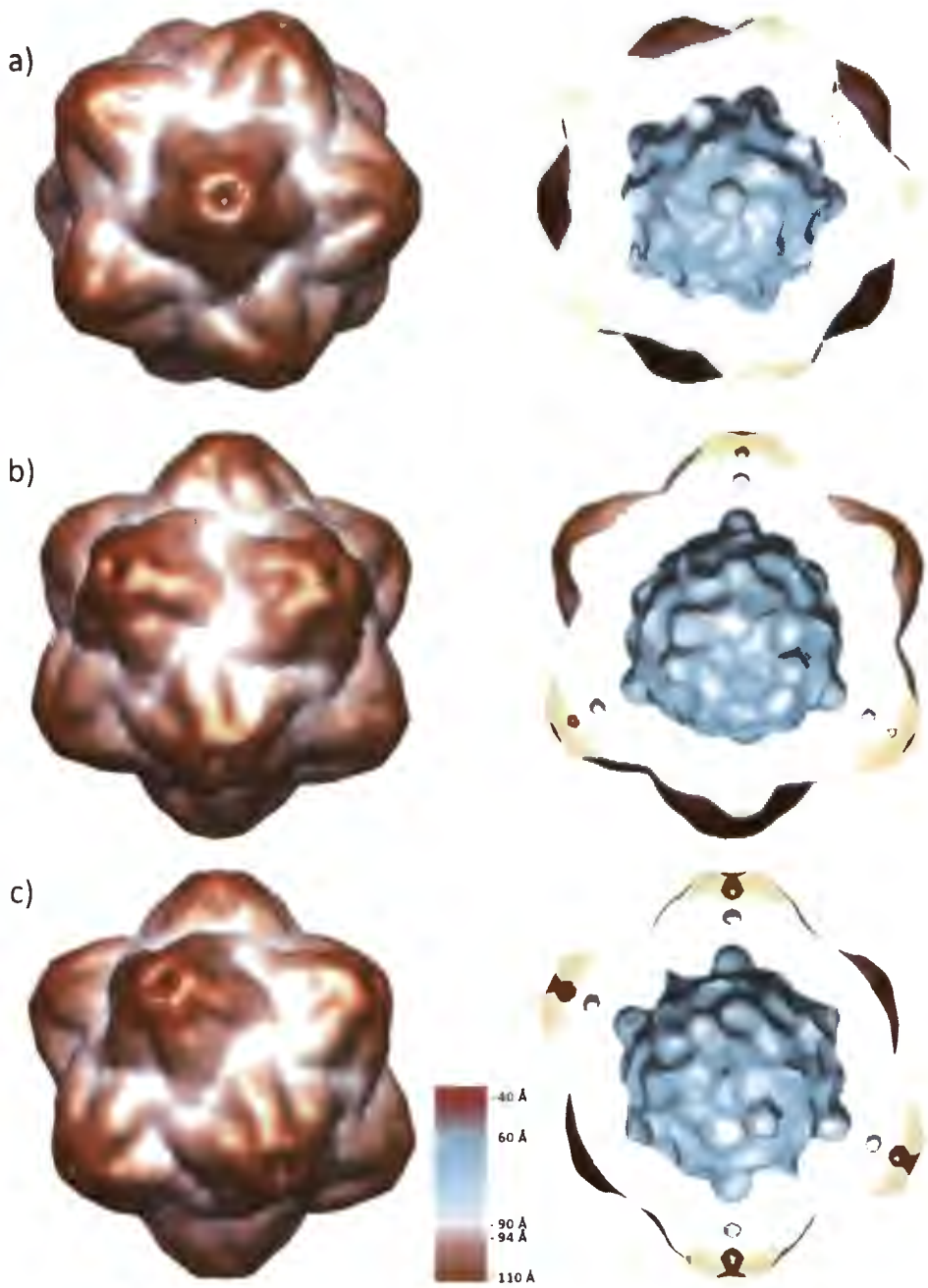

Figure 3-9 3D reconstruction of the monopartite (icosahedral, 532) MSV capsid. The reconstructed volume is rendered as isosurface and coloured according to radius (refer to key). a) 5-fold, b) 3-fold, and c) 2-fold symmetric views are shown (left). a/b/c, right) Equivalent views to a/b/c, left) but only the rear half of the capsid is shown. The pentameric morphological units of the capsid are clearly defined, and a 'pore' is visible at the 5-fold symmetry axes. At the resolution of the reconstruction, the sliced views reveal that the pore is not continuous through to the center of the particle.

3.2.3 Negative-stain geminate reconstruction

Two 3D reconstructions were calculated from the 10878 image negative-stain dataset from two independent starting-models using iterative-projection-matching (Figure 3-11). Figure 3-12 shows that the two reconstructions show good convergence to two independent solutions that correlate almost identically (global average cross-correlation co-efficient) with the underlying dataset. The reconstructions both converged within 5 iterations. Increasing $\Delta\theta$ to 3° (429 projections, and highest theoretical resolution of ~15 Å according to the Crowther criterion) did not result in an improvement of resolution, but did increase the average correlation of images with each model, resulting in an increase in average correlation (data not shown). This showed that the images themselves were resolution limiting and not the geometry for back-projection. Importantly, 98% of images showed convergence to a single reference (view) during the course of the refinement, and the cross-correlation scores showed a Gaussian distribution (Figure 3-13, a), improving our overall confidence in the validity of the alignment. Inspection of volumes calculated showed good agreement as expected, and Fourier shell correlation suggested that structure factors agreed to a FSC 0.5 resolution of 1/26 $Å^{-1}$.

As expected, the map represents a bipartite structure made up of two $T = 1$ incomplete icosahedra which are apparently offset from one another by a well-defined value; any flexibility or ensemble of offsets would result in smearing during 2D averaging and 3D reconstruction. The map shows clear pentameric capsomeres, which have a defined orientation within the context of the icosahedral heads. Although at the resolution achieved, these do not show high-resolution features that might be compared with the 532 reconstruction directly. Thresholding the reconstruction at different levels was informative revealing that densities at the local axes of 3-fold symmetry are somewhat lower than the remainder of the map (not shown). Indeed, the densities appear concentrated at the capsomeres, and the connections between these structural units are strongest across interactions at each local 2-fold axis of icosahedral symmetry. This is a significant difference in comparison to the 532 reconstruction that shows well-defined interactions at both the 3-fold and 2-fold axes.

In comparison to the 532 reconstruction, no clear evidence for packaged genetic material is apparent, even though we can be sure that the particles had a contents (based on the appearance of the images in negative-stain). The inner surface of the particle is of a different quality to that of the negatively stained monopartite (icosahedral) reconstruction (Figure 3-11). Pronounced circular "dimples", corresponding to the local 5-fold axes of symmetry are visible, and it would appear that the map accounts for only the β-barrel domains (the cores of which appear to

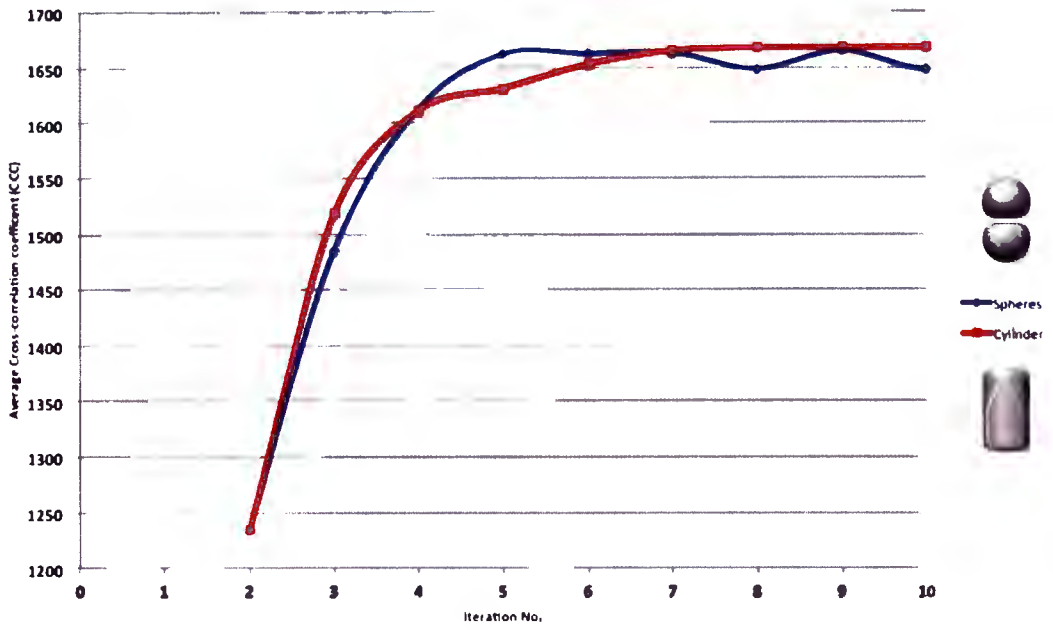

Figure 3-12 Convergence of negative-stain geminate 3D reconstructions. Two 3D reconstructions were carried out by 3D projection matching against two independent starting models (model of spheres, and cylinder). In both cases the average cross-correlation for the whole dataset (global) was calculated and is plotted for each as a function of the round of projection matching. It may be observed that both reconstructions converged to very similar results.

correspond with the "bumps" on the internal surface, as there is no density present which can be assigned to the either the N-terminal arms or any genomic material.

A mentioned previously, the equator of the particle is thought to be the site of conformational variability in the coat protein subunits, and is the area where quasi-equivalent interactions must, therefore, occur. In the case of the negative-stain reconstruction the equator shows only one type of interaction across the groove. Each one of these interactions coincides with one of the global 2-fold symmetry axes of the D5 geminate particle. Furthermore, and of great importance, these global 2-fold axes appear to correspond with what would be the 2-fold symmetry axes of the icosahedral particle at the site were the interaction is observed to occur. Following the global 2-fold axes to the opposite side of the particle, we observed the global axis coincides with the 3-fold symmetry axis of the icosahedral particle – although no interactions are observed at these sites across the equator. Figure 3-11, d allows these interactions at the equator to be evaluated.

3.2.4 Cryo-EM geminate reconstruction

The cryo-EM dataset was initially reconstructed using the same procedures as the negative-stain dataset, but the reconstruction (along with dataset parameters) were subsequently inputted into FREALIGN (Grigorieff, 2007) so that CTF correction could be applied. 3D reconstruction from 4000 single-particles yielded a map at an estimated resolution of ~ 23 Å (Figure 3-14). The statistics associated with the reconstruction were similar to those of the negative-stain reconstruction, suggesting that the reconstruction was valid (data not shown). That the dataset was smaller than that used for the negative-stain reconstruction was not desirable as the signal-to-noise ratio of the final reconstruction would be relatively low given the low SNR of the cryo-EM images. CTF correction and independent refinement of particle parameters allowed information in the images to be extended past the first zero of the microscope CTF, however, the FSC curve reveals that the dataset was apparently not ideal with respect to the defocus coverage; this is revealed by the oscillations of the CTF curve that show 'troughs' of low correlation. Hence it may be concluded, that the performance of CTF correction step was not as effective as was initially hoped, and that the dataset was not large enough to overcome noise inherent to the unstained cryo-EM experiment.

The cryo-EM reconstruction of the geminate capsid has important similarities with that of the NS reconstruction. A bipartite structure consisting of two $T = 1$ heads is observed, and by inspection alone, it is evident that these are offset at a similar angle. Notably the heads of the geminate assembly are connected by 10 interactions of two types, rather than 5 interactions of one type as observed in the NS reconstruction. These interactions were observed in the previous cryo-EM reconstruction of MSV (Zhang *et al.*, 2001), but once again the interactions observed in this

reconstruction are not cylindrical in morphology and for this reason it was deemed unlikely that these are produced α-helices.

The maps show considerable differences with regards to arrangement of internal densities. The interpretation of these is limited by the resolution available, but comparison reveals that the capsid shell is considerably thicker in the case of the cryo-EM reconstruction; this possibly owing to the visibility of nucleic acid associated with the inner surface of the capsid. In additional support of this assertion, the inner surface of the shell has a different appearance to the NS reconstruction, and in addition there is a second (inner) shell of density. The inner shell is coloured red in Figure 3-14 and Figure 3-15.

Further analysis of the inner surface of the cryo-EM reconstruction reveals previously unseen indentations at the local axes of 3-fold symmetry. The indentations are triangular in shape (despite 3-fold averaging not occurring in these regions), but are difficult to observe in Figure 3-14 and Figure 3-15 because of a cylindrical motif lying immediately below them. That these regions are so regularly triangular suggests that a protein component must be ordered below the capsomeres at each axis of local 3-fold symmetry. In the case of STNV it is known that N-terminal arms descent into the capsid at these sites and make interactions with one another. The observations made of this EM map suggest similar behaviour for MSV. The indentations at the local 5-fold are circular in comparison, and do not inform us of any possible structural arrangement in these sites. While no such indentations are observed at the 3-fold axes of the negative-stain geminate reconstruction, these are remarkable pronounced, and lie immediately beneath (from the inside looking outward) tubular densities that at first thought appeared to be nucleic acid. These tubular densities. The circular indentations characterizing the 5-fold axes are not as pronounced as those visible in the negative-stain geminate reconstruction, but are comparable to those of the icosahedral 532 reconstruction suggesting that these corresponding to a layer of nucleic acid in these instances.

The tubular densities of the inner shell are shown in a range of different views in Figure 3-15, and might represent single-stranded DNA adopting a specific arrangement around these areas in a pattern that is influenced by the symmetric averaging on the D5 object. Alternatively, the tubes may also represent regions of double-stranded DNA, however, at the resolution presently available no grooves for the RNA are apparent, and based on their diameter, it appears more likely that ssDNA segments may be ordered to low resolution by interactions with the N-terminal arms of the CPs of the MSV capsid.

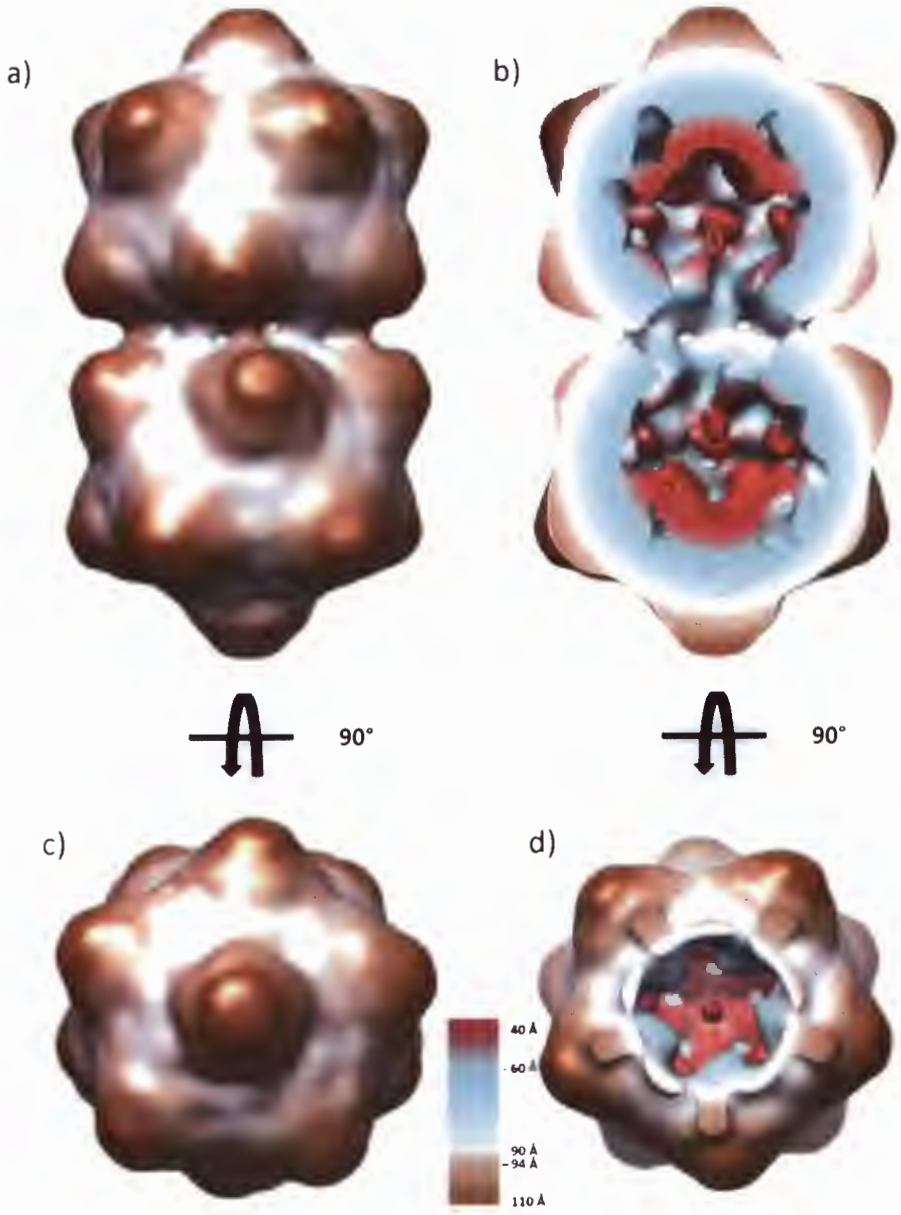

Figure 3-14 The cryo-EM reconstruction of the MSV geminate capsid. The reconstructed volume is rendered as isosurface coloured by radius (refer to key). A selection of views are provided along two different axes of global symmetry. A side view looking down an axis of 2-fold symmetry, reveals the twin $T = 1$ incomplete icosahedral architecture characteristic of the geminate morphology. b) An identical view in which only the rear half of the capsid is shown. Unlike the negative-stain reconstruction (**Figure 3-11**), the inside of the capsid is rich with features which may correspond to regions of the genome or regions of the CP N-terminal arms which may be sufficiently ordered at this resolution. a and b) Show that five interactions of two types may be observed to take place across the equator of the geminate particle. c) A view down the 5-fold axes of global symmetry. d) Same view as in c) but the top half of the capsid is clipped away revealing only the bottom half of the capsid. The reconstruction clearly reveals morphological units that consistently demonstrate 5-fold symmetry or 5-fold pseudo-symmetry (in equatorial regions) as expected.

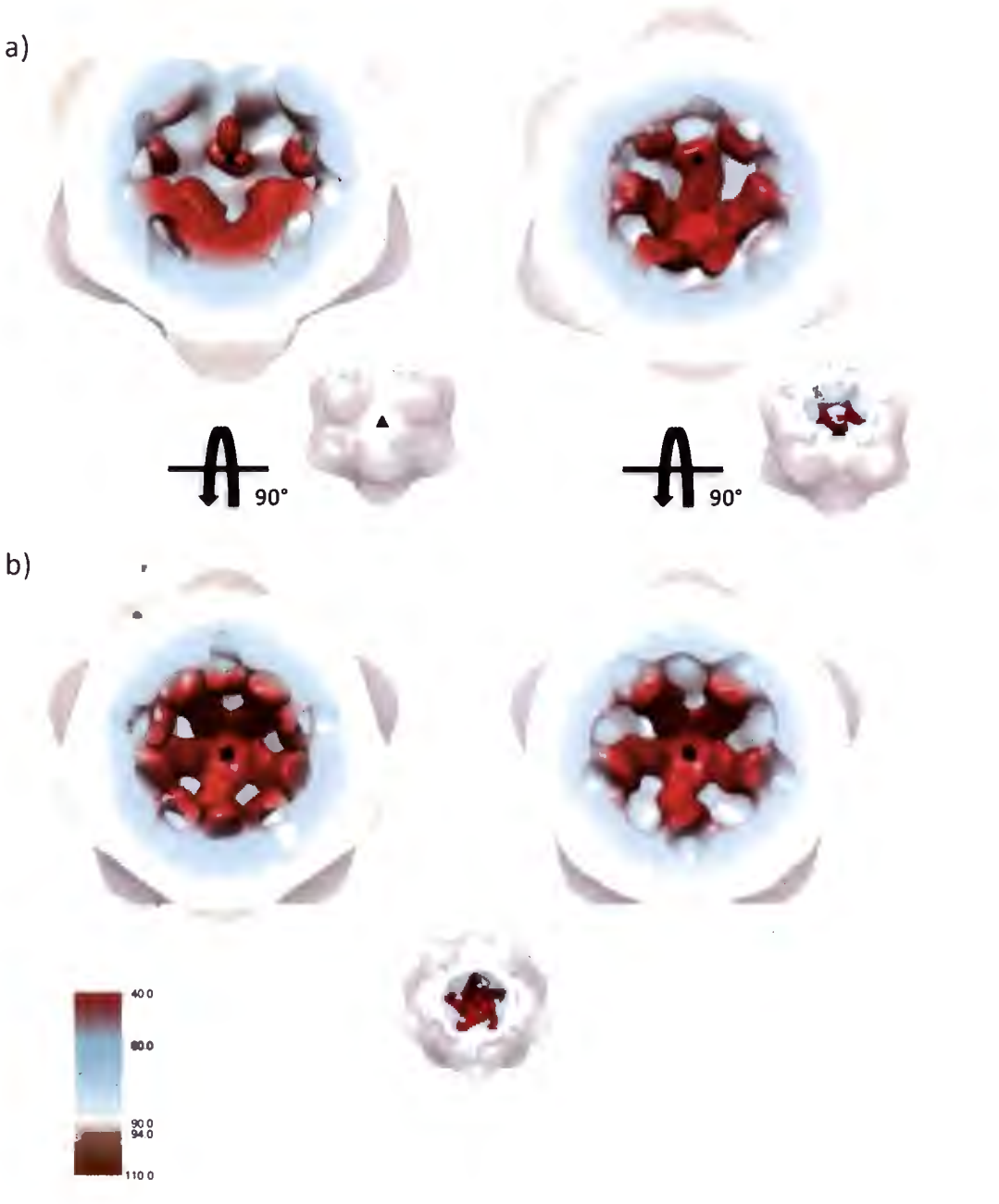

Figure 3-15 Internal structure visible in the cryo-EM reconstruction. A single head of cryo-EM geminate reconstruction is viewed with the front half clipped away. Views corresponding to a) a view along an axis of local 3-fold rotational symmetry, and b) a view along the global axis of 5-fold rotational symmetry. Tubular densities (colour red) are apparent on the inside of the structure at radii that suggest that they are unlikely to be contributed by the core domains (β-barrel) of the coat protein subunits. It is uncertain whether this density can be attributed to the N-terminal arms of the CP subunits, or to the MSV genome or a complex of both which is inappropriately averaged during 3D reconstruction.

A possible explanation for the increased thickness of the primary layer of density is the association of dsDNA (possibly stem-loops) onto the inside layer of the capsid. This may be averaged away in the negative-stain bipartite reconstruction owing to variable staining (depth and penetration of the particles), but is represented in the cryo-EM reconstruction because the image is produced independently of stain presence. The representation of these features afforded by D5 averaging is not to be considered wholly accurate with regard to their precise structure, but is instead informative regarding their spatial/radial distribution.

3.2.5 Determination of handedness, and comparison of 3D reconstructions

Knowledge of the correct hand of a chiral structure is crucial for accurate interpretation. If sufficient resolution is available this can be confirmed by inspection of α-helices (which are right-handed), however, in cases where resolution is limiting, cross-validation approaches must be applied. Owing to time constraints, as well as difficulties with the goniometer of the Tecnai F20 microscope, this study did not record untilted-tilted projection pairs of the geminate particles so that handedness could be assigned to the reconstructions by cross-correlation (cross-validation with information derived from experiment). Instead the hand of each 3D reconstruction was assigned by visual comparison of these reconstructions to the original reconstruction carried out by (Zhang *et al.* 2001), which did include hand determination by tilting, and also by cross-correlation fitting of the STNV capsomer into the 532 reconstruction. The relatively high-resolution information present in the monopartite (532) reconstruction revealed the "precession" of β-barrels (which, in the context of the pentameric capsomer, contribute to a motif that is not unlike a Catharine wheel in appearance), and this quantitative docking provided another line of evidence that could be used to bolster confidence in the assignment of hand to each reconstruction. Using these two pieces of information, visual comparison of the orientation and inter-capsomere interactions in each respective map could be used to an unambiguously assign the accurate hand in each case by mirroring the maps when necessary. This assignment was later supported by agreement of docking (model fitting) parameters.

Figure 3-19 shows a superimposition of the two density maps for the geminate reconstructions. Despite different surface appearances of the capsomeres, it is important to note that the overall dimensions of each reconstruction appear to be almost identical. This is surprising given that dehydration during negative-staining is known to "flatten" macromolecular complexes in most instances, causing an over estimation of the particle dimensions by as much as 20% in many cases. Thus, that the dimensions of the particles are so similar, can be taken as a strong indication of rigidity of the complex. Additionally, good agreement in alignment between capsomers of the heads of the geminate structures suggests that the offset between incomplete $T = 1$ icosahedra is

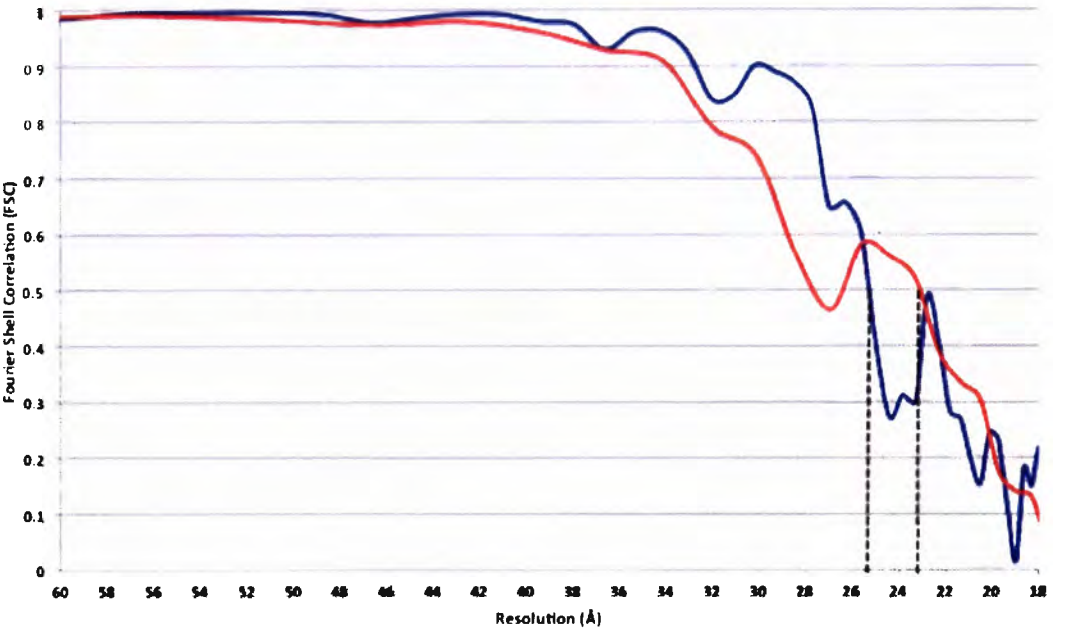

Figure 3-16 Fourier shell correlation curves for cryo-EM geminate reconstructions. FSC curves area shown for the cryo-EM reconstructions in which uncorrected (blue), and CTF corrected images were reconstructed. The curves suggest that (according to the FSC 0.5 criterion) the resolution of the maps is ~ 26 and 23 Å, respectively.

also in very close agreement, again despite extreme forces experienced by the particle during dehydration. To truly confirm whether this is the case, image classification algorithms could be used to attempt to sort the images out into classes corresponding to populations varying in offset. During this study this was not attempted owing to time constraints, as well as the limited sizes of each dataset, however, such an analysis might be useful for future studies. Certainly the resolution of the maps limits our confidence in the stability of the particle, but the agreement observed, appears to suggest that the particle is immensely stable, and this can be rationalized in light of many protein-protein and protein-ssDNA interactions that contribute to the assembly of the particle. Additionally, we report the first evidence that divalent ions (e.g. Ca^{2+}) carry out a structural role in the capsid, and these are known to further stabilise such assemblies (Persson *et al.*, 2008).

Central slices taken from the two geminate maps appear to provide complementary information regarding the organization of the interior of the MSV virion (Figure 3-17, top and Figure 3-18, top). Additionally, certain of the central slices shown in Figure 3-17 provide excellent views of the pentameric capsomeres, and reveal that these agree very well between the reconstructions. Inspection of central slices of the NS-reconstruction provides further evidence that the negative-stain reconstruction corresponds to the core β-barrel domains of the virion capsid only, with genome and CP N-terminal arms not appearing in the reconstruction. Features such as these are rarely visible in a NS-reconstruction because of inadequate stain penetration; the stain cannot enter the particle, or enters the particle in a variable way, meaning that these features are not reinforced during averaging. Additionally, the genome is possible disordered and certainly does not correspond to the D5 symmetry imposed during reconstruction. Observations mentioned in the previous sections have suggested that the MSV capsid is at least partially permeable to the stain, and that at least some stain enters the particle to also encase the genome. The false colour representation (Figure 3-18, right) is surprisingly informative regarding the possible arrangement and location of the genome, as this depiction (unlike the isosurface representation) allows us to observe very weak densities within the maps even in the case of the NS-reconstruction. Inspection of the slice taken from the NS map will show that the very central densities of the map are very weak, comparable to the map background, while there is some indication of internal densities nearer to the capsid interior surface. The outline thus created by this differential contrast appears to suggest that the genome is tightly compacted near the capsid interior (preventing any stain penetration), and this is likely to be a result of extensive secondary structure as well as regions in tight association with the N-terminal arms of the CP subunits. On the other hand, the core of the cryo-EM reconstruction is filled with densities, which are not as intense as those of the CP lattice of the capsid itself but are nevertheless worth analysing. The cryo-EM map is reconstructed using

information arising from scattering of electrons by the protein and nucleic acid itself (and not an envelope of stain) and hence will provide complementary information. Coinciding structural features will be reinforced during reconstruction and averaging, and in this case, it is the presence of nucleic acid immediately adjacent to (and in probable association with) the coat protein layer that is highlighted by reconstruction. It is the repeated interaction of these components with the inside of the capsid which produces this result, however, the map provides no evidence that the CP layer constrains the orientation of the nucleic acid by specific interactions at the resolution achieved in this study. In addition to the primary layer of nucleic acid, a second shell of density is visible in the quasi-icosahedral heads. This is of comparatively much weaker density, and this is understandable given that no rigid interactions (such as those with the capsid) are constraining the precise nature of the structural features present in these regions. As mentioned already, it is likely that the N-terminal arms contribute to the density visible in these areas.

Visual analysis of the geminate capsids revealed that the offset between heads is ~ -11°. The capsid geometry of isometric viruses is often analysed with respect to a hexagonal lattice, and it makes sense to expand this representation to incorporate the geminiviruses so that the implications of the head offset can be evaluated more easily. A means of representing the geminate capsid is to combine two separate hexagonal lattices constructed according to the $T = 1$ arrangement and this is shown in Figure 3-20. Each lattice is modified such that the asymmetric unit is modified near the equatorial regions to represent the loss of 5 monomers (a single pentamer), and joined in such a way that the offset of -11° is represented with the correct handedness. Axes of global symmetry have been marked with appropriate symbols. Notably the representation reveals that the offset appears to bring residues that would contribute to the formation of interfaces at the 2-fold and 3-fold axes of icosahedral symmetry into close proximity. This is a significant observation which reveals that the assembly has an arrangement that appears to conserve the constituent interacting surfaces of the quasi-icosahedral heads (i.e. those also forming the icosahedral monopartite species), or that independent surfaces are in close proximity to one another.

3.2.6 Conclusions

Single-particle electron microscopy has been used successfully to carry out the 3D reconstruction of two of the MSV capsid species (monopartite and bipartite). The maps provide insight into the architecture of protein subunits as well as the arrangement of the genome, although these insights are limited by the imposition of symmetry, as well as the limited resolution of the reconstructions. The specimen provided unexpected challenges, but nevertheless the image processing strategy

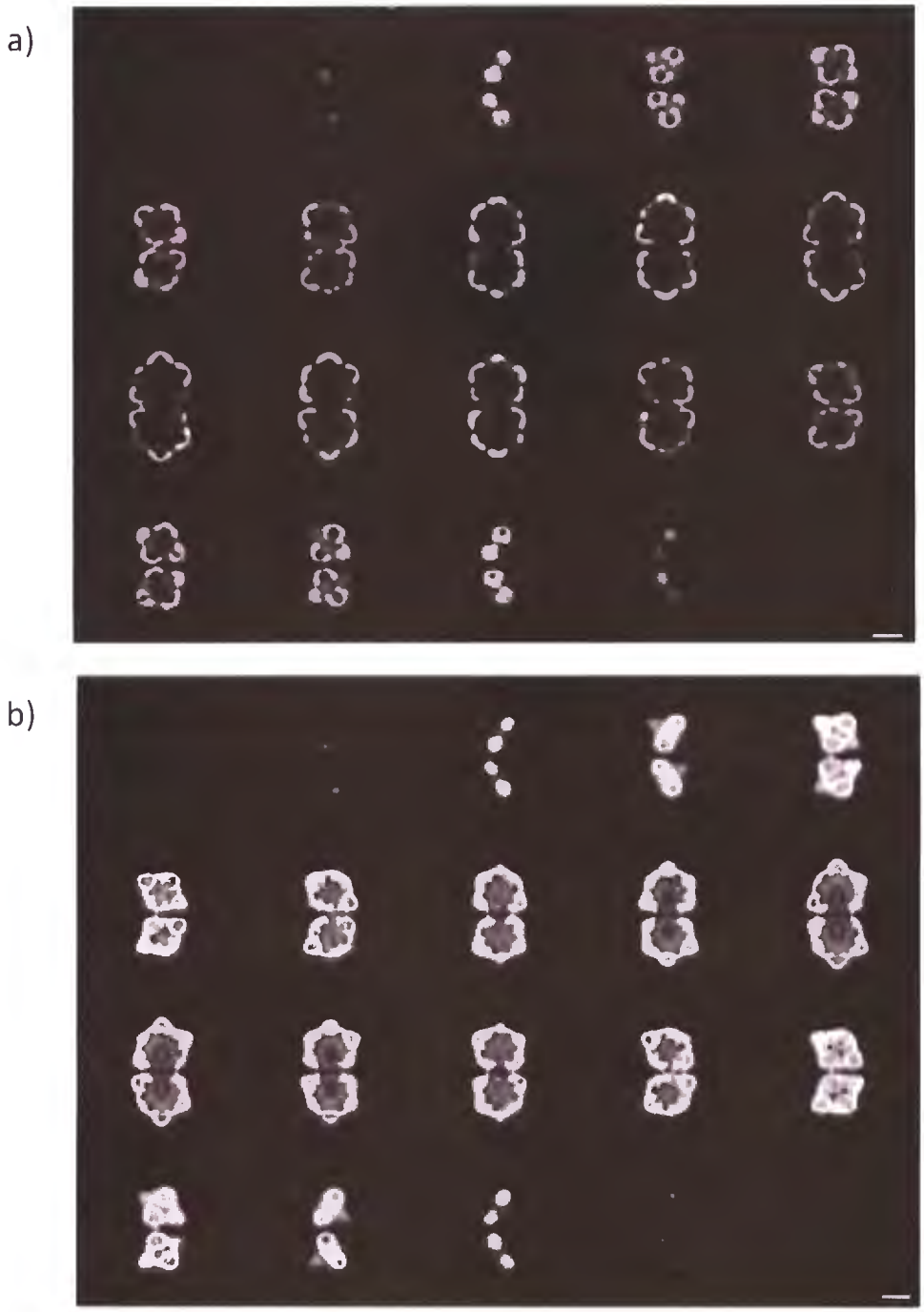

Figure 3-17 Slice views of the geminate reconstructions. Peripheral slices (shown in first and last rows) reveal the general morphology of the pentameric capsomers, as well as allowing their interconnectivity to be assessed. The comparison also reveals that internal densities vary greatly between the two capsid structures derived from a) negative-stain and b) cryo-EM datasets. Furthermore the centers of the capsomers in the negative-stain reconstruction appear "hollow". Scale bars represent 10 nm.

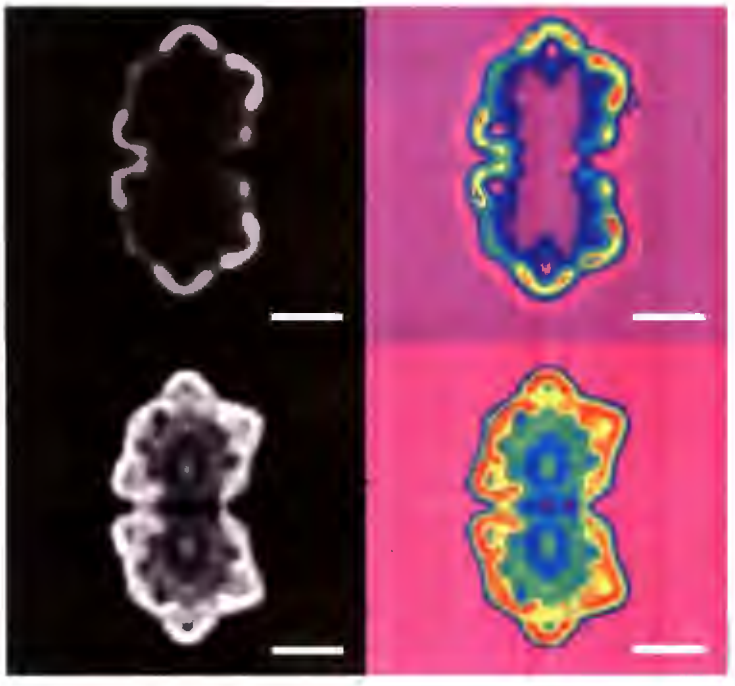

Figure 3-18 Central slices of 3D reconstruction of geminate capsids. The negative-stain reconstruction (top), and cryo-EM reconstruction (bottom), are shown as grey-scale (left) and false colour (right) respectively. The maps are clearly similar, however, the cryo-EM reconstruction reveals internal details in addition to the coat protein lattice that forms the primary layer of protein (red, in false colour) of the capsid. The false colour images reveal that the presence of the genome causes stain to be excluded (top right), resulting in little or no contrast for the genome in the negative-stain reconstruction. Scale bars represent 10 nm.

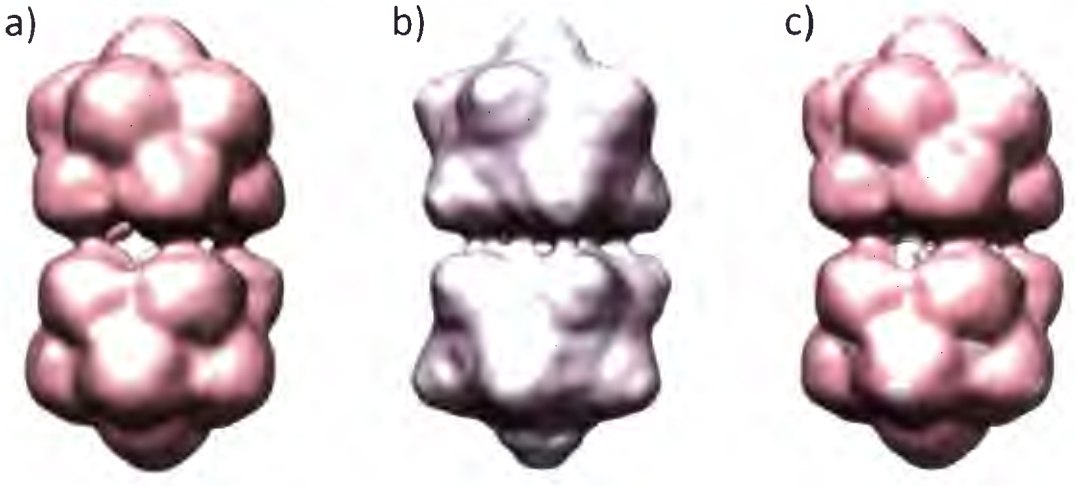

Figure 3-19 Comparison and superimposition of geminate 3D reconstructions. The a) negative-stain, and b) cryo-EM reconstructions are c) superimposed revealing that their dimensions are roughly equivalent, despite differences in surface appearance.

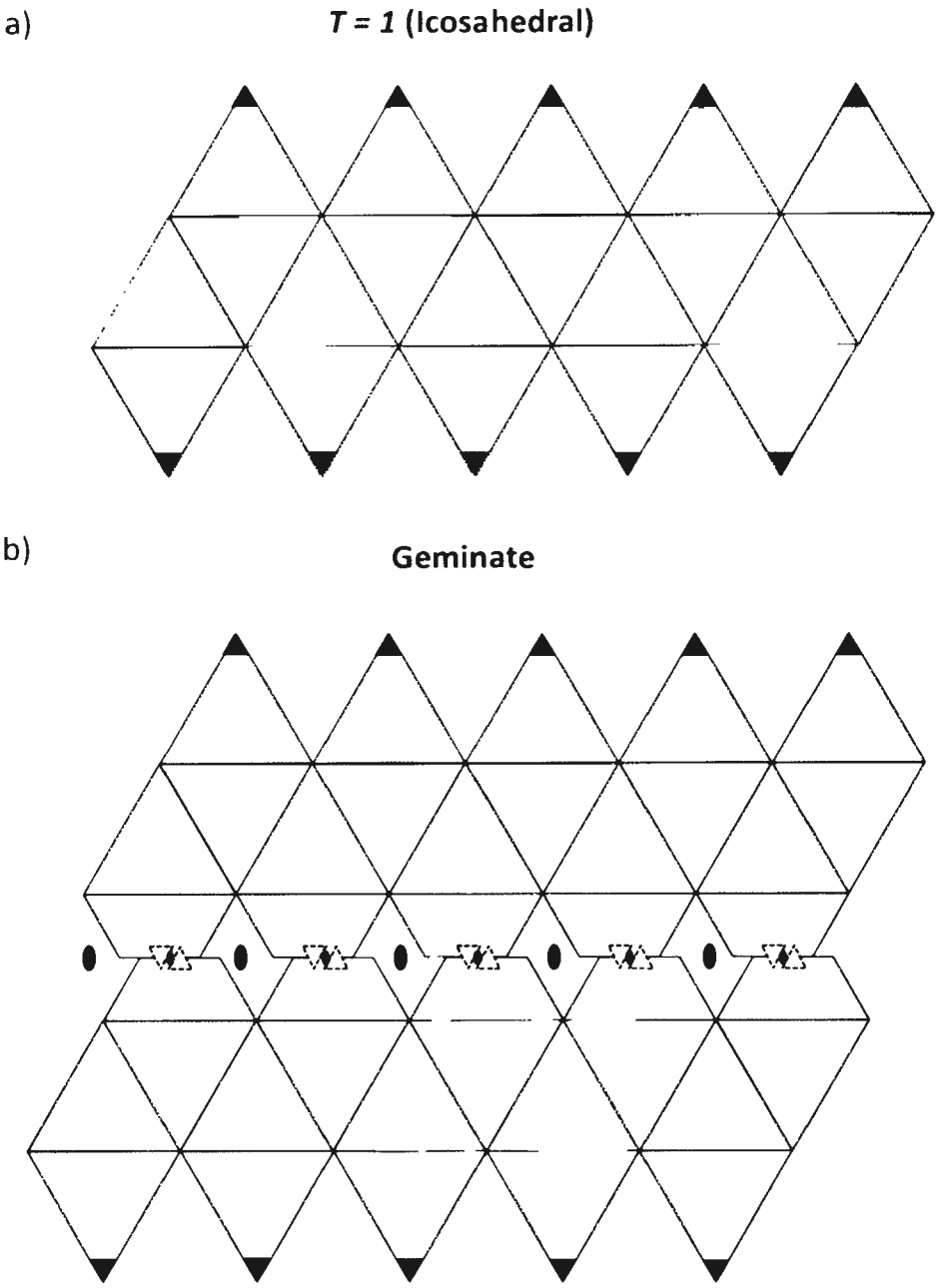

Figure 3-20 Representation of geminivirus capsid architectures by hexagonal lattice. The views (a and b) are equivalent to "unfolding/opening" the capsids onto a flat surface in which the capsid is viewed from the inside. a) The icosahedral particle determined by negative-staining is shown as a $T = 1$ lattice. b) The bipartite capsid is shown here as two incomplete $T = 1$ lattices in which 10 subunits have been removed from each. The lattices are displaced by an amount equivalent to the ~10° offset observed in the experimental structures. The schematic illustrates how two closed lattices can associate in a way compatible with global 2-fold symmetry while maintaining largely quasi-equivalent conformations between protein subunits of the capsid. Positions of global symmetry axes are marked with black shapes with grey linings. Offset maximizes interactions (conserved interactions) between components of the two halves.

undertaken has provided a set of experimentally determined structures that are able to inform modelling efforts and these are described in the following section.

3.3 Structural modelling

The 3D reconstructions described are experimentally determined structures of MSV capsid at relatively low resolutions (~15 Å for the icosahedral (532) reconstruction and 23 $Å^{-1}$ and 26 $Å^{-1}$ for the geminate reconstructions). While easily inspected using computer based molecular graphics software - e.g. UCSF Chimera (Pettersen *et al.*, 2004) - a more complete understanding of the capsid components and their interactions must rely on higher-resolution information and analysis. Direct assessment of atomic interactions would require that the virus be visualized at atomic ($\leq$ 1.2 Å) or near-atomic (< 3.5 Å) resolution where constraints of peptide geometry allow the peptide backbones and side-chains to be confidently modelled within the electron density. Given that EM has a long history of providing structural information of difficult to crystallize viruses at low or intermediate resolutions, it is not surprising that methods for addressing this resolution discrepancy have been developed.

For complex viruses that encode very large capsids, it is often the case that individual components can be crystallized, and then used to model the whole capsid by fitting them into an EM density map. Minor components, which may be difficult to crystallize can potentially be solved to high-resolution using NMR, and then also fitted into appropriate regions of the capsid. Such experimental structures have allowed the placement of residues to be interrogated at interfaces between capsomeres, and have also allowed interactions between different components of the virion to be probed.

For this study, no experimental structure is available for a coat protein (CP) of a member of the *Geminiviridae*, and therefore the CP sequence of MSV must be modelled using a structural template that *is* derived from experimental sources. This approach is known as homology modelling, and has been developed extensively in the group of Andrej Sali (University of California, San Francisco). In principal, homology modelling would allow an approximation of the pentameric capsomer (CP_5) to be constructed, and this model could then be used to probe the EM data by quantitative approaches for fitting of atomic models. In cases where sequence similarity is high between template (experimental structure) and query sequence (in this case the MSV CP), prediction of the tertiary structure of the CP sequence will allow us to infer the tertiary structure of the query sequence with very high confidence. However, owing to the high rates of mutation for both ssDNA and ssRNA viruses, it is not surprising that the sequence similarity between the CPs of MSV and STNV has been reported to be very low. Low sequence similarity challenges the applicability of homology modelling to such problems, as the quality of the

sequence-structure alignment becomes a great challenge. In this study, we have applied modern methods for sequence threading that allow alignment ambiguities to be overcome by application of novel scoring functions that assess the likelihood that specific sequences are compatible with specific protein conformations.

The aim of this section is to report on the construction of such an atomic model for the very specific purpose of suggesting which regions of the CP amino-acid sequence contribute to the interfaces which stabilize the geminate architecture, as well as how the quaternary architecture has changed with the evolution of the viruses CP. The knowledge-based modelling approach known as homology modeling was chosen over *ab initio* approaches due to its general reliability, computational tractability, and applicability proven during previous investigations of members of the *Geminiviridae* (Bottcher *et al.*, 2004; Zhang *et al.*, 2001).

To summarise the findings of this study, structural bioinformatics approaches were found to have an important role to play in understanding the arrangement of CP_5 structural units in the MSV capsid, as well as understanding how the MSV coat protein has diverged in comparison to STNV. We found that while the divergence of MSV from STNV is significant, we can still be certain that these two viruses share a common ancestor. Quantitative fitting approaches led to construction of both icosahedral and geminate models that shared the same geometric parameters describing construction of the capsids from the CP_5 model. That geometrical arrangement of CP_5 units agrees between the icosahedral and geminate capsids is an important result, as this shows that no 'global' conformational rearrangement or transition is required to form the geminate morphology.

3.3.1 Primary and secondary structure

The CP of MSV is 244 amino acids long, and has a number of features in common with other plant viruses. The N-terminal arm (residues 1 - 37), contains 12 positively charged residues (lysine and arginine residues) which are likely to be involved in the recruitment and neutralization of the phosphate backbone charges of the ssDNA genome during packaging and assembly. The structure is predicted by the PSIPRED webserver (McGuffin *et al.*, 2000) to have eight β-strands with high confidence values (80-90%), which is consistent with the expected jelly-roll antiparallel β-barrel architecture common to simple spherical viruses. While STNV has four short α-helices, MSV is predicted to have none. Given the significant divergence of the sequence, and the general lack of abundance of structural templates, it is possible that automated methods are not able to accurately predict α-helical regions in MSV especially since those in STNV take place at interacting surfaces (hence of possibly induced secondary structure). For this reason all the α-helices of STNV were maintained during homology modelling of MSV.

Ultimately, experimental visualization by X-ray crystallography or high-resolution single-particle EM will be required to verify the secondary structure of the MSV CP, however, secondary-structure prediction has nevertheless been an essential source of information in the structural bioinformatics study described in the following paragraphs.

3.3.1.1.1 Predicted nuclear localization signals (NLS)

Prediction of nuclear localization signals (NLS) that might account for the karyophilic properties of MSV CP (Liu *et al.*, 1999) were investigated using the cNLS Mapper webserver (Kosugi *et al.*, 2009). A bipartite NLS was predicted to span the first 25 amino acids of MSV CP (as shown in Figure 3-21), whereas for the STNV CP no NLS were predicted as would be expected for an RNA virus, which does not enter the cell nucleus in order to replicate (in all cases except some complex viruses such as members of the *Poxviridae* or *Orthomyxoviridae*). The first half of the bipartite NLS accounts for the difference in lengths between the MSV CP and the STNV CP, and inspection of Figure 3-35, shows that this arrangement is likely to be a conserved feature of the Geminivirus family. Notably, the N-terminal α-helix does not contribute to the NLS, and it may be speculated that this motif is likely to play an exclusively structural role; as is the case for STNV (Jones and Liljas, 1984a).

An important goal from the outset of this study was the localization and modelling of this protein segment within the geminate assembly, as the structure and function of the N-terminus is essential at many different stages of the viral life-cycle.

3.3.2 Fold prediction, tertiary structure and template-sequence alignment

Accurate sequence alignment was critical to our tertiary structure predictions, and hence our ability to assess how quaternary structure arises within the *Geminiviridae*. Fold prediction, according to mGenTHREADER (Jones, 1999a), showed that the coat protein of STNV was the only atomic structure with sufficiently high structural homology to warrant use as a template for homology modelling. According to the alignment suggested by mGenTHREADER, the sequence identity between the STNV and MSV CP sequences is only 14.1%, which implied that the sequence alignment provided by the threading algorithm was likely prone to ambiguity. Subsequent analysis showed that this was certainly the case, requiring that considered manual adjustments be made.

It must be noted that the mGenTHREADER sequence alignment served as an excellent starting point, and highlighted some important properties of the MSV CP that were not detected by alignment using CLUSTAL W (Thompson *et al.*, 1994) or MUSCLE (Edgar, 2004). However, as already mentioned, careful inspection in consideration of secondary-structure predictions resulted

in a number of modifications (by hand) in regions where loop insertions, and generally low sequence similarity in the MSV CP had caused mGenTHREADER to fail. Inspection of the sequence alignment (Figure 3-21) shows that most sequence identity between MSV and STNV CPs is limited primarily to the β-strands that make up the core of the fold. In the majority, the length and relative positions of the β-strands was found to be conserved, given our understanding that the core of jelly-roll topology is generally structurally conserved, it was deemed likely that the STNV CP could still be useful as a template for the core of MSV CP. However, it must be stressed, that our confidence in the conformations assigned to N-terminal, C-terminal residues and insertions between the β-strands must be low, as either no template for the backbone conformation is available for these regions, or the sequence identity is so low that modelling is certainly prone to significant error. Additionally, it is possible that different modes of quaternary association would cause differences in conformation in regions engaged in these interactions making STNV and inappropriate template for modelling these flexible portions of the MSV CP.

With respect to the size and presence of β-loop insertions, the new alignment allowed significant progress in understanding the size and occurrence of these. The alignment revealed that four major (> 4 amino acids), and 4 minor (≤ 2 amino acids) insertions exist in comparison of the two CPs. Three out of four insertions occur in β-loops that are presented onto the surface of the capsid, and are presumably involved in virion-cell or virion-vector interactions. These insertions are marked by red arrow, and are illustrated in Figure 3-21 and Figure 3-22. The remaining major insertion occurs in a region between an α-helix and β-strand F; this region is referred to as [E:F][3] and highlighted in the sequence alignment figures using a red dashed outline for each of identification. This region of the βE:βF loop is involved in formation of both pentameric (CP_5), and 2-fold symmetric quaternary contacts in STNV, and is likely relevant to the same interactions of MSV.

As revealed by mGenTHREADER, the most notable (but nevertheless provisional) feature revealed by the alignment is the conservation of residues involved in the calcium-binding site of STNV. As shown in Figure 3-21, the amino acid residues involved in coordinating a calcium ion at the 5-fold axis of the pentameric unit of STNV are conserved as a 5 amino acid stretch (L179-S183) in which T180 and S183 of STNV are identical in the MSV CP (this site is the β-hairpin of strands βF and βG in STNV). In MSV CP, this motif is flanked by two insertions of 8 amino acids each, both occurring between β-strands βF and βG. Two proline residues occur on either side suggesting that the peptide backbone makes significant deviations in its path in this region, suggesting that not one, but two loop structures exist at this region while simultaneously preserving the conformation of the calcium coordinating β-hairpin. This is a significant finding

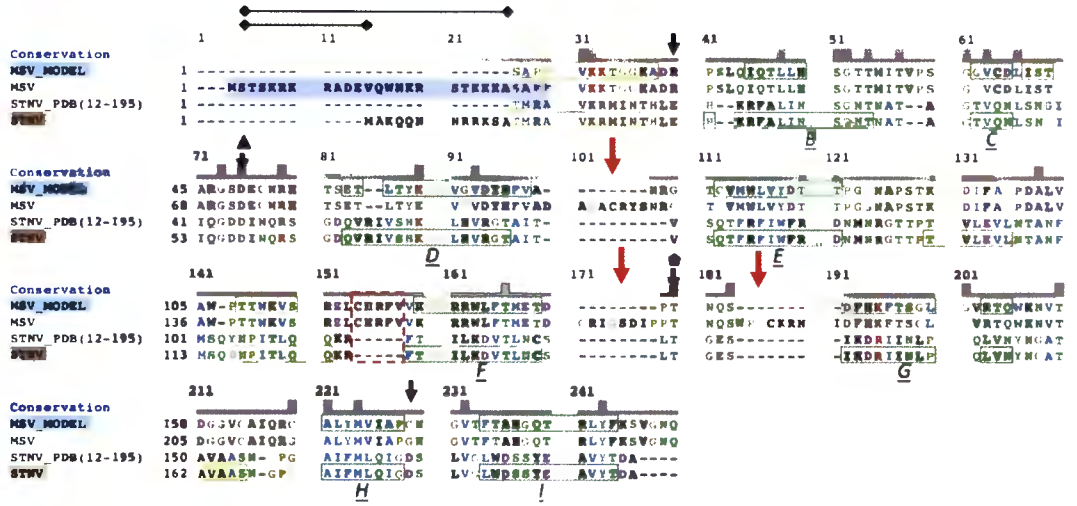

Figure 3-21 Multiple sequence alignment used for comparative modeling. The sequence of the MSV coat protein (MSV) is shown aligned to that of the STNV coat protein (STNV). Sequences of atomic models involved in the process of modeling are shown (MSV_MODEL, and STNV_PDB) are also shown. The N-terminus of STNV was not resolved in the crystal structure, but nevertheless the alignment reveals that various positively-charged residues are conserved in this region between MSV and STNV. Black lines above the N-terminal region indicate the predicted components of the nuclear localization signal encoded in this region. Comparison of the sequences of the MSV CP (full-length and model), will reveal β-loops that were removed from the MSV sequence prior to modeling; these are marked by red arrows.

because all previous alignments of geminivirus CPs have modelled this insertion as one large loop that contributes to the surface reactivity of the virion only; eliminating the calcium-binding site.

The findings of this sequence alignment study provide the first evidence suggesting that, like many plant viruses, MSV may co-ordinate at least one divalent ion (e.g. Ca^{2+}) at each local 5-fold axis of each CPs, and that the region between strands βF and βG, is in fact multifunctional, as it probably contributes to the quaternary stability of the particle, as well as the particles surface properties (responsible for vector interaction in the case of MSV). Having made this observation at the sequence level, the question became whether the experimental results of the EM reconstructions provided information to support this hypothesis. This was the case and is described in Section 3.3.4.1.

3.3.3 Comparative modelling of the pentameric capsomer

3.3.3.1 Modifications to the sequence alignment for homology modelling

A sequence alignment was constructed for sole purpose of probing the cryo-EM density in a manner as accurate as possible. Consequently, the sequence of the MSV CP was modified to remove loops that could not be meaningfully modelled using the template available, and the multiple sequence alignment modified by hand to ensure that no insertions interrupted secondary structure elements (particularly for β-strands βE and βD). The most significant regions excluded where the N-terminal 23 amino acids of the MSV CP, as well as "crowning" loops between strands βF:βG and βD:βE. These measures improved the identity between query sequence and template from the 14.1 % to 17.2 %; this difference is small, but further supports the assignment of the STNV β-barrel fold to the MSV CP.

3.3.3.2 Description of the homology model

The monomeric comparative model is shown annotated in Figure 3-22. The β-barrel fold is "wedge" shaped, and this is well represented by the homology model. All secondary-structure features are present as expected, and the model was observed to superimpose well with the Cα backbone of the template structure.

The homology model was modelled not as a monomer, but as a pentamer and is shown in Figure 3-23. For the simple plant viruses the sides of the wedge form the 5-fold interfaces that stabilize pentamer formation. Three loops at the base of wedge and four loops at the top of wedge allow functional divergence, hence facilitating interactions of the capsomer. The loops of greatest relevance to this study are those involved in quaternary contact formation and are shown labelled in Figure 3-22, b; the βC:βD, βG:βH, and βE:βF loops. During modelling, these regions are assigned the backbone conformation of the STNV derived template structure, and whether these

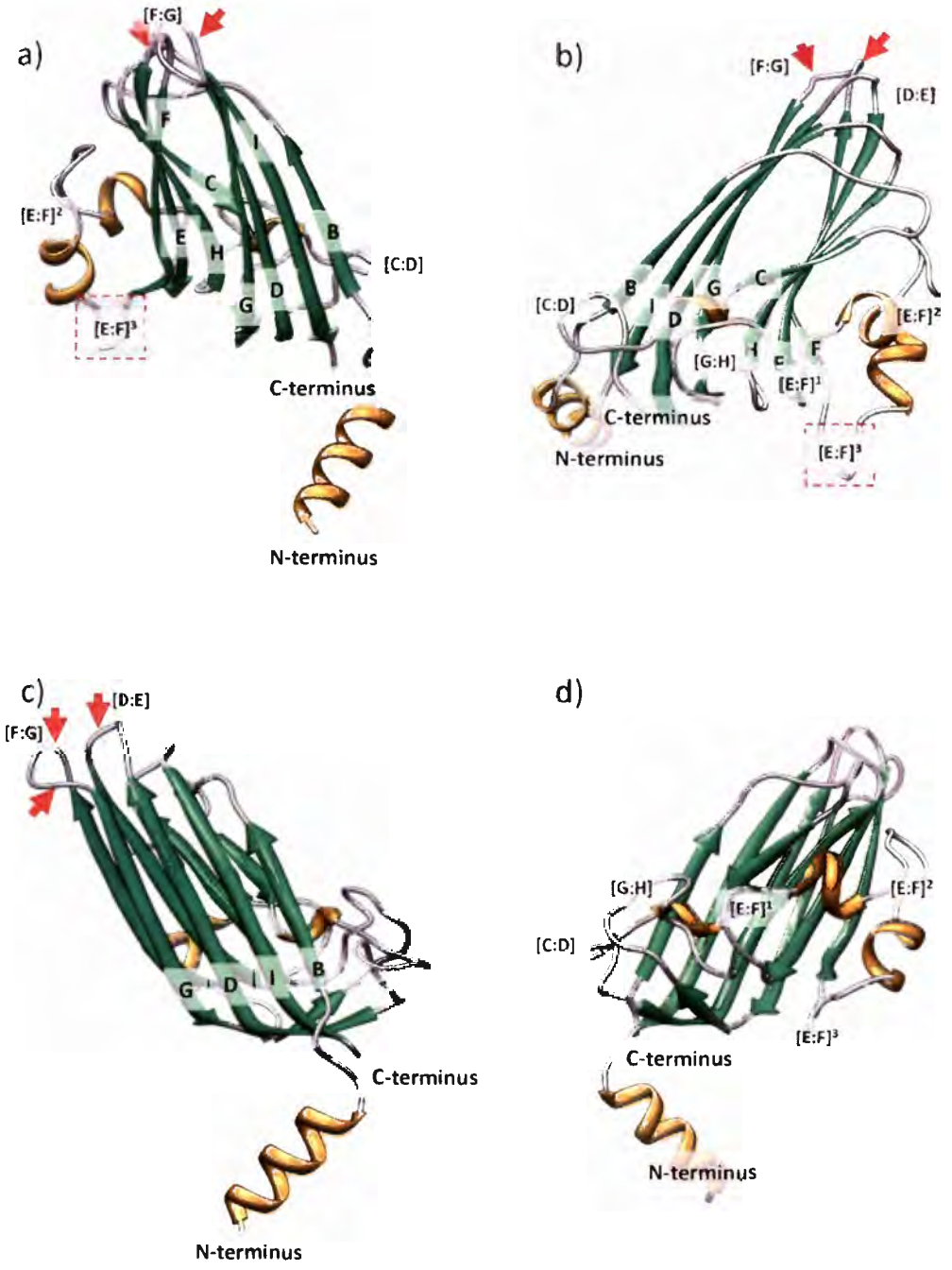

Figure 3-22 Comparative model of the MSV coat protein monomer. In a-d) the CP monomer is viewed from different angles in which β-strands (βB:βH) and their respective loop regions (annotated with as [βX:βX]) are annotated. Location of the single loop insertion that was modeled during this study is shown using a grey rectangular box with a red dashed outline. The location of unmodeled loops is shown with red arrows.

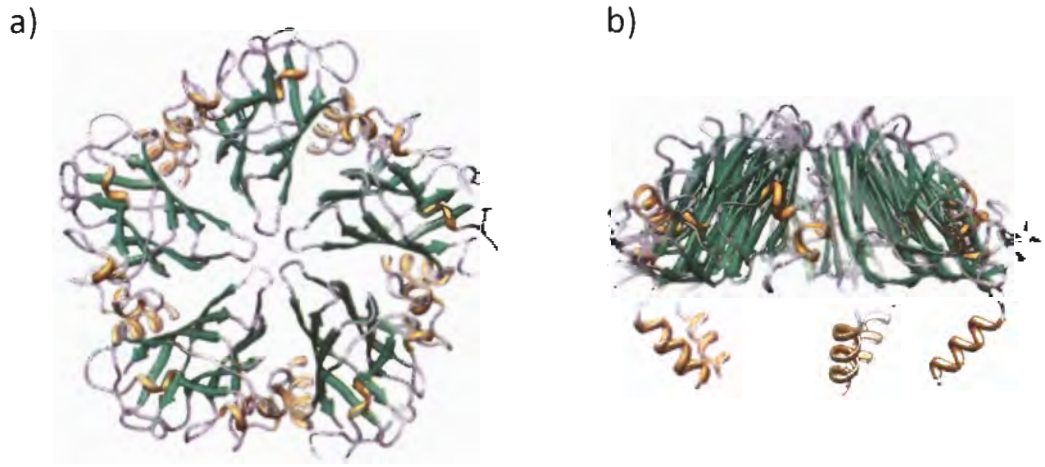

Figure 3-23 Comparative model of the pentameric MSV assembly unit. Modelling of the MSV CP was carried out using the pentamer of STNV. Putative metal ions were not included in modelling, and the validity of this approach was evident during the pseudo-atomic modelling studies, as described in overview in the next Figure. Excellent agreement between the pentameric model, and the pentameric capsomere of the experimental densities suggested the fold and pentamer forming interface are conserved. A strong indication of an ancestral relationship between MSV and STNV.

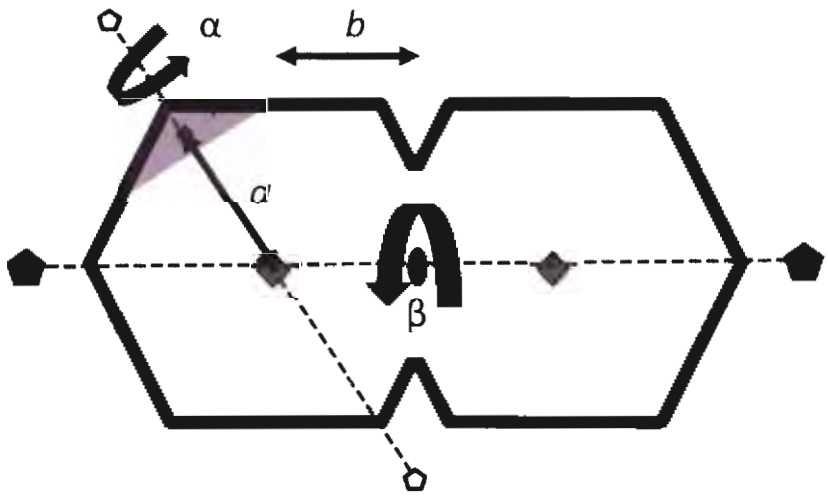

Figure 3-24 A schematic model of parameters used to construct the geminate models. The displacement of the pentamer (grey triangle) center of gravity from the icosahedral center (a), as well as the angular offset around each axis of 5-fold symmetry from the alignment according to the models principal axes (α). The displacement (b) and angular offset (β) between heads, could then be determined using the model of a single head.

conformations are valid in the context of the MSV capsid must be considered in relation to the experimental information available in the EM maps. The pentameric model also preserves the general features of the calcium-binding site of STNV.

3.3.4 Pseudo-atomic modelling of the MSV capsid by computational docking

Construction of a homology model for the pentameric capsomer (CP_5) of MSV does not inform us about the organization of these within the entire capsid itself. To answer this question, a series of capsid models were created and compared with the EM maps in order to identify parameters that produced the "best fitting" model. This was carried out in two stages, the first to identify the orientation of the CP_5 in the context of the quasi-icosahedral head, and the second to identify the relation of the two geminate heads to one another (refer to Figure 3-24). This involved selecting the model of highest correlation from a set of models created in which the parameters ('**α**', '**a**' and '**β**', '**b**') were systematically varied (refer to Figure 3-24).

3.3.4.1 The geometric construction of the quasi-icosahedral geminate head

Within the context of the quasi-icosahedral symmetry of each of the geminate heads, or the icosahedral symmetry of monopartite capsid species, each CP_5 is located on one of six possible 5-fold axes of rotational symmetry. All CP_5 capsomers (11 in a geminate head, and 12 in a monopartite capsid) will share a rotation about these axes, as well as a translation along the axis from the capsid centre; both parameters (referred to as '**α**', and '**a**' herein) influence the diameter of the assembled head as well as the placement of potentially interacting residues in the atomic models. That these parameters are the same for each CP_5 was assumed during this study (i.e. these parameters were not varied locally owing to time constraints involved in such a search), and has the biological implication that all CP_5:CP_5 interactions making up the geminate head are identical (i.e. that they correspond to a $T = 1$ icosahedral spherical lattice). Later analysis of the computational docking results suggested that this assumption was likely to be the case.

In this study, the geometric parameters are determined in relation to a CP_5 model that has been aligned along its principal axes and centred according to its centre of gravity (COG). Hence radii determined in fitting refer to the radius to the CP_5 COG and not the true radius of the geminate head. Correlation scores were calculated by quantitative fitting of a series of capsid models in which these parameters are varied, and these results are shown in Figure 3-25. The C5 redundancy (5-fold rotational symmetry) of the CP_5 model, limited the search space substantially, which expedited this process considerably.

The precise nature of the fitting result (i.e. best position, orientation and score) may vary depending on the type of density map being used for the 6D cross-correlation search (three

translation and three rotational degrees of freedom). In the case of the cryo-EM map, the variation in voxel densities should directly reflect the variation in protein densities of the specimen. This is not the case for reconstructions from negative-stain data, because contrast arises from scattering of heavy atoms surrounding the protein, and not the atoms of the protein itself. For these reasons the correlation search in a negative-stain map is only sensitive to the "boundary" of the molecule at the density threshold chosen at the outset of the cross-correlation search. This had broad implications for this study because no attempt was made to model certain regions of the CP, or any associated nucleic acid. Furthermore, the internal densities of the maps varied (a layer of nucleic acid was present in the monopartite but not bipartite negative-stain reconstructions) meaning that the maps would correlate to different extents with the N-terminal α-helices of the CPs that would be located in these regions. In addition to this, the homology model did not include residues contributing the "crowning" loop insertions of the MSV CP, implying that radii of CP_5 could be overestimated.

For these reasons, the atomic fitting into negative-stain maps did not produce scores that were consistent regarding the diameter of the quasi-icosahedral head (parameter '**a**'). Based on visual inspection of the fitted capsid models, qualitative features (described in the following paragraph) could be compared to show that the particle radii estimated from each map did in fact agree with the cryo-EM density map (parameter a = 84 Å), and the quality of this fit is shown in Figure 3-26. Notably, for all three maps (monopartite, and two bipartite) the angular orientation of the pentamer was also determined to be the in agreement ($\alpha = \sim 43.5°$). The monopartite map produced the sharpest correlation peak, while the geminate maps showed a much broader distribution of correlation values for this parameter. This was attributed to the significantly greater resolution of the 532 reconstruction. This is an important observation because, firstly, it shows that by independent determination, the maps are in agreement, and secondly that the conformation of residues at interfaces between pentamers are likely to not be altered at the time the bipartite capsid forms. This is significant because from the outset of this study, it was speculated that it might be possible that $T = 1$ assembly is held together by flexible interfaces that are able swivel into a different conformations (the "pentamer-swivel" model), thus facilitating assembly of the geminate capsid form.

Notably, inspection of the quality of the fit reveals excellent agreement in features between the CP_5 model and the capsomere densities of the EM maps (Figure 3-26). The contours of the cryo-EM density appear to follow the borders of the CP_5 model most closely at the surface in the cryo-EM map, while the density of the capsomeres appears somewhat inflated in the negative-stain maps. This may be caused by a degree of positive staining of the capsid surfaces, alternatively

variable distortion of the pentamers depending on whether or not they are pressed up against the carbon support, or not, would (under dehydration) produce in average a somewhat distorted result. The NS bipartite reconstruction does not reveal internal densities, and consequently also shows the best agreement with the CP_5 model at the interior surface. The other two maps include an additional layer of density that is most likely DNA, and at this resolution the boundary between protein and nucleic acid is not resolved.

All three maps reveal important density variations along the 5-fold axes of rotations symmetry for each pentameric capsomere at each location in the quasi-icosahedron (only one of which is shown in Figure 3-26). This variation is most pronounced and easily analysed in the monopartite icosahedral map (532, which is the highest resolution and hence reveals the most detail). The fit of the CP_5 model is informative in the analysis of which structural features likely contribute to the regions along the axes. Proceeding from features from the outside moving inwards, all three maps reveal unfilled density for the crowning loops that were not modelled in the homology model. In the case of the 532 map, this density has clearly defined properties showing that the crowning loop insertions are likely to adopt a conformation which produces a cavity that is only resolved in the 532 map. In consideration of the fitted CP_5 model, it is clear that the base of this cavity is formed by residues that appear to be in a similar conformation to that the β-hairpin structures responsible for coordinating the bound Ca^{2+} ion in STNV capsid at the 5-fold axes of symmetry. That this conformation is conserved is quite remarkable given the relatively large loop insertions that exist on either side of it; to reiterate, no attempt was made to model these loop insertions during this study owing to their size and lack of a suitable template. To our knowledge, this result represents the first empirical information to support the hypothesis that MSV binds calcium ions to stabilize its capsid. Immediately below these β-hairpins (βF:βG) is a low-density region that appears to be unoccupied density. This region lies above the layer of density thought to be DNA, and is likely unoccupied because it is simply sterically inaccessible to nucleic acid molecule, or the capsid has low affinity for nucleic acid in this region of the CP_5. This is further explored in Section 3.3.6.

The capsid models of highest correlation (or best fit) are shown in Figure 3-27, Figure 3-29, and Figure 3-30. What is quite striking regarding the fit of the quasi-icosahedral models is the amount of space that exists between the CP_5 units. While this was initially thought to be a result of the flattening in the case of the negative-stain reconstruction, the equivalent fit into the cryo-EM map

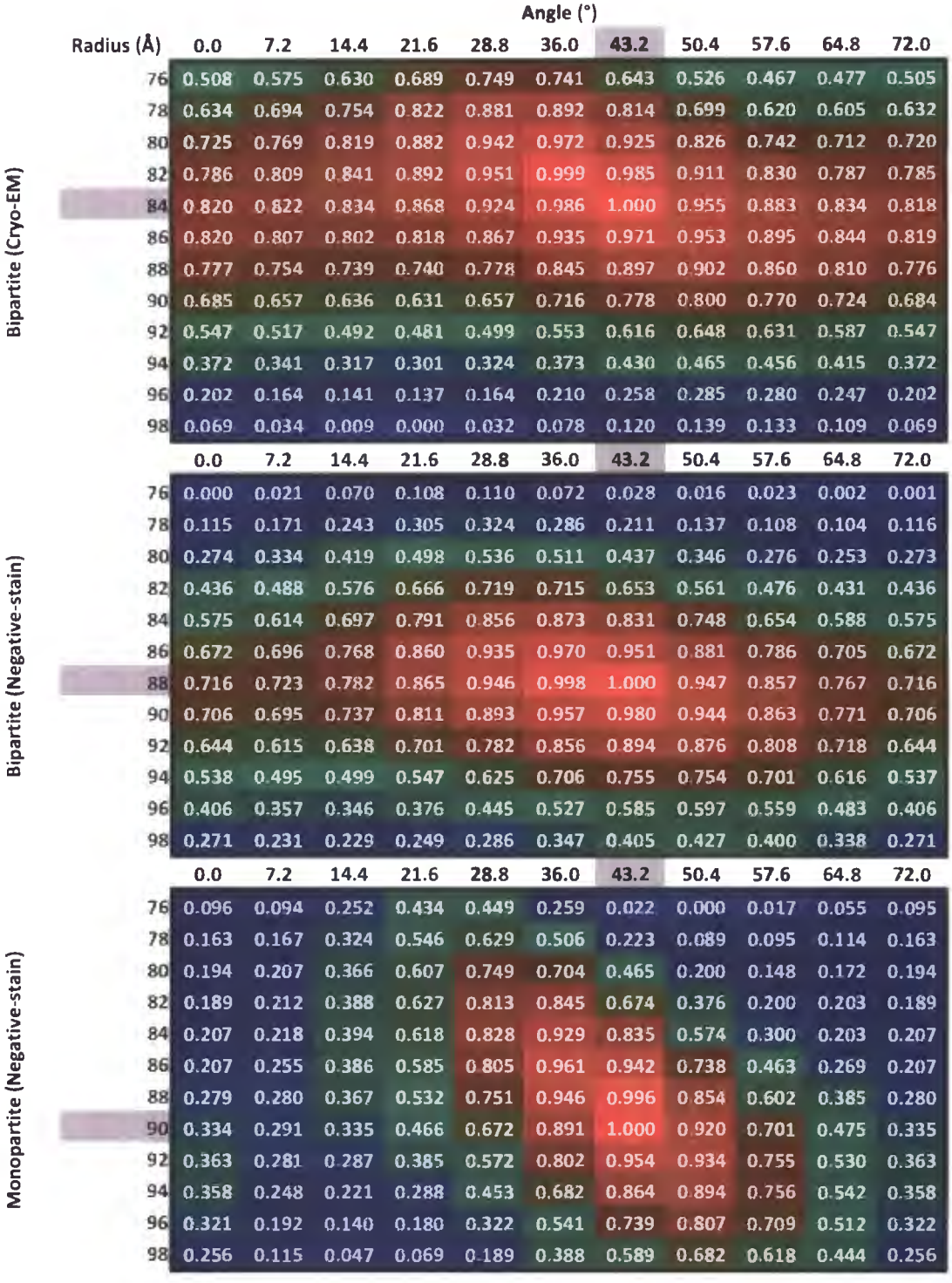

Figure 3-25 Fitting scores determined during first stage of capsid modelling. Normalized cross-correlation scores are tabulated with rows representing the parameter 'a'- the radial displacement of the CP_5 centre-of-gravity, and columns representing the parameter 'α' the axial rotation applied to the CP_5 before construction of the icosahedral capsid model.

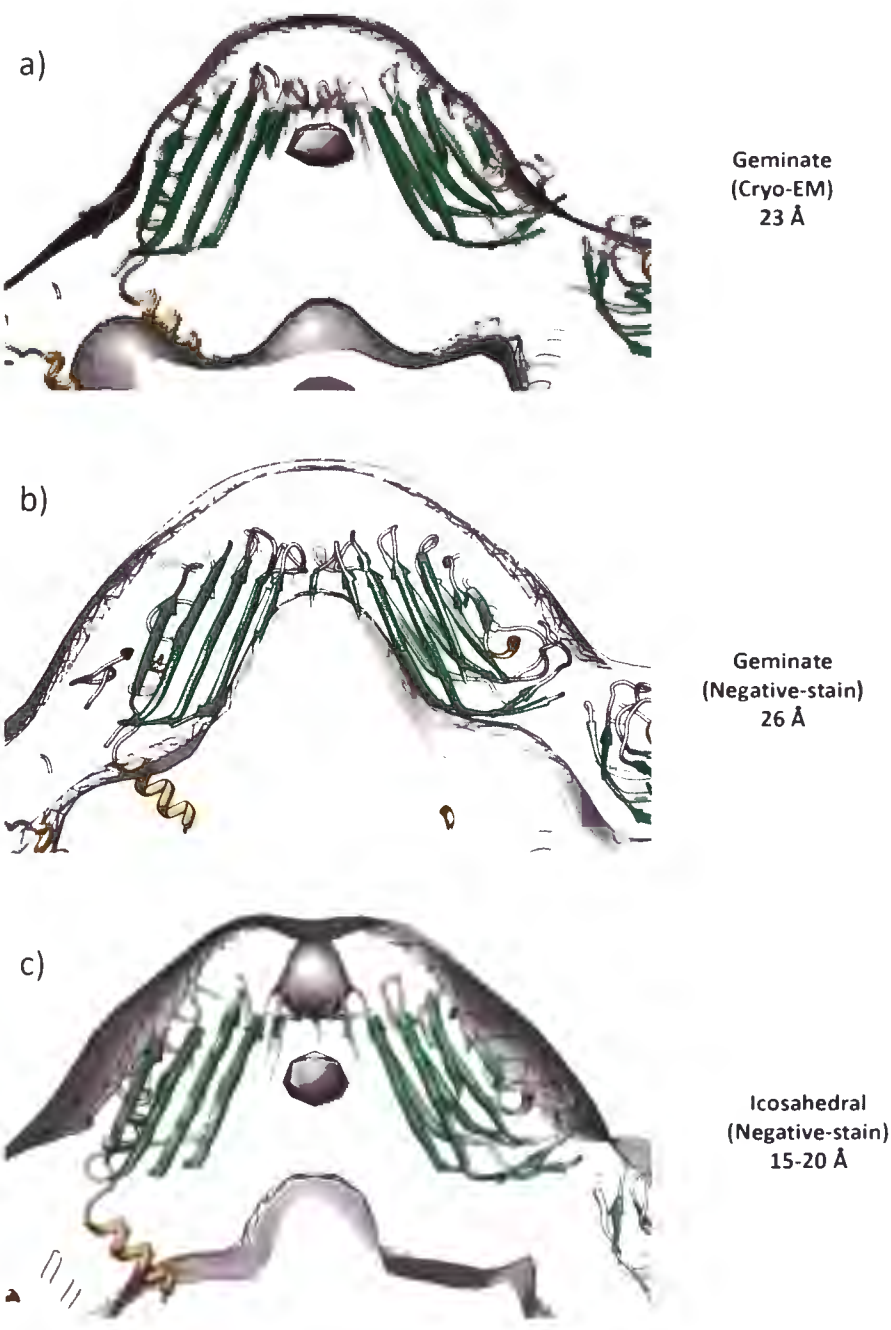

Figure 3-26 Fit of pentameric comparative model into the apical capsomeres of the experimental density maps. 3D volumes derived from electron microscopy are rendered as grey isosurfaces and sliced such that a 20 Å thick central slab is visible. The comparative model is rendered as ribbon in which β-strands are coloured green, and α-helices coloured yellow. Each model is fit into the maps at the same radius from the icosahedral center. Even though the inner structures of the density maps vary, it provides different structural features which support that the fit is accurate, and that the radius of the pentameric units are the same between the units irrespective of how they were prepared for EM.

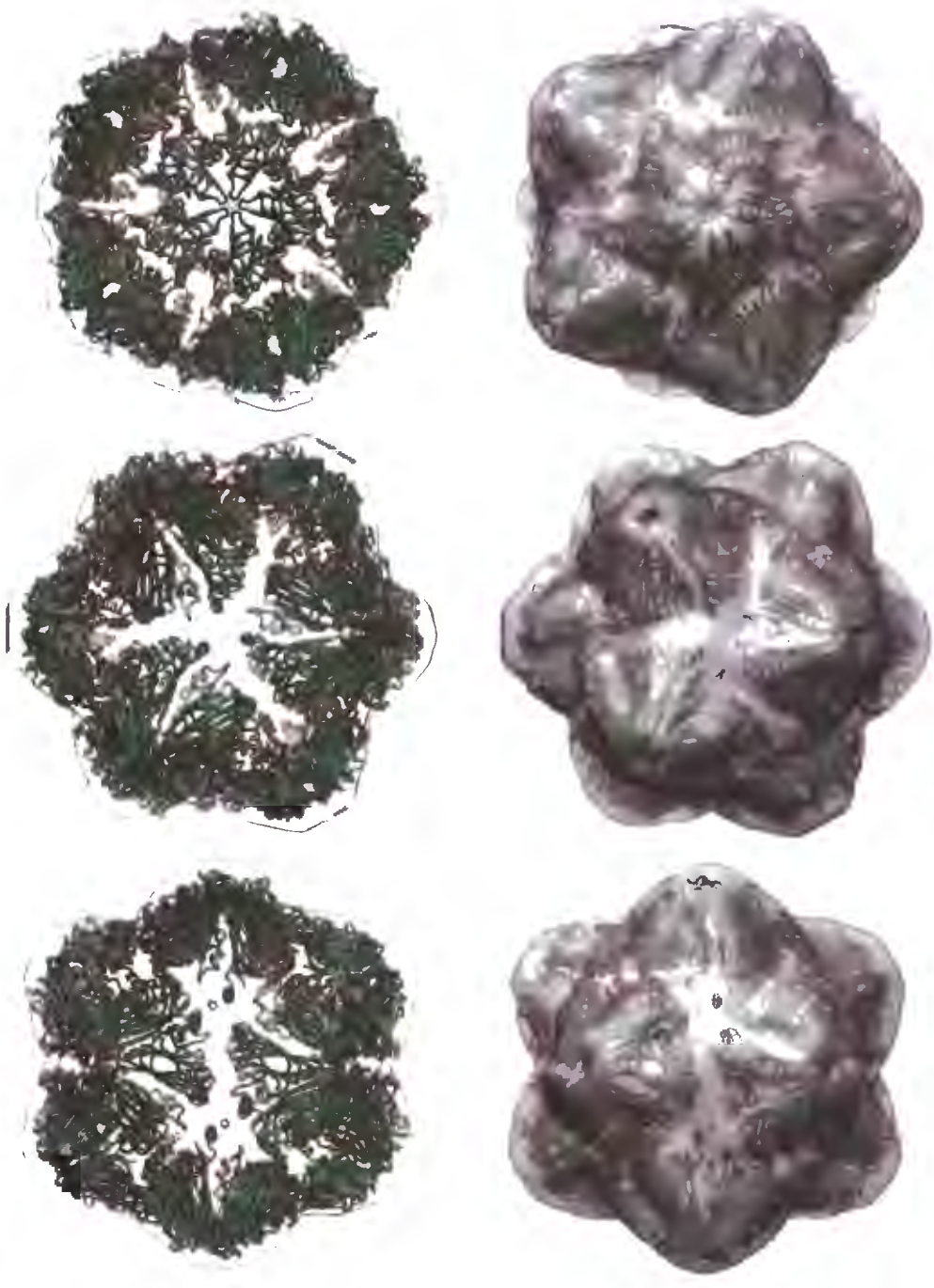

Figure 3-27 Fitting of the icosahedral (532) comparative model into the 3D reconstruction of monopartite MSV capsids from negative-stain data. 5-fold, 3-fold, and 2-fold views of the icosahedral particle are shown in which, left) the EM density is rendered transparent and only an outline is shown, and, right) the EM density is shown as semi-transparent and the location of pentameric units with respect to the map is shown.

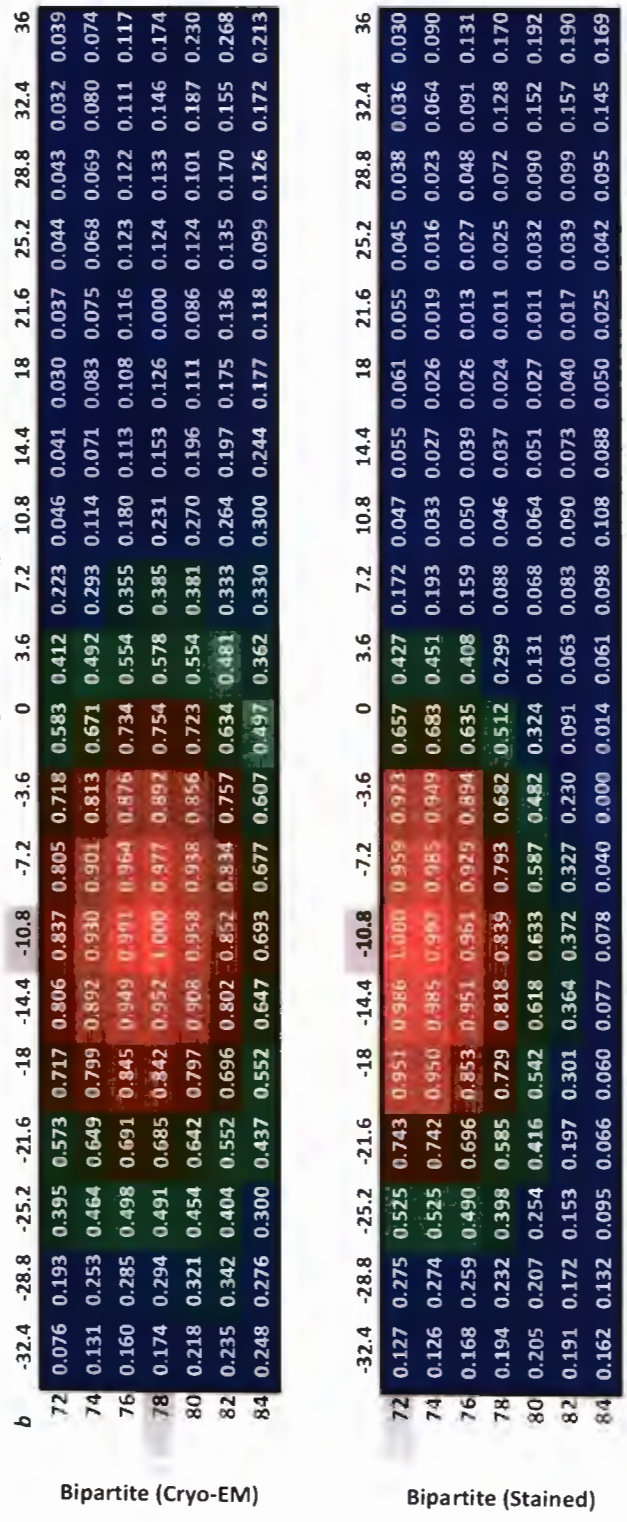

Figure 3-28 Fitting scores determined during second stage of capsid modelling. Normalized cross-correlation coefficients are coloured according to the quality of the fit (red – high correlation, blue – low correlation). The angular offset (parameter 'β') is seen to be identical between the negative-stain and cryo-EM reconstructions, while the displacement between heads (parameter 'b') is seen to vary. Refer to text for discussion.

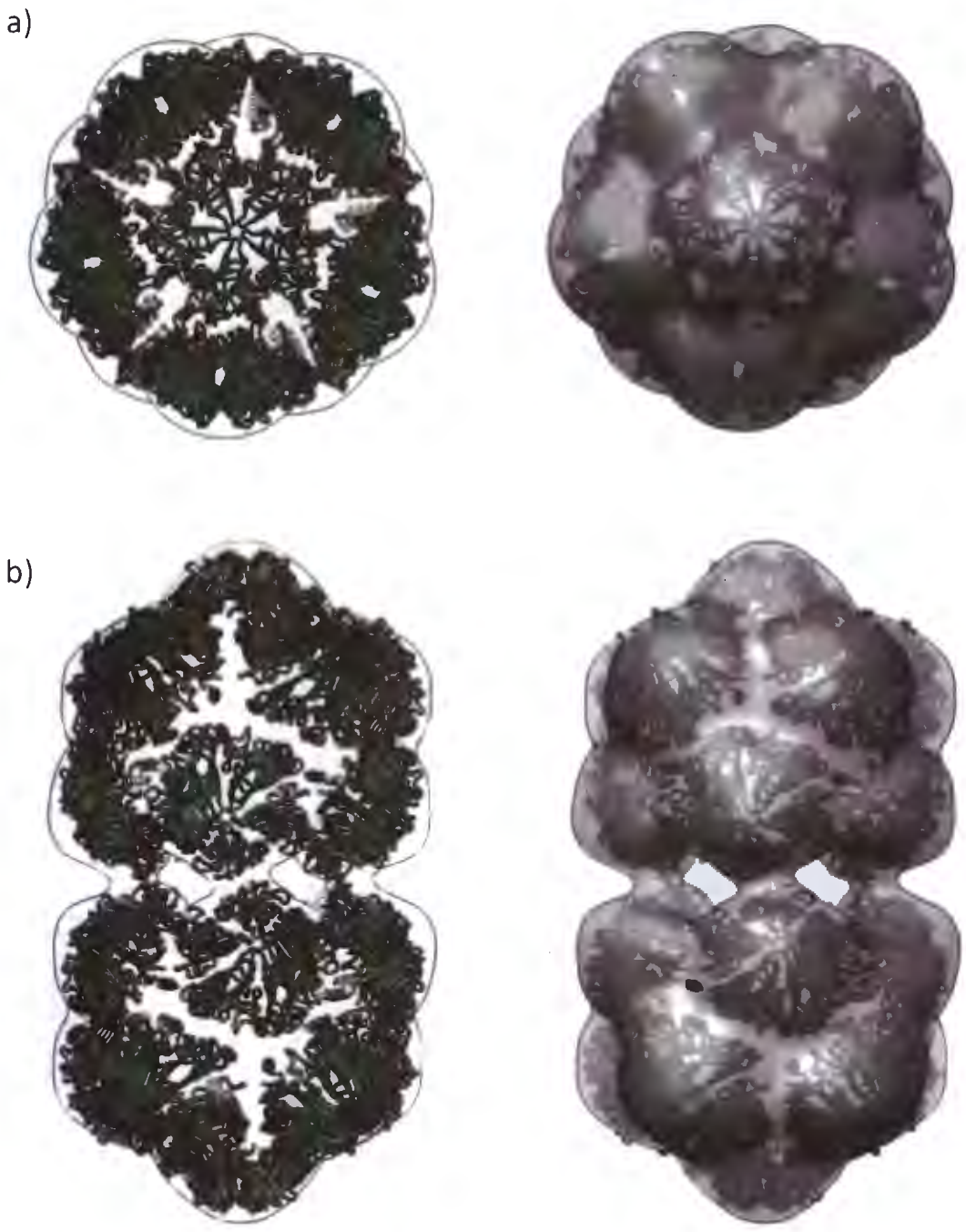

Figure 3-29 Fit of the geminate (D5) comparative model into negative-stain 3D reconstruction. The reconstruction is viewed in a) from a 5-fold axis of global symmetry, and b) from a 2-fold axis of global symmetry in which, left) the EM density is rendered transparent and is shown only an outline, and, right) the EM density is rendered semi-transparent.

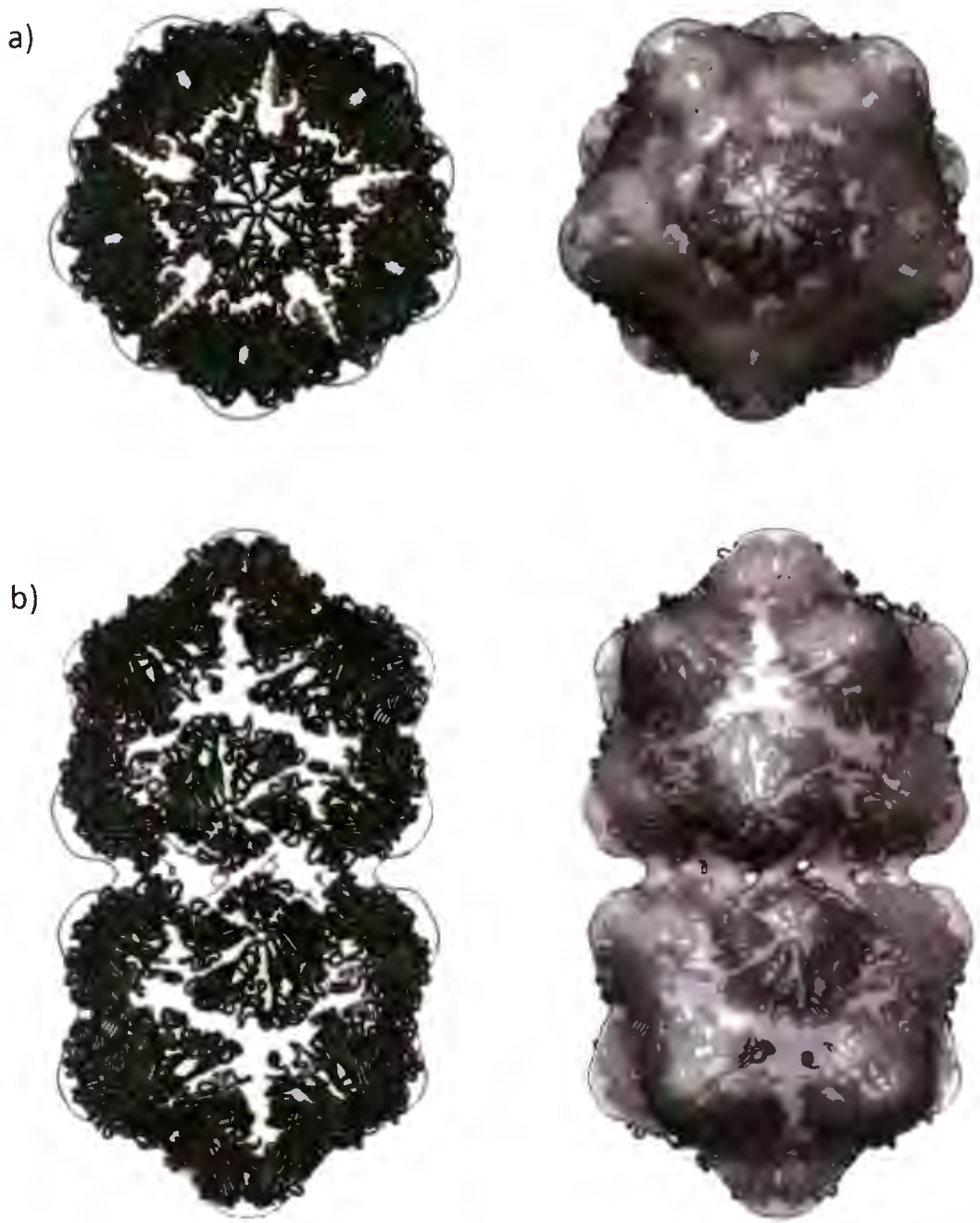

Figure 3-30 Fit of the geminate (D5) comparative model into 3D reconstruction derived from cryo-EM. The reconstruction is viewed in a) from a 5-fold axis of global symmetry, and b) from a 2-fold axis of global symmetry in which, left) the EM density is rendered transparent and is shown only as an outline, and, right) the EM density is rendered semi-transparent. Overall the fit of the pentameric units is satisfactory, aside from loops involved in the formation of local 2-fold symmetric contacts. The loops exist in a retracted conformation which is appropriate to the STNV capsid, but presumably exist in an extended conformation in order to mediate contacts between capsomers in the geminivirus capsid.

suggested that this was a significant observation. The implications of this are discussion in Section 3.3.4.3 regarding interactions that stabilize the assembly.

3.3.4.2 Geometric relation between bipartite heads of the MSV virion

In the second stage of capsid modelling, incomplete icosahedra (11 CP_5 units) were generated using geometric parameters described in the previous section and combined to create geminate capsid models. These were then fitted into the experimental densities of the geminate reconstructions, just as described in the previous section.

Parameters under investigation were the angular offset between the heads (parameter '**β**') as well as the distances between the head centres (parameter **b**); these parameters are represented schematically in Figure 3-24. Both negative-stain and cryo-EM reconstructions were then probed with a series of models in which these parameters were varied, and normalized scores are shown in Figure 3-28. Both maps correlated highest with a geminate model constructed with an offset of ~11° (β= -11°). The translational parameter (b) describing the separation of the heads was found to vary between the two maps, just as parameter '**a**' had varied in the previous section. Inspection of the quality of the fit revealed that once again, the model itself as well as the properties of the negative-stain reconstruction influenced the accuracy of the fitting. Based on inspection of the maps, if modelling was carried out to not include the N-terminal arms of the CP units, we would expect a more accurate result in the case of the NS-reconstruction that did not include any internal details. However, because the study did include this feature, parameter '**b**' was overestimated in these cases. Qualitative inspection showed that the parameter assigned to the cryo-EM was the most correct, and subsequent fitting of this result into both maps produced the result shown in Figure 3-29, Figure 3-30, and Figure 3-38.

As described in the previous section, the geometric agreement between the maps is important because it increases our confidence in the overall result. In addition, that the geminate architecture is seemingly unaltered by negative staining also indicates that negative-stain EM can be a valuable tool for the study of other geminiviruses that are difficult to isolate at concentrations required for cryo-EM.

3.3.4.3 Analysis of interacting surfaces

The structures of the monopartite and bipartite MSV capsids described in the previous sections, are "quasi-atomic models" that incorporate information from the EM maps. As a result these allow the information present in the EM maps to be probed at higher resolution, and thus can potentially shed light on which regions of the CP are likely to interact to stabilize the protein-protein contacts between capsomers (CP_5 units). While the model will not provide precise information about loop

conformations or side-chain interactions between components, we can be confident that the orientation and placement of core β-barrel domains are well approximated and hence the relative arrangement of interacting regions is likely reflected in the modelled structure.

As described in Chapter 1, there are different local (structural) environments within the geminate capsid; the precise details of which have fundamentally motivated this study. The most interesting of these environments are the contacts between subunits at the equator of the assembly. However, to understand the significance of these contacts, we must also consider their counterparts that occur many times within the spherical construct ($T = 1$ lattice) of the geminate quasi-icosahedral head. Inspection of the EM maps showed that the equatorial contacts of two types (primary and secondary) are likely to be formed by residues that contribute to the 2- and 3-fold interactions of the CPs in each head. As a result, in this section the contacts formed by the MSV CP are analysed within the head by reference to the icosahedral reconstruction first, and subsequently the analysis is expanded to describe contacts made within the geminate assembly.

3.3.4.3.1 CP:CP interfaces within the icosahedral monopartite capsid

Interactions between CPs, whether around 2-fold, 3-fold, or 5-fold axes, are shown for both MSV and STNV in Figure 3-31. It should be noted that these same interactions are involved in formation of the quasi-icosahedral heads of the geminate capsids (as well as high order species that encapsidate larger ssDNA molecules).

The outcome of the relative ~10° rotation of the CP_5 (parameter '**α**') in consideration of the 2-fold (dimer-like) interaction, is to align the two CP subunits, bringing a different set of loops (in comparison to STNV) into register such that dyad symmetry is preserved at this interface; a notable observation. At the radial placement of subunits in STNV, such an orientation would not be possible owing to steric collisions. To assess whether the 532 symmetric EM map could provide additional information to support the role of the β-loops in mediating this interface, the fit of the pseudo-atomic model was assessed at a higher isosurface threshold for the EM map. The effect of this is to render the map using a more restrictive contour, and can be used to assess where the densest regions of the protein reside. This analysis is shown in Figure 3-32, and shows that at higher threshold the map is indeed informative, revealing that the interactions at these sites are indeed "discrete", i.e. mediated by precise locations that agree well with the modelled placement of loops which are themselves discrete features of the CP. The loops mediating the 2-fold interactions within the context of the icosahedral architecture of the geminate head are annotated in detail in Figure 3-34, a. The regions involved in formation of the 2-fold interfaces are those between β-strands βC:βD, βG:βH, and βE:βF^3, and useful comparison may be made between Figure 3-22 and Figure 3-34 a, regarding these sites. The effect of the rotation of the CP_5 by 10°

is that the interaction at the 2-folds is formed by three discrete regions, as opposed to only two in STNV. By doing so the surface area of the 2-fold site is effectively increased meaning that there is more opportunity for the divergence and adaption of additional modes of interactions at this site. The "new" interface could in fact be described as being more "versatile". Based on this observation, it could be proposed that the evolutionary remodelling of the icosahedral architecture in this way, was driven by the requirements of geminate capsid formation.

Analysis of the sequence alignment regarding these regions reveals that these sites are composed of residues with either charged or hydrophilic nature, and that a number of glycine residues are also present. Hydrophilicity is expected for a surface (solvent) exposed regions. Additionally, it would appear that each of these regions have pairs of oppositely charged residues that might be involved in forming complementary intermolecular interactions; as would be expected for a dyad interface. Notably the $\beta E{:}\beta F^3$ region implicated in forming an interaction, is actually an insertion in the MSV CP in comparison to the STNV CP (Section 3.3.3.2), and is clearly an adaptation by MSV to mediate formation of this particular architectural arrangement.

For interactions occurring around the 3-fold axes, the effect of the rotation ('**α**'), and increased radius ('**a**') is most pronounced, as a large separation opens up between models at this location. For the 2-fold interaction it can be reasonably proposed that the loops in this region, existing in a folded-back conformation in STNV, extend outwards in MSV meeting each other halfway across the separation between CP_5 capsomers to form the 2-fold interaction. Such a proposition can be made for the 3-fold interaction, but not with as much confidence in precisely which protein segments a rearranged to make the interaction. Inspection of the inset shown in Figure 3-32, b will show that the icosahedral map (at increased threshold) shows that a well-defined interaction is occurring at the 3-fold axes. It appears that for amino-acid residues to occupy this space, the βC:βD loop would need to extend significantly outwards, also disrupting the α-helix in this region in order to extend the backbone sufficiently far outwards to meet around the 3-fold axes. In previous geminivirus structural studies, this space has been modelled by adjusting the orientation of the N-terminal α-helix. This study does not support such a conclusion based on a number of observations. Firstly, the density in our 532 symmetric map does not agree well with placement of three α-helices in this location. The density in this region is seemingly more flat, and would appear to be contributed by a planar loop-like conformation at this resolution. Secondly, careful inspection of side-views of the CP_5 model fitted into the capsomeres of cryo-EM reconstruction (Figure 3-26 and Figure 3-38) will show the indentations at the 3-fold axes on the inside of the capsid, are well accounted for by the placement of N-terminal α-helices in these internal regions. It must be stated, however, there can be no certainty as to which residues occupy the regions

Figure 3-31 A comparison of CP:CP interactions between MSV and STNV. CP subunits engaged in interactions around sites with either local or global symmetry are shown for the 2-fold (top), 3-fold (middle) and 5-fold symmetry (bottom) sites. The location of calcium ions in the STNV structure (right) are shown as green spheres. Atomic models are rendered as ribbon diagrams with green β-strands, and yellow (MSV) and blue (STNV) α-helices. The figure illustrates the effects 10° difference in orientation of the pentameric capsomer (CP_5), as well as the differences in separation between CP_5s.

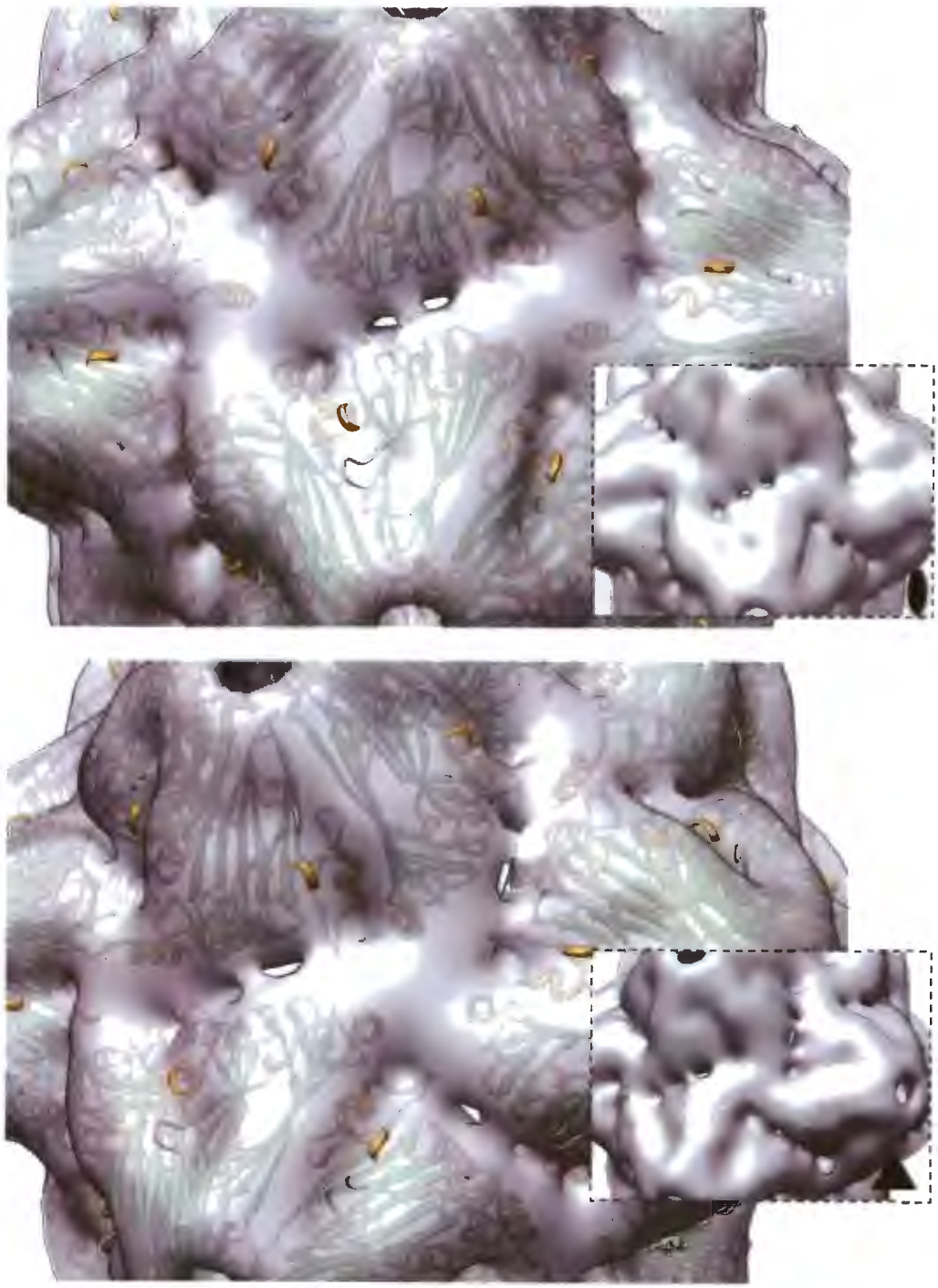

Figure 3-32 Analysis of 2-fold and 3-fold interactions in the monopartite (532) MSV capsid reconstruction. The monopartite negative-stain reconstruction is rendered a higher isosurface density threshold and viewed along a 2-fold (top) and 3-fold (bottom), axis of symmetry. The 3D map, rendered in grey, is transparent revealing the placement of the highest cross-correlating pseudo-atomic model. The insets show the map without transparent rendering.

interacting across the equator are vastly different for those interacting within the icosahedral heads. Inspection of interacting regions in the STNV structure reveals that the 2-fold contacts are fewer (as already mentioned), but also involve regions required for formation of the intra-CP_5 contacts, and are hence expected to be considerably less flexible. In the case of MSV all interacting regions appear to be more extended, and are likely to be more flexible. These features are further evidence of how the quaternary remodelling demonstrated by MSV serves as the structural basis for formation of the geminate architecture.

A detailed analysis of loop interactions mediating the primary interactions across the equator are shown in Figure 3-34, b. Notably, a unique set of interactions are seen to characterise this 2-fold symmetric "2-fold-like" CP:CP interaction. In comparison to the icosahedral 2-fold contacts, the interacting loops are offset, or shifted by a step of one in the direction of the geminate offset. The most notable of these are interactions is between the $\beta E:\beta F^3$ and $\beta G:\beta H$ loops across the interface. $\beta E:\beta F^3$ also plays a crucial role in formation of the standard icosahedral 2-fold interaction (refer previous section). Inspection of the sequence alignment shows that the $\beta E:\beta F^3$ loop includes a 5 amino acid insertion that is highly conserved in geminiviruses (Figure 3-35). In total this region is 7 amino acids long and includes two charged residues of opposite charge. The $\beta G:\beta H$ loop region is rich in glycine residues, and has just two charged residues, an aspartate and a lysine, also of opposite charge – suggesting compatibility with the requirements of a 2-fold symmetric (dyad) interaction. The pseudo-atomic model produced by atomic docking provides a strong case for assigning a role to this sequence insertion with respect to STNV, and suggests an essential role in formation of the capsid architecture of geminiviruses.

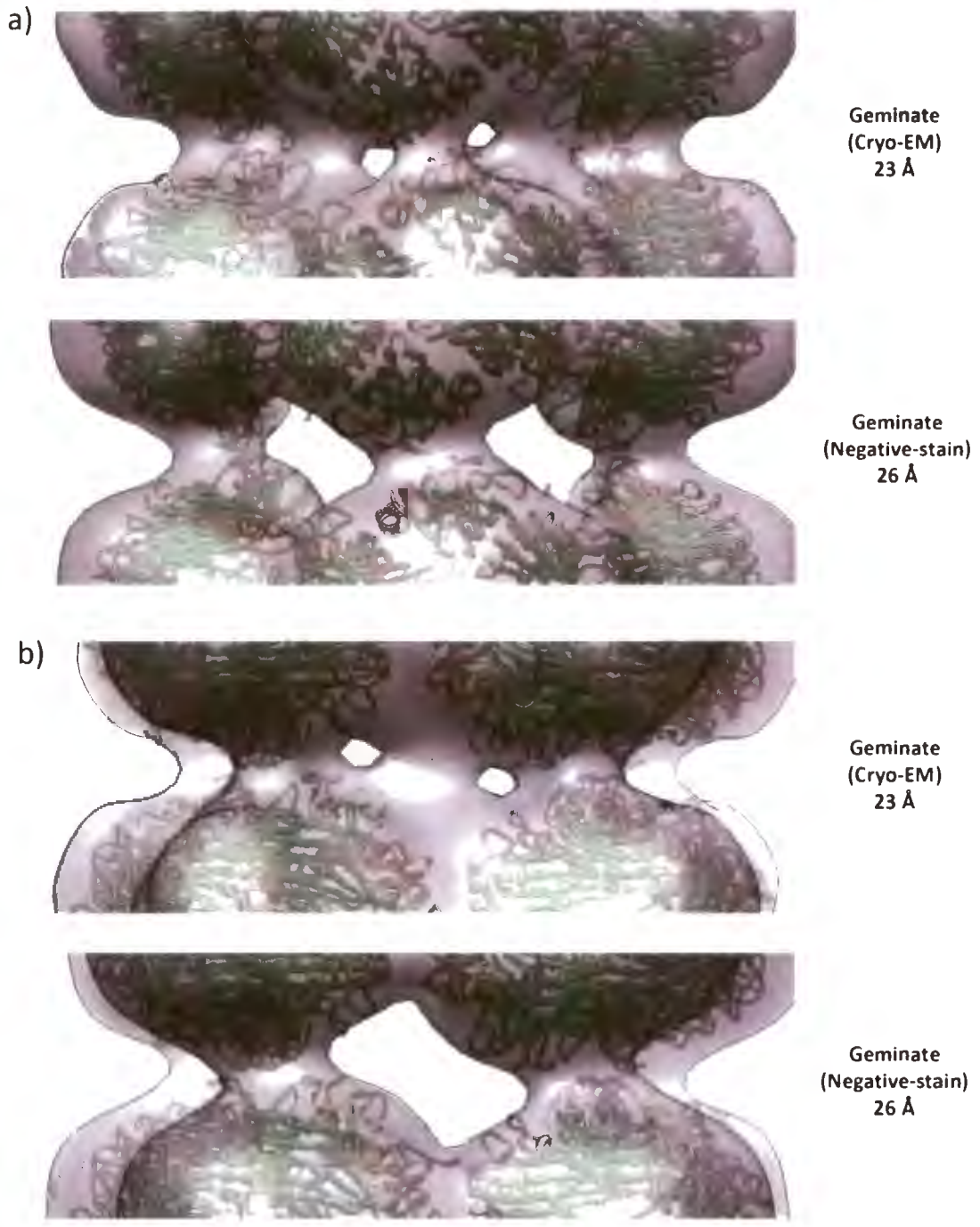

Figure 3-33 Equatorial interfaces of the geminate models viewed from either side of a global two-fold axis of symmetry. a) A global axis of 2-fold symmetry viewed from the "front" which reveals the equatorial interaction that is best accounted for by the comparative model. b) The same axis as a), but viewed from the opposite side of the particle.

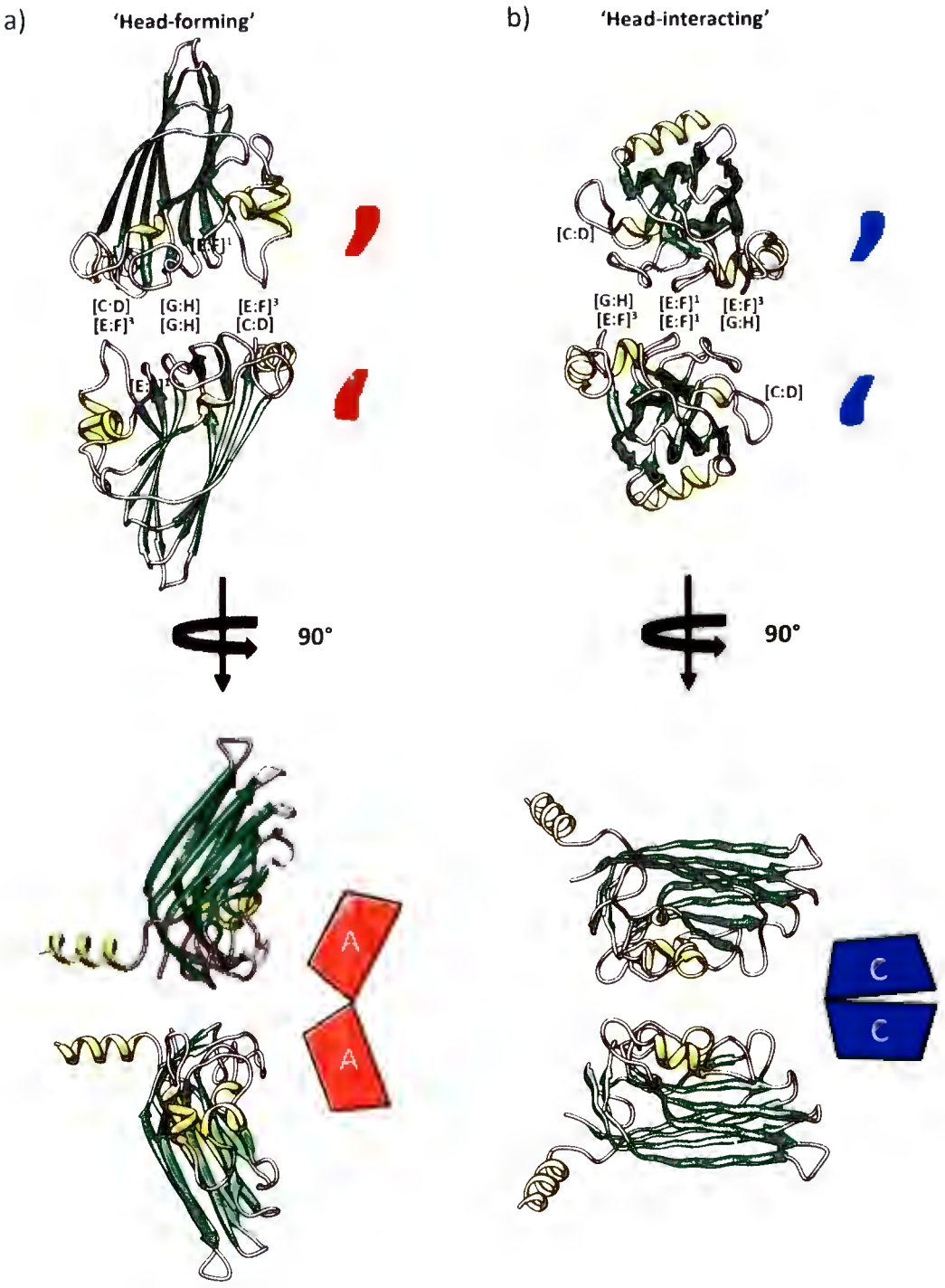

Figure 3-34 CP:CP interactions at axes of 2-fold rotational symmetry. MSV CPs forming interactions across the 2-fold axes of the icosahedral monopartite capsid, or local 2-fold symmetric axes of quasi-icosahedral geminate head, are shown in a), and are termed 'heading-forming' interactions. Shown in b) are the primary 2-fold globally symmetric interactions that allow formation of the geminate capsid, and are described as 'head-interactions'). The colour scheme of geometric schematics (apostrophe and wedge 'models') isin agreement with those used in **Figure 3-39**, and is intended to reflect either one of the three putative MSV CP conformers.

3.3.5 Evaluating the possible structural divergence of geminivirus capsids

The homology modelling and pseudo-atomic modelling study described in the previous sections was further elaborated by using the alignment as a starting point to carry out additional alignment of other geminivirus CP sequences. The primary question was to determine how the CPs have been adapted or diverged, and furthermore, are these differences likely to result in significant differences in capsid architecture as well as function during each of the stages of the virus life-cycle. Such an alignment should shed light on whether the calcium-binding site identified (refer Section 3.3.4.1) is a conserved feature across the family, and whether this sequence level information suggested that there were indeed distinct structural classes of geminate architecture (that vary geometrically). The latter point is suggested by our observation that the offset between geminate halves is different between MSV and ACMV; -11° and to 20°, respectively.

The sequence-structure alignment shown in Figure 3-21, served as a means of accurately aligning other geminivirus coat protein (CP) sequences using the MAFFT webserver (Katoh *et al.*, 2002). This was used to produce a multiple sequence alignment (MSA) that included some of the prototypical species of the *Geminiviridae*, and this is shown in Figure 3-35. Just as the original sequence-structure alignment revealed important insertions in the CP sequence of MSV CP, so does the MSA reveal important insertions in many other geminivirus CPs that are in many instances in different locations of the β-barrel. These are discussed in the following paragraphs, and an effort to summarise the presence or absence of these insertions is made in Figure 3-36. What is remarkable about the insertions (especially those contributing to quaternary interfaces) is that when they are present they are of similar sequence, even across genera.

The calcium-binding site consisting of five amino acid residues between β-strands F and G (βF:βG), appears to show some degree of conservation in a large number of geminivirus CPs. While the sequences in this region are not identical, the threonine, whose backbone carbonyl oxygen atom is responsible for co-ordinating a calcium ion in STNV on the 5-fold axes of symmetry, appears to be maintained in most cases, but in two cases this residue is substituted with either a serine or a tyrosine (possibly neutral substitutions). While the sequence alignment suggests that MSV may also preserve calcium ion coordination on the 3-fold axes of symmetry, it would appear that the third and 'general site' occurring at a 7 Å displacement from the 3-fold axes is not conserved. Inspection of the remainder of the geminivirus CP sequences in these regions shows significant sequence divergence suggesting that these capsids may have evolved a separate mechanism for regulating the stability of the capsid in environments where genome release is appropriate.

With respect to possible effects on geometric architecture, as well as functionalization for interaction with specific insect vectors, additional sequence insertions are observed in all regions intervening β-strands of the STNV β-barrel, but not all of these are manifest in any one geminivirus CP at the same time. Interestingly, we see that in some regions two separate loop-insertions occur, each with a length of typically 4-6 amino acids on each side of the β-hairpin. The pattern of loop insertions varies among geminiviruses CP sequences, but it is clear that CPs can be grouped according to similar patterns. Such regions of the β-barrel, by conventional wisdom, would be expected to show high rates of divergence. However, inspection of the geminivirus sequences will show that these are conserved amongst genera, suggesting that these regions have highly specific functional roles that would likely constrain these rates. Assessing the phylogenetics and insertion patterns we can see that the geminivirus CP sequences can be grouped in corresponding ways. Similarities and differences are discussed briefly in the paragraphs that follow.

BCTV and TpCTV, occur in a similar region geographically, and show a similar, but not identical pattern of loop insertions. These two viruses, despite being of different genera, are grouped by the phylogenetic tree suggesting that they may have diverged from a common ancestor relatively recently. ACMV, ToLCV, and BGYMV are all Begomoviruses and show precisely the same pattern of loop insertions, and are grouped closely by the phylogenetic analysis (not surprisingly). Their pattern of loop insertions differs significantly to that of MSV. TpCTV (a Topocuvirus) shows some similarities to the Begomovirus sequences with respect to loops involved in quaternary contacts, and TpCTV also shares a loop insertion at the βC:βD loop insertion.

Relevantly, all geminiviruses possess the βE:βF[3] loop-insertion, suggested to be involved in quaternary contacts between CP_5 capsomers. That the length, sequence and presence of this insertion is highly conserved in geminiviruses supports the hypothesis that it may be involved in formation of the geminate morphology specifically. While the βE:βF[3] loop-insertion shows high conservation, the remaining loops proposed to be involved in formation of the 2-fold and 2-fold-like ('head-interacting') interactions are not conserved. Furthermore, a loop insertion in the βC:βD region, possibly involved in forming 3-fold ('head-forming') and 3-fold-like ('head-interacting') contacts also varies across geminiviruses with begomoviruses and the topocuvirus showing a similar pattern, different to that of the Mastrevirus and Curtovirus. A similar situation exists for the βG:βH loop, with mastrevirus, curtovirus and topocuvirus showing a similar pattern, different to that of the begomoviruses. MSV shows a pattern of loop insertion presence that is identical to BCTV (a curtovirus), and consequently we might expect this virus to show an identical geometric architecture to MSV. In consideration of the other differences it would appear

possible, perhaps even likely, that the geminate architectures could have diverged at the sequence level to result in assemblies which vary geometrically. Structural analysis of the topocuviruses, as a starting point, will be most informative in this regard as TpCTV exhibits all loop insertions at relevant sites, and must also package the largest genome (2850 nt); important differences with respect to other geminiviruses which might motivate differences in architecture.

The βF:βG insertions show high similarity between these viruses, and probably reflects that these regions of the CP are responsible for interacting with leaf hopper (vectors) proteins. TpCTV is next most similar in this region, with the implication that the treehopper vector is more similar to leafhoppers, than the whitefly vector of begomoviruses.

3.3.6 Electrostatic surface potential of interior of capsid

The electrostatic potential of a molecular surface will influence how it interacts with other molecules in its vicinity (Baker *et al.*, 2001). In this analysis we wished to determine whether the MSV CP is likely to interact strongly (or weakly) with the its packaged genome, whether these interactions are specific, and any other properties of the site(s) of (possible) ssDNA recognition that might inform our understanding of the interior arrangement of ssDNA in the infectious virion. Analysis of the predicted electrostatic potential of MSV CP is made in reference to the predicted potential of STNV CP. The validity of this comparison is supported by the apparent structural and functional homology of these proteins, as well as the numerous studies available describing the structure of interacting RNA motifs with have associated with STNV CP (Bunka *et al.*, 2011; Ford *et al.*, 2013).

Inspection of Figure 3-37 will show that like STNV, MSV also exhibits a predominantly positively charged inner surface. The electropositive surface is continuous and produced by six residues located on inward presented β-strands (βF, βG, and βD, with respect to the capsid interior) as well as three residues located on the capsomer 5-fold axis facing side of the N-terminal α-helix. The electrostatic surfaces are rendered at an equivalent threshold, and this reveals that MSV has a stronger potential in this region of the CP when compared to STNV. That the electropositive surface is continuous (or electrostatically 'unstructured', provides some evidence to suggest that the inner surface does not impose a degree of selection on the orientation and shape of DNA which binds with it. This analysis is by no means conclusive, as the mechanisms by which such recognition events occur can be complex, and in some cases even mediated by layers of structured water molecules (Lin *et al.*, 2003).

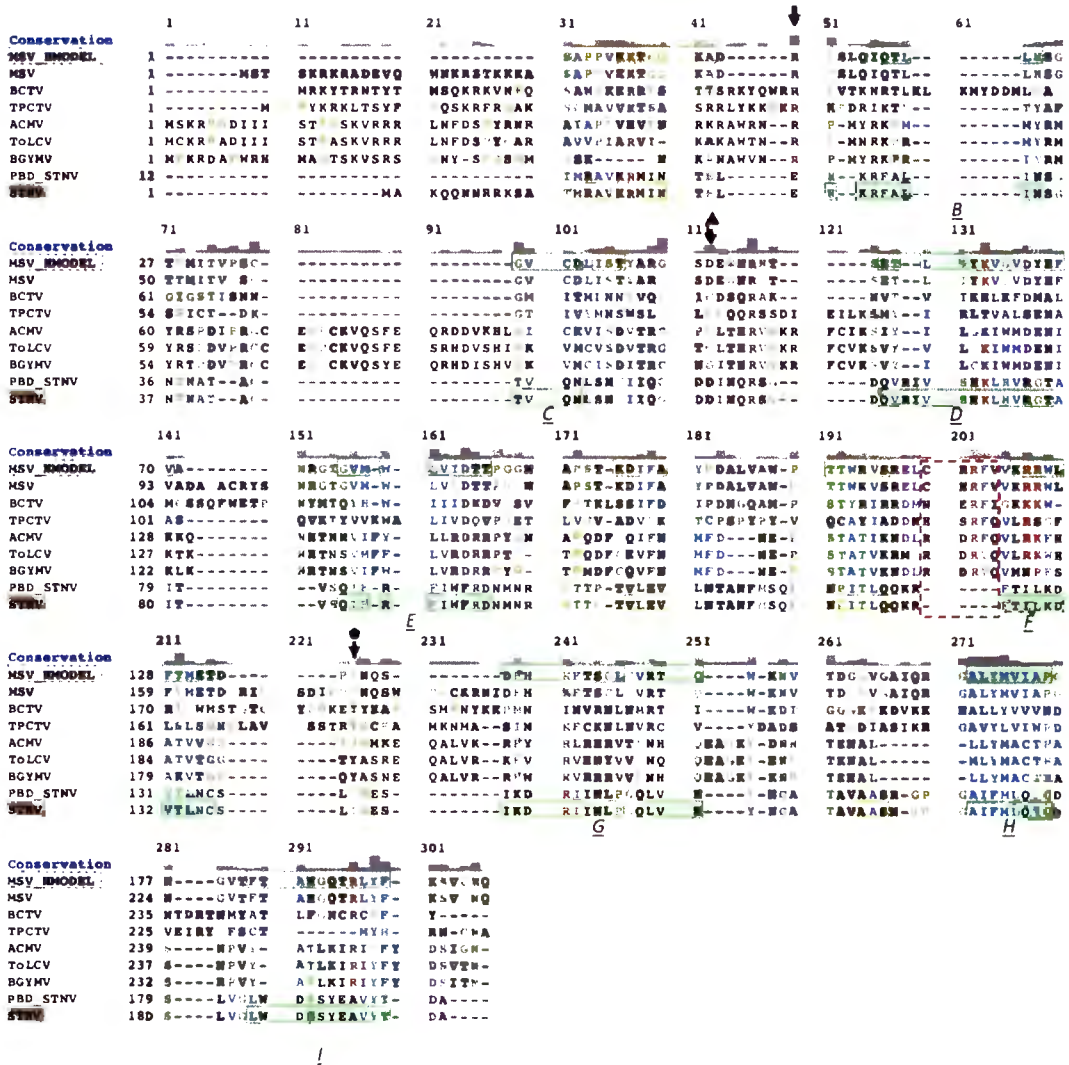

Figure 3-35 Multiple sequence alignment of geminivirus coat proteins against STNV. A multiple sequence alignment was constructed using a number of full-length geminivirus CP sequences and the structural information derived from sequence-structure threading. MSV, BCTV, TpCTV and BGYMV represent the major geminivirus genera. For information these are also aligned to the full-length and PDB structure sequences of STNV; which differs only at the N-terminus.

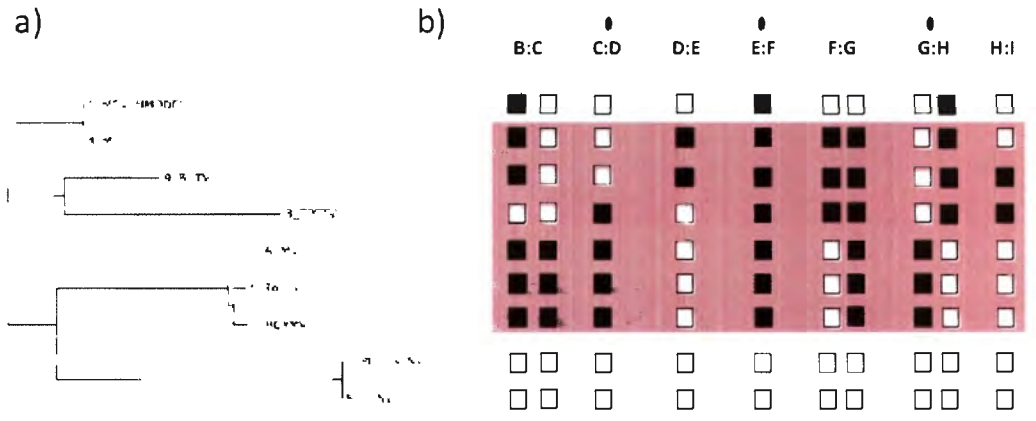

Figure 3-36 Divergence of viral CPs and presence of loop insertions. a) The evolutionary distances of geminiviral CPs is shown by phylogenetic tree. As expected the STNV sequences form their own group that is distantly related to the remainder of the sequences. b) Presence of β-loops is summarised by table schematic. Columns, as labelled, represent each β-loop position in sequence from left to right (N-terminus, to C-terminus). If a loop insertion is present in a viral CP, the box is filled black, if the no insertion exists with respect to STNV then the box remains unfilled (white). Loop insertion in the vicinity of CP_5 interactions across the particle 2-fold axes are marked by ellipse. Full length geminiviral CP sequences are highlighted by the red-shaded box.

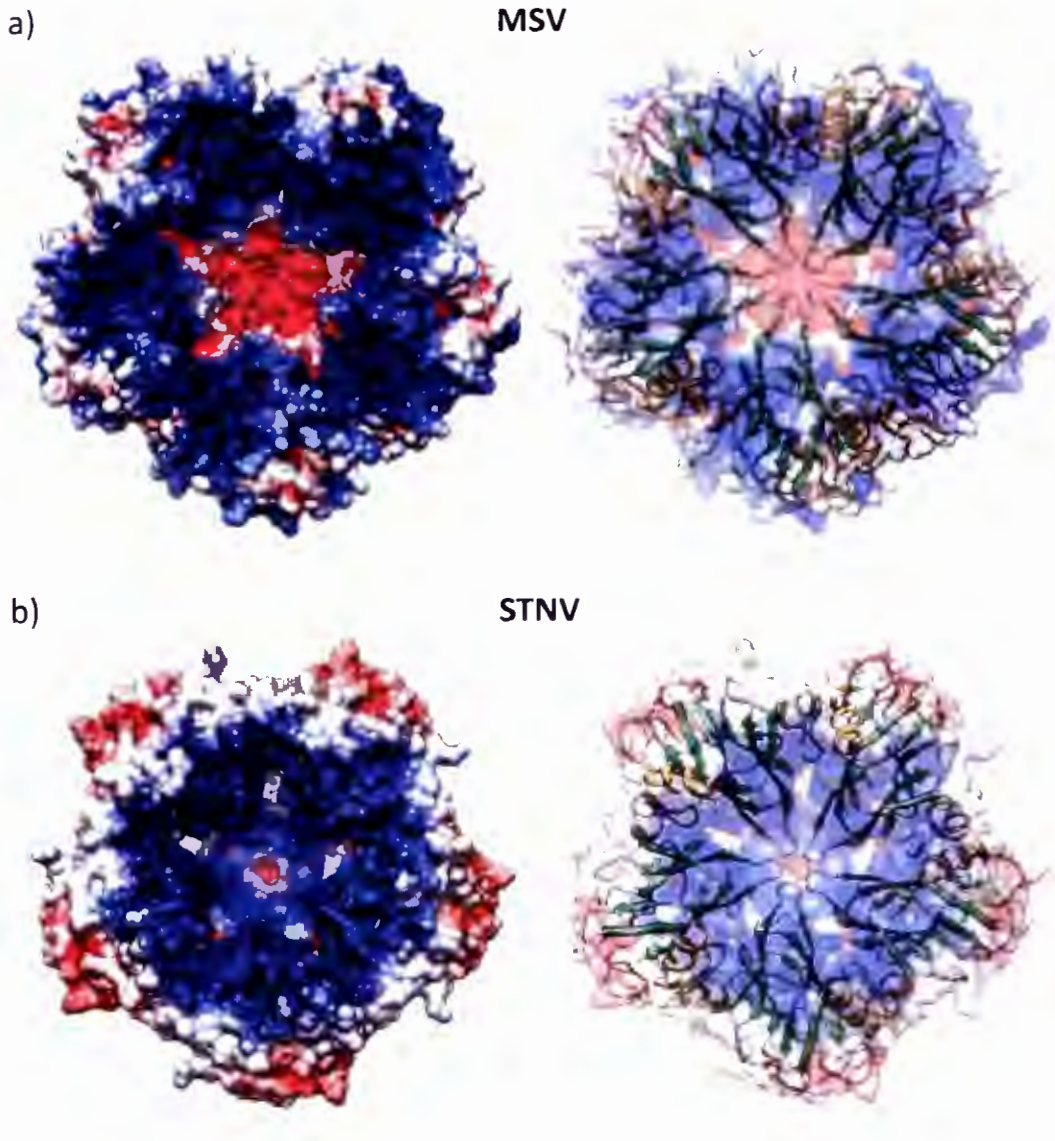

Figure 3-37 Electrostatic potential maps of the inner surface of the pentameric capsomers. The a) MSV comparative model, and b) STNV capsids are viewed from the interior of the capsid. Strongly positive electrostatic potential is seen to cover the inner surface of the STNV capsomer. This observation suggests that residues involved in binding RNA are distributed regularly over this surface. For MSV this property is generally preserved, however at the very center of the capsomer the potential is largely electronegative, suggesting that DNA does not bind to this region. This agrees with the EM reconstructions (refer Figure 3-26).

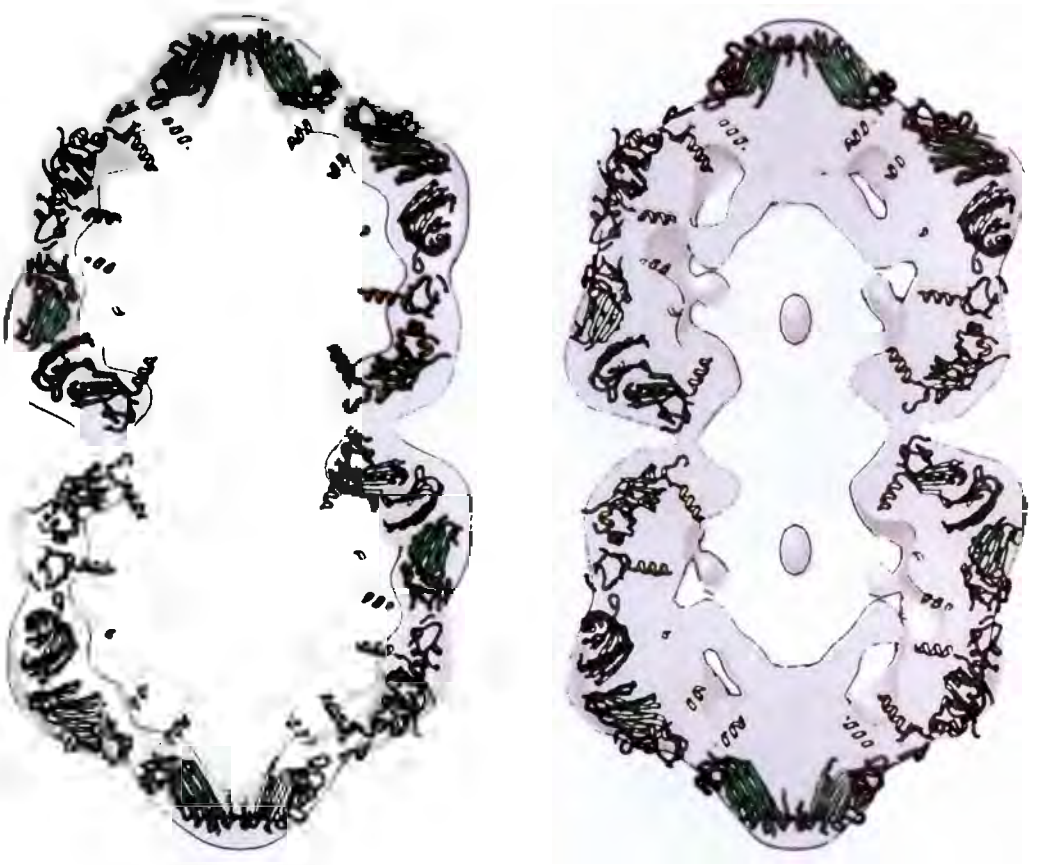

Figure 3-38 Internal densities of the EM reconstructions assessed in light of the comparative model. Central slabs of the bipartite reconstructions are shown as transparent grey in which the geminate atomic models are rendered as ribbon and coloured as previously described. It is clear that the N-terminal arms are not accounted for in the map determined by negative-stain electron microscopy, however, the appearance of the pentameric units of assembly (capsomeres) is in excellent agreement with the model at the resolution available. The cryo-EM reconstruction provides details which support the conformation assigned to the N-terminal α-helix. These α-helices can be seen to organize around 3-fold symmetric indentations on the interior of the virion, an observation that supports the assignment of these as structurally important features of the capsid interior.

3.4 Discussion

Each 3D reconstruction has been shown to be geometrically consistent irrespective of whether the reconstruction was derived from negative-stain or cryo-EM. This is particularly surprising because some degree of distortion is expected for dehydrated and stained particles. Instead, we observe that the dimensions of the particles are very similar; this within the limits of accuracy made available by the resolution of the map, as well as the fitting parameters explored. Even more remarkable is that the offset between the geminate particles is also in good agreement. This observation confirms the previous finding that the geminate assembly is a rigid assembly, without flexibility at the equator (Zhang *et al.*, 2001). This stability also implies that well-defined and specific interactions characterize interactions between the CP_5 units. As already revealed in the previous sections, this was confirmed by inspection of pseudo-atomic models, and the implications of the fitting study on our understanding of CP_5:CP_5 interactions and conformational switching during assembly are further elaborated in the following sections.

Assembly and formation of the virion will have an influence on three inter-dependent aspects of the virions structure, these are: adoption of quasi-equivalent conformers, the organization of the genome as well as the assembly pathway itself. Based on available information from previous studies, as well as inspection and consideration of the present data, hypotheses regarding the details of these processes are derived such that these might inform future experiments.

3.4.1 The potential nature of quasi-equivalence in MSV

The geometric consistency between the reconstructions was also important because it provided a solid foundation for comparative analysis of the interfaces predicted by pseudo-atomic model. Arguably, the pseudo-atomic models should provide sufficient information to speculate about the nature and number of the various structural environments within the capsid, and how the CP units relate to these.

Within the context of the quasi-icosahedral head, the fitting study revealed that the CP_5 is rotated by ~10° in comparison to STNV. This implies that the quaternary interactions between pentameric capsomers are not conserved between these viruses, despite the apparent high degree of conservation of each CPs tertiary and quaternary structure; i.e. the conserved pentamer forming interaction. This difference allows for interactions between different (loop) regions of the β-barrel, and in fact, it would appear that amino acids that extend some distance from the core β-barrel structure mediate all CP_5:CP_5 interactions. It may be reasonable to assert that during the divergence of the geminiviruses, capsids able to package more nucleic acid were selected over those with reduced radii and able to package less nucleic acid. However, an additional reason for

the observed geometric differences, may relate more to local properties allowing for formation of quasi-equivalent conformers and interactions required by the geminate morphology itself. Specifically, it is the opinion of the author, that the structural flexibly in these regions must be related to the ability to form the geminate virions.

The geminate model shows that the primary 2-fold interaction (occurring across the equator of the particle) is mediated by a single CP on each side that is engaged in an icosahedral "2-fold-like" interaction, but with very important differences. Firstly, different pairs of loops are in association at these sites, and secondly the dihedral angle between interacting β-barrels are different by approximately 60°; this because the edges of two spherical incomplete quasi-icosahedral heads meet at these locations. These observations strongly imply that the CPs in these locations are in quasi-equivalent conformations in order to make such contacts at such remarkably different geometries of association. That the offset between heads arises owing to a different set of interactions between loops is extremely relevant because it reveals that at least for this site, true quasi-equivalent interactions (varying only subtly in geometry between the same interacting pairs of residues) is unlikely to be case. Instead, the MSV CP appears to have engaged in a separate (distinct) interface, made of a unique set of interacting residues, in order to mediate the associations between incomplete icosahedral environments. It is likely that similar behaviour occurs at the secondary equatorial interfaces, but such information will only be revealed by a future study of greatly improved resolution.

Predictions of the number and location of actual CP conformers can be reduced to consideration of the location of each respective CPs 2-fold and 3-fold interfaces within the context of the whole assembly; i.e. whether these regions contribute to formation of the quasi-icosahedral head, or to the interface between quasi-icosahedral heads. This is a coarse approach, appropriate to the resolution available, but should capture the general properties of capsid. As is apparent in Figure 3-39, three conformers are predicted by this method and are termed A (red), B (green), and C (blue). The A conformers are most abundant and form the majority of structural units of the quasi-icosahedral heads. The B, and C conformers mediate interactions between heads ('inter-head' contacts), and these conformers are the true basis for the term 'quasi-icosahedral' in reference to the structural properties of each half of the geminate capsid. The C conformer (mediating the primary interaction across the equator) is the most structurally deviant in that both of its 2-fold and 3-fold interfaces appear to adopt alternative conformations to mediate contacts between the geminate halves. The B conformer makes 2-fold contacts with an A conformer with an adjacent CP_5 (within the head, 'intra-head'), and 3-fold contacts along with its pentameric partner CP_5 C

conformer, across the equator ('inter-head'), with a B conformer and C conformer from another quasi-icosahedral head – thus forming the secondary type of equatorial interaction.

In comparison to $T = 3$ viruses, the A, B, and C conformers demonstrate some architectural properties in common, and these similarities were used as the basis for assigning these labels (A, B and C) to each. These properties are mainly in reference to the global symmetry operators of each architecture and are as follows: The A conformer is located around the global 5-fold axes of symmetry in both the geminate $T = 1$ bipartite structure, as well as in the $T = 3$ architectures. Also in both architectures, the C conformer is found to mediate global 2-fold interactions (the primary equatorial interaction in MSV). Hence, to construct the bipartite architecture from 110 CPs, important geometrical properties are shared with the icosahedral architectures that most often consist of multiples of a multiple of 60 subunits.

3.4.2 Arrangement of ssDNA on the interior of the virion

MSV, and other geminiviruses, possesses circular genomes that presumably self-interact in a sequence specific manner to adopt secondary structure. That the genome will fold further to adopt a tertiary structure implies that the conformation of the genome may be restricted, and these constraints probably facilitate self-assembly of the virion by compacting the genome prior to association of the coat protein pentamers. Additionally such folding may present structural motifs that are responsible for recruiting the CP_5 units, however, little is known about the structural organization of the MSV genome beyond computational secondary structure predictions. Additionally, precisely how secondary/tertiary structure are influenced during virion assembly are important questions to address. Related to these is whether the packaged viral genomes adopt multiple conformations or just one conformation in the symmetric environment of the shell?

How does the geminate capsid compare to other $T = 3$ architectures with respect to nucleic acid packaging capacity and density? Calculations regarding available volume for nucleic acid packaging, as well as estimated packaging densities are provided in Appendix A. These were calculated according the scheme set out by Speir and Johnson (Speir and Johnson, 2012), and are discussed here in reference to an improved understanding of packaging in the case of MSV in relation to other similar viruses. As may be seen, the column showing relative capsid volumes reveals that in general, $T = 3$ architectures provide 1.2-1.5 times the volume for packaging than does the bipartite $T = 1$ architecture. Hence we might say that the geminate architecture provides an architecture for nucleic acid encapsidation that *approaches* that of other $T = 3$ viruses and is closest to CCMV, a plant virus which we have already discussed as being similar to in some behavioural respects to the geminiviruses. CCMV has a multipartite genome, each segment

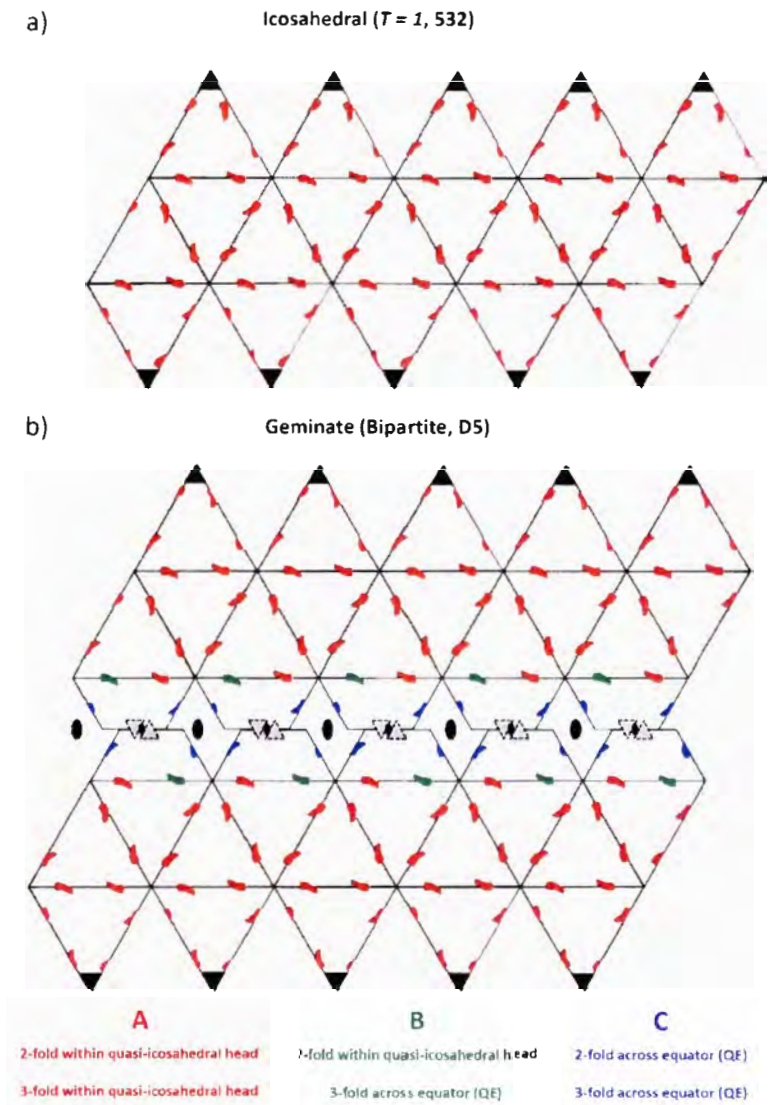

Figure 3-39 Quasi-equivalent conformers of the MSV CP plotted on the hexagonal lattice representation of the architecture. The geminate capsid architectures are represented by hexagonal lattices as in **Figure 3-20**. a) Coat protein units are represented by semi-colons (') and are coloured red indicating that a single conformer is present in 60 equivalent environments, and adopt the A conformer in all instances. b) The bipartite capsid is shown here as two incomplete $T = 1$ lattices in which 10 subunits have been removed from each. The lattices of each head are displaced by an amount equivalent to the ~ -11° offset observed in the experimental EM maps. The schematic illustrates how two closed lattices can associate in a way compatible with global 2-fold symmetry while maintaining largely quasi-equivalent conformations between protein subunits of the capsid. Positions of global symmetry axes are marked with black shapes with grey linings. Three conformers are predicted according to this scheme, with two conformers (C - blue and B - green) unique to the bipartite capsid and higher species, which mediate interactions across the equatorial region(s) of the assembly. It is clear that the offset maximizes the number of interactions taking place between components of the two halves, while preserving interacting regions used to form each head itself. See text for further discussion, and refer to Figure 3-40 for additional information.

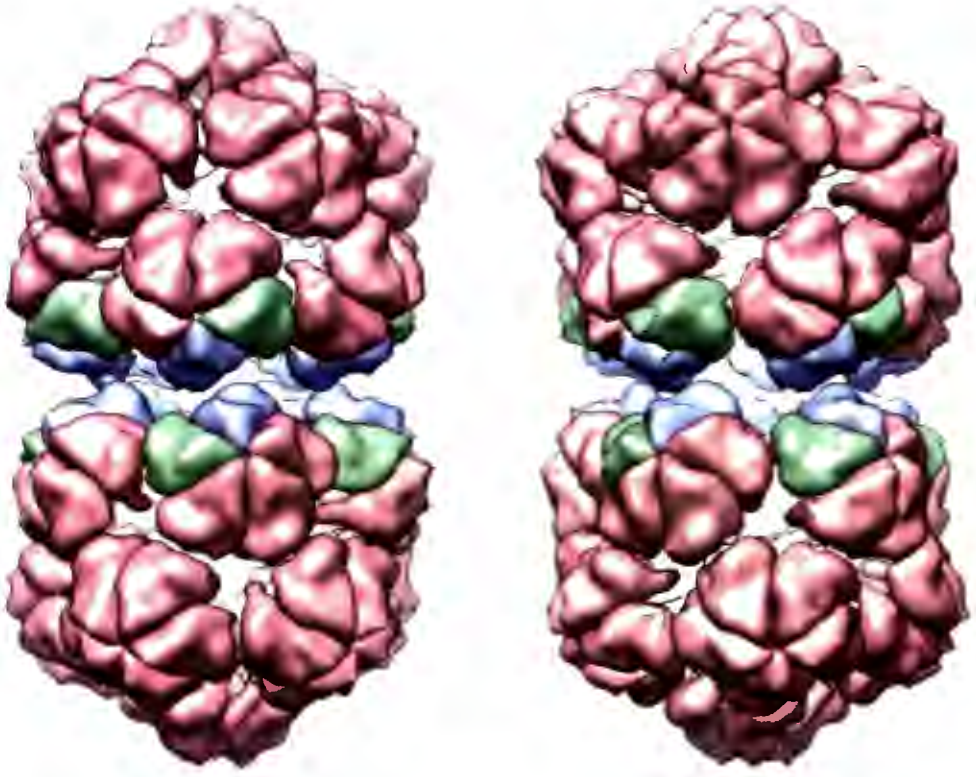

Figure 3-40 Location of quasi-equivalent conformers in the geminate pseudo-atomic model. The pseudo-atomic model is rendered as a surface representation at intermediate resolution and predicted conformers coloured according to the scheme described in Figure 1-4 and Figure 3-39. The model is viewed from along both directions (from the front and from behind) of a typical global axis of 2-fold symmetry. Geminate heads were displaced by 5 Å in order to render the figure and also to improve clarity; blue CP subunits (C conformers) are in close contact in the actual pseudo-atomic model.

packaged independently into its own capsid, with the largest segment being 3.2 knt in size. behaviours analogous to the begomoviruses, for example.

From analysis of the values provided in Appendix B it is also apparent that $T = 3$ viral genome sizes, for the most part, also reflect the distribution of capsid volumes. And most importantly, inspection of the resultant packaging densities, as well as percentage of the capsid volume predicted to be occupied, suggests that viral genome sizes of the geminate and $T = 3$ viruses appear to vary in size so as not to exceed a certain packaging density (~4 A^3/Da). One possible explanation for this behaviour is that the energetics of packaging associated with spontaneous events of macromolecular association and folding (outlined as the 'concerted mechanism' previously), can only achieve a limited concentration of nucleic acid owing to the persistence length of these molecules, as well as charge repulsions, which would otherwise require energy to overcome. Poliovirus, is known to couple replication and packaging of viral genomes in a poorly understood process, however, the relatively large genome is thought to be driven into a preformed capsid under the same force required to execute the replication itself; hence it is reasonable to suggest that high packing densities can be achieved.

As may be seen in Appendix A, the $T = 1$ viruses also achieve high packing densities, similar to that of poliovirus. For viruses of this class the adoption of tight associations of duplex RNA motifs at icosahedral 2-fold vertices, induces significant ordering in the molecule which can help rationalize how this might be possible. In such cases the genome is largely constrained to interactions with protein surface, and is not diffusely organized within the capsid. What these previous studies have revealed is that the genome is not equally distributed throughout the inner volume of the capsid, but is structured by adoption of its own secondary structures as well as associations made with the protein shell itself. Do the 3D reconstructions determined during this study support this assertion in the case of MSV?

The appearance of protein-associated density in the monopartite and cryo-EM (bipartite) maps suggests that nucleic-acid is bound to the protein layer with high occupancy. This is supported by the electrostatic potential prediction, which showed a strong electropositive surface formed by the underside of each β-barrel domain. At this resolution it is not possible to comment on the structure or arrangement of this protein-associated layer of DNA, beyond the simple observation that the ssDNA in this region probably adopts duplex secondary structure arrangements (according to the radial dimensions of the layer itself). The association of stem-loops in such regions has been demonstrated for a number of similar simple viruses, and is appears to be a strong possibility in this case as well. This can only be confirmed by high-resolution study, and it can be expected that such features (e.g. stem-loops themselves) will only be visible at higher

resolutions if they adopt an arrangement that is regular and consistent with the symmetry of the capsid.

A schematic representation of the organization of nucleic acid is shown in Figure 3-41. This is a qualitative model, showing the hypothesized general location of DNA in the virion. The model does not speculate about the true secondary or tertiary structure of the genome, as there is very little experimental information available beyond radial density measurements provided by cryo-EM. The model shows that nucleic acid is organized in two shells. The protein-associated layer discussed in the previous paragraph, as well as inner shell which is likely in association with the N-terminal arms of the CPs. Analysis of the cryo-EM map suggested that these regions of nucleic acid could be single-stranded, and that the reason for this is the binding of the N-terminal in an extended (disordered) conformation prevents Watson-Crick associations by binding along the length of ssDNA. Alternatively, the N-terminal arms could bind duplex secondary structure regions, and assist compaction by neutralizing charges along the length of these structures. In the model of assembly formulated in the follow section, such binding could occur during rolling-circle replication, or soon after its completion. Analysis of the cryo-EM reconstruction with the geminate model fitted, provides some indication of where the NLS is located in the assembled virion, although there is some ambiguity. The cryo-EM reconstruction agrees with the predicted placement of the N-terminal α-helix predicted by homology modelling. The location of the additional 23 N-terminal amino acids (containing the NLS) is open to speculation, however, the core of the geminate heads consists of a relatively dense arrangement of both DNA and protein was suggested by the negative-stain map that showed that stain failed to penetrate into the inner regions of the virion.

3.4.3 Possible models for assembly and disassembly of the geminate virion

Viral particles assemble from an initiation (or nucleation) complex, and proceed to the end state via a series of intermediate species. Understanding the details of a virus's assembly pathway offers the potential that this might be interfered with, providing an avenue for development of highly specific therapeutics. Determining such a pathway is challenging owing to the transient nature of the intermediate species, however, an attempt is made in this section to formulate a hypothetical model which might inform future experiments which attempt to observe these intermediates. The hypothesis relies on the assumption that the basic unit of assembly (whether CP monomer, dimer or pentamer), as well the interactions of this with the viral genome will influence how and in what order these intermediate species arise.

Given the intricate nature of atomic interactions and buried surface area between CPs at the pentameric interfaces of the CP_5, it has been argued previously that this is most likely the unit of

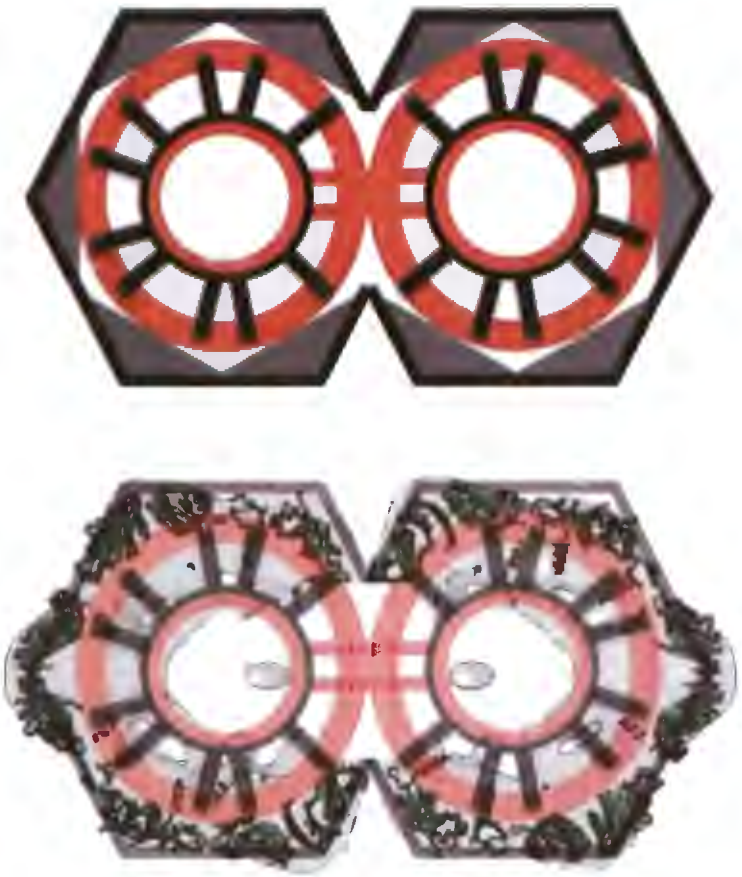

Figure 3-41 A hypothetical and qualitative schematic of MSV genome organization informed by the cryo-EM single-particle study. Top, the coat protein subunits (black/grey) contribute to the formation of two primary layers of DNA within the virion (red). The outermost layer binds to the inward facing surface of the β-barrel domains, and the inner-layer binds to the N-terminal arms (black cylinders and inner black circles). The genome and N-terminal arms are likely to be heavily disordered (residues 1-23 binding to the asymmetric and possibly unstructured genome). Bottom, the general qualitative agreement of this model with the experimental map is shown by superimposition. It should be noted that the schematic does not reflect the secondary-structure expected of ssDNA molecule such as the viral genome. In the outmost layer, it might be expected that the genome to adopt stem-loop arrangements, while in the inner regions secondary structure is expected, although this may be reduced by association of the positively charged N-terminal arms of the viral CPs. At the chosen threshold, the map reveals densities on the interior of the virion (grey central ellipsoids, as well as internal tubes) that expected to be contributed by genome as well as N-terminal arms.

the most obvious example of properties directly related to the behaviour of the CP at different stages of the virus life-cycle, but other – less obvious – constraints may also exist.

It has been suggested in the previous paragraph that the $T = 1$ lattice forming functionality of the CP is merely adapted to form the bipartite structure, this in favour of the alternative of redesigning the CP from scratch to form a quasi-symmetric spherical shell of larger radius. This adaptability leads us to the question of whether the geminate geometrical parameters vary among geminivirus genera? This is a question of interest because we have seen that interacting regions are composed of discrete interfaces encoded by residues on highly variable β-loops, and the CP amino acid sequence analysis described in this study showed that different patterns of insertion characterize the CPs of different geminivirus genera in CP_5:CP_5 interacting regions of the core domain. These observations suggest that geminate capsids varying in geometric properties may indeed exist - certainly for the geminate coat to function adequately, the most important property is that the heads interact, and not the precise details of how they interact. However, it will be interesting to see whether the quasi-icosahedral geometry of the heads places a constraint of the geometry of the D5 bipartite structure by requiring that the 3-fold interfaces are brought into conjunction in a manner that adheres to the chirality of the heads, as is the case for MSV. If such a constraint exists, all capsids are expected to show the same geometrical properties. Negative-stain has proved immensely valuable in this study, providing information of about the MSV particle that with respect to the particle geometry rivaled the 'gold-standard' near-native visualization provided by cryo-EM. Cryo-EM is of course, however, required to probe the conformation of interior structures such as the genome and the N-terminal packaging arms, but these results show convincingly that other relatively low abundance (low 'titre') geminiviruses could be reliably studied by negative-stain EM to investigate possible differences in architecture.

Analysis of the results revealed important details about MSV structure, but failed to critically answer important questions that arose from the analysis such as what is the precise conformation of interacting loops, and which residues are interacting and how. These limitations necessitate further study by an integrated structural/molecular approach described in the following paragraphs. The important questions about MSV can be separated into two categories: the fine structure of the capsid, and the structure and arrangement of the genome and N-terminal arms. These questions can only be addressed by determining the structure of the capsid to near-atomic resolution, and it is important to note that the backbone and side-chain conformations of all loop regions in the context of 'head-forming' interactions will likely require empirical information arising from a crystallographic study of the monopartite capsid (discussed below). As discussed in Chapter 3, the present study motivates attempts to crystallize the icosahedral form of MSV, as

confirmed that amino acid insertions in regions intervening conserved β-strands (referred to as β-loop insertions) suggest that the geminivirus CP is functionalized within the context of each viruses host and vector using modification of the domain in these regions. This property is common to many other viruses encoding the same fold (the β-barrel), but has important implications for geminate structure as discussed in the next paragraph.

It has been asserted herein that the bipartite geminate capsid represents an alternative, but nearly equivalent packaging solution to that of $T = 3$ quasi-symmetric viruses. Hence, while most isometric viral capsids have evolved to adhere to the theory of quasi-equivalence and triangulation, the geminiviruses are another example showing that these entities are by no means obligated to follow these principles to the full extent. Put another way, from an evolutionary perspective, quasi-equivalence is clearly a highly successful geometrical solution to capsid construction, however, it would appear that CP subunits can evolve to adhere to other geometric solutions which are equally valid – given the requirements of the virus, as well as the 'evolutionary potential' of the packaging machine (discussed further below). To elaborate, and based on the results of this study, as well as the body of structural information available for $T = 3$ viruses, we might speculate as to why the geminate morphology emerged at all, and also why it is only observed for the geminiviruses? The coat protein of STNV has been noted amongst satellite viruses as having particularly pronounced and intricate pentamer forming interfaces. Indeed, this study has furthered our understanding of just how well this interface and the tertiary structure of the MSV CP appears to be conserved with respect to STNV. It might be asserted that unlike $T = 3$ viruses which are able to form both pentavalent and hexavalent assembly intermediates from dimeric assembly units, the geminiviruses are unable to form hexavalent assemblies because they are constructed from a strongly associating pentameric building block. We might speculate that mutations leading to disruption of the pentamer interface would require a major remodelling of the β-barrel domain, and hence would be evolutionarily unfavourable. From this point of view the geminate morphology might be considered a solution to packaging a larger ($T = 3$ near-equivalent) genome that is closer in 'sequence-structure space', as only modifications at the CP_5:CP_5 interfaces would be required, and occurring only at highly variable peripheral regions intervening the core β-strands; mutations which may be considered largely neutral with respect to evolutionary and folding fitness of the CP. Virus CPs must fulfil many requirements, not least that the genomes be packaged, and in addition to the previous argument, it must be kept in mind that the CP 'packaging machine' must be able to serve the contextual requirements the virion with regards to both assembly and disassembly, and it may be that the ancestral $T = 1$ architecture offered the best solution with respect to the particular host organism in a number of other ways. The calcium-ion binding sites identified during this study are possibly

also suggested that conformational switching of CP units is not directly (or locally) regulated by a specific switching mechanism, but rather arises as an emergent event when two spherical CP lattices interact, and is thus indirectly regulated by the tertiary structure of packaged genome.

The structure and role of the MSV genome are important considerations from the perspective of viral assembly and evolution, as influenced by packaging. In review of the findings of this study: the negative-stain reconstructions did not reveal many details about the genome organization or internal structure of the MSV capsid; as was expected. However, the cryo-EM reconstruction revealed layers of internal density (in addition to the protein capsid), which despite being obscured by symmetry averaging, were able to inform us of possible arrangements of internal components (i.e. the N-terminal arms of the MSV CP, as well as the MSV genome). The strongly positive electrostatic potential of the inner surface of the protein capsid suggests that ssDNA associates at these sites to produce a layer of protein-associated DNA, however, neither the cryo-EM nor the structural bioinformatics study suggested that the ssDNA adopts a preferred secondary structure (e.g. duplex or hairpin) to form this interaction, and in fact suggested that it is non-specific binding of phosphate backbone charges in these regions that accounts for the observed density in the EM map. It may be that the 'true' structure of the genome and N-terminal arms may never be known because in every virion the precise details are different, but experimentation must be carried out to investigate the structure and possible ordering of this genome – as this may provide information about a structural feature of these viruses of fundamental importance to their evolution. Such an investigation can only proceed by finding a means of asymmetrically labelling a component of the virion. Because the protein shell is symmetric, and furthermore, because reconstitution of the virion from disassembled components has never been demonstrated, a method that focuses on the (asymmetric) genome must be developed. Labelling a single-stranded DNA oligonucleotide of complementary sequence to a region of the genome with an electron dense tag might allow an asymmetric signal to be introduced into a virion preparation; whether sufficiently high signal-to-noise ratios for the label can be achieved in the images would have to be assessed, with the approach being heavily reliant on an ordered genome in the first place.

With respect to divergence of the geminivirus coat protein, sequence analysis after multiple sequence alignment of geminivirus CPs (against that of STNV) showed that the geminivirus CPs have diverged in important ways, but also appear to conserve features such as the putative calcium ion-binding site identified in this study. With respect to divergence, it is clear that all geminiviruses encode large N-terminal arms which all include nuclear localization signals. These regions of the CPs are all densely populated with positively charged residues, and are presumed to interact with the phosphate backbone of the viral genome. Additionally, the sequence analysis

level, quasi-equivalence describes requirements for conformational switching that allow for the formation of unique structural and chemical environments within an icosahedral assembly consisting of more than 60 subunits – resulting in the construction of asymmetric units consisting of more than just one CP subunit. We must expect that this requirement of a set of unique quasi-equivalent conformers also exists for the geminiviruses, as in no other way could such an assembly form from coat protein subunits of identical amino acid sequence. Specifically, we must ask what is the number of CP conformers that are required to construct the bipartite (or higher-order) virion? Given that the interacting loops of the CP_5 are likely to be flexible, it is probable that a number (probably three or four) of different CP conformers contribute to formation the structure in a similar way that multiple conformations contribute to quasi-icosahedral architectures (e.g. those of $T \geq 3$ viruses). Based on analysis of the EM maps we propose that three conformers contribute to the bipartite assembly. Consequently, we can expect that 60 A conformers (i.e. 12 CP_5 units) self-assemble to form the icosahedral particle. While for the geminate capsid, the A conformer occurs a total of 90 times to form heads of the geminate assembly by assembling to form 12 locally symmetric CP_5s, and 10 locally asymmetric CP_5s at the equatorial regions. To form an equator, 10 sets of CPs (5 sets of 2 on each side) adopt both the B and C conformers (i.e. 20 CPs in total) to mediate interactions between independent quasi-icosahedral $T = 1$ shells. It must be stressed that these conclusions were informed by the experiments described in this thesis, but are nevertheless to be considered hypothetical and will require validation by high-resolution structure determination. Future studies will not only discern whether there are truly three or four conformers, but will also describe the precise conformational differences, and structural roles of each conformer.

With regards to assembly of the particle, the observations made during this study, as well as those reported in the literature, strongly suggest that it is the genome, and in particular the size of the genome, that governs the conformational switching required to bring about formation of the geminate morphology and equatorial CP:CP interactions. So versatile is this switching, that properties (yet to be determined) of the genome, allow multiple switching events, to occur in a single quasi-icosahedral head. More specifically, the tripartite and higher species observed by EM (refer to Figure 3-2) show that the B/C conformers can exist to bridge multiple heads at up to two separate pentameric sites in a single quasi-icosahedral unit. While these observations are intriguing, the clear implication is that the MSV CP alone does not encode the information required to form the geminate (or higher) assembly, and that if the genome of MSV was smaller it would always self-assemble to form icosahedral shells (as is the case for sub-genomic fragments). In consideration of these observations, this study has informed a model of geminate assembly that relies on formation of incomplete and independently assembled $T = 1$ lattices, and

set of CP_5:CP_5 interactions at the 2-fold and 3-fold CP interfaces in comparison to STNV. The experimental reconstructions determined during this study suggest that these interactions are largely "discrete", occurring between specific amino acid segments (referred to as β-loops) between the βC:βD, βG:βH and βE:βF β-strands – and that where necessary, sequence insertions have occurred to mediate the formation and stabilization of these interfaces required for assembly formation. Differences in capsomer surface appearance (i.e. a pronounced crown on the MSV CP_5 unit) were noted to be contributed by insertions between βF:βG and βD:βE strands. These insertions on the surface of the capsid are thought to mediate vector interactions, and as such, were not the focus of this study, and consequently not modeled.

As described above, while the core of the β-barrel domain is apparently well conserved, the β-loops adopt completely different conformations in the context of the MSV architecture – and this is particularly relevant at the equator of the particle. In review of interactions involved in construction of the bipartite and higher order capsid species (i.e. tripartite capsids): geometric analysis of the reconstructions revealed the offset between the interacting incomplete quasi-icosahedral heads to be ~ -11°, and analysis suggested that this geometry arose in order to preserve CP_5:CP_5 interacting regions, implying that interactions between heads appeared to 'mimic' interactions required to form each head. However, detailed analysis suggested that separate sets of interacting residues had evolved in these regions, suggesting that discrete sets of interactions are involved in "head-forming" and "head-interacting" associations. The pseudo-atomic modelling, allowed a different set of interacting β-loops to be identified as those responsible for mediating the primary-equatorial interactions between heads, and it appeared to be the properties of these interfaces that directly caused the offset between heads to arise.

An overall outcome of this study was the advancement of our understanding of geminate structure and architecture. This study has proposed interfaces based on the 2-fold and 3-fold interfaces that stabilize the head, and makes strides towards a pseudo-atomic view of the virion consisting of a defined number of CP conformers which interact with one another with highly specific and distinct interfaces. The role of CP conformers proposed to contribute to the geminivirus capsids are now reviewed in consideration of the principles set out by the quasi-equivalence theory of Casper and Klug (1962). From a geometrical perspective, quasi-equivalence pertains directly to the placement of pentavalent assembly units in hexavalent lattices, and thus describes how a planar lattice can be folded to create a closed spherical lattice. Structural investigations such as this one, have shown that the geminiviruses do not exhibit hexavalent assembly intermediates, and as a consequence cannot form closed spheres of a T number greater than 1 ($T = 1$). For these reasons the theory of quasi-equivalence does not pertain to them directly. However, at a molecular

projection images) as well as the imaging parameters. For the cryo-EM reconstruction (23 Å resolution), virions were observed in ice of suitable thickness (~ 100 nm), however, the presence of the thin-carbon support necessitated far-from-focus imaging in order to produce sufficient image contrast; this unfortunately severely attenuated amplitudes of high-resolution structure factors in the few thousand images which were recorded. Ideally, a dataset consisting of 10,000 – 30,000 images should have been recorded but this was not possible in the time frame available given the difficulties with specimen preparation which entailed many attempts at optimization. In the case of the negative-stain reconstruction (26 Å), ~10,000 images were available, however, these were recorded at high defocus values and were hence limited by the first zero of the microscope CTF. Despite these limitations, useful reconstructions were nevertheless calculated, and all showed clear pentameric capsomeres (CP_5) that could be probed by docking at low resolution. The experimental EM densities were therefore used as constraints that allowed parameters describing the geometric construction of the capsid to be determined. The outcome of this process were "pseudo-atomic" capsids models constructed using an atomic model of the CP_5 produced by homology modelling. That docking into each EM map (each produced from a different virion preparation (plant infection), and prepared for imaging using different specimen preparation technique (cryo-EM or negative-staining) produced very close estimates of the geminate geometric parameters is an important result as it improves our confidence in the final result, as each model correlated with regard to probable CP_5:CP_5 contacts. Additionally, by comparison of the structures of both icosahedral and geminate particles, these provided complementary information about the nature of the different structural and chemical environments within the different capsid species; which allowed fundamental insight with regard to the capsid construction from quasi-equivalent conformers to be gleaned. These are reviewed with regard to virion assembly in the following paragraphs.

In review of the principle interactions involved in construction of the icosahedral and quasi-icosahedral geminate heads: the intermediate resolution of the icosahedral (532) reconstruction, as well as lower resolution geminate maps strongly supported the assignment of the β-barrel fold and pentameric capsomer (CP_5) of STNV to MSV. This was confirmed by careful analysis of the fit of the CP_5 capsomer at the resolution available in the maps. The EM maps confirmed that regions intervening β-strands are highly divergent resulting in pronounced overall morphological differences between the two virus capsids. The capsids of STNV and MSV have differing radii, as well as a very pronounced differences in appearance. In the case of MSV, β-loops mediating CP_5 association must be in an extended conformation, and associate in such a way that the capsomer is rotated 10° around each 5-fold (and local 5-fold) axis of symmetry of each $T = 1$ lattice of CP_5 units. Specifically, the difference in geometry appears to arise owing to a different

4 Summary and conclusions

Structural molecular studies are primarily limited by the availability of sufficient quantities of material for study, as well as the whether these can be prepared for the experiment. X-ray crystallography of mammalian viruses was hampered by the requirement of large quantities of virus for crystallization (e.g. grams). As a result, initial progress in the field was made with (+) ssRNA plant viruses (e.g. TMV and TBSV) which could be harvested from plant material at high yields. This study of MSV used cryo-EM because sufficient material (e.g. 1 mg) could be harvested from a reasonable amount of plant matter (1 kg), and in an achievable amount of time. An attempt to crystallize the virus was inappropriate given the time frame available, and also that there was no guarantee that the virus can be crystallized or that sufficient material for crystallization attempts could be produced.

Even in the absence of the crystallization requirement, the study encountered significant difficulties during specimen preparation. MSV virions are prone to aggregation at high concentrations. This is a phenomenon that appears to reflect their aggregative behaviour previously observed within the plant cell nucleus (Hatta and Francki, 1979), which has not been explained, but is likely determined by the electrostatic surface potential of the capsid. These properties made typical cryo-EM specimen preparation impossible, as within the timeframe of blotting to produce a thin-film, and plunge-freezing, 99% of virions had clumped together. The only means of overcoming this behaviour was to prepare the virions in the presence of a planar thin-carbon support (5 nm thick) that had been glow-discharged to induce electrostatic and hydrophilic properties. While able to adsorb the virions, and hence sequester them from any potential aggregation, the support served to limit contrast in the images necessitating high-defocuses in order to produce sufficient contrast. Within the time-frame allowed, cryo-EM specimen preparation could not be optimized further, e.g. by attempting to prepare grids with thinner ice/carbon, or merely by imaging closer to focus (higher resolution, reduced contrast), and attempting to work with the images despite extremely low contrast. However, an important result of this study, and informative for future work, is the confirmation that MSV structure can only be probed by cryo-EM successfully when a thin-carbon overlay is used.

Despite these difficulties, this study resulted in three single-particle EM reconstructions that have informed our understanding of MSV structure and architecture. The resolution of the icosahedral monopartite reconstruction was (by inspection) better than 20 Å, and calculated from just a few hundred projections recorded at close-to-focus by application of icosahedral symmetry. The resolution of the geminate structures was limited by the availability of data (i.e. the number of

3.5 Conclusions

Single-particle electron microscopy in combination with structural bioinformatics methods was successfully applied to produce experimentally informed pseudo-atomic models of two different capsid species (icosahedral and geminate) of MSV. Detailed analysis of each of the models has improved our understanding of protein-protein, and protein-nucleic acid interactions required to form and stabilize the MSV virion, as well as provided insight allowing a model for virion assembly to be formulated. This study will hopefully inform the design of experiments which may expand our understanding of specific aspects of the structure and function of this class of viruses, and also provides motivation for future high-resolution structural investigation using X-ray crystallography and cryo-EM. In the next chapter the findings of this study are summarized, and in light of possible evolutionary forces that underpin the structural properties of the geminiviruses, speculations regarding various properties of the viruses are developed with regard to formulation of testable hypotheses for future experiments.

a separate set of interacting residues (distinct interfaces) would imply that global factors (incidentally determined by genome structure) determine conformational switching by CP subunits, and that it is not specific association of nucleic acid motifs that induce conformational switches in the CPs, causing them to adopt either of three conformations. In the case of the *Nodaviridae*, association of the genome with CP dimers directly determines the curvature of this unit of assembly (a 'local' effect directly dictating structural properties of the shell). To reiterate, for the geminiviruses the influence of the genome would appear to be far more indirect, although this remains a hypothesis at this time.

The resolution of the 532 map in particular, revealed features that support the notion that the calcium-ion binding site of the CP_5 is structurally conserved between MSV and STNV. This is an important result because it reveals that divalent ions (e.g. Mg^{2+} and Ca^{2+}) probably serve roles in regulating the stability of the MSV capsid, and that this is likely to be a conserved feature of the geminiviruses. Given the observation that the MSV geminate capsid is apparently extremely stable, it is important that we consider mechanisms by which this stability is contextually regulated to allow for genome egress or uncoating at the appropriate time and place. Because the modelling in this study suggests that the N-terminal arms are buried deeply within the particle it is reasonable to assert that for the viral genome to enter the cell nucleus it is likely that some degree of disassembly must occur, and in the low calcium environment of the host cell cytoplasm it is likely that disassembly is triggered by removal of calcium ions when virions enter the host cells cytoplasm, as has been hypothesize for many plant viruses (Durham, 1977). Dissociation of just one CP_5 will result in externalization of N-terminal arms, thereby initiating nuclear trafficking as the NLS becomes exposed to host cell factors. Alternatively, loss of calcium ions may cause the virion to dissociate to become more of a "diffuse" assembly. Such an assembly would be held together by many DNA-protein interactions. Some of these contacts may be disrupted, but many could also be maintained, particularly those of the N-terminal arms. Another alternative is that dissociation of pentamers renders a protease cleavage site immediately downstream of the NLS available for processing. Multiple cleavage events would effectively 'strip' the genome of the CP_5 units leaving the NLS associated with regions of the genome. These would remain associated with the genome, and by interaction with host factors, facilitate nuclear entry. Indeed, this may be the reason these functional sites of the CP overlap in sequence. The precise mechanisms and cellular location where the capsid becomes destabilised, and how this relates to presentation of the N-terminal nuclear-localization sequences will, however, require further investigation.

assembly for STNV (Ban *et al.*, 1995). The results from this study show that this interface shows high structural homology, indicating that the CP_5 may be the unit of assembly for MSV as well. In support of this assertion, disassembly studies of ACMV have shown that CP_5 appears to dissociate from the capsid as a unit (Kittelmann and Jeske, 2008). Expression of the coat protein late in the viral infectious cycle must result in accumulation of these inside the nucleus of the host cell, where they become available to sequester nascent or newly synthesized viral genomes. Whether these accumulate as CP_5 units, has not been determined but given that the pentameric assembly is expected to be the most energetically favourable state, a CP_5 unit of assembly is assumed in this section.

The viral assembly initiation complex (nucleating further addition of structural units by protein-protein interactions) may be protein or nucleic acid, depending on the mechanism by which nucleic acid is packaged. For MSV, that empty virions are not observed, suggests that the nucleation complex must involve the ssDNA genome. Whether this involves binding to a specific sequence (e.g. the stem-loop origin of replication) requires further investigation, but certainly previous work has shown that the CP will bind either ds- or ssDNA indiscriminately. How then is selective packaging of nucleic acid achieved? Two options exist, first that the CP_5 shows high affinity to a specific sequence/structural motif of the genome (as already mentioned), or that the CP_5 has affinity for the viral Rep, and is localized to the site of viral replication by interaction with these complexes. Indeed, such an interaction between Rep and CP has been demonstrated for *Mung bean yellow mosaic India virus* (Malik *et al.*, 2005). With this co-localization established, it is possible that the long N-terminal arms of the CP extend outwards, and subsequently interact with the nascent ssDNA effectively ensuring that this is sequestered from conversion to a double-stranded form as it is synthesized. This Rep:CP interaction was originally postulated to possibly regulate viral replication (Malik *et al.*, 2005), but in the model described here, the replicative process (occurring on the double-stranded viral minichromosome) is unaltered, while the fate of the newly synthesized ssDNA genome is dramatically altered by the recruitment of CP_5 capsomers. This role of the N-terminal arms is experimentally supported by a study carried out on ACMV, which showed that N-terminal truncation CP mutants which removed parts of the N-terminal NLS (overlapping with known DNA binding regions) prevented functional assembly to form geminate virions (Unseld *et al.*, 2004). This is an important result, highlighting that CP_5:CP_5 interactions do not readily form on their own, but require a common 'anchor', i.e. the encapsidated genome, that serves to raise of 'local concentration' of these components sufficiently that stable associations are able to form. It worth considering that because CP_5:CP_5 interactions are mediated by such flexible regions (as revealed by this study) that it is just this property of the CPs that prevents empty capsid formation, because inter-capsomeric

interfaces are largely undefined until these regions of the CP_5 are brought into close proximity by the N-terminal 'anchors', and also by association with ssDNA at the electropositive inward facing surfaces of each CP_5. In other words, this model asserts that it is the N-terminal arms binding to the ssDNA that occurs as a first step, with subsequent orientation of the CP units onto the DNA by virtue of the positively charged underside of the CP_5 itself. Once in close proximity a closed network of many weak interactions is able to take place at inter-capsomeric interfaces, which in addition to the protein-nucleic acid interactions, results in formation of a robust particle.

It has been noted previously, that natural selection is an important determinant in which viral recombinants are observed empirically (Martin *et al.*, 2011). It is quite possible that the packaging events described in the previous section, as well as the influence of genomic structure (secondary and tertiary) on the outcome of these, contribute to yet another mechanism of selection. This asserts that folding of the single-stranded genome is required for the nucleic acid to be compatible with geminate CP packaging. Given the unusual shape of the infectious virion (i.e. geminate) such a proposition is perhaps not unreasonable, and experimental evidence that genome folding is highly relevant for these viruses has been reported previously (Shepherd *et al.*, 2006). For the mastreviruses, as well as many other geminiviruses, this is particularly relevant because CP oligomerization and assembly are required for systemic viral movement (Pooma *et al.*, 1996). Formation of stem-loop secondary structure is clearly relevant at the origin of replication for gemini-viral replication, but such features may also be important for viral assembly, as for STNV (Ford *et al.*, 2013). Such stem loops would not need to be as conserved at a sequence level, and such features would only exist in the genome if a stem-loop CP_5 interaction takes place. This study provides no clear evidence to suggest that there are such interactions, however future studies should probe this question further to establish whether structural homology between MSV and STNV extends to their encapsidated genomes as well.

To further elaborate on the possible structural roles of the viral genome, it is also possible that the tertiary structure of the genome also has a large role to play in formation of virions. That 'important' secondary structures exists (Martin *et al.*, 2011; Shepherd *et al.*, 2006), suggests that the ssDNA genome may fold to adopt 'tertiary modules', of a predetermined size to fit within the $T = 1$ quasi-icosahedral heads; i.e. the modules have evolved somewhat independently to be compatible with $T = 1$ packaging. If this is the case, recombinant variants disrupting the secondary-structure, and hence tertiary structure formation could be selected out by unsuccessful packaging, at least partially explaining the observations reported by Martin and co-workers (Martin *et al.*, 2011).

Describing a conception for the sequential addition of CP_5s to the nascent genome, it seems likely that as CP_5 units are recruited during replication they would be in position to begin interacting with each other, and in so doing begin the construction of *T = 1* lattice shells. Depending on the tertiary structure of ssDNA genome, and its compactness, CP_5 units would be excluded at certain positions on the lattice by the genome itself, creating a location where two *independently* assembled *T = 1* shells (onto the same genome) could be brought into contact – leading to equatorial or "head-interacting" associations. It is important to stress that *T = 1* lattice forming (or head-forming) interactions are separate from 'head-interacting' interactions, with 'heading-forming' interactions thought to take priority during initial stages of assembly. The fact that the unit of assembly is pentameric is fundamental to this behaviour, as it ensures that *T = 1* lattices are formed instead of *T = 3* or higher lattices. The formation of independent *T = 1* lattices also provides some explanation for how the two separate interface types (intra-head and inter-head) could have evolved. The ancestral CP common to both geminiviruses and STNV would most likely been capable of forming only *T = 1* organizations, and would have been continued to function this way (at the time that it was recruited by the ancestral geminivirus), coating single-stranded genomes to form separate lattice shells which may have been loosely associated. Generation of variants by CP sequence divergence would have resulted in capsids that differed in how CP_5 units interacted, each with consequently different environmental finesses, leading to the eventual emergence of the 'true' geminate architecture we see today owing to natural selection of the most environmentally appropriate packing solution. The observation that multipartite capsid species exist, packaging concatamers of the genome, is consistent with this hypothesis, as the contacamers would likely preserve the 'tertiary-folded modules' required for packing and would still allow meeting of multiple independent *T = 1* lattices at the 'inter-head' interfaces. It is worth noting that these assembly requirements (CP and genome structure, as well as interactivity), must have evolved interdependently as a functional whole to result in these behaviours. Hence while each viral component is a module with some evolutionary independence, the infectious viral strain itself is essentially a 'product' of these, showing properties that are more than merely the sum of each of these modules.

This study provided evidenced that allows us to discern between two different hypotheses regarding, the conformational switching required at the level of each CP subunit. The "pentamer swivel" model predicted changes in the orientation of the CP_5 to facilitate formation of the geminate particle. However, what we observe instead is that all interactions between CP_5s occur using distinct interfaces that are apparently maintained during assembly; i.e. this study suggests that no structural transition at a global level is required to facilitate formation of the geminate particle. The association of independently assembled shells to form the equatorial regions using

a separate set of interacting residues (distinct interfaces) would imply that global factors (incidentally determined by genome structure) determine conformational switching by CP subunits, and that it is not specific association of nucleic acid motifs that induce conformational switches in the CPs, causing them to adopt either of three conformations. In the case of the *Nodaviridae*, association of the genome with CP dimers directly determines the curvature of this unit of assembly (a 'local' effect directly dictating structural properties of the shell). To reiterate, for the geminiviruses the influence of the genome would appear to be far more indirect, although this remains a hypothesis at this time.

The resolution of the 532 map in particular, revealed features that support the notion that the calcium-ion binding site of the CP_5 is structurally conserved between MSV and STNV. This is an important result because it reveals that divalent ions (e.g. Mg^{2+} and Ca^{2+}) probably serve roles in regulating the stability of the MSV capsid, and that this is likely to be a conserved feature of the geminiviruses. Given the observation that the MSV geminate capsid is apparently extremely stable, it is important that we consider mechanisms by which this stability is contextually regulated to allow for genome egress or uncoating at the appropriate time and place. Because the modelling in this study suggests that the N-terminal arms are buried deeply within the particle it is reasonable to assert that for the viral genome to enter the cell nucleus it is likely that some degree of disassembly must occur, and in the low calcium environment of the host cell cytoplasm it is likely that disassembly is triggered by removal of calcium ions when virions enter the host cells cytoplasm, as has been hypothesize for many plant viruses (Durham, 1977). Dissociation of just one CP_5 will result in externalization of N-terminal arms, thereby initiating nuclear trafficking as the NLS becomes exposed to host cell factors. Alternatively, loss of calcium ions may cause the virion to dissociate to become more of a "diffuse" assembly. Such an assembly would be held together by many DNA-protein interactions. Some of these contacts may be disrupted, but many could also be maintained, particularly those of the N-terminal arms. Another alternative is that dissociation of pentamers renders a protease cleavage site immediately downstream of the NLS available for processing. Multiple cleavage events would effectively 'strip' the genome of the CP_5 units leaving the NLS associated with regions of the genome. These would remain associated with the genome, and by interaction with host factors, facilitate nuclear entry. Indeed, this may be the reason these functional sites of the CP overlap in sequence. The precise mechanisms and cellular location where the capsid becomes destabilised, and how this relates to presentation of the N-terminal nuclear-localization sequences will, however, require further investigation.

3.5 Conclusions

Single-particle electron microscopy in combination with structural bioinformatics methods was successfully applied to produce experimentally informed pseudo-atomic models of two different capsid species (icosahedral and geminate) of MSV. Detailed analysis of each of the models has improved our understanding of protein-protein, and protein-nucleic acid interactions required to form and stabilize the MSV virion, as well as provided insight allowing a model for virion assembly to be formulated. This study will hopefully inform the design of experiments which may expand our understanding of specific aspects of the structure and function of this class of viruses, and also provides motivation for future high-resolution structural investigation using X-ray crystallography and cryo-EM. In the next chapter the findings of this study are summarized, and in light of possible evolutionary forces that underpin the structural properties of the geminiviruses, speculations regarding various properties of the viruses are developed with regard to formulation of testable hypotheses for future experiments.

4 Summary and conclusions

Structural molecular studies are primarily limited by the availability of sufficient quantities of material for study, as well as the whether these can be prepared for the experiment. X-ray crystallography of mammalian viruses was hampered by the requirement of large quantities of virus for crystallization (e.g. grams). As a result, initial progress in the field was made with (+) ssRNA plant viruses (e.g. TMV and TBSV) which could be harvested from plant material at high yields. This study of MSV used cryo-EM because sufficient material (e.g. 1 mg) could be harvested from a reasonable amount of plant matter (1 kg), and in an achievable amount of time. An attempt to crystallize the virus was inappropriate given the time frame available, and also that there was no guarantee that the virus can be crystallized or that sufficient material for crystallization attempts could be produced.

Even in the absence of the crystallization requirement, the study encountered significant difficulties during specimen preparation. MSV virions are prone to aggregation at high concentrations. This is a phenomenon that appears to reflect their aggregative behaviour previously observed within the plant cell nucleus (Hatta and Francki, 1979), which has not been explained, but is likely determined by the electrostatic surface potential of the capsid. These properties made typical cryo-EM specimen preparation impossible, as within the timeframe of blotting to produce a thin-film, and plunge-freezing, 99% of virions had clumped together. The only means of overcoming this behaviour was to prepare the virions in the presence of a planar thin-carbon support (5 nm thick) that had been glow-discharged to induce electrostatic and hydrophilic properties. While able to adsorb the virions, and hence sequester them from any potential aggregation, the support served to limit contrast in the images necessitating high-defocuses in order to produce sufficient contrast. Within the time-frame allowed, cryo-EM specimen preparation could not be optimized further, e.g. by attempting to prepare grids with thinner ice/carbon, or merely by imaging closer to focus (higher resolution, reduced contrast), and attempting to work with the images despite extremely low contrast. However, an important result of this study, and informative for future work, is the confirmation that MSV structure can only be probed by cryo-EM successfully when a thin-carbon overlay is used.

Despite these difficulties, this study resulted in three single-particle EM reconstructions that have informed our understanding of MSV structure and architecture. The resolution of the icosahedral monopartite reconstruction was (by inspection) better than 20 Å, and calculated from just a few hundred projections recorded at close-to-focus by application of icosahedral symmetry. The resolution of the geminate structures was limited by the availability of data (i.e. the number of

projection images) as well as the imaging parameters. For the cryo-EM reconstruction (23 Å resolution), virions were observed in ice of suitable thickness (~ 100 nm), however, the presence of the thin-carbon support necessitated far-from-focus imaging in order to produce sufficient image contrast; this unfortunately severely attenuated amplitudes of high-resolution structure factors in the few thousand images which were recorded. Ideally, a dataset consisting of 10,000 – 30,000 images should have been recorded but this was not possible in the time frame available given the difficulties with specimen preparation which entailed many attempts at optimization. In the case of the negative-stain reconstruction (26 Å), ~10,000 images were available, however, these were recorded at high defocus values and were hence limited by the first zero of the microscope CTF. Despite these limitations, useful reconstructions were nevertheless calculated, and all showed clear pentameric capsomeres (CP_5) that could be probed by docking at low resolution. The experimental EM densities were therefore used as constraints that allowed parameters describing the geometric construction of the capsid to be determined. The outcome of this process were "pseudo-atomic" capsids models constructed using an atomic model of the CP_5 produced by homology modelling. That docking into each EM map (each produced from a different virion preparation (plant infection), and prepared for imaging using different specimen preparation technique (cryo-EM or negative-staining) produced very close estimates of the geminate geometric parameters is an important result as it improves our confidence in the final result, as each model correlated with regard to probable CP_5:CP_5 contacts. Additionally, by comparison of the structures of both icosahedral and geminate particles, these provided complementary information about the nature of the different structural and chemical environments within the different capsid species; which allowed fundamental insight with regard to the capsid construction from quasi-equivalent conformers to be gleaned. These are reviewed with regard to virion assembly in the following paragraphs.

In review of the principle interactions involved in construction of the icosahedral and quasi-icosahedral geminate heads: the intermediate resolution of the icosahedral (532) reconstruction, as well as lower resolution geminate maps strongly supported the assignment of the β-barrel fold and pentameric capsomer (CP_5) of STNV to MSV. This was confirmed by careful analysis of the fit of the CP_5 capsomer at the resolution available in the maps. The EM maps confirmed that regions intervening β-strands are highly divergent resulting in pronounced overall morphological differences between the two virus capsids. The capsids of STNV and MSV have differing radii, as well as a very pronounced differences in appearance. In the case of MSV, β-loops mediating CP_5 association must be in an extended conformation, and associate in such a way that the capsomer is rotated 10° around each 5-fold (and local 5-fold) axis of symmetry of each $T = 1$ lattice of CP_5 units. Specifically, the difference in geometry appears to arise owing to a different

set of CP_5:CP_5 interactions at the 2-fold and 3-fold CP interfaces in comparison to STNV. The experimental reconstructions determined during this study suggest that these interactions are largely "discrete", occurring between specific amino acid segments (referred to as β-loops) between the βC:βD, βG:βH and βE:βF β-strands – and that where necessary, sequence insertions have occurred to mediate the formation and stabilization of these interfaces required for assembly formation. Differences in capsomer surface appearance (i.e. a pronounced crown on the MSV CP_5 unit) were noted to be contributed by insertions between βF:βG and βD:βE strands. These insertions on the surface of the capsid are thought to mediate vector interactions, and as such, were not the focus of this study, and consequently not modeled.

As described above, while the core of the β-barrel domain is apparently well conserved, the β-loops adopt completely different conformations in the context of the MSV architecture – and this is particularly relevant at the equator of the particle. In review of interactions involved in construction of the bipartite and higher order capsid species (i.e. tripartite capsids): geometric analysis of the reconstructions revealed the offset between the interacting incomplete quasi-icosahedral heads to be ~ -11°, and analysis suggested that this geometry arose in order to preserve CP_5:CP_5 interacting regions, implying that interactions between heads appeared to 'mimic' interactions required to form each head. However, detailed analysis suggested that separate sets of interacting residues had evolved in these regions, suggesting that discrete sets of interactions are involved in "head-forming" and "head-interacting" associations. The pseudo-atomic modelling, allowed a different set of interacting β-loops to be identified as those responsible for mediating the primary-equatorial interactions between heads, and it appeared to be the properties of these interfaces that directly caused the offset between heads to arise.

An overall outcome of this study was the advancement of our understanding of geminate structure and architecture. This study has proposed interfaces based on the 2-fold and 3-fold interfaces that stabilize the head, and makes strides towards a pseudo-atomic view of the virion consisting of a defined number of CP conformers which interact with one another with highly specific and distinct interfaces. The role of CP conformers proposed to contribute to the geminivirus capsids are now reviewed in consideration of the principles set out by the quasi-equivalence theory of Casper and Klug (1962). From a geometrical perspective, quasi-equivalence pertains directly to the placement of pentavalent assembly units in hexavalent lattices, and thus describes how a planar lattice can be folded to create a closed spherical lattice. Structural investigations such as this one, have shown that the geminiviruses do not exhibit hexavalent assembly intermediates, and as a consequence cannot form closed spheres of a T number greater than 1 ($T = 1$). For these reasons the theory of quasi-equivalence does not pertain to them directly. However, at a molecular

well as a high-resolution EM study to finally reveal at near-atomic resolution the structural basis for formation of the geminate particle. A crystallography study of the icosahedral form could potentially be carried out to atomic resolution (~1.2 Å), revealing arrangements of loops, as well as DNA-protein binding motifs.

Alternatively, preparations containing both monopartite and bipartite capsids could be imaged using a direct electron detector (DED). The DEDs represent a new generation of detectors that vastly improve image quality by reducing detector contributed noise during imaging, and furthermore allow "exposure movies" to be recorded which allow beam-induced motion to be quantified and corrected. The implication of these detectors is that doses used for imaging can be reduced to achieve same quality as previous detectors, or maintained to achieve even greater SNR. Recent studies of the ribosome and proteasome have resulted in determination of these asymmetric structures to 4.5 and 3.3 Å resolution; a resolution previously only available for large and highly symmetric objects such as rotavirus and 2D crystals. For the first time, these macromolecular assemblies (typically very difficult to crystallize) are seen at a resolution where atomic positions can be modelled with high accuracy for a range of conformations that occur outside of crystallographic lattice. Clearly, studies utilizing any degree of symmetric (for instance the D6 object determined to 3.3 Å) benefit in that even fewer images need to be recorded than ever before suggesting that data-collection times are reduced, or that extended data-collection times should yield sufficient information to confidently determine many structural intermediates.

To summarise finally, the MSV CP has been described as a multifunctional protein, carrying out important roles in the viral infectious cycle including: ssDNA accumulation, protection of these genomes from nuclease activity, cell-to-cell movement, vector-mediated transmission and nuclear import of the viral genome. Many of these functions rely on the ability of the CP to self-assemble, functionality that is in turn related to interactions with ssDNA molecules. These interactions (both protein-protein and protein-nucleic acid) can be considered a small fraction of the entire intra-genome interaction network required for persistence of this pathogen in the environment. Efforts to characterize the viral interaction network are important, because if quantified at the level of amino acid and nucleotide interaction, could be considered part of a model of all 'information' encoded by the virus itself, and would thus provide a foundation from which anti-viral strategies could be developed.

This study makes some progress in showing that discrete regions of the folded CP protein carry out the functions of the MSV CP. Within the context of the functioning CP subunit, the core β-barrel domain clearly functions to arrange functional segments and sites in three-dimensional space so that they are able to fulfil their roles. This study has also revealed that the quaternary

well as a high-resolution EM study to finally reveal at near-atomic resolution the structural basis for formation of the geminate particle. A crystallography study of the icosahedral form could potentially be carried out to atomic resolution (~1.2 Å), revealing arrangements of loops, as well as DNA-protein binding motifs.

Alternatively, preparations containing both monopartite and bipartite capsids could be imaged using a direct electron detector (DED). The DEDs represent a new generation of detectors that vastly improve image quality by reducing detector contributed noise during imaging, and furthermore allow "exposure movies" to be recorded which allow beam-induced motion to be quantified and corrected. The implication of these detectors is that doses used for imaging can be reduced to achieve same quality as previous detectors, or maintained to achieve even greater SNR. Recent studies of the ribosome and proteasome have resulted in determination of these asymmetric structures to 4.5 and 3.3 Å resolution; a resolution previously only available for large and highly symmetric objects such as rotavirus and 2D crystals. For the first time, these macromolecular assemblies (typically very difficult to crystallize) are seen at a resolution where atomic positions can be modelled with high accuracy for a range of conformations that occur outside of crystallographic lattice. Clearly, studies utilizing any degree of symmetric (for instance the D6 object determined to 3.3 Å) benefit in that even fewer images need to be recorded than ever before suggesting that data-collection times are reduced, or that extended data-collection times should yield sufficient information to confidently determine many structural intermediates.

To summarise finally, the MSV CP has been described as a multifunctional protein, carrying out important roles in the viral infectious cycle including: ssDNA accumulation, protection of these genomes from nuclease activity, cell-to-cell movement, vector-mediated transmission and nuclear import of the viral genome. Many of these functions rely on the ability of the CP to self-assemble, functionality that is in turn related to interactions with ssDNA molecules. These interactions (both protein-protein and protein-nucleic acid) can be considered a small fraction of the entire intra-genome interaction network required for persistence of this pathogen in the environment. Efforts to characterize the viral interaction network are important, because if quantified at the level of amino acid and nucleotide interaction, could be considered part of a model of all 'information' encoded by the virus itself, and would thus provide a foundation from which anti-viral strategies could be developed.

This study makes some progress in showing that discrete regions of the folded CP protein carry out the functions of the MSV CP. Within the context of the functioning CP subunit, the core β-barrel domain clearly functions to arrange functional segments and sites in three-dimensional space so that they are able to fulfil their roles. This study has also revealed that the quaternary

interactions of the CP_5 unit have heavily diverged, and that these are mediated by apparent flexible regions that intervene β-strands, and provides evidence to suggest that it is the 'hinge-like' properties of these regions that allows multipartite capsids to form from a series of independently assembled lattice intermediates. These observations have important implications, because while structural virology may have a tendency to describe the structures of viral capsids as precisely defined containers, 'designed' to encapsidate a specific volume of nucleic acid, it seems that in some cases this may simply not be the case. For small viruses with limited coding capacity, the requirement of genetically efficient solutions to the packaging problem appears to motivate the emergence of alternatives. This is especially the case for the geminiviruses, as the CP in these cases appears to have evolved as a nucleic acid "coat" rather than a strictly defined "container". The distinction is perhaps subtle, but important, because in the case of such coat, it is the package being coated that governs the appearance of the final assembly product. A container, on the other hand, would be unaltered by the act of packaging, whereas for geminiviruses such as MSV it is the interplay between the genome and its coat that appears to govern not only morphology of the capsid, but also a mechanism of selectivity; not only of genome size, but also 'shape'. Put simply, by way of folding, not all viral variants may be 'tailored' to the viral packaging coat (CP), and will consequently demonstrate different finesses with respect to mechanisms for viral movement that rely on formation of stable capsids by spontaneous processes of assembly.

The geminiviruses continue to provide a rich source of fundamental questions for future structural and molecular investigation. Clearly a thorough understanding of these viruses will be challenging, but it is likely that these challenges can be overcome with improved technologies, appropriate experiment design, as well as integration of information from multiple techniques and methods of inquiry.

5 References

Abrescia, N., Bamford, D., Grimes, J., Stuart, D., 2012. Structure unifies the viral universe. Annu. Rev. Biochem. 81, 795–822.

Abrescia, N., Grimes, J., Fry, E., Ravantti, J., Bamford, D., Stuart, D., 2010. What does it take to make a virus: The concept of the viral "self", in: Stockley, P.G., Twarock, R. (Eds.), Emerging Topics in Physical Virology. Imperial College Press, pp. 35–58.

Adrian, M., Dubochet, J., Lepault, J., McDowall, A.A.W., 1984. Cryo-electron microscopy of viruses. Nature 308, 32–36.

Bai, X.-C., Fernandez, I.S., McMullan, G., Scheres, S.H., 2013. Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. Elife 2, e00461.

Baker, N.A., Sept, D., Joseph, S., Holst, M.J., McCammon, J.A., 2001. Electrostatics of nanosystems: application to microtubules and the ribosome. Proc. Natl. Acad. Sci. 98, 10037–41.

Baker, T.S., Olson, N.H., Fuller, S.D., 1999. Adding the third dimension to virus life cycles: three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs. Microbiol. Mol. Biol. Rev. 63, 862–922.

Bakker, S.E., Ford, R.J., Barker, A.M., Robottom, J., Saunders, K., Pearson, A.R., Ranson, N.A., Stockley, P.G., 2012. Isolation of an asymmetric RNA uncoating intermediate for a single-stranded RNA plant virus. J. Mol. Biol. 417, 65–78.

Baltimore, D., 1971. Expression of animal virus genomes. Bacteriol. Rev. 35, 235–41.

Bamford, D., 2002. Evolution of viral structure. Theor. Popul. Biol. 61, 461–470.

Bamford, D.H., Gilbert, R.J., Grimes, J.M., Stuart, D.I., 2001. Macromolecular assemblies: greater than their parts. Curr. Opin. Struct. Biol. 11, 107–13.

Bamford, D.H., Grimes, J.M., Stuart, D.I., 2005. What does structure tell us about virus evolution? Curr. Opin. Struct. Biol. 15, 655–63.

Ban, N., Larson, S.B., McPherson, A., 1995. Structural comparison of the plant satellite viruses. Virology 214, 571–583.

Bancroft, J.B., Hiebert, E., 1967. Formation of an infectious nucleoprotein from protein and nucleic acid isolated from a small spherical virus. Virology 32, 354–356.

Basavappa, R., Syed, R., Icenogle, J.P., Filman, D.J., Hogle, J.M., 1994. Role and mechanism of the maturation cleavage of VP0 in poliovirus assembly: Structure of the empty capsid assembly intermediate at 2.9 A resolution. Protein Sci. 3, 1651–1669.

Baumeister, W., Grimm, R., 1999. Electron tomography of molecules and cells. Trends Cell Biol. 9, 81–85.

Bennett, A., McKenna, R., Agbandje-McKenna, M., 2008. A Comparative Analysis of the Structural Architecture of ssDNA Viruses. Comput. Math. Methods Med. 9, 183–196.

Biagini, P., 2004. Human circoviruses. Vet. Microbiol. 98, 95–101.

Bleker, S., Sonntag, F., Kleinschmidt, A., 2005. Mutational Analysis of Narrow Pores at the Fivefold Symmetry Axes of Adeno-Associated Virus Type 2 Capsids Reveals a Dual Role in Genome Packaging and Activation of Phospholipase A2 Activity. J. Virol. 79, 2528–2540.

Bloomer, A.C., Champness, J.N., Bricogne, G., Staden, R., Klug, A., 1978. Protein disk of tobacco mosaic virus at 2.8 A resolution showing the interactions within and between subunits. Nature 276, 362–368.

Bock, K., Guthrie, E., Meredith, G., Bakker, H., 1977. RNA and protein components of maize streak and cassava latent viruses. Ann. Appl. Biol. 85, 305–308.

Bonneau, R., Baker, D., 2001. Ab intio protein structure prediction: progress and progres. Annu. Rev. Biophys. Biomol. Struct. 30, 173–189.

Borodavka, A., Tuma, R., Stockley, P.G., 2012. Evidence that viral RNAs have evolved for efficient, two-stage packaging. Proc. Natl. Acad. Sci. 109, 15769–74.

Bostina, M., Levy, H., Filman, D.J., Hogle, J.M., 2011. Poliovirus RNA is released from the capsid near a twofold symmetry axis. J. Virol. 85, 776–83.

Bottcher, B., Unseld, S., Ceulemans, H., Russell, R.B., Jeske, H., Bo, B., 2004. Geminate structures of African cassava mosaic virus. J. Virol. 78, 6758–6765.

Bottcher, B., Wynne, S., Crowther, R., 1997. Determination of the fold of the core protein hepatitis B virus by electron cryomicroscopy. Nature 386, 88–91.

Boulton, M., Buchholz, W., Marks, M., Markham, P., Davies, J., 1989a. Specificity of Agrobacterium-mediated delivery of maize streak virus DNA to members of the Gramineae. Plant Mol. Biol. 12, 31–40.

Boulton, M.I., 2002. Functions and interactions of mastrevirus gene products. Physiol. Mol. Plant Pathol. 60, 243–255.

Boulton, M.I., Pallaghy, C.K., Chatani, M., MacFarlane, S., Davies, J.W., 1993. Replication of maize streak virus mutants in maize protoplasts: evidence for a movement protein. Virology 192, 85–93.

Boulton, M.I., Steinkellner, H., Donson, J., Markham, P.G., King, D.I., Davies, J.W., 1989b. Mutational analysis of the virion-sense genes of maize streak virus. J. Gen. Virol. 70, 2309–2323.

Brenner, S., Horne, R.W., 1959. A negative staining method for high resolution electron microscopy of viruses. Biochim. Biophys. Acta 34, 103–110.

Briddon, R.W., Markham, P.G., 1995. Geminiviridae, in: Murphy, F.A., Fauguet, C.M., Bishop, D.H.L., Ghabrial, S.A., Jarvis, A.W., Martelli, G.P., Mayo, M.A., Summers, M.D. (Eds.), Virus Taxonomy. Sixth Report of the International Commitee on Taxonomy of Viruses. Wien:Springer-Verlag, pp. 158–165.

Briddon, R.W., Pinner, M.S., Stanley, J., Markham, P.G., 1990. Geminivirus coat protein gene replacement alters insect specificity. Virology 177, 85–94.

Brumfield, S., 2004. Heterologous expression of the modified coat protein of Cowpea chlorotic mottle bromovirus results in the assembly of protein cages with altered architectures and function. J. Gen. Virol. 85, 1049–1053.

Bubeck, D., Filman, D.J., Hogle, J.M., 2005. Cryo-electron microscopy reconstruction of a poliovirus-receptor-membrane complex. Nat. Struct. Mol. Biol. 12, 615–8.

Bunka, D.H.J., Lane, S.W., Lane, C.L., Dykeman, E.C., Ford, R.J., Barker, A.M., Twarock, R., Phillips, S.E. V, Stockley, P.G., 2011. Degenerate RNA packaging signals in the genome of Satellite tobacco necrosis virus: implications for the assembly of a T=1 capsid. J. Mol. Biol. 413, 51–65.

Casado, C.G., Javier Ortiz, G., Padron, E., Bean, S.J., McKenna, R., Agbandje-McKenna, M., Boulton, M.I., 2004. Isolation and characterization of subgenomic DNAs encapsidated in "single" T = 1 isometric particles of Maize streak virus. Virology 323, 164–171.

Caspar, D.L.D., 1956. Structure of bushy stunt virus. Nature 177, 475–476.

Caspar, D.L.D., Klug, A., 1962. Physical principles in the construction of regular viruses. Cold Spring Harb. Symp. Quant. Biol. 27, 1–24.

Castellano, M.M., Sanz-Burgos, A.P., Gutierrez, C., 1999. Initiation of DNA replication in a eukaryotic rolling-circle replicon: identification of multiple DNA-protein complexes at the geminivirus origin. J. Mol. Biol. 290, 639–652.

Chan, K.-Y., Gumbart, J., McGreevy, R., Watermeyer, J., Sewell, B., 2012. Symmetry-restrained flexible fitting for symmetric EM maps. Structure 19, 1211–1218.

Chang, J., Weigele, P., King, J., Chiu, W., Jiang, W., 2006. Cryo-EM asymmetric reconstruction of bacteriophage P22 reveals organization of its DNA packaging and infecting machinery. Structure 14, 1073–82.

Chang, J.T., Schmid, M.F., Haase-pettingell, C., Weigele, P.R., King, J.A., Chiu, W., 2011. Visualizing the structural changes of bacteriophag epsilon15 and its Salmonella host during infection. J. Mol. Biol. 402, 731–740.

Chang, W.-H., Chiu, M.T.-K., Chen, C.-Y., Yen, C.-F., Lin, Y.-C., Weng, Y.-P., Chang, J.-C., Wu, Y.-M., Cheng, H., Fu, J., Tu, I.-P., 2010. Zernike phase plate cryoelectron microscopy facilitates single particle analysis of unstained asymmetric protein complexes. Structure 18, 17–27.

Chapman, M., Rossmann, M., 1993. Structure, sequence, and function correlations among Parvoviruses. Virology 194, 491–508.

Chapman, M.S., Rossmann, M.G., 1995. Single-stranded DNA-protein interactions in canine parvovirus. Structure 3, 151–62.

Chen, J.Z., Grigorieff, N., 2007. SIGNATURE: A single-particle selection system for molecular electron microscopy. J. Struct. Biol. 157, 168–173.

Cheng, R.H., Kuhn, R.J., Olson, N.H., Rossmann, M.G., Choi, H.K., Smith, T.J., Baker, T.S., 1995. Nucleocapsid and glycoprotein organization in an enveloped virus. Cell 80, 621–30.

Cheng, R.H., Reddy, V.S., Olson, N.H., Fisher, a J., Baker, T.S., Johnson, J.E., 1994. Functional implications of quasi-equivalence in a T = 3 icosahedral animal virus established by cryo-electron microscopy and X-ray crystallography. Structure 2, 271–82.

Clamp, M., Cuff, J., Searle, S.M., Barton, G.J., 2004. The Jalview Java alignment editor. Bioinformatics 20, 426–427.

Crick, F.H.C., Watson, J.D., 1956. Structure of small viruses. Nature 177, 473–475.

Crowther, R.A., Amos, L.A., Finch, J.T., De Rosier, D.J., Klug, A., 1970. Three dimensional reconstructions of spherical viruses by fourier synthesis from electron micrographs. Nature 226, 421–425.

Davies, J.W., Stanley, J., Donson, J., Mullineaux, P.M., Boulton, M.I., 1987. Structure and replication of geminivirus genomes. J. Cell Sci. 7, 95–107.

De Carlo, S., Adrian, M., Kälin, P., Mayer, J., Dubochet, J., 1999. Unexpected property of trehalose as observed by cryo-electron microscopy. J. Microsc. 196, 40–45.

De Rosier, D.J., Klug, A., 1968. Reconstruction of three dimensional structures from electron micrographs. Nature 217, 130–134.

Dickinson, V.J., Halder, J., Woolston, C.J., 1996. The product of maize streak virus ORF V1 is associated with secondary plasmodesmata and is first detected with the onset of viral lesions. Virology 220, 51–59.

Dill, K.A., Ozkan, S.B., Weikl, T.R., Chodera, J.D., Voelz, V.A., 2007. The protein folding problem: when will it be solved? Curr. Opin. Struct. Biol. 17, 342–6.

Dokland, T., McKenna, R., Ilag, L.L., Bowman, B.R., Incardona, N.L., Fane, B.A., Rossmann, M.G., 1997. Structure of a viral procapsid with molecular scaffolding. Nature 389, 308–313.

Douglas, T., Young, M., 1998. Host-guest encapsulation of materials by assembled virus protein cages. Nature 393, 152–155.

Dubochet, J., Adrian, M., Chang, J.J., Homo, J.C., Lepault, J., McDowall, a W., Schultz, P., 1988. Cryo-electron microscopy of vitrified specimens. Q. Rev. Biophys. 21, 129–228.

Duda, R.L., Conway, J.F., 2008. Asymmetric EM reveals new twists in phage phi29 biology. Structure 16, 831–2.

Duffy, S., Shackelton, L.A., Holmes, E.C., 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet. 9, 267–276.

Durham, A.C., 1977. Do viruses use calcium ions to shut off host cell functions? Nature 267, 375–376.

Durham, A.C.H., Hendry, D.A., von Wechmar, M.B., 1977. Does calcium ion binding control plant virus disassembly? Virology 533, 524–533.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–7.

Erdmann, J.B., Shepherd, D.N., Martin, D.P., Varsani, A., Rybicki, E.P., Jeske, H., 2010. Replicative intermediates of maize streak virus found during leaf development. J. Gen. Virol. 91, 1077–81.

Erickson, H.P., Klug, A., 1970. The Fourier Transform of an Electron Micrograph: Effects of Defocussing and Aberrations, and Implications for the Use of Underfocus Contrast Enhancement. Berichte der Bunsengesellschaft für Phys. Chemie 74, 1129–1137.

Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.Y., Pieper, U., Sali, A., 2007. Comparative protein structure modeling using MODELLER. Curr. Protoc. Protein Sci. Chapter 2, Unit 2 9.

Faruqi, a R., Henderson, R., 2007. Electronic detectors for electron microscopy. Curr. Opin. Struct. Biol. 17, 549–55.

Fernández, J.J., Li, S., Crowther, R. a, 2006. CTF determination and correction in electron cryotomography. Ultramicroscopy 106, 587–96.

Ford, R.J., Barker, A.M., Bakker, S.E., Coutts, R.H., Ranson, N. a, Phillips, S.E. V, Pearson, A.R., Stockley, P.G., 2013. Sequence-specific, RNA-protein interactions overcome electrostatic barriers preventing assembly of satellite tobacco necrosis virus coat protein. J. Mol. Biol. 425, 1050–64.

Forterre, P., 2010. Defining life: the virus viewpoint. Orig. Life Evol. Biosph. 40, 151–60.

Fraenkel-Conrat, H., Williams, R., 1955. Reconstitution of active tobacco mosaic virus from its inactive protein and nucleic acid components. Proc. Natl. Acad. Sci. 41, 690–698.

Francki, R.I., Hatta, T., Boccardo, G., Randles, J.W., 1980. The composition of Chloris striate mosaic virus, a geminivirus. Virology 101, 233–241.

Frank, J., 2010. The ribosome comes alive. Isr. J. Chem. 50, 95–98.

Frank, J., Radermacher, M., Penczek, P., Zhu, J., Li, Y., Ladjadj, M., Leith, A., 1996. SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. J. Struct. Biol. 116, 190–199.

Franklin, R., 1956. Location of the ribonucleic acid in the Tobacco mosaic virus particle. Nature 177, 929–930.

Frischmuth, T., Ringel, M., Kocher, C., 2001. The size of encapsidated single-stranded DNA determines the multiplicity of African cassava mosaic virus particles. J. Gen. Virol. 82, 673–676.

Fu, C., Johnson, J.E., 2011. Viral life cycles captured in three-dimensions with electron microscopy tomography. Curr. Opin. Virol. 1, 125–133.

Fuller, S.D., Butcher, S.J., Cheng, R.H., Baker, T.S., 1996. Three-dimensional reconstruction of icosahedral particles - the uncommon line. J. Struct. Biol. 116, 48–55.

Gilbertson, R.L., Sudarshana, M., Jiang, H., Rojas, M.R., Lucas, W.J., 2003. Limitations on geminivirus genome size imposed by plasmodesmata and virus-encoded movement protein: insights into DNA trafficking. Plant Cell 15, 2578–2591.

Goddard, T.D., Huang, C.C., Ferrin, T.E., 2007. Visualizing density maps with UCSF Chimera. J. Struct. Biol. 157, 281–287.

Grigorieff, N., 2007. FREALIGN: high-resolution refinement of single particle structures. J. Struct. Biol. 157, 117–125.

Grigorieff, N., Harrison, S.C., 2011. Near-atomic resolution reconstructions of icosahedral viruses from electron cryo-microscopy. Curr. Opin. Struct. Biol. 21, 265–73.

Grimsley, N., Hohn, B., Ramos, C., Kado, C., Rogowsky, P., 1989. DNA transfer from Agrobacterium to Zea mays or Brassica by agroinfection is dependent on bacterial virulence functions. Mol. Gen. Genet. 217, 309–316.

Grunewald, K., Cyrklaff, M., 2006. Structure of complex viruses and virus-infected cells by electron cryo tomography. Curr. Opin. Microbiol. 9, 437–442.

Harauz, G., van Heel, M., 1986. Exact filters for general geometry three-dimensional reconstruction. Optik (Stuttg). 73, 146–156.

Harkins, G.W., Delport, W., Duffy, S., Wood, N., Monjane, A.L., Owor, B.E., Donaldson, L., Saumtally, S., Triton, G., Briddon, R.W., Shepherd, D.N., Rybicki, E.P., Martin, D.P., Varsani, A., 2009. Experimental evidence indicating that mastreviruses probably did not co-diverge with their hosts. Virol. J. 6, 104.

Harrison, S., Olson, A., Schutt, C., Winkler, F., Bricogne, G., 1978. Tomato bushy stunt virus at 2.9 Å resolution. Nature 276, 368–373.

Harrison, S., Skehel, J., Wiley, D., 1996. Virus Structure, in: Fields, B., Knipe, D., Howley, P. (Eds.), Fields Virology. Lippincott - Raven Publishers, Philadelphia.

Harrison, S.C., 1980. Protein interfaces and intersubunit bonding. The case of tomato bushy stunt virus. Biophys. J. 32, 139–53.

Hatta, T., Francki, R.I.B., 1979. The fine structure of chloris striate mosaic virus. Virology 435, 428–435.

Henderson, R., 1995. The potential and limitations of neutrons , electrons and X-rays for atomic resolution microscopy of unstained biological molecules. Q. Rev. Biophys. 2, 171–193.

Henderson, R., Sali, A., Baker, M., Carragher, B., Devkota, B., Downing, K., Egelman, E., Feng, Z., Frank, J., Grigorieff, N., Jiang, W., Ludtke, S., Medalia, O., Penczek, P., Rosenthal, P., Rossmann, M., Schmid, M., Schroder, G., Steven, A., Stokes, D., Westbrook, J., Wriggers, W., Yang, H., Young, J., Berman, H., Chiu, W., Kleywegt, G., Lawson, C., 2012. Outcome of the first electron microscopy validation task force meeting. Structure 20, 205–214.

Hendrickson, W.A., Horton, J.R., LeMaster, D.M., 1990. Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. EMBO J. 9, 1665–72.

Hogle, J.M., Chow, M., Filman, D.J., 1985. Three-dimensional structure of poliovirus at 2.9 A resolution. Science (80-.). 229, 1358–1365.

Huang, Z., Baldwin, P.R., Mullapudi, S., Penczek, P. a, 2003. Automated determination of parameters describing power spectra of micrograph images in electron microscopy. J. Struct. Biol. 144, 79–94.

Huiskonen, J.T., Overby, A.K., Weber, F., Grünewald, K., 2009. Electron cryo-microscopy and single-particle averaging of Rift Valley fever virus: evidence for GN-GC glycoprotein heterodimers. J. Virol. 83, 3762–9.

Humphrey, W., Dalke, A., Schulten, K., 1996. VMD: visual molecular dynamics. J. Mol. Graph. 14, 33–8, 27–8.

Jeske, H., Lütgemeier, M., Preiss, W., 2001. DNA forms indicate rolling circle and recombination-dependent replication of Abutilon mosaic virus. EMBO J. 20, 6158–67.

Jiang, W., Chiu, W., 2007. Cryoelectron Microscopy of Icosahedral Virus Particles, in: Kuo, J. (Ed.), Electron Microscopy: Methods and Protocols. Humana Press Inc., pp. 345–363.

Johnson, J.E., 1996. Functional implications of protein-protein interactions in icosahedral viruses. Proc. Natl. Acad. Sci. 93, 27–33.

Jones, D.T., 1999a. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J. Mol. Biol. 287, 797–815.

Jones, D.T., 1999b. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195–202.

Jones, T.A., Liljas, L., 1984a. Structure of satellite tobacco necrosis virus after crystallographic refinement at 2.5 Å resolution. J. Mol. Biol. 177, 735–767.

Jones, T.A., Liljas, L., 1984b. Crystallographic Refinement of Macromolecules having Non-crystallographic Symmetry. Acta Crystallogr. Sect. D A40, 50–57.

Jovel, J., Preiss, W., Jeske, H., 2007. Characterization of DNA intermediates of an arising geminivirus. Virus Res. 130, 63–70.

Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., Xu, J., 2012. Template-based protein structure modeling using the RaptorX web server. Nat. Protoc. 7, 1511–1522.

Katoh, K., Frith, M.C., 2012. Adding unaligned sequences into an existing alignment using MAFFT and LAST. Bioinformatics 28, 3144–6.

Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–66.

Kelley, L.A., Sternberg, M.J.E., 2009. Protein structure prediction on the web: a case study using the Phyre server. Nat. Protoc. 4, 363–71.

Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H., Phillips, D.C., 1958. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. Nature 181, 662–666.

Kittelmann, K., Jeske, H., 2008. Disassembly of African cassava mosaic virus. J. Gen. Virol. 89, 2029–2036.

Koonin, E. V, Ilyina, T. V, 1992. Geminivirus replication proteins are related to prokaryotic plasmid rolling circle DNA replication initiator proteins. J. Gen. Virol. 73, 2763–2766.

Kosugi, S., Hasebe, M., Tomita, M., Yanagawa, H., 2009. Systematic identification of cell cycle-dependent yeast nucleocytoplasmic shuttling proteins. Proc. Natl. Acad. Sci. 106, 10171–10176.

Krupovic, M., Bamford, D.H., 2010. Order to the viral universe. J. Virol. 84, 12476–12479.

Krupovic, M., Ravantti, J.J., Bamford, D.H., 2009. Geminiviruses: a tale of a plasmid becoming a virus. BMC Evol. Biol. 9, 112.

Laemmli, U.K., 1970. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 227, 680–685.

Lattman, E., 1985. Use of the rotation and translation functions. Methods Enzymol. 115, 55–77.

Lazarowitz, S.G., 1988. Infectivity and complete nucleotide sequence of the genome of a South African isolate of maize streak virus. Nucleic Acids Res. 16, 229–249.

Lazarowitz, S.G., Pinder, A.J., Damsteegt, V.D., Rogers, S.G., 1989. Maize streak virus genes essential for systemic spread and symptom development. EMBO J. 8, 1023–1032.

Li, X., Mooney, P., Zheng, S., Booth, C.R., Braunfeld, M.B., Gubbens, S., Agard, D.A., Cheng, Y., 2013. Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. Nat. Methods 10, 584–590.

Liljas, L., 2004. The role of disordered segments in viral coat proteins., in: Conformational Proteomics of Macromolecular Architecture. pp. 53–77.

Lin, T., Cavarelli, J., Johnson, J.E., 2003. Evidence for assembly-dependent folding of protein and RNA in an icosahedral virus. Virology 314, 26–33.

Liu, H., 1997. Molecular biology of maize streak virus movement in maize. PhD Thesis, Univ. East Angl.

Liu, H., Boulton, M.I., Davies, J.W., 1997a. Maize streak virus coat protein binds single- and double-stranded DNA in vitro. J. Gen. Virol. 78, 1265–1270.

Liu, H., Boulton, M.I., Oparka, K.J., Davies, J.W., 2001a. Interaction of the movement and coat proteins of Maize streak virus: implications for the transport of viral DNA. J. Gen. Virol. 82, 35–44.

Liu, H., Boulton, M.I., Thomas, C.L., Prior, D.A., Oparka, K.J., Davies, J.W., 1999. Maize streak virus coat protein is karyophyllic and facilitates nuclear transport of viral DNA. Mol. Plant-Microbe Interact. 12, 894–900.

Liu, H., Lucy, A.P., Davies, J.W., Boulton, M.I., 2001b. A single amino acid change in the coat protein of Maize streak virus abolishes systemic infection, but not interaction with viral DNA or movement protein. Mol. Plant Pathol. 2, 223–8.

Liu, L., Davies, J.W., Stanley, J., 1998. Mutational analysis of bean yellow dwarf virus, a geminivirus of the genus Mastrevirus that is adapted to dicotyledonous plants. J. Gen. Virol. 79, 2265–2274.

Liu, L., van Tonder, T., Pietersen, G., Davies, J.W., Stanley, J., 1997b. Molecular characterization of a subgroup I geminivirus from a legume in South Africa. J. Gen. Virol. 78 (Pt 8), 2113–2117.

Lucy, A., Boulton, M., Davies, J., Maule, A., 1996. Tissue specificity of Zea mays infection by maize streak virus. Mol. Plant-Microbe Interact. 9, 22–31.

Ludtke, S.J., Baldwin, P.R., Chiu, W., 1999. EMAN: semiautomated software for high-resolution single-particle reconstructions. J. Struct. Biol. 128, 82–97.

Malhotra, A., Harvey, S.C., 1994. A quantitative model of the Escherichia coli 16 S RNA in the 30 S ribosomal subunit. J. Mol. Biol. 240, 308–340.

Malik, P., Kumar, V., Bagewadi, B., Mukherjee, S., 2005. Interaction between coat protein and replication initiation protein of Mung bean yellow mosaic India virus might lead to control of viral DNA replication. Virology 337, 273–283.

Martin, D., Lefeuvre, P., Varsani, A., Hoareau, M., Semegni, J.-Y., Dijoux, B., Vincent, C., Reynaud, B., Lett, J.-M., 2011. Complex recombination patterns arising during geminivirus coinfections preserve and demarcate biologically important intra-genome interaction networks. PLoS Pathog. 7, e1002203.

Martin, D., van der Walt, E., Posada, D., Rybicki, E., 2005. The evolutionary value of recombination is constrained by genome modularity. PLoS Genet 1, e51.

Martin, D.P., Willment, J.A., Billharz, R., Velders, R., Odhiambo, B., Njuguna, J., James, D., Rybicki, E.P., 2001. Sequence diversity and virulence in Zea mays of Maize streak virus isolates. Virology 288, 247–255.

Martí-Renom, M.A., Stuart, A.C., Fiser, A., Sánchez, R., Melo, F., Šali, A., 2000. Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. 29, 291–325.

McGuffin, L.J., Bryson, K., Jones, D.T., 2000. The PSIPRED protein structure prediction server. Bioinformatics 16, 404–405.

McKenna, R., Xia, D., Willingmann, P., Ilag, L.L., Krishnaswamy, S., Rossmann, M.G., Olson, N.H., Baker, T.S., Incardona, N.L., 1992. Atomic structure of single-stranded DNA bacteriophage [Phi]X174 and its functional implications. Nature 355, 137–143.

Miller, J.L., Woodward, J., Chen, S., Jaffer, M., Weber, B., Nagasaki, K., Tomaru, Y., Wepf, R., Roseman, A., Varsani, A., Sewell, T., 2011. Three-dimensional reconstruction of Heterocapsa circularisquama RNA virus by electron cryo-microscopy. J. Gen. Virol. 92, 1960–70.

Mindell, J. a, Grigorieff, N., 2003. Accurate determination of local defocus and specimen tilt in electron microscopy. J. Struct. Biol. 142, 334–47.

Moffat, A.S., 1999. Geminiviruses Emerge as Serious Crop Threat. Sci. 286 , 1835.

Monjane, A.L., Harkins, G.W., Martin, D.P., Lemey, P., Lefeuvre, P., Shepherd, D.N., Oluwafemi, S., Simuyandi, M., Zinga, I., Komba, E.K., Lakoutene, D.P., Mandakombo, N., Mboukoulida, J., Semballa, S., Tagne, A., Tiendrébéogo, F., Erdmann, J.B., van Antwerpen, T., Owor, B.E., Flett, B., Ramusi, M., Windram, O.P., Syed, R., Lett, J.-M., Briddon, R.W., Markham, P.G., Rybicki, E.P., Varsani, A., 2011. Reconstructing the history of maize streak virus strain a dispersal to reveal diversification hot spots and its origin in southern Africa. J. Virol. 85, 9623–36.

Monjane, A.L., Pande, D., Lakay, F., Shepherd, D.N., van der Walt, E., Lefeuvre, P., Lett, J.-M., Varsani, A., Rybicki, E.P., Martin, D.P., 2012. Adaptive evolution by recombination is not associated with increased mutation rates in Maize streak virus. BMC Evol. Biol. 12, 252.

Morris-Krsinich, B.A., Mullineaux, P.M., Donson, J., Boulton, M.I., Markham, P.G., Short, M.N., Davies, J.W., 1985. Bidirectional transcription of maize streak virus DNA and identification of the coat protein gene. Nucleic Acids Res. 13, 7237–7256.

Mullineaux, P.M., Donson, J., Morris-Krsinich, B.A., Boulton, M.I., Davies, J.W., 1984. The nucleotide sequence of maize streak virus DNA. EMBO J. 3, 3063–3068.

Murata, K., Liu, X., Danev, R., Jakana, J., Schmid, M.F., King, J., Nagayama, K., Chiu, W., 2010. Zernike phase contrast cryo-electron microscopy and tomography for structure determination at nanometer and subnanometer resolutions. Structure 18, 903–912.

Nagar, S., Pedersen, T.J., Carrick, K.M., Hanley-Bowdoin, L., Robertson, D., 1995. A geminivirus induces expression of a host DNA synthesis protein in terminally differentiated plant cells. Plant Cell 7, 705–719.

Nagayama, K., Danev, R., 2008. Phase contrast electron microscopy: development of thin-film phase plates and biological applications. Philos. Trans. R. Soc. London (Series B) 363, 2153–62.

Nagayama, K., Danev, R., 2009. Phase-plate electron microscopy: a novel imaging tool to reveal close-to-life nano-structures. Biophys. Rev. 1, 37–42.

Namba, K., Pattanayek, R., Stubbs, G., 1989. Visualization of protein-nucleic acid interactions in a virus. Refined structure of intact tobacco mosaic virus at 2.9 A resolution by X-ray fiber diffraction. J. Mol. Biol. 208, 307–325.

Nawaz-ul-Rehman, M.S., Fauquet, C.M., 2009. Evolution of geminiviruses and their satellites. FEBS Lett. 583, 1825–32.

Nugent, C.I., Johnson, K.L., Sarnow, P., Kirkegaard, K., 1999. Functional coupling between replication and packaging of Poliovirus replicon RNA. J. Virol. 73, 427–435.

Ohi, M., Li, Y., Cheng, Y., Walz, T., 2004. Negative staining and image classification - powerful tools in modern electron microscopy. Biol. Proced. Online 6, 23–34.

Orlova, E. V, Saibil, H.R., 2011. Structural analysis of macromolecular assemblies by electron microscopy. Chem. Rev. 111, 7710–7748.

Padidam, M., Sawyer, S., Fauquet, C.M., 1999. Possible emergence of new geminiviruses by frequent recombination. Virology 265, 218–225.

Palmer, K.E., Rybicki, E.P., 1998. The molecular biology of mastreviruses. Adv. Virus Res. 50, 183–234.

Penczek, P. a, Grassucci, R. a, Frank, J., 1994. The ribosome at improved resolution: new techniques for merging and orientation refinement in 3D cryo-electron microscopy of biological particles. Ultramicroscopy 53, 251–70.

Penczek, P., Zhu, J., Frank, J., 1996. A common-lines based method for determining orientations for N > 3 particle projections simultaneously. Ultramicroscopy 63, 205–218.

Persson, M., Tars, K., Liljas, L., 2008. The capsid of the small RNA phage PRR1 is stabilized by metal ions. J. Mol. Biol. 383, 914–22.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., Ferrin, T.E., 2004. UCSF Chimera - a visualization system for exploratory research and analysis. J. Comput. Chem. 25, 1605–12.

Pilartz, M., Jeske, H., 1992. Abutilon mosaic geminivirus double-stranded DNA is packed into minichromosomes. Virology 189, 800–802.

Pinner, M.S., Medina, V., Plaskitt, K.A., Markham, P.G., 1993. Viral inclusions in monocotyledons infected by maize streak and related geminiviruses. Plant Pathol. 42, 75–87.

Pooma, W., Gillette, W.K., Jeffrey, J.L., Petty, I.A.N.T.D., 1996. Host and viral factors determine the dispensability of coat Protein for bipartite geminivirus systemic movement. Virology 268, 264–268.

Preiss, W., Jeske, H., 2003. Multitasking in replication is common among geminiviruses. J. Virol. 77, 2972–2980.

Qin, S., Ward, B.M., Lazarowitz, S.G., 1998. The bipartite geminivirus coat protein aids BR1 function in viral movement by affecting the accumulation of viral single-stranded DNA. J. Virol. 72, 9247–9256.

Radermacher, M., Wagenknecht, T., Verschoor, A., Frank, J., 1987. Three-dimensional reconstruction from a single-exposure, random conical tilt series applied to the 50S ribosomal subunit of Escherichia coli. J. Microsc. 146, 113–136.

Raoult, D., Forterre, P., 2008. Redefining viruses: lessons from Mimivirus. Nat. Rev. Microbiol. 6, 315–9.

Remmert, M., Biegert, A., Hauser, A., Soding, J., 2012. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods 9, 173–175.

Robbins, J., Dilworth, S.M., Laskey, R.A., Dingwall, C., 1991. Two Interdependent Basic Domains in Nucleoplasmin Nuclear Targeting Sequence : Identification of a Class of Bipartite Nuclear Targeting Sequence. Cell 64, 615–623.

Robinson, I., Harrison, S., 1982. Structure of the expanded state of tomato bushy stunt virus. Nature 297, 563–568.

Robinson, C. V, Sali, A., Baumeister, W., 2007. The molecular sociology of the cell. Nature 450, 973–982.

Roseman, A.M., 2000. Docking structures of domains into maps from cryo-electron microscopy using local correlation. Acta Crystallogr. Sect. D Biol. Crystallogr. 56, 1332–1340.

Rosenthal, P.B., Henderson, R., 2003. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. J. Mol. Biol. 333, 721–745.

Rossmann, M., Blow, D., 1962. The detection of sub-units within the crystallographic asymmetric unit. Acta Crystallogr. Sect. F 15, 24–31.

Rossmann, M.G., 1960. The accurate determination of the position and shape of heavy-atom replacement groups in proteins. Acta Crystallogr. 13, 221–226.

Rossmann, M.G., Johnson, J.E., 1989. Icosahedral RNA virus structure. Annu. Rev. Biochem. 58, 533–573.

Rybicki, E.P., 1994. A phylogenetic and evolutionary justification for three genera of Geminiviridae. Arch. Virol. 139, 49–77.

Saccardo, F., Cettul, E., Palmano, S., Noris, E., Firrao, G., 2011. On the alleged origin of geminiviruses from extrachromosomal DNAs of phytoplasmas. BMC Evol. Biol. 11, 185.

Sainsbury, F., Canizares, M.C., Lomonossoff, G.P., Cañizares, M.C., 2010. Cowpea mosaic virus: the plant virus-based biotechnology workhorse. Annu. Rev. Phytopathol. 48, 437–55.

Sali, A., Blundell, T.L., 1993. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779–815.

Šali, A., Potterton, L., Yuan, F., van Vlijmen, H., Karplus, M., 1995. Evaluation of comparative protein modeling by MODELLER. Proteins Struct. Funct. Bioinforma. 23, 318–326.

Saunders, K., Lucy, A., Stanley, J., 1991. DNA forms of the geminivirus African cassava mosaic virus consistent with a rolling circle mechanism of replication. Nucleic Acids Res. 19, 2325–2330.

Schalk, H.J., Matzeit, V., Schiller, B., Schell, J., Gronenborn, B., 1989. Wheat dwarf virus, a geminivirus of graminaceous plants needs splicing for replication. EMBO J. 8, 359–364.

Scheres, S.H.W., Valle, M., Nuñez, R., Sorzano, C.O.S., Marabini, R., Herman, G.T., Carazo, J.-M., 2005. Maximum-likelihood multi-reference refinement for electron microscopy images. J. Mol. Biol. 348, 139–49.

Schrodinger LLC, 2010. The PyMOL Molecular Graphics System, Version 1.3r1.

Scola, B. La, Audic, S., Robert, C., Jungang, L., de Lamballerie, X., Drancourt, M., Birtles, R., Claverie, J.-M., Raoult, D., 2003. A Giant Virus in Amoebae. Sci. 299 , 2033.

Shepherd, D.N., Mangwende, T., Martin, D.P., Bezuidenhout, M., Kloppers, F.J., Carolissen, C.H., Monjane, A.L., Rybicki, E.P., Thomson, J.A., 2007. Maize streak virus-resistant transgenic maize: a first for Africa. Plant Biotechnol. J. 5, 759–767.

Shepherd, D.N., Martin, D.P., McGivern, D.R., Boulton, M.I., Thomson, J.A., Rybicki, E.P., 2005. A three-nucleotide mutation altering the Maize streak virus Rep pRBR-interaction motif reduces symptom severity in maize and partially reverts at high frequency without restoring pRBR-Rep binding. J. Gen. Virol. 86, 803–813.

Shepherd, D.N., Martin, D.P., Van Der Walt, E., Dent, K.C., Varsani, A., Rybicki, E.P., 2010. Maize streak virus: an old and complex "emerging" pathogen. Mol. Plant Pathol. 11, 1–12.

Shepherd, D.N., Martin, D.P., Varsani, A., Thomson, J.A., Rybicki, E.P., Klump, H.H., 2006. Restoration of native folding of single-stranded DNA sequences through reverse mutations: an indication of a new epigenetic mechanism. Arch. Biochem. Biophys. 453, 108–122.

Speir, J.A., Johnson, J.E., 2012. Nucleic acid packaging in viruses. Curr. Opin. Struct. Biol. 22, 65–71.

Speir, J.A., Munshi, S., Wang, G., Baker, T.S., Johnson, J.E., 1995. Structures of the native and swollen forms of cowpea chlorotic mottle virus determined by X-ray crystallography and cryo-electron microscopy. Structure 3, 63–78.

Steinmetz, N.F., Evans, D.J., 2007. Utilisation of plant viruses in bionanotechnology. Org. Biomol. Chem. 5, 2891–902.

Stockley, P.G., Rolfsson, O., Thompson, G.S., Basnak, G., Francese, S., Stonehouse, N.J., Homans, S.W., Ashcroft, A.E., 2007. A simple, RNA-mediated allosteric switch controls the pathway to formation of a T=3 viral capsid. J. Mol. Biol. 369, 541–52.

Stockley, P.G., Twarock, R., Bakker, S.E., Barker, A.M., Borodavka, A., Dykeman, E., Ford, R.J., Pearson, A.R., Phillips, S.E. V, Ranson, N.A., Tuma, R., 2013. Packaging signals in single-stranded RNA viruses: nature's alternative to a purely electrostatic assembly mechanism. J. Biol. Phys. 39, 277–87.

Stuart, D.I., Grimes, J.M., Burroughs, N., Mertens, P.P.C., 1999. Crystallography and the atomic anatomy of viruses. Microbiol. Today 26, 59–61.

Subramaniam, S., Bartesaghi, A., Liu, J., Bennett, A.E., Sougrat, R., 2007. Electron tomography of viruses. Curr. Opin. Struct. Biol. 17, 596–602.

Tang, L., Johnson, J.E., 2002. Structural biology of viruses by the combination of electron cryomicroscopy and X-ray crytallography. Biochemistry 41, 11517–11524.

Tang, L., Johnson, K.N., Ball, L.A., Lin, T., Yeager, M., Johnson, J.E., 2001. The structure of pariacoto virus reveals a dodecahedral cage of duplex RNA. Nat. Struct. Biol. 8, 77–83.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–80.

Thon, F., 1971. Phase contrast electron microscopy. Electron Microsc. Mater. Sci. 571–625.

Timmermans, M.C., Das, O.P., Messing, J., 1992. Trans replication and high copy numbers of wheat dwarf virus vectors in maize cells. Nucleic Acids Res. 20, 4047–4054.

Topf, M., Lasker, K., Webb, B., Wolfson, H., Chiu, W., Sali, A., 2008. Protein structure fitting and refinement guided by cryo-EM density. Structure 16, 295–307.

Toropova, K., Basnak, G., Twarock, R., Stockley, P.G., Ranson, N.A., 2008. The three-dimensional structure of genomic RNA in bacteriophage MS2: implications for assembly. J. Mol. Biol. 375, 824–836.

Townsend, R., Watts, J., Stanley, J., 1986. Synthesis of viral DNA forms in Nicotiana plumbaginifolia protoplasts inoculated with cassava latent virus (CLV); evidence for the independent replication of one component of the CLV genome. Nucleic Acids Res. 14 , 1253–1265.

Unseld, S., Frischmuth, T., Jeske, H., 2004. Short deletions in nuclear targeting sequences of African cassava mosaic virus coat protein prevent geminivirus twinned particle formation. Virology 318, 90–101.

Valegård, K., Liljas, L., Fridborg, K., Unge, T., 1990. The three-dimensional structure of the bacterial virus MS2. Nature 345, 36–41.

Valegård, K., Murray, J., Stockley, P.G., Stonehouse, N.J., Liljas, L., 1994. Crystal structure of an RNA bacteriophage coat protein-operator complex. Nature 371, 623–626.

Van den Worm, S.H.E., Koning, R.I., Warmenhoven, H.J., Koerten, H.K., van Duin, J., 2006. Cryo electron microscopy reconstructions of the Leviviridae unveil the densest icosahedral RNA packing possible. J. Mol. Biol. 363, 858–65.

Van Heel, M., 1987. Angular reconstitution: a posteriori assignment of projection directions for 3D reconstruction. Ultramicroscopy 21, 111–123.

Van Heel, M., Schatz, M., 2005. Fourier shell correlation threshold criteria. J. Struct. Biol. 151, 250–62.

Van Heel, M., Schatz, M., 2009. Multivariate Statistical Analysis in Single Particle (Cryo) Electron Microscopy, in: Verkley, A., Orlova, E. V (Eds.), Electron Microscopy in Life Science.

Wade, R.H., 1992. A brief look at imaging and contrast transfer. Ultramicroscopy 46, 145–156.

Waterhouse, A.M., Procter, J.B., Martin, D.M. a, Clamp, M., Barton, G.J., 2009. Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. Bioinformatics 25, 1189–91.

Watson, J.D., Crick, F.H.C., 1953. A structure for deoxyribose nucleic acid. Nature 171, 737–738.

White, T.A., Bartesaghi, A., Borgnia, M.J., Meyerson, J.R., de la Cruz, M.J. V, Bess, J.W., Nandwani, R., Hoxie, J.A., Lifson, J.D., Milne, J.L.S., Subramaniam, S., 2010. Molecular architectures of trimeric SIV and HIV-1 envelope glycoproteins on intact viruses: strain-dependent variation in quaternary structure. PLoS Pathog. 6, e1001249.

Woodward, J., Wepf, R., Sewell, B., 2009. Three-dimensional reconstruction of biological macromolecular complexes from in-lens scanning electron micrographs. J. Microsc. 234, 287–292.

Wriggers, W., 2010. Using Situs for the integration of multi-resolution structures. Biophys. Rev. 2, 21–27.

Xia, Y., Levitt, M., 2004. Simulating protein evolution in sequence and structure space. Curr. Opin. Struct. Biol. 14, 202–7.

Xiang, Z., 2007. Advances in homology protein structure modeling. Curr. Protein Pept. Sci. 7, 217–227.

Zhang, W., Olson, N.H., Baker, T.S., Faulkner, L., Agbandje-McKenna, M., Boulton, M.I., Davies, J.W., McKenna, R., 2001. Structure of the Maize streak virus geminate particle. Virology 279, 471–477.

Zhou, S., Standring, D.N., 1992. Hepatitis B virus capsid particles are assembled from core-protein dimer precursors. Proc. Natl. Acad. Sci. 89, 10046–50.

Zhou, Z.H., Hardt, S., Wang, B., Sherman, M.B., Jakana, J., Chiu, W., 1996. CTF determination of images of ice-embedded single particles using a graphics interface. J. Struct. Biol. 116, 216–22.

6 Appendix A – Virion properties of a selection of simple viruses.

Table 2 Virion properties of a selection of simple viruses. These properties are shown alongside those MSV, and were predicted according to the methods set out by Speir and Johnson, 2012. An average RNA nucleotide mass of 340 Da, and an average DNA nucleotide mass of 330 Da were assumed. Capsid inner radii were evaluated coarsely using UCSF Chimera (Pettersen *et al.*, 2004)

Virus	Family	Architecture	Genome type	Inner radius (Å)	Volume capsid (Å^3)	Relative volume	Genome size (nt)	Genome weight (Da)	Vm (Å^3/Da)	Fraction Filled
STNV	*Satellite*	*T = 1*	ssRNA	60	9.04E+05	0.2	1200	4.08E+05	2.22	0.86
CV	*Circoviridae*	*T = 1*	ssDNA	70	1.44E+06	0.4	2000	6.60E+05	2.18	0.91
MSV	*Geminiviridae*	*T = 1 (bipartite)*	ssDNA	76	3.68E+06	1.0	2700	8.91E+05	4.13	0.48
CCMV	*Bromoviridae*	*T = 3*	ssRNA	102	4.40E+06	1.2	3200	1.09E+06	4.04	0.47
ϕCB5	*Leviviridae*	*T = 3*	ssRNA	112	5.90E+06	1.6	3760	1.28E+06	4.62	0.41
MS2	*Leviviridae*	*T = 3*	ssRNA	110	5.60E+06	1.5	3659	1.24E+06	4.50	0.42
FHV	*Nodaviridae*	*T = 3*	ssRNA	113	6.00E+06	1.6	4500	1.53E+06	3.92	0.49
PV	*Nodaviridae*	*T = 3*	ssRNA	113	6.00E+06	1.6	4300	1.46E+06	4.10	0.47
PoV	*Picornaviridae*	*T = 3*	ssRNA	108	5.27E+06	1.4	7500	2.55E+06	2.07	0.92

7 Appendix B – SPIDER scripts

reconAPSH.kcd

```
md
vb off

md
set mp
(4)

; create global symbols
fr g
[inputstack]../../output_stk_190_invert

fr g
[stack_doc]../../output_stk_doc; SPIDER document containing image info (defocus,
micrograph number, etc.)

fr g
[apref]./apref/apref

fr g
[vol]./vol/vol

fr g
[refprojs]./proj/proj

fr g
[bpangles]./bprojangles/bprojangles

fr g
[projangles]./projangles

fr g
[classavg]./classavg/class

fr g
[frealign]./frealign/spider_output

fr l
[class_arr]_10

; *** PARAMATERS FOR EACH ITERATION ***
[x92] = 21                              ; index of starting iteration
[x93] = 100                             ; index of last iteration
[x94] = 190                             ; outer ring for image alignment (radius in
Angstroms)
[x95] = 190                             ; input image dimensions
[x96] = 4                               ; sampling of input stack (A/pixel)
[x97] = 1                               ; decimation factor (for speed purposes): 1 = no
decimation
[x98] = [x96]*[x97]                     ; sampling after decimation
[x99] = 1                               ; set 1 if resolution estimation is required for
each round
[x100] = [x95]/[x96]                    ; image dimensions after decimation
[x101] = [x94]/[x98]                    ; outer ring for image alignment (radius in pixels)
[x102] = 15                             ; resolution of filtered output
[x103] = 50                             ; resolution to filter smooth mask
[x104] = 0                              ; smooth mask interior
[x105] = 0                              ; exploit the mirror relationship
[x106] = 0                              ; create output for FREALIGN

; create angles document
; *** Implement detection of mirroring etc in this part.
vm
rm [projangles].dat

vo ea
(3)
(0,90)
(0,72)
```

```
[projangles]

[x90] = 10878(4559) ; number of images to reconstruct,  first [x90] images will be
loaded into memory for iteration(s)

ud n [x91]                              ; set number of reference projections
[projangles]

ud e                                    ; close document

ms                                      ; create inline stack to store images for
iteration(s)
_1@
([x95],[x95],1)                         ; dimensions
([x90])                                 ; number of images

; load input stack into core
do lb9 [x19]=1,[x90]
        ;fq
        ;[inputstack]@{*****[x19]}
        ;_2
        ;(3)
        ;(

        ;dc
        ;[inputstack]@{*****[x19]}
        ;_1@{*****[x19]}
        ;([x97],[x97])

        cp
        [inputstack]@{*****[x19]}
        _1@{*****[x19]}
lb9

do lb1 [x11] = [x92],[x93]              ; major loop iterating project matching
        [x61] = 0                       ; initialize image counter
        [x12] = [x11]-1

        cp
        ./vol/vol{***[x12]}
        _7

        ; *** CALCULATE REFERENCE PROJECTIONS ***
        pj 3q
        _7                              ; input volume of previous iteration

        (50)                            ; radius

        (1-[x91])                       ; angle numbers

        [projangles]                    ; Euler angles used for project

        [refprojs]{***[x11]}@***        ; template for 2-d projections


        ;vm
        ;proc2d proj/proj{***[x11]}.dat proj/projn{***[x11]}.dat spider norm=0,1


        vm

        echo "Aligning images against reference projections:" {***[x11]} ; date


        ap sh                           ; match original stack against reference
projections of previous model

        [refprojs]{***[x11]}@****       ;template for reference images
        (1-[x91])                       ;selection
        (5,1)                           ;translation search range
        (5,[x101],1,1)
        [projangles]
        _1@*****
```

```
        (1-[x90])
        *
        (0.0,8)
        ([x105])

        [apref]{***[x11]}


        vm

        echo "Applying in-plane rotation and translation parameters and generating class
averages for " {*****[x90]} "images on iteration:" {***[x11]}


        @gen_classaverages[[x90],[x91],[x11],[x61],[x105],[x106]]

        vm

        echo "Back projecting {*****[x61]} images..." {***[x11]}

        if([x99].eq.1) then                    ; perform odd even reconstructions
                bp 32f
                [inputstack]_bpstk@*****
                (1-[x61])
                [bpangles]{***[x11]}
                sym
                _8
                [vol]{***[x11]}_fh             ; 3D model from first half
                [vol]{***[x11]}_sh             ; 3D model from first half
        else

                bp 3f
                [inputstack]_bpstk@*****
                (1-[x61])
                [bpangles]{***[x11]}
                sym
                _8
        endif

        cp                                              ; output volume for next iteration.
        _8
        ./vol/vol{***[x11]}

        vm
        proc3d ./vol/vol{***[x11]}.dat ./vol/vol{***[x11]}n.dat spidersingle norm=0,1
apix={***[x98]}

        ; tight mask volume to exclude noise surrounding structure
        @mask_vol[[x95],[x98],[x103],[104]]
        ./vol/vol{***[x11]}n

        vm
        cp ./vol/vol{***[x11]}n_masked.dat ./vol/vol{***[x11]}.dat

        ; create filtered output for inspection purposes
        fq
        ./vol/vol{***[x11]}.dat
        ./vol/vol{***[x11]}_fq.dat
        (3)
        ([x98]/[x102])                                  ; sampling/resolution

        vm

        echo "Generating statistics for iteration:" {***[x11]}

        @manage_stats[[x90],[x91],[x11],[x99]]


lbl
```

manage_stats.kcd

```
([x92],[x91],[x11],[x99])

; create histogram for correlation scores

vm
echo "Creating histogram of correlation scores..."
hd d
[apref]{***[x11]}
(11)
(50)
./stats/cc_dist{***[x11]}

; create histogram for reference projection frequencies
vm
echo "Creating histogram of reference projection frequencies..."
hd d
[apref]{***[x11]}
(4)
([x91])
./stats/ref_dist{***[x11]}

; update convergence measure
vm
echo "Updating convergence stats..."
ud n [x50]

[apref]{***[x11]}

vm
echo {*****[x50]}

[x60]=0
[x61]=0

; loop through each record
do lbl [x12] = 1,[x50]

        ud                                                                                                  ic
[x12],[x13],[x14],[x15],[x16],[x17],[x18],[x19],[x20],[x21],[x22],[x23],[x24],[x25],[x26
]

        [apref]{***[x11]}


        [x60] = [x60] + [x23]                          ; total of correlation scores
        if([x11].gt.2) then
                [x13]=[x11]-1
                ud n [x51]
                [apref]{***[x13]}

                if([x51].eq.[x50]) then
                        ud                                                                                  ic
[x12],[x33],[x34],[x35],[x36],[x37],[x38],[x39],[x40],[x41],[x42],[x43],[x44],[x45],[x46
]

                        [apref]{***[x13]}

                        if([x16].ne.[x36]) then
                                [x62]=[x62]+1
                        endif
                endif
        endif


lbl


[x61] = [x60]/[x50]


sd [x11],[x61]
```

```
./stats/averagecc

sd [x11],[x62]
./stats/ref_changes

ud ice
[apref]{***[x11]}

if([x11].gt.2) then
        ud ice
        [apref]{***[x13]}
endif

sd e
./stats/averagecc

sd e
./stats/ref_change

vm
echo "Calculating resolution statistics..."

if([x99].eq.1) then
        rf 3            ; Reliability Factor
        [vol]{***[x11]}_fh      ; first
        [vol]{***[x11]}_sh      ; second
        (0.5)                   ; ring width
        (0.6,1.2)               ; scale factor
        W                       ; missing Cone or Wedge
        (30)
        (3.0)
        ./stats/fscoutput_{***[x11]}
endif

re
```

mask_vol.kcd

```
([x95], [x98], [x103]) ; square (3D vol) dimensions, sampling, mask filter resolution,
fr
?vol?[vol]

; CONSIDERATIONS
; Smooth masking should not alter any of the ACTUAL structure's densities! (Hence the
volume must remain untouched within the masked region). We simply want to exclude
densities surrounding the particles boundary. Hence, it is important that we use a
conservative estimate of the this boundary in order to prevent masking of meaningful
information

vm
rm temp.dat

ip
[vol]
_21
([x95],[x95],[x95])

; CREATE MASK
; threshold volume to exlude noise
th                                    ; threshold
_21                                   ; input
_28                                   ; output
B                                     ; alter values (B)elow, (A)bove or (C) on both
sides.
(1.5)                                 ; units are sigma if the volume has been normalized
to mean of zero, std dev. of 1

cp
_28
[vol]_mask_th

; filter thresholded mask to increase boundary
fq                                    ; filter volume to a low resolution
_28
_29
(3)
([x98]/[x103])

cp
_29
[vol]_mask_th_fq

; create binary mask. Value = 1 where greater than equal 1.5, 0 where below 1.5
th m
_29
_22
(B)
(1.6)

cp
_22
[vol]_mask_thm

; filter binary mask to create smooth mask between 0 and 1.
fq
_22
_23
(3)
([x98]/[x103])

cp
_22
[vol]_mask_thm_fq

; Smooth masking
mm c
_22
_21
(0.0)   ;remove this if nec.
```

```
cp
_21
[vol]_masked

re
```