Life on the Edge: Structural Studies of the Extremophilic Viruses P23-77 and STIV2

Lotta J. Happonen

Institute of Biotechnology and
Department of Biosciences, Division of Genetics
Faculty of Biological and Environmental Sciences
University of Helsinki

and

Viikki Doctoral Programme in Molecular Biosciences
University of Helsinki

ACADEMIC DISSERTATION

To be presented for public examination with the permission of the Faculty of Biological and Environmental Sciences, University of Helsinki, in the auditorium 1041 of Biocenter 2, Viikinkaari 5, Helsinki, on June 8th 2012 at 12 o’clock noon.

HELSINKI 2012
The pursuit of truth and beauty is a sphere of activity in which we are permitted to remain children all our lives.

— *Albert Einstein*
ORIGINAL PUBLICATIONS

The thesis is based on the following articles, which are referred to in the text by their Roman numerals, and which are reprinted at the end of this thesis with permission from the publishers.


* These authors contributed equally.

Unpublished data will also be presented.
ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D</td>
<td>two-dimensional</td>
</tr>
<tr>
<td>3D</td>
<td>three-dimensional</td>
</tr>
<tr>
<td>3DEM</td>
<td>electron cryo-microscopy and three-dimensional image reconstruction</td>
</tr>
<tr>
<td>ADP</td>
<td>adenosine diphosphate</td>
</tr>
<tr>
<td>AFV</td>
<td>Acidianus filamentous virus</td>
</tr>
<tr>
<td>AMP</td>
<td>adenosine monophosphate</td>
</tr>
<tr>
<td>AMPPCP</td>
<td>adenylylmethylenediphosphonate</td>
</tr>
<tr>
<td>ATP</td>
<td>adenosine triphosphate</td>
</tr>
<tr>
<td>ATV</td>
<td>Acidianus two-tailed virus</td>
</tr>
<tr>
<td>bp</td>
<td>basepair</td>
</tr>
<tr>
<td>BTV</td>
<td>Bluetongue virus</td>
</tr>
<tr>
<td>cfu</td>
<td>colony forming unit</td>
</tr>
<tr>
<td>CIV</td>
<td>Chilo Iridescent Virus</td>
</tr>
<tr>
<td>cryoEM</td>
<td>electron cryo-microscopy</td>
</tr>
<tr>
<td>cryoET</td>
<td>electron cryo-tomography</td>
</tr>
<tr>
<td>ds</td>
<td>double-stranded (nucleic acid)</td>
</tr>
<tr>
<td>EMDB</td>
<td>electron microscopy data bank (<a href="http://www.ebi.ac.uk/pdbe/emdb/">http://www.ebi.ac.uk/pdbe/emdb/</a>)</td>
</tr>
<tr>
<td>FSC</td>
<td>Fourier shell correlation</td>
</tr>
<tr>
<td>gp</td>
<td>gene product</td>
</tr>
<tr>
<td>HHPV</td>
<td>Haloarchula hispanica pleomorphic virus</td>
</tr>
<tr>
<td>HRPV</td>
<td>Halorubrum pleomorphic virus</td>
</tr>
<tr>
<td>kDa</td>
<td>kilodalton</td>
</tr>
<tr>
<td>MDa</td>
<td>megadalton</td>
</tr>
<tr>
<td>NCLDV</td>
<td>Nucleo-Cytoplasmic Large DNA Virus</td>
</tr>
<tr>
<td>nm</td>
<td>nanometer</td>
</tr>
<tr>
<td>NMR</td>
<td>nuclear magnetic resonance spectroscopy</td>
</tr>
<tr>
<td>NTP</td>
<td>nucleotide triphosphate</td>
</tr>
<tr>
<td>ORF</td>
<td>open reading frame</td>
</tr>
<tr>
<td>p.i.</td>
<td>post-infection</td>
</tr>
<tr>
<td>PBCV-1</td>
<td>Paramecium bursaria Chlorella Virus 1</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>PDB</td>
<td>protein data bank (<a href="http://www.rcsb.org/pdb/home/home.do">http://www.rcsb.org/pdb/home/home.do</a>)</td>
</tr>
<tr>
<td>pfu</td>
<td>plaque forming units</td>
</tr>
<tr>
<td>RMSD</td>
<td>root-mean-square deviation</td>
</tr>
<tr>
<td>rpm</td>
<td>rounds per minute</td>
</tr>
<tr>
<td>S</td>
<td>Svedberg units</td>
</tr>
<tr>
<td>SIRV</td>
<td>Sulfolobus islandicus rod-shaped virus</td>
</tr>
<tr>
<td>ss</td>
<td>single-stranded (nucleic acid)</td>
</tr>
<tr>
<td>SSV</td>
<td>Sulfolobus spindle shaped virus</td>
</tr>
<tr>
<td>STIV</td>
<td>Sulfolobus turreted icosahedral virus</td>
</tr>
<tr>
<td>T</td>
<td>triangulation number</td>
</tr>
<tr>
<td>TLC</td>
<td>thin layer chromatography</td>
</tr>
<tr>
<td>TMV</td>
<td>Tobacco mosaic virus</td>
</tr>
<tr>
<td>w/v</td>
<td>weight per volume</td>
</tr>
<tr>
<td>VP</td>
<td>virion protein</td>
</tr>
<tr>
<td>Å</td>
<td>ångström</td>
</tr>
<tr>
<td>σ</td>
<td>standard deviation</td>
</tr>
</tbody>
</table>
SUMMARY

Viruses are the most abundant replicating entities on Earth, and they infect cells from all three domains of life - where there are cells, there are viruses. Extremophilic organisms and viruses thrive in hostile environments including hot, acidic springs, oceanic hydrothermal vents, and salt lakes. Due to their adaptation to extreme environments, these organisms and their viruses have been exploited for enzymes useful for industrial and biotechnological applications. Such enzymes include starch processing, cellulose degrading, proteolytic and DNA-processing enzymes. The latter ones are used in molecular biology applications such as polymerase chain reactions and DNA-sequencing.

The aim of this study was to characterize novel, extremophilic viruses living in hot springs. I solved the three dimensional structure of two such viruses using electron cryo-microscopy and three dimensional image reconstruction, and explored the presence of extremophilic enzymes based on their genome sequence. One of the viruses characterized in this study is P23-77 that infects the thermophilic bacterium Thermus thermophilus living in alkaline hot springs. P23-77 has been proposed to belong to the Tectiviridae family of viruses characterized by an internal lipid bilayer surrounded by an icosahedral protein capsid. The structure of the icosahedral P23-77 was initially solved to 1.4 nm resolution, and subsequently to 1.0 nm resolution. The reconstruction, together with thin-layer chromatography, confirmed the presence of an internal lipid bilayer composed of neutral lipids. Analysis of the P23-77 protein profile revealed it to have 10 structural proteins, two of which were major ones based on their abundance in SDS-PAGE gels. These proteins were suggested to form the capsomers with hexameric bases of the P23-77 T = 28d capsid lattice. Surprisingly, P23-77 closely resembles the haloarchaeal virus SH1, both of which are suggested to have single β-barrel major capsid proteins, and together forming a novel viral lineage.

The other virus characterized in this study is the Sulfolobus turreted icosahedral virus 2 (STIV2) infecting the crenarchaeon Sulfolobus islandicus that lives in acidic hot springs. The genome of STIV2 was sequenced, and some of its structural proteins were determined by mass-peptide fingerprinting. The structure of STIV2 was solved to 2.0 nm resolution. The genome sequence and the structure of STIV2 revealed it to resemble most closely STIV, infecting S. solfataricus. Like P23-77, both STIV and STIV2 have an outer protein capsid surrounding the internal lipid bilayer and the double-stranded (ds) DNA genome. The most striking difference between STIV and STIV2 resides in the host-cell recognition and attachments structures, which in STIV2 lacks the petal-like appendages present in STIV. Based on difference imaging, homology modeling and comparison to STIV, a model for the organization of the STIV2 virion was proposed. Furthermore, based on sequence data and homology modeling I identified the postulated genome packaging NTPase B204 of STIV2. I expressed and purified B204, and studied the nucleotide hydrolysis catalyzed by it. I furthermore solved four structures of B204 – more precisely, in complex with a sulphate ion, adenosine monophosphate, the product adenosine diphosphate, and the substrate analogue adenylylmethylenediphosphonate. B204 is the first genome packaging NTPase of a membrane-containing virus for which the structure has been solved. Based on the structure of B204, comparison to other known DNA-translocating enzymes, and other genome packaging NTPases of dsDNA and dsRNA viruses, I propose a model for the genome packaging of STIV2.
Contents

ORIGINAL PUBLICATIONS i
ABBREVIATIONS ii
SUMMARY iii

1. INTRODUCTION ............................................................................................................. 1
  1.1 The three ages of phage .......................................................................................... 2
  1.2 Viral capsid morphology ......................................................................................... 4
  1.3 The evolution of protein structure .......................................................................... 6
    1.3.1 The evolution of the active site ....................................................................... 7
  1.4 Protein folds in viral capsids .................................................................................. 9
  1.5 Architecture of icosahedrally-ordered capsids ...................................................... 11
    1.5.1 Quasi-equivalence in icosahedrally-ordered capsids .................................... 14
  1.6 Virus structure determination ................................................................................ 15
  1.7 Genome packaging into preformed capsids ............................................................ 20
    1.7.1 Common motifs in genome packaging ATPases ............................................. 20
    1.7.2 Genome packaging in dsDNA tailed bacteriophages ...................................... 20
    1.7.3 Genome encapsidation in the dsRNA virus φ12 .............................................. 22
    1.7.4 PRD1 genome packaging ............................................................................... 22
    1.7.5 The driving force behind genome translocation ......................................... 24
  1.8 Life at extremes ........................................................................................................ 24
    1.8.1 Applications of extremophiles and their proteins ......................................... 25
    1.8.2 Extremozymes from viruses .......................................................................... 26
    1.8.3 What stabilizes a thermophilic protein? ......................................................... 27
  1.9 Viruses from extreme environments ........................................................................ 28
    1.9.1 Viruses infecting Thermus species .................................................................. 28
    1.9.2 Sulfolobus Turreted Icosahedral Virus (STIV) ............................................. 29
    1.9.3 The haloarchaeal virus SH1 ........................................................................... 36

2. AIMS OF THE PRESENT STUDY .................................................................................. 39

3. MATERIALS AND METHODS ..................................................................................... 40
  3.1.1 Virus propagation and purification for cryoEM studies (I, II) ......................... 40
  3.1.2 CryoEM (I, II) ................................................................................................. 40
  3.1.3 CryoEM image processing (I, II, unpublished) .............................................. 41
  3.1.4 Structural docking, map segmentation and difference imaging (I, II, unpublished) 41
  3.1.5 DNA cloning (II, III) ..................................................................................... 42
  3.1.6 STIV2 genome annotation (II) ....................................................................... 42
  3.1.7 Expression and purification of recombinant B204 (III) ................................... 43
  3.1.8 Protein analysis (I, II, III) ............................................................................ 44
4. RESULTS AND DISCUSSION .......................................................................................... 47

4.1 Extremophilic viruses within a single β-barrel lineage ........................................... 47

4.1.1 Comparison of Thermus phages (I) ........................................................................ 47

4.1.2 Biochemical characterization of bacteriophage P23-77 (I) ...................................... 48

4.1.3 Structure of P23-77 to 1.4 nm (I) ........................................................................ 49

4.1.4 Structure of P23-77 to 1.0 nm (unpublished) ......................................................... 52

4.1.5 Homology modeling of P23-77 proteins (unpublished) .......................................... 56

4.2 Structural studies on the hyperthermo-acidophilic virus STIV2 ................................. 58

4.2.1 The STIV2 genome (II) ...................................................................................... 58

4.2.2 Identification of structural STIV2 proteins (II, III) .............................................. 58

4.2.3 Homology modeling of structural STIV2 proteins (II) ......................................... 59

4.2.4 Structure of the archaeal virus STIV2 and suggested virion arrangement (II, III) .... 59

4.2.5 Life cycle of P23-77 and STIV2 (I, II) ................................................................... 61

4.2.6 Isolation and stability of B204 (III, unpublished) ................................................ 61

4.2.7 B204 enzymatic activity (III) ............................................................................. 63

4.2.8 Structure of the STIV2 ATPase B204 (III) ............................................................ 64

4.2.9 Proposed catalytic cycle of B204 (III) ................................................................. 65

4.2.10 B204 DNA-binding and proposed DNA-binding site (III) .................................... 66

4.2.11 Proposed model for STIV2 genome packaging using B204 (III) ......................... 66

4.2.12 Implications for the PRD1-like viruses (III) ....................................................... 67

4.2.13 Proposed factors contributing to the thermostability of B204 (III) ....................... 67

4.3 The origin of viruses and viral lineages .................................................................... 69

4.3.1 Evolution of viruses ............................................................................................ 70

5. CONCLUSIONS AND FUTURE PROSPECTS ................................................................ 73

ACKNOWLEDGEMENTS .................................................................................................... 75

REFERENCES .................................................................................................................... 77
1. INTRODUCTION

Viruses are the most abundant replicating entities on Earth (Bergh et al., 1989), and are generally conceived as pathogens causing disease in animals and plants. Typical human illnesses caused by viruses are the common cold, influenza, cold sores, chickenpox, acquired immune deficiency syndrome, polio, and measles. The current estimated number for viruses in the biosphere is $10^{31}$ (Fischetti, 2011), with marine environments and especially marine sediments being the most prevalent location (Fuhrman, 1999).

Viruses are obligate parasites infecting cells from all three domains of life – Eukarya, Bacteria and Archaea. In principle, all cellular organisms are susceptible to viral infection, often by more than one type of virus. Viruses are small, independent particles with a typical size range from 20 to 400 nm in diameter – however, the largest virus described to date, Mimivirus, has a diameter of 750 nm and is thus bigger than some of the smallest bacteria (La Scola et al., 2003). Viruses consist of genetic material (DNA or RNA), which is surrounded by a protein capsid and a lipid membrane in enveloped viruses (Figure 1). The lipid membrane can reside outside or inside of the protein capsid. Viruses are parasites and need the host cell to replicate. Viruses contact their host cells by passive diffusion and adsorb to a susceptible host cell’s surface to initiate infection. Upon infection, viruses can leave their protein capsid on the outside of the cell and inject only their genome, something which is typical of tailed viruses and was initially described in 1952 (Hershey and Chase, 1952). Alternatively, the whole infectious virus particle – the virion – engages with the host cell and enters through a series of steps, such as the endocytic pathway, leading to progressive breakdown of the virion and delivery of the genome to the site of replication, such as the nucleus for transcription and translation. This pathway is typical for viruses infecting eukaryotes (Marsh and Helenius, 2006).

Figure 1. Schematic representations of viruses. (a) A simple icosahedral virus with a protein capsid (black icosahedron), host-attachment structures (black lines), and an internal nucleic acid (black spiral) (see sections 1.2 and 1.5 for details). A viral membrane (grey icosahedrons) can reside either outside (b) or inside (c) the protein capsid.

Viruses mainly reproduce in three ways. Firstly, in lytic infection, the viral genome injected into the host cell reprograms the cell into a virus-producing cellular factory. The viral genome is transcribed and translated, and directs the assembly of new viral particles that at the end of the lytic cycle break the host cell. The released progeny viruses reinitiate the lytic cycle by infecting the next susceptible
host cell. Secondly, in lysogenic infection, the viral genome is integrated into the host genome as a provirus or prophage. The viral genome is replicated as a part of the host genome. The excision of the viral genome can be induced, typically by stress such as temperature change or UV-irradiation, and switch the life cycle into a lytic one. Lastly, in chronic infection, the viruses reproduce inside the host cell. As opposed to lytic infection, the host cell does not burst upon progeny virus release, but remains intact as the progeny viruses are released in a controlled fashion via extrusion or budding. Two less well known reproduction strategies utilize pseudo-lysogeny and transposition. In pseudo-lysogeny, the viral nucleic acid resides within a starved host cell in an inactive, unstable state (Ripp and Miller, 1997; Ripp and Miller, 1998). Once the cells energy and nutrient requirements are met, the pseudo-lysogeny is resolved and the virus can continue its lysogenic or lytic life cycle. Pseudo-lysogeny is thought to be prevalent in natural ecosystems where it allows the virus to stay viable despite unfavorable conditions to the host cells (Ripp and Miller, 1997; Ripp and Miller, 1998). A few viruses, such as bacteriophage Mu, use transposition in their reproduction cycle. Mu uses transposition both when integrating into its host genome during its lysogenic life cycle and when replicating its genome during its lytic life cycle (Chaconas, Kennedy, and Evans, 1983; Haniford and Chaconas, 1992; Harshey, 1984). Similarly, retroviruses such as human immunodeficiency virus use transposition during their life cycle. The viral RNA genome is reverse transcribed into complementary DNA prior to integration into the host cell genome via a transposition based mechanism. The virus genome is subsequently transcribed and translated for the production of progeny viruses, much like in the case of a lysogenic virus.

1.1 The three ages of phage

The three ages of phage, or bacteriophage – viruses infecting bacteria – can roughly be divided as follows: the first age of phage concerns the discovery of the bacteriophages and their potential applications as therapeutic agents; the second age of phage bacteriophage research that led to some major discoveries on basic biochemical phenomena and supported the development of modern molecular biology; the third age of phage is the metagenomic age of phage, with the notion that phages might represent the largest gene pool in the biosphere.

Stepping back in time to the discovery of viruses, the idea of a disease causing agent smaller than bacteria was described independently by Dmitri Iwanowski (1864-1920) and Martinus Beijerinck (1851-1931). Both studied a tobacco disease causing spotted leaf coloring and showed that the disease-causing agent retained “its infectious properties even after filtration through Chamberland filter candles” (Iwanowski, 1892). Beijerinck continued these studies and showed that the agent can reproduce itself and do so only in living tissue, ruling out the possibility of the agent being a toxin (Beijerinck, 1898; Levine, 2001). They had unknowingly discovered tobacco mosaic virus. The question of the nature of the infectious agent continued for 25 years, until Twort (1915) and d’Herelle (1918) described the first viruses of bacteria – bacteriophages. d’Herelle subsequently demonstrated that the first step of a viral infection is the attachment of the virus to its host cell and described cell lysis and the release of progeny viruses by studying feces from patients suffering from dysentery caused by *Shigella* (d’Herelle, 1917; d’Herelle, 1921; d’Herelle, 1926; Twort, 1915). Indeed, during the first age of phage, the initial interest in viruses actually lay in their potential use for therapeutic applications. At that time d’Herelle noted that these disease-causing bacteria can be fought by viruses infecting them (d’Herelle, 1917) – a term coined phage therapy. Subsequently,
during the second age of phage, phage research supported the development of modern molecular biology, with some major findings in this field coming from the study of phages during 1940-1970, including the discovery of the nature of DNA as the genetic material (Hershey and Chase, 1952) and on gene regulation at almost every level (transcription, protein synthesis, and processing) (reviewed by [Levine, 2001; Mann, 2005]).

We are currently living in the third age of phage. On the one hand, it has been recognized that viruses might represent the largest gene pool in the biosphere and that phages might play a major role in the great planetary biogeochemical cycles (for example, nutrient and carbon cycling in the oceans) (Fuhrman, 1999; Mann, 2005; Wilhelm and Suttle, 1999). On the other hand, the third age of phage can also be seen as the metagenomic age of phage (Comeau et al., 2008). Metagenomics deals with the isolation of samples from the environment and instead of culturing the microbes and their viruses, the genomes in the samples are sheared, amplified, cloned, sequenced, and analyzed for content. This approach has great advantages in phage description, as identifying new viruses based on cultivation and physical isolation alone is inefficient and biased, since less than one percent of their possible microbial hosts have been cultivated (Edwards and Rohwer, 2005). What makes viral metagenomics difficult is the lack of a gene common to all phages – something corresponding to the 16S ribosomal DNA used in the reconstruction of phylogenies of cellular organisms (Rohwer and Edwards, 2002).

The most common way of classifying uncultured viruses is to use a single gene locus for a polymerase chain reaction (PCR), such as a gene encoding a viral capsid protein or a DNA polymerase (Rohwer and Edwards, 2002), but this approach only works for specific groups of viruses (Edwards and Rohwer, 2005) with conserved sequences for the target genes amplified. Alternatively, sequences for clusters of regularly interspaced short palindromic repeats residing within bacterial and archaeal genomes can be used, as they are often derived from viruses and can be used as a record of the viruses that have replicated within the cells (Snyder et al., 2010). Furthermore, as approximately 3 % of bacterial genomes consist of proviruses, the isolation of new, uncultured microbe metagenomes provides access to these viral genomes (Edwards and Rohwer, 2005). In addition to describing the viral diversity in various environments, viral metagenomics is a proven source for the identification of novel enzymes with possible biotechnological applications (Schmitz, Schuch, and Fischetti, 2010; Schoenfeld et al., 2009).
1.2 Viral capsid morphology

The viral protein shell is built of capsid proteins arranged into capsomers further assembled into the viral capsid enclosing the genome. The purpose of the capsid is to protect the nucleic acid against the environment while the virus is extracellular. This requires the capsid to be stable. It has frequently been observed that the most effective ways of protecting the viral genome is to package it into an icosahedrally-ordered protein shell or to cover it with a helical arrangement of capsomers (see section 1.5). These are the prevalent architectures utilized by the majority of bacterial and eukaryal viruses that have been described. In archaeal viruses, however, a striking diversity of different morphotypes have been identified, such as the spindle-shaped fuselloviruses and the bottle-shaped ampullaviruses, both of which are thus far absent from both bacterial and eukaryal hosts (Prangishvili, Forterre, and Garrett, 2006; Rachel et al., 2002) (Figure 2). Despite overall architecture and stability, and indifferent of the morphology, the robust capsid must be unstable enough to undergo conformational changes in order to deliver the viral genome into the host during infection.

Viruses are usually classified based on their nucleic acid type or virion morphology (Baltimore, 1971; Carstens, 2009; Fauquet and Fargette, 2005). Based on the nucleic acid type, Baltimore proposed six different classes for animal viruses, namely viruses with i) dsDNA genomes; ii) ssDNA genomes; iii) dsRNA genomes; iv) ssRNA [(+ sense) genomes; v) ssRNA [(- sense) genomes and; vi) viruses with ssRNA [(+] sense] genomes with a DNA intermediate, i.e. retroviruses (Baltimore, 1971). Furthermore, as described in the paragraph above, viruses can be divided into three broad classes based on morphology: i) viruses with icosahedral symmetry; ii) viruses with helical symmetry; and iii) viruses that do not belong to either of these prevalent classes and display no clear symmetry (Figure 2).
Figure 2. Examples of morphotypes of bacteriophages and archaeal viruses. The morphotypes are arranged according to their respective nucleic acid. For some viral morphologies example viruses are named. The abbreviations of the viral names are listed at the beginning of this thesis. Redrawn based on (Ackermann, 2007) and (Pietilä et al., 2009) with permission from the publishers. The structures are not drawn to scale.
1.3 The evolution of protein structure

Proteins are constructed from 20 amino acids in the order encoded by the organisms DNA sequence. The primary protein structure (amino acid sequence) folds into secondary structures (α-helices and β-strands) and further into the final tertiary three-dimensional structure. These tertiary structures can further assemble into quaternary structures forming multi-subunit protein complexes. Parts of the tertiary structures can form domains. These domains can be seen as individual building blocks and the primary units of evolution (Chothia and Gough, 2009) that exist and function independently. Typical domain sizes are between 50 and 200 amino acids (Chothia and Gough, 2009).

Based on the presence of one or more domains, proteins are divided into families. Other ways of classifying proteins are based on sequence similarity; for instance, proteins with > 40 % sequence identity are grouped together as families, whereas more distant proteins with < 30 % sequence similarity are joined in superfamilies (broader evolutionary families) (Orengo and Thornton, 2005). Family members are perceived as homologous descending from a common ancestor (Worth, Gong, and Blundell, 2009). Currently, protein families are believed to arise in two ways – via orthologs or paralogs. Orthologs are genes that have persisted through evolution in the species compared and share ancestry (Koonin, 2005). A crucial property of ortholog genes is equivalent functions in the organisms. Paralogs, on the other hand, are genes related via duplication (Koonin, 2005). Complicating the distinction between orthologs and paralogs is horizontal gene transfer between organisms (Koonin, 2005).

As more and more protein structures are being solved (approximately 77,000 entries in PDB, February 2012), there is an increasing interest in determining how many protein folds (tertiary structures) are used in nature (Coulson and Moult, 2002). Often the folding cores of evolving proteins remain similar with the majority of the changes occurring in loop-regions and peripheral domains or subdomains. Whilst many relationships between proteins can be deduced based on sequence similarity, a significant portion of relationships can only be detected based on comparison of the three-dimensional structures, since all significant sequence similarity has vanished during the course of evolution (Orengo et al., 2001). Programs such as Pfam (Punta et al., 2012) and DALI (Holm and Rosenström, 2010), and databases such as CATH (Pearl et al., 2005) have been created for structural based functional assignment of proteins. As of January 2012, the CATH database has a total of 152,920 domains, excluding nucleic acids and theoretical models. As reported by Marsden et al. in 2006 (Marsden et al., 2006), the number of domains has increased in six years (2000 to 2006) from 67,054 to 152,920 and the fold groups from 900 to 1,282, highlighting the rapid data accumulation in the field. As sequence data in public databases is simultaneously accumulated – there are a total of 14,734,050 reads in the reference sequence collection at the National Center for Biotechnology Information as of March 2012 – worldwide structural genomics initiatives have been launched in order to map the known protein sequences into structure based families (Brenner, 2001; Todd et al., 2005). The aim of these initiatives is to increase the coverage of protein fold space by solving structures of selected family representatives and to build homology models for other known, related proteins (Brenner, 2001; Todd et al., 2005). A reasonable goal for the structural genomics ambition is to solve at least one structure of each sequence family (Orengo et al., 2001).
Despite the fact that there are practically an unlimited number of possible protein sequences, the number of folds seems finite and the number of different folds is suggested to be around 10,000 (Koonin, Wolf, and Karev, 2002). It seems certain that the great majority of protein families belong to approximately 1,000 common folds. What is still disputed is the number of unifolds – folds found only in a single, narrow sequence family (Coulson and Moult, 2002). Moreover, the distribution of proteins among fold groups is highly non-homogeneous. Some folds are extremely abundant, like the P-loop nucleotide triphosphate hydrolases, and some protein superfamilies are thus highly occupied (Koonin, Wolf, and Karev, 2002). Currently, there are three favored possible explanations for the biased distribution of the fold and family sizes (Koonin, Wolf, and Karev, 2002). Firstly, certain folds might be frequently represented in the protein universe due to their topological properties, such as the high symmetry observed in TIM-barrels (Koonin, Wolf, and Karev, 2002). It has been postulated that certain protein families evolving independently might converge to the same fold if the sequences forming that fold are favored for some reason. Alternatively, if a protein family forms a highly symmetric and simple fold, it might be able to diverge widely resulting in a large number of families sharing that fold (Goldstein, 2008). Secondly, certain common and important biochemical activities, such as nucleoside 5’-triphosphate hydrolysis, are in greater demand than other highly specialized activities, which might lead to the observed bias of their abundance (Koonin, Wolf, and Karev, 2002). Finally, domain birth and death models include the elementary process of family growth via domain birth (gene duplication) and domain death as a result of inactivation via the accumulation of mutations and possible domain loss (Karev et al., 2002). Furthermore, innovation is an efficient way of spreading a protein and its fold. Such innovation processes involve gene acquisition via horizontal gene transfer or the emergence of a new gene from a non-coding sequence (Karev et al., 2002).

Domains with generic or especially useful functions are of most use in many organisms (Apic, Gough, and Teichmann, 2001). One such domain is the aforementioned P-loop nucleotide triphosphate (NTP) hydrolase. The P-loop NTPases represent an extremely diverse and abundant protein superfamily with characteristic Walker A and Walker B sequence motifs involved in nucleotide and divalent cation binding, as described for the genome packaging ATPases of icosahedrally-ordered viruses (see section 1.7.1). Detection of these proteins by sequence similarity is straightforward (Kinch and Grishin, 2002) and a monophyletic origin for these NTPases has been proposed (Iyer et al., 2004a; Iyer et al., 2004b). Representatives for most of the P-loop NTPases have been determined and all of these structures have an αβα-sandwich architecture, with the central β sheet being highly conserved. However, the locations of the Walker A and B sequence motifs differ in the various structures (Iyer et al., 2004b; Kinch and Grishin, 2002).

1.3.1 The evolution of the active site

Enzymes are proteins catalyzing essential biochemical reactions. An important aspect in the study of enzyme function is to understand how they use the limited set of available amino acid side chains to perform the reactions (Gutteridge and Thornton, 2005). Of the 20 side chains available, 11 polar and charged residues are generally found in catalytic sites (Bartlett et al., 2002). Furthermore, metal ions, cofactors, and water molecules are used to aid catalysis. As catalytic residues are not always consistently defined in the literature, Bartlett et al. (Bartlett et al., 2002) defined a list of rules for the description of residues involved in catalysis, according to which the residue has to: i) be directly
involved in the catalytic mechanism (for instance, act as the nucleophile); ii) exert an effect on another residue or on a water molecule, which is directly involved in the catalytic mechanism; iii) stabilize a proposed transition-state intermediate; or iv) exert an effect on a substrate or a cofactor which aids catalysis. Residues that merely bind substrate, cofactor or metal are not included in the definition of catalytic residues (Bartlett et al., 2002).

Using the rules above, Bartlett et al. deduced that 65 % of catalytic residues are the charged groups of histidine, arginine, lysine, glutamic acid, and aspartate residues and 27 % of catalytic residues are provided by the polar groups of glutamine, threonine, serine, asparagine, cysteine, and tyrosine residues (Bartlett et al., 2002). 18 % of all catalytic residues are provided by histidines (Bartlett et al., 2002). Histidine is particularly suitable for carrying out the catalytic reaction steps, as it can act as the nucleophile, the acid or base, or stabilize the transition state (Bartlett et al., 2002). While the aforementioned residues participate in the catalysis via side chain interactions and constitute 92 % of the interactions, 8 % of the catalytic interactions are provided by main chain interactions. 44 % of the main chain interactions are provided by glycine residues (Bartlett et al., 2002). Furthermore, 50 % of catalytic residues reside in coil regions (Bartlett et al., 2002). Catalytic residues are more conserved than average residues both in the one-dimensional amino acid sequence as measured by conservation of residues away from active site residues and in the three-dimensional structure as measured by conservation within a given radii around the active site (Bartlett et al., 2002). This indicates that catalytic residues are under strong selection pressure (Bartlett et al., 2002).

Most enzymes analyzed by Bartlett et al. have their catalytic residues located within one subunit (Bartlett et al., 2002). Furthermore, 60 % of the analyzed proteins contain more than one domain, with the catalytic residues split between more than one domain in 20 % of the enzymes (Bartlett et al., 2002). This poses an intriguing question about the evolution of enzymes. If one considers the domain as a structural and evolutionary unit, as mentioned above, how can it be explained that the active site is split between different domains or subunits? Two explanations have been provided. Firstly, one might consider that a primitive enzyme has catalyzed the initial steps on one domain. Later on, subsequent steps might have evolved residing on another domain, with residues well positioned to speed up, stabilize intermediates, or provide other selective advantages for the initial reaction (Bartlett et al., 2002). Secondly, convergent evolution of two distinct functional domains might have occurred to form an enzyme with an adapted function (Bartlett et al., 2002).
1.4 Protein folds in viral capsids

There are four prevalent folds in the major capsid proteins of viruses. These folds are the β-barrel fold, the HK97-fold, the BTV-like fold, and the four-helix bundle. These folds are shared by viruses belonging to different domains of life and are suggested to be the basis of viral lineages with their members possibly evolutionarily related (Abrescia et al., 2010; Bamford, 2003; Bamford, Burnett, and Stuart, 2002; Benson et al., 2004; Goulet et al., 2009; Krupovic and Bamford, 2008a).

The eight stranded β-barrel is the most common fold in viral major capsid proteins (Chapman and Liljas, 2003; Krupovic and Bamford, 2011; Rossmann and Johnson, 1989). It consists of two four-stranded β-sheets facing one another, with the ends almost closing in forming a barrel-like structure. All of the strands run antiparallel to their neighbors (Figure 3a). There are two different ways in which viruses utilize this fold in their capsid assembly. The prevalent version is to line the β-barrels tangentially along the surface, such as in picornaviruses (Hogle, Chow, and Filman, 1985; Khayat and Johnson, 2011; Rossmann and Johnson, 1989). Currently, omitting the results published in this thesis, the only virus known to arrange its proposed single β-barrel vertically to the capsid is the haloarchaeal virus SH1 (Jäälinoja et al., 2008), although the minor capsid proteins forming the pentamers sitting at the vertices of PRD1, PM2 and adenovirus are all single β-barrel proteins (Abrescia et al., 2004; Abrescia et al., 2008; Huiskonen et al., 2004; Rydman et al., 1999; Zubieta et al., 2005). In some viruses, the gene encoding the β-barrel has possibly been duplicated or fused, giving rise to the double β-barrel fold (Jäälinoja et al., 2008) (Figure 3b). Common to all the viruses utilizing this double β-barrel fold is that they align the barrels vertically towards the viral surface, as in SH1. This way of assembling capsids is seen in viruses proposed to belong to the double β-barrel virus lineage containing viruses such as PRD1, STIV, PM2, vaccinia, PBCV-1, and adenovirus (Figure 2; PRD1, STIV, PM2) (Abrescia et al., 2008; Athappilly et al., 1994; Bahar et al., 2011; Benson et al., 1999; Benson et al., 2002; Khayat et al., 2005; Nandhagopal et al., 2002).

Furthermore, dsRNA viruses with segmented genomes, such as the bluetongue virus (BTV) and the φ-viruses, organize their core particle on a T = 2 lattice (Grimes et al., 1998; Huiskonen et al., 2006; Jäälinoja, Huiskonen, and Butcher, 2007). The major capsid protein of these particles is shown to have a similar fold, suggesting a common ancestor (Bamford, Grimes, and Stuart, 2005) (Figure 3c).

Based on electron micrographic studies, tailed bacteriophages are the most abundant virus species (Ackermann, 2007). Although these viruses include myoviruses with contractile tails, siphoviruses with long, non-contractile tails, and podoviruses with short, non-contractile tails (Figure 2), they all seem to share the same HK97-fold in their major capsid protein (Figure 3d) – named after the HK97 siphovirus. Examples of viruses proposed to belong to the HK97-like viruses are P22, T4, and φ29.

Lastly, rod-shaped viruses with linear, single-stranded genomes have a four-helix bundle fold in their capsid protein (Figure 3e). These viruses include the well characterized TMV and the archaeal rudi- and lipothrixviruses, such as AFV1 and SIRV (Figure 2) (Goulet et al., 2009; Szymczyna et al., 2009).
Figure 3. Folds of viral major capsid proteins. (a) The single β-barrel of the picornavirus HPV16 (Protein data bank (PDB) accession code: 1dzl) (Chen et al., 2000). The β-barrel is highlighted in red. (b) The double β-barrel of the tectivirus PRD1 (PDB: 1hx6) (Benson et al., 2002). (c) The BTV capsid protein (PDB: 2btv) (Grimes et al., 1998). (d) The HK97-fold of the siphovirus HK97 (PDB: 1ohg) (Helgstrand et al., 2003). (e) Capsid protein of TMV (PDB: 1ei7) (Bhyravbhatla, Watowich, and Caspar, 1998). The figure was rendered in PyMOL.
1.5 Architecture of icosahedrally-ordered capsids

The discussion on the structure of spherical viruses began in the 1950’s with Crick and Watson’s theoretical speculations on the geometry of small, isometric viruses (Crick and Watson, 1956). They proposed that the icosahedral symmetry with its 60 asymmetric subunits is the most likely arrangement for these viral capsids (as opposed to octahedral or tetrahedral symmetry) as it allows for a larger volume of nucleic acid to be encapsitated (Crick and Watson, 1956). Caspar and Klug (Caspar and Klug, 1962) continued building on this idea and proposed a mathematical model for how larger viruses with additional subunits – more than 60 – can utilize icosahedral symmetry in their capsid architecture. In the simplest viruses with icosahedral symmetry and 60 subunits in their capsid lattice, all subunits have chemically equivalent environments; i.e. the chemical bonding pattern between protein subunits is identical (Caspar and Klug, 1962). As was predicted by Caspar and Klug, for viruses with more than 60 subunits in their capsid lattice, the bonding pattern for all subunits is not equivalent but rather quasi-equivalent (Caspar and Klug, 1962) (see section 1.5.1).

An icosahedron is an isometric geometrical body with 12 pentagonal vertices and 20 triangular faces. Each icosahedron has defined symmetry elements, described by the six five-fold axes through the 12 vertices, ten three-fold axes through the 20 facets, and 15 two-fold axes through the edges between the facets (Figure 4a). As will be discussed in section 1.6, these symmetry axes gives rise to 37 self common lines within each icosahedral particle; each five-fold axis contributes to two lines, each three-fold axis to one line, and each two-fold axis to one line (2 × 6 + 10 + 15 = 37) (Crowther, 1971; Fuller et al., 1996). The icosahedral lattice in viruses is built of major capsid protein subunits organized into defined capsomers. The pentameric capsomers at the vertices are symmetric pentavalent subunit clusters whereas the hexameric capsomers are symmetric hexavalent subunit clusters (Caspar and Klug, 1962). Furthermore, in capsids where all capsomers are formed of pentameric subunit clusters, such as polyoma simian virus 40, their environments are either pentavalent or hexavalent (Baker, Drak, and Bina, 1988).

The icosahedrally-symmetric capsid size is dependent on the number of capsomers in the lattice. The size of a viral capsid is usually referred to using a triangulation (T) number. The T-number is described using the rule $T = h^2 + hk + k^2$, where $h$ and $k$ are the integers describing the capsid lattice (Caspar and Klug, 1962). Furthermore, the T-number describes the number of capsid protein subunits in the icosahedral asymmetric unit (1/60th of an icosahedron). The simplest viruses with icosahedral symmetry have 60 copies of a single capsid protein subunit arranged in equivalent positions (Caspar and Klug, 1962). This arrangement has a T-number of 1 and the capsomers are actually ordered on a dodecahedral lattice as opposed to an icosahedral one (both with 532 space group symmetry). An example of such a T = 1 virus is the microvirus φX174 (McKenna et al., 1992) (Figure 2). There are a certain number of “allowed” T-numbers, with $T = 3$ being the next allowed value ($h = 1, k = 1$) followed by $T = 4$ ($h = 2, k = 0$) (Figure 4b) (Caspar and Klug, 1962). As the numbers of capsomers in the lattice increase, the T-number grows and the number of quasi-equivalent bonding patterns between subunits increases. Thus, for example, for a virus with a $T = 3$ capsid arrangement, there are 180 subunits ($60 \times 3$) with three different types of bonds or three distinct environments with respect to icosahedral symmetry axes. Furthermore, in a $T = 4$ capsid arrangement there are 240 subunits ($60 \times 4$) with four different environments. Examples of $T = 3$ and
$T = 4$ viruses are the levivirus MS2 and Sindbis virus, respectively (Golmohammadi et al., 1993; Paredes et al., 1993) (Figure 2, MS2). Some of the $T$-numbers are skewed or handed, such as $T = 7$, 13, 19, and 21 (Caspar and Klug, 1962) (Figure 4c). Interestingly, Caspar and Klug (Caspar and Klug, 1962) predicted that nature is unlikely to use the skewed $T$-numbers in virus structures due to possible mistakes during assembly leading to defective particles. As we know today, there are several viruses with skewed $T$-numbers, ranging from the rather small polyoma simian virus 40, with a $T = 7d$ capsid arrangement (Baker, Drak, and Bina, 1988), to the large PpV01 virus with an impressive $T$-number of 219d (Yan et al., 2005).

Other unique $T$ numbers include the $T = 2$ capsomer arrangement, as described briefly in section 1.4, and the pseudo- $T$ capsid lattices. The $T = 2$ capsomer arrangement is most prevalent in dsRNA viruses, such as reoviruses, cystoviruses, and partiviruses (Grimes et al., 1998; Huiskonen et al., 2006; Jäälinoja, Huiskonen, and Butcher, 2007; Ochoa et al., 2008). In the $T = 2$ capsid lattice, 120 copies (rather than 60 copies) of a single capsid protein subunit are arranged on a $T = 1$ lattice. In pseudo- $T$ capsid lattices, such as the pseudo- $T = 25$ of the tectivirus PRD1 or adenovirus (Athappilly et al., 1994; Benson et al., 1999; Benson et al., 2002; San Martin et al., 2001), where the capsid protein subunits forming the lattice are not hexameric but trimeric, the arrangement is referred to as a pseudo-equivalence rather than a quasi-equivalence.
Figure 4. Geometric principles of constructing icosahedrally-ordered virus capsids of defined triangulation ($T$) numbers. (a) An icosahedron. One facet is colored in light grey and an asymmetric unit within that facet is colored darker grey. The icosahedral symmetry axes are indicated with a black ellipse (two-fold), triangle (three-fold), and pentagon (five-fold). (b) Capsid arrangements for $T = 3$ and $T = 4$ lattices. The $h$ and $k$ lattice vectors are indicated. The shaded triangle represents one facet of the icosahedral lattice. The icosahedral closed surface is generated from the hexameric planar sheet by introducing pentamers at the indicated five-folds (corners of grey triangle). (c) Capsid arrangement for the skewed or handed $T = 7$ lattice. The grey triangle corresponds to the *laevo* ($l$) lattice and the triangle outlined in dotted line the *dextro* ($d$) lattice. The figure is adapted from Baker *et al.* (Baker, Olson, and Fuller, 1999) with permission from the publisher.
1.5.1 Quasi-equivalence in icosahedrally-ordered capsids

Viruses are self-assembling macromolecular complexes. This self-assembly requires built-in directions on how to assemble into the correct architecture. These directions are found in the structure subunits – the capsid proteins – with inherent sets of specific bond sites (Caspar and Klug, 1962). As high-resolution electron cryo-microscopy (cryoEM) reconstructions and X-ray crystallographic models of viral particles have become available, it has been possible to analyze the bonding patterns of the major capsid proteins in detail. As was predicted by Caspar and Klug, for viruses with more than 60 subunits in their capsid lattice, the bonding pattern for all subunits is not equivalent but rather quasi-equivalent (Caspar and Klug, 1962).

The quasi-equivalence of the bonding pattern of the major capsid protein subunits is prevalent and concerns all viruses with a $T$-number larger than 1. As this thesis describes structural analysis of icosahedral, membrane-containing viruses, the type virus PRD1 of the tectivirus family will be used as an example. In addition to having a $T$-number larger than 1, the PRD1 major capsid protein P3 subunits are arranged on a pseudo-$T = 25$ capsid lattice (Butcher, Bamford, and Fuller, 1995). The structure of PRD1 has been studied in detail (Butcher, Bamford, and Fuller, 1995; Rydman et al., 1999) and atomic models exist for the particle as well as individual proteins (Abrescia et al., 2004; Benson et al., 1999; Benson et al., 2002; Cockburn et al., 2004; Merckel et al., 2005; Xu et al., 2003). Furthermore, there are cryoEM and small angle X-ray scattering data available describing the molecular interactions of the host-attachment structures of PRD1 (Huiskonen, Manole, and Butcher, 2007b; Sokolova et al., 2001). Briefly summarizing the overall PRD1 structure (Butcher, Manole, and Karhu, 2012), the major capsid protein of PRD1 is the homotrimeric double β-barrel protein P3. The cementing minor capsid protein P30 locks the trimeric P3 in position (Abrescia et al., 2004; San Martin et al., 2002). The icosahedral vertices are occupied by the pentameric P31, the spike forming P5 and the receptor binding P2 proteins (Huiskonen, Manole, and Butcher, 2007b; Merckel et al., 2005; Sokolova et al., 2001). The genome of PRD1 is encapsidated via a dedicated packaging vertex containing the packaging machinery consisting of at least P6 and the ATPase P9. These are anchored in position in the viral membrane via the integral membrane proteins P20 and P22 (Gowen et al., 2003; Strömsten, Bamford, and Bamford, 2003; Strömsten, Bamford, and Bamford, 2005).

Initially, the quasi-equivalence of the PRD1 P3 was analyzed by generating a quasi-atomic model of the particle by fitting the P3 X-ray model into the PRD1 cryoEM density (San Martin et al., 2001). Based on this model, it was proposed that the P3 trimers make five different types of trimer-trimer contacts with varying types of amino acid residues at the interfaces (San Martin et al., 2001). Subsequently, when the X-ray structure of the PRD1 particle became available, it was shown that the N- and C-termini of P3 take different conformations based on the location of the subunit in the capsid lattice (Abrescia et al., 2004). A total of 12 different environments for P3 have been observed, with the N-termini taking eight different conformations and the C-termini taking nine (Abrescia et al., 2004). As a whole, the X-ray crystallographic structure of the PRD1 particle, together with results from Raman spectroscopy, illustrate a complex network of interactions between the P3 termini attaching to the membrane, to other minor proteins, and to each other stabilizing the capsid and assisting in capsid assembly (Abrescia et al., 2004; Tuma et al., 1996).
The adenovirus hexon and the PRD1 P3 have been shown to be strikingly similar suggesting an evolutionary relationship (Benson et al., 1999; Rux, Kuser, and Burnett, 2003). In addition, there has been significant advancement in the atomic modeling of adenovirus (Liu et al., 2010; Reddy et al., 2010), and therefore it is relevant to mention a few words about the quasi-equivalence observed in the adenovirus hexon. In PRD1 P3, the N- and the C-termini have been shown to be disordered in comparison to the subunits seen in the X-ray crystallographic study on the particle (Abrescia et al., 2004; Benson et al., 1999; Benson et al., 2002). This same phenomenon has been observed for the adenovirus hexon (Liu et al., 2010; Reddy et al., 2010; Rux, Kuser, and Burnett, 2003). Each asymmetric unit in the adenovirus is built of twelve hexon monomers (Burnett, 1985). These twelve monomers exhibit five different N-terminal and six different C-terminal conformations within the asymmetric unit (Liu et al., 2010). Additionally, as in PRD1 where P3 is shown to interact with the minor capsid protein P30, several interactions between the adenovirus hexon and minor capsid proteins could be discerned in the atomic resolution structures of the viral particles (Liu et al., 2010; Reddy et al., 2010).

1.6 Virus structure determination

There are five main methods yielding structural information and used in virus research, namely cryoEM, electron cryo-tomography (cryoET), X-ray crystallography, nuclear magnetic resonance, and small angle X-ray scattering. Furthermore, mass-spectrometric methods yield valuable information on virus assembly and assembly intermediates as well as protein interaction sites. Even though all methods are suitable for the study of isolated viral proteins, the methods usually used for the study of icosahedrally-symmetric viral particles are limited to cryoEM, cryoET, and X-ray crystallography.

Electron microscopy using either negative stained or cryo-preserved virus samples has traditionally been used for generating moderate to low resolution (10-30 Å) reconstructions of viruses. X-ray crystallography has further been used in order to achieve higher resolution structures of 3 Å or more. These two methods can be used in the studies of icosahedral and helical viral particles. However, due to the absence of symmetry in pleomorphic viruses, such as influenza, measles, and human immunodeficiency virus, their overall structures cannot be subjected to similar averaging operations as for symmetrical objects and they are thus usually studied by cryoET (Fontana et al., 2012; Liljeroos et al., 2011; Liu, Wright, and Winkler, 2010). Additionally, both cryoEM and cryoET have several advantages over X-ray crystallography when it comes to studies on large macromolecular complexes such as viruses (Tang and Johnson, 2002). Firstly, there is no special requirement for the purity of the sample. Secondly, there is no need to produce crystals of the sample. Thirdly, only a small amount of sample is needed, as typical sample concentrations range between 0.05-5 mg ml⁻¹ and only 2-5 µl of sample is used per grid (Baker, Olson, and Fuller, 1999). Finally, the structure is maintained in its native form by the rapid vitrification, preserving the state of the macromolecule in the solution (Adrian et al., 1984; Dubochet et al., 1988). As both cryoEM and X-ray crystallography have been used in this thesis, the general processes behind both methods are summarized in Figures 5 and 6. Only cryoEM was, however, used in the structural determination of viral particles and will be described in more detail below.

Summarizing briefly the basis of electron cryo-microscopy and three-dimensional image reconstruction (3DEM) of icosahedral viruses (Figure 5), an aqueous virus suspension is applied to a
holey-carbon grid, plunged rapidly into liquid ethane using a guillotine-like device to vitrify the water, and kept at liquid nitrogen (or helium) temperatures (below -150 °C). The grid containing the virus particles is transferred into the electron microscope in a cryo-cooled sample holder and two-dimensional projection images of the three-dimensional virus particles are normally collected either on electron sensitive film or on a charge-coupled device (CCD) camera using low-dose conditions in order to avoid radiation damage (Adrian et al., 1984; Baker, Olson, and Fuller, 1999; Dubochet et al., 1988; van Heel et al., 2000). The virus particles are assumed to be randomly oriented in a layer of vitreous water on the grid. For a three-dimensional (3D) reconstruction, anything between tens to tens of thousands of particles are used, depending on the research question put forward and the material at hand – usually a higher number of particles produce a higher resolution reconstruction with finer details discerned. At the start, the virus particles are selected and boxed out from the electron micrographs, the resulting images normalized and filtered, and the initial particle centers \((x,y)\) and orientations \((\theta, \phi, \omega)\) determined (Figure 5). This orientation determination is based on the symmetry elements within the icosahedral shape of the particle and can be determined either using the 37 common lines described in section 1.5 or by projection matching to a known 3D structure (Baker and Cheng, 1996; Crowther, 1971; Fuller et al., 1996). When describing the orientations using polar Fourier coordinates, as in Figure 5, the icosahedral two-fold axis lies along the z-axis, the adjacent three-folds along the x-axis, and the adjacent five-folds along the y axis. A three-fold view, for instance, in these coordinates is \(\theta = 69.09°\) and \(\phi = 0°\) (Fuller et al., 1996). The \(\omega\) angle describes the rotation of the particle in plane of the projection (Figure 5). The process of determining the centers and orientations is an iterative process and typically tens to hundreds of iterations are needed for the values to converge. There are several dedicated software programs to carry out these iterations – AUTO3DEM (Yan, Sinkovits, and Baker, 2007), EMAN (Ludtke, Baldwin, and Chiu, 1999), FREALIGN (Grigorieff, 2007), and IMAGIC (van Heel et al., 1996) – to name a few. The steps of determining the origin and orientation of each particle are critical as they determine the quality of the final reconstruction (Crowther, 1971; Fuller et al., 1996).

There are three main methods used to generate a 3D model of the two-dimensional (2D) projection images of icosahedrally-ordered particles. The reconstruction process can either be done using a model-based approach (projection matching) or \textit{ab initio} using the common-lines methods (Baker and Cheng, 1996; Crowther, 1971; van Heel, 1987). Furthermore, the common-lines method can either be applied in real space using angular reconstitution (van Heel, 1987) or in Fourier space (Crowther, 1971; Fuller et al., 1996). The common-lines method in Fourier space was the first one implemented for the reconstruction of icosahedrally-ordered viruses (Crowther, 1971). Based on the central section theorem, the parameters for aligning 2D particles with known origins and orientations to yield a 3D representation can be determined from the intersecting common lines in Fourier space. As described in section 1.5, there are 37 pairs of self common lines within each particle and 60 cross common lines between any two particles. These common lines can be used to find the axes of symmetry in the individual projection images (Fuller et al., 1996). A 3D representation of the viral particles imaged can be generated by “filling-up” Fourier space by adding particles with different orientations to the data set (Crowther, 1971). In order to generate a 3D model of the viral particle in real space (for visualization and interpretation), an inversion of the Fourier transform is calculated. For this to work, the 2D Fourier transforms of the particles need to be regularly spaced, which is rarely the case with real data. These missing data points are interpolated by, for instance, a cylindrical expansion using the Fourier-Bessel inversion in polar space.
coordinates (Crowther, 1971). The coefficients for the corresponding expansion in real space are subsequently determined and this polar expansion is interpolated into Cartesian coordinates to complete the reconstruction (Crowther, 1971; Fuller et al., 1996).

Alternatively, and the method used in this thesis, is the reconstruction of the 3D volume using projection matching based on an existing model (Baker and Cheng, 1996; Ji et al., 2006; Yan, Sinkovits, and Baker, 2007). This polar Fourier transform method used to rely on the availability of a good enough starting model; for instance, a similarly sized and shaped icosahedrally-ordered virus (Baker and Cheng, 1996). However, with recent software modifications such a model is no longer needed, but is generated ab initio from the particle projection images (Yan et al., 2007; Yan, Sinkovits, and Baker, 2007). In the projection matching method, the initial 3D model is projected into 2D images and these images are used as references when determining origins and orientations of each particle (Figure 5). Once the orientations of the 2D particle images have been determined to satisfaction, a discrete Fourier transform is calculated for each image (Marinescu and Ji, 2003). These discrete Fourier transforms together with the knowledge of the particle orientation are subsequently used to interpolate the values of the 3D discrete Fourier transform, which is inverse transformed in order to generate a real space model (Marinescu and Ji, 2003).

Recently, 3DEM has advanced towards atomic resolution, as demonstrated for many icosahedral viruses. During the last five years, several virus structures have been solved to below 5 Å resolution allowing for the main polypeptide chain to be traced and for the bulkier amino acid side chains to be discerned. These viruses include Venezuelan equine encephalitic virus (Zhang et al., 2011a), cytoplasmic polyhedrosis virus (Cheng et al., 2011), the structures of bacteriophages P22 (Chen et al., 2011) and ε15 (Jiang et al., 2008), aquareovirus infectious subvirion particle (Zhang et al., 2010), rotavirus (Zhang et al., 2008), and adenovirus (Liu et al., 2010). Importantly, the stage for such high-resolution structures was set by Böttcher and Conway with colleagues (Böttcher, Wynne, and Crowther, 1997; Conway et al., 1997) who were able to trace the main chain of the hepatitis B core protein.

Entire viral particle structures to atomic resolution are still mainly solved by X-ray crystallography. Summarizing briefly, the virus to be crystallized is needed in mg quantities and has to produce diffraction quality crystals. Once these crystals are readily available, a native dataset is collected and processed; i.e. the diffraction spots are indexed (Figure 6). In an X-ray crystallographic experiment one only measures the intensities of Bragg reflections which are proportional to the square of the complex structure factors. However, to calculate the electron density by inverse Fourier transform the phases of the structure factors are also required, but they cannot be directly measured (summarized by Oksanen and Goldman [Oksanen and Goldman, 2010]). Therefore, diffraction data collected on isolated proteins is phased either using: i) homologous structures; ii) using a selenemethionine labeled derivate of the protein; or iii) using heavy-metal labeled derivates of the protein (Figure 6). Diffraction data from viral particles is typically phased using an existing structure obtained either from cryoEM or X-ray crystallographic studies (Abrescia et al., 2004; Abrescia et al., 2008; Grimes, Fuller, and Stuart, 1999; Reddy et al., 2010).
Figure 5. A schematic diagram of the 3D image reconstruction process. Modified from Baker et al. (Baker, Olson, and Fuller, 1999) with permission from the publisher.
Figure 6. A schematic representation of the various steps included in structure determination using X-ray crystallography. The figure was redrawn from a schematic by Oksanen (unpublished). Se-Met: selenomethionine; MIR: multiple isomorphous replacement; SIRAS: single isomorphous replacement with anomalous scattering; MAD: multiple wavelength anomalous dispersion.
1.7 Genome packaging into preformed capsids

Viral genomes are enclosed inside a protein capsid for protection against the environment. In rod-shaped viruses, such as the ssDNA virus M13 (Figure 2), the capsid proteins are usually assembled around the viral genome. Thus, the size of the virus is proportional to the length of the virus’ single-stranded genome. However, all viruses with double-stranded genomes package their nucleic acids into a preassembled procapsid (Guo and Lee, 2007). Packaging of the genomic nucleic acid into procapsids is a complex process that involves the injection of the genome into the procapsid via ATP (or NTP) powered hydrolysis, machinery to utilize this energy, structural arrangement of the encapsidated genome, and its transformation into a highly condensed, nearly crystalline state, concomitant with the maturation of the capsid (Petrov and Harvey, 2007). The vast majority of our current knowledge on genome packaging into preformed procapsids stems from studies on tailed dsDNA bacteriophages such as φ29 and T4 (reviewed by [Rao and Feiss, 2008]). Furthermore, cryoEM and X-ray crystallographic studies on the icosahedral dsRNA bacteriophages φ6, φ8, and φ12 and their RNA genome packaging NTPase P4 has shed light on the encapsidation of ssRNA by these dsRNA viruses (Butcher et al., 1997; Huiskonen et al., 2006; Huiskonen et al., 2007a; Jäälinoja, Huiskonen, and Butcher, 2007; Kainov et al., 2004; Kainov et al., 2003; Kainov, Tuma, and Mancini, 2006; Lisal et al., 2004; Mancini et al., 2004).

1.7.1 Common motifs in genome packaging ATPases

Despite the variety of mechanisms used in genome translocation, the common theme to the motor proteins generating energy via the hydrolysis of ATP is that the active site residues that catalyze this reaction are highly conserved. Summarizing briefly to enlighten the discussion to follow, the ATP-binding domain of these genome packaging enzymes contains the phosphate-binding loop (P-loop) (also known as the Walker A sequence motif) as well as the Walker B motif (Hanson and Whiteheart, 2005; Iyer et al., 2004b; Walker et al., 1982). The consensus sequences for the respective motifs are GXXXGK(T/S) and hhhhDE, where X denotes any amino acid and h any hydrophobic amino acid (Hanson and Whiteheart, 2005; Walker et al., 1982). Amino acids from both sequence motifs take part in nucleotide binding and hydrolysis: most importantly the conserved lysine in the Walker A motif is responsible for nucleotide binding and the conserved glutamate in the Walker B motif is responsible for activation of water for the hydrolysis reaction (Goetzinger and Rao, 2003). The Mg\(^{2+}\) ion required for ATP hydrolysis can be coordinated either by the conserved aspartate in the Walker B domain (Hanson and Whiteheart, 2005; Iyer et al., 2004a; Walker et al., 1982) or by a conserved serine in the Walker A domain (Massey et al., 2006; Subramanya et al., 1996). So-called arginine fingers facilitate the formation of the transition state (Nadanaciva et al., 1999) and are inserted into the catalytic site in response to a conformational change prior to the catalytic step. It has also been suggested that the arginine-finger is involved in the propagation of conformational changes within the multimer (Mancini et al., 2004). The arginine finger can be located either on the same (Sun et al., 2007) or on a separate subunit (Mancini et al., 2004) as the Walker A and B motifs.

1.7.2 Genome packaging in dsDNA tailed bacteriophages

As summarized by Rao and Feiss (Rao and Feiss, 2008), the key components in the genome encapsidation of tailed bacteriophages, such as φ29, P22, HK97 or T4, are the ring-forming portal
through which the genome is translocated and the translocating motor. A narrow stalk-like domain on the connector protrudes on the outside of the prohead, providing a platform for assembly of the translocating motor. The translocating motor – the terminase – is usually hetero-oligomeric with a small subunit involved in the recognition of the incoming DNA and a large subunit containing the ATPase domain, a motif for docking at the portal vertex, and in some cases, an endonuclease cutting the bacteriophage genome upon headful packaging. Typically, the small subunit recognizes a specific sequence on the phage genome, such as the pac site on P22 (Wu et al., 2002). Additionally, in φ29 one of the genomic DNA packaging enzymes is a hexameric packaging RNA (pRNA) (Guo et al., 1987a; Guo, Erickson, and Anderson, 1987b; Shu et al., 2007).

One of the most studied genome encapsidation machineries is that of the dsDNA bacteriophage φ29. The structure of the podovirus φ29 is well characterized, with the first description on its morphology published already in 1966 (Anderson, Hickman, and Reilly, 1966). Almost 40 years later, a nanometer resolution reconstruction has been published showing that the fold of the φ29 major capsid protein (gp8) is similar to that of HK97 and P22 (Morais et al., 2005) (Figure 3d). Although the genome packaging process of φ29 is well studied with the involved key molecules identified and the machinery well utilized in various in vitro assays, the only atomic resolution structure currently available is for the connector protein (Simpson et al., 2000). The φ29 genome is linear dsDNA with covalently linked terminal proteins (gp3) – much like the PRD1 genome with its covalently linked P8 protein (Bamford and Mindich, 1984). The φ29 procapsids consist of the aforementioned connector protein (gp10) (Simpson et al., 2000), the scaffolding protein (gp7) (Morais et al., 2003), the major capsid protein (gp8) (Morais et al., 2005), the head fiber protein (gp8.5) (Morais et al., 2005), and the phage-encoded pRNA molecule (Tao et al., 1998) (Figure 7a). The energy for genome packaging is provided by the ATPase (gp16), which has been reported to bind DNA non-specifically, as opposed to what has been reported for P22 (Lee and Guo, 2006; Wu et al., 2002). In φ29 it has been shown that the presence of pRNA or DNA stimulates the ATPase activity of gp16, as does the presence of procapsids containing pRNA (Guo, Peterson, and Anderson, 1987c; Guo, Erickson, and Anderson, 1987b). Furthermore, it has been shown that every ATP molecule hydrolyzed translocates 2 bp of DNA (Guo, Peterson, and Anderson, 1987c). Additionally, in φ29 genome packaging machinery has been studied in detail using single molecule experiments. As an example, using optical tweezers it has been shown that the initial rate of φ29 genome translocation is 100 bp s⁻¹ and the machinery can package against a pressure of 57 pN (Smith et al., 2001).

The structure of the T4 gp17 ATPase is the only available atomic model for a dsDNA viral genome packaging ATPase (Sun et al., 2008; Sun et al., 2007) (Figure 7b). In T4, the packaging machinery consists of the dodecameric gp20 portal protein, the gp17 large terminase subunit, and the gp16 small terminase subunit (Rao and Black, 2010). Five copies of gp17 bind to the portal (Sun et al., 2008). The N-terminal domain of gp17 contains the ATPase domain and the C-terminal domain the endonuclease cleaving the genome upon headful packaging. Simplifying vastly, binding of T4 genomic DNA to the C-terminus of gp17 and an ATP-molecule to the N-terminus of the same gp17 is proposed to induce a conformational shift, resulting in the insertion of the arginine-finger residue R162 into the active site to trigger ATP-hydrolysis. Subsequently, electrostatic forces are proposed to realign the N- and C-terminal domains causing the translocation of two base pairs of DNA. The DNA binding site of gp17 is suggested to flex between “relaxed” and “tensed” (DNA bound) states, with one monomer of the pentamer being in the tensed conformation at any given time point (Sun et al.,
Furthermore, it has been shown that the portal gp20 of T4 is not required for the suggested packaging mechanism. Rather, it functions as a nucleation site for capsid assembly, as a channel for DNA passage (Sun et al., 2008), and additionally as a platform for tail assembly.

### 1.7.3 Genome encapsidation in the dsRNA virus φ12

The hexameric dsRNA genome packaging ATPase P4 of φ12 functions rather differently (Mancini et al., 2004) (Figure 7c). In contrast to T4, φ12 lacks a tail, packages ssRNA into a procapsid consisting of only four proteins, and after synthesis of minus strand RNA, is subsequently enveloped before host cell lysis (Poranen and Tuma, 2004). The P4 monomer is composed of three domains. The central core together with the C-terminal domain forms the canonical Rossmann-type nucleotide binding site, with a twisted eight-stranded β-sheet. The nucleotide binding site lies at the interface between two subunits in the hexamer. Residues in the nucleotide binding pocket are involved in catalysis, stabilization of substrate binding, and as γ-phosphate sensors. The amino acids K136 and T137 of the Walker A motif are part of the P-loop in φ12 P4 and interact with the α- and β-phosphate groups of substrate and product. D189 coordinates the metal ion used in catalysis and E160 activates the water molecule for nucleophilic attack on the γ-phosphate. The arginine finger in φ12 P4 is contributed by R272 and R279 (Kainov et al., 2008; Mancini et al., 2004).

Again, simplifying vastly, binding of an ATP molecule seems to lock the P4 in the “up” conformation, making it conformationally susceptible to interact with the incoming ssRNA (Mancini et al., 2004). Upon ATP hydrolysis, the P4 conformation is swiveled to the “down” conformation. The P-loop of φ12 P4 is likened to a piston that drives a helix (α6) down and with it the bound RNA (Mancini et al., 2004). As in T4 gp17, the ATP hydrolysis and genome translocation are proposed to induce conformational changes in adjacent subunits of the multimeric motor (Kainov et al., 2008; Mancini et al., 2004). In φ12 P4, it has been suggested that one catalytic circuit around the hexameric ring would consume six ATP molecules and translocate eleven bases of RNA (Mancini et al., 2004). This is in line with what has been described for φ29 gp16 and T4 gp17, in which one ATP molecule hydrolyzed translocates two base pairs of dsDNA (Guo, Peterson, and Anderson, 1987c; Sun et al., 2008). Calculated to a catalytic circuit around the pentameric T4 gp17 ring, this corresponds to five ATP molecules per 10 base pairs of translocated DNA.

### 1.7.4 PRD1 genome packaging

The current knowledge on the packaging of PRD1 comes mainly from in vitro studies using cell lysates, as the P9 ATPase has been recalcitrant to structural characterization (Strömsten, Bamford, and Bamford, 2005; Žiedaite et al., 2009). The membrane proteins P20 and P22 of PRD1 can be likened to the connector protein of tailed viruses. Based on in vitro studies using the PRD1 genome, empty procapsids, and P9, it has been concluded that the packaging rate of the PRD1 machinery is 340 bp s⁻¹ by observing the appearance of the first infectious particles (Žiedaite et al., 2009). This value is similar to what has been established for φ29 (100 bp s⁻¹) or T4 (700 bp s⁻¹) as determined by optical tweezer experiments (Fuller et al., 2007; Smith et al., 2001). It has been furthermore demonstrated that the PRD1 genome terminal protein P8 is required for genome packaging, as neither naked PRD1 DNA, or DNA from Bam35 or PM2 was incorporated into the PRD1 capsid (Žiedaite et al., 2009).
Figure 7. Genome packaging machineries of bacteriophages. (a) A montage representation of the φ29 genome packaging machinery calculated from difference maps. The connector is shown in green, the pRNA in magenta, and the ATPase gp16 in blue (A-D). In (C) a dsDNA molecule is depicted in the central channel of the motor complex. The complex is shown in relation to the φ29 prohead in (D). Reprinted from (Morais et al., 2008) with permission from the publisher. (b) Structure of the T4 gp17 (colored) (PDB: 3dpe) and the N-terminal ATPase domain in complex with ADP (grey, ADP indicated with black arrow) (PDB: 2o0j) (Sun et al., 2008; Sun et al., 2007). The C-terminal DNA-binding domain is highlighted with a black square. (c) Structure of φ12 P4 (PDB: 1w4b) (Mancini et al., 2004). The ADP molecule is indicated with a black arrow and the α6-helix is shown in dark blue. Panels (b) and (c) were rendered in PyMOL.
1.7.5 The driving force behind genome translocation

There are several competing models for the mechanism of viral genome translocation, as reviewed by Guo and Lee (Guo and Lee, 2007). One of these models – the gyrase-driven packaging model – suggests that the terminase domain of T4 gp17 and φ29 gp16 introduces supercoiling into the genome to be packed near the apex of the portal. According to this model, the supercoiling would compress the DNA to be packed, which after encapsidation would be relaxed by introducing nicks into the genomic DNA (Grimes and Anderson, 1997; Guo and Lee, 2007). The shortcoming of this hypothesis is that the φ29 genome packaging machinery has been shown to package nicked DNA (Moll and Guo, 2005). Secondly, the motor-ratchet model suggests that the difference between the low pressure inside the procapsid and the high external pressure generates osmotic pressure that drives translocation (Serwer, 2003). Thirdly, as many connector proteins have been suggested or shown to exist as hexamers or dodecamers and to be located on a five-fold axis of symmetry in the virus’ icosahedral capsid shell, this symmetry mismatch leading to the rotation of the portal has been suggested to be the driving force of genome translocation (Hendrix, 1978). This idea has, however, been challenged in T4 using an in vitro assay (Baumann, Mullaney, and Black, 2006). Lastly, according to the supercoiled DNA wrapping model, the genome wraps around the portal vertex and the presumed rotation of the connector allows DNA passage into the procapsid (Guo and Lee, 2007). This model assumes that the genome does not enter the procapsid via the channel in the connector, but rather on the outside – this model is, however, in contrast to what has been described for the φ29 connector showing DNA positioned in the central channel (Simpson et al., 2000) (Figure 7a).

1.8 Life at extremes

The term “extremophile” was first coined in 1974 to describe organisms that thrive in extreme conditions from the point of view of man. Most of these extremophiles are micro-organisms. The currently known upper temperature for life is 121°C for archaea (Kashefi and Lovley, 2003), 95°C for bacteria, and 62°C for single-celled eukaryotes (Gerday and Glansdorff, 2007). For multicellular eukaryotes the upper temperature for life is 50°C (Gerday and Glansdorff, 2007). The majority of extremophiles identified to date – approximately 300 species – belong to the archaeal domain (Egorova and Antranikian, 2007). Extremophilic organisms can be described using a variety of terms, some of which are presented in Table 1. In accordance with their harsh living environments, most of the enzymes isolated from hyperthermophiles have enhanced thermostability as compared to their mesophilic homologues (Rees and Adams, 1995). This enhanced thermostability has been utilized both in studies concerning fundamental biologic questions, such as factors contributing to protein structure and stability, as well as for the development of technological applications that require protein stability at high temperatures (Rees and Adams, 1995). Furthermore, when analyzing 16S ribosomal DNA sequences for phylogenetic typing, it has been observed that in the phylogenetic tree, hyperthermophiles occupy the deepest and shortest branches, indicating that the last universal common ancestor (LUCA) was a thermo- or hyperthermophile (Di Giulio, 2000; Schwartzman and Lineweaver, 2004).
Table 1. Characteristics of extremophiles with descriptions. Adapted from (Gerday and Glansdorff, 2007).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>acidophile</td>
<td>an organism with an optimal growth rate at pH values of 3 or below</td>
</tr>
<tr>
<td>alkaliphile</td>
<td>an organism with an optimal growth rate at pH values of 9 or above</td>
</tr>
<tr>
<td>halophile</td>
<td>an organism that requires at least 0.2 M salt for growth</td>
</tr>
<tr>
<td>piezophile</td>
<td>an organism that grows optimally at high hydrostatic pressure</td>
</tr>
<tr>
<td>psychrophile</td>
<td>an organism with a maximum growth temperature of 20°C (and an optimal</td>
</tr>
<tr>
<td></td>
<td>growth temperature of 15°C or lower)</td>
</tr>
<tr>
<td>thermophile</td>
<td>an organism that thrives at temperatures between 60 and 80°C</td>
</tr>
<tr>
<td>hyperthermophile</td>
<td>an organism that has an optimal growth temperature of at least 80°C and has</td>
</tr>
<tr>
<td></td>
<td>a maximum growth temperature of over 90°C</td>
</tr>
<tr>
<td>endolith</td>
<td>an organism that lives within stony substances</td>
</tr>
<tr>
<td>metallotolerant</td>
<td>an organism that tolerates high levels of heavy metals</td>
</tr>
<tr>
<td>toxitolerant</td>
<td>an organism that tolerates high levels of toxic agents, such as toluene or</td>
</tr>
<tr>
<td></td>
<td>benzene</td>
</tr>
<tr>
<td>radioresistant</td>
<td>an organism that resists high levels of ionizing radiation</td>
</tr>
</tbody>
</table>

1.8.1 Applications of extremophiles and their proteins

As described above, the extremophiles colonize extreme environments. Due to their harsh habitat they possess unusual properties, which are potentially valuable resources in the industry and in the development of novel biotechnological processes. The extremophiles, their cellular components, and the proteins they encode are used in waste treatment (sulfur oxidizing extremophiles), in decontamination of polluted soil, in pulp bleaching (xylanases, hemicellulases), in cellulose degradation (endoglucanases, β-glucosidases), in baking and brewery (amylases), as detergents (proteases, lipases), in pharmaceuticals (lipids and liposomes, especially those from archaea [archaeosomes]), and in molecular biology (polymerases, ligases, alkaline phosphatases, restriction endonucleases) (Antranikian and Egorova, 2007; Vieille and Zeikus, 2001).

To highlight a few success stories that have been achieved using extremophilic organisms and their proteins, the invention of PCR and the determination of the structure of the ribosome will be presented briefly. Although there is some controversy over who invented PCR, the honor is usually given to Kary Mullis, who received the Nobel Prize in chemistry 1993 for his invention. PCR is used today in a wide variety of applications, such as DNA cloning and sequencing, the diagnostics of hereditary diseases and in forensic sciences and paternity testing for the identification of genetic fingerprints. Simplifying, the PCR reaction uses two short oligonucleotides (primers), deoxynucleosides and a polymerase enzyme to amplify the desired DNA region. The amplification is achieved by cycling three different temperatures: i) the denaturation step, where the primers dissociate from the DNA to be amplified; ii) the annealing step, where the primers bind to the DNA to be amplified; and iii) the elongation step during which the polymerase uses the deoxynucleosides to build a new DNA strand. Initially, the Klenow fragment of *Escherichia coli* DNA polymerase I was used in the elongation step (Mullis and Faloona, 1987; Saiki *et al.*, 1985). This enzyme was, however, destroyed during the denaturation step and had to be added freshly to the reaction during each cycle. PCR technology became much more feasible when the heat-stable *Taq* polymerase, isolated
from *Thermus aquaticus* (Chien, Edgar, and Trela, 1976), was harnessed to perform the elongation step. Currently, the selection of commercial heat-stable polymerases engineered for PCR is vast.

Another Nobel Prize winning research accomplishment was the determination of an atomic structure of the ribosome. The atomic structures of the archaeal 50S and bacterial 30S ribosome subunits were published almost simultaneously by two groups in 2000 and a year later the structure of the whole bacterial ribosome was published (Ban et al., 2000; Wimberly et al., 2000; Yusupov et al., 2001). Importantly, the knowledge about the structure of the ribosome is essential for our understanding on how this complex machinery functions in protein translation. Furthermore, the paper by Wimberly et al. (Wimberly et al., 2000) was accompanied by a paper describing the interactions of the 30S subunit with three different antibiotics, highlighting how they interact with the bacterial ribosome and providing directions on how to design novel antibiotics that target bacterial protein synthesis (Carter et al., 2000). What is common to all of these structures is that they were achieved using ribosomes from extremophilic organisms; the structure of the archaeal 50S subunit was achieved using *Haloarcula marismortui* as a source (Ban et al., 2000) and that of the 30S subunit and 70S ribosome using *Thermus thermophilus* as a source (Carter et al., 2000; Wimberly et al., 2000; Yusupov et al., 2001).

### 1.8.2 Extremozymes from viruses

Although many of the extremozymes – enzymes from extremophilic organisms or viruses – have been isolated from cells, there are also examples of viruses providing a source for these enzymes. In 1967, N. Welker (Welker, 1967) purified a peptidase from a thermophilic bacteriophage, becoming the first one to isolate a thermophilic protein from a virus. The maximum lytic activity (degradation of the cell wall peptidoglycan of *Bacillus stearothermophilus* was observed at pH 6.3 and 55°C. The peptidase was suggested to be a valuable tool in the study of the cell walls of thermophiles.

During the last 10 years, several extremophilic viral enzymes with possible biotechnical applications have been characterized. Two of these enzymes and their potential applications will be described below. Blondal et al. (Blondal et al., 2003; Blondal et al., 2005) isolated and characterized the RNA ligase 1 from two different thermophilic bacteriophages, RM378 and TS2126. Both of these ligases work on both RNA and ssDNA, and have been suggested to have applications in ligase-mediated rapid amplification of cDNA ends as well as other RNA and DNA ligation applications in which high-temperatures are required. The enzymes ability to ligate ssDNA might be useful in solid-phase, single-stranded gene synthesis technology. Furthermore, as the enzyme ligates ssDNA and RNA it could also be used to circularize these molecules for enhanced protection against nucleases. Song and Zhang (Song and Zhang, 2008) were the first ones to isolate and characterize a novel, thermostable, non-specific, nuclease from the thermophilic bacteriophage GBSV1. This new nuclease, GBSV1-NSN, was shown to be active at temperatures ranging from 20-80°C, with an optimum at 60°C, and to have a pH optimum of 7. Furthermore, GBSV1-NSN was shown to hydrolyze linear and circular dsDNA, ssDNA, and RNA. The enzyme is suggested to be useful in molecular biology applications performed at high temperatures, such as the determination of nucleic acid structure and the removal of nucleic acids during protein purification.
1.8.3 What stabilizes a thermophilic protein?

Engineering proteins for enhanced thermostability would have applications in many industrial processes. To the disappointment of many, even though some general trends for enhanced thermostability can be seen when comparing mesophilic and thermophilic proteins, no new amino acids, covalent modifications or structural motifs have been identified that would explain the ability of thermozymes to function at higher temperatures than mesozymes (Fields, 2001). The methods to analyze changes contributing to increased thermostability are most often either sequence or structure based comparisons between thermophilic and mesophilic homologous proteins (Haney et al., 1999; Querol, Perez-Pons, and Mozo-Villarias, 1996; Vogt, Woell, and Argos, 1997). Alternatively, protein engineering, usually via site-directed mutagenesis, is the preferred method when experimentally analyzing enhanced thermal stability (Heikinheimo et al., 1996).

The most extensive study on thermal protein stability so far has been done by Haney et al. (Haney et al., 1999), who performed sequence based comparison between 115 hyperthermophilic and mesophilic proteins. According to their results, hyperthermophilic proteins tend to have a reduced number of serine, asparagine, glutamine, threonine, and methionine residues and an increased number of isoleucine, arginine, glutamic acid, lysine, and proline residues. The reduced numbers of asparagine and glutamine residues can be explained by their deamination at high temperatures leading to protein inactivation, as observed in α-amylase (Tomazic and Klibanov, 1988). Additionally, the increased number of proline residues decreases the entropy of unfolding (Vieille and Zeikus, 2001), as it restricts the conformations allowed for the preceding residue. Furthermore, specific amino acid replacements observed were serine to lysine or alanine and lysine to arginine, when moving from a mesophilic protein to a hyperthermophilic one. The stabilizing effect of the lysine to arginine substitution, as well as the glycine to alanine substitution, have been reported by Querol et al. (Querol, Perez-Pons, and Mozo-Villarias, 1996). Overall, the strongest correlations for increased thermal stability relating to residue composition can be summarized as follows: i) decrease in uncharged polar residues, with an increase in charged residues; ii) increased residue hydrophobicity; and c) increased residue volume. The effect of these changes has been summarized by Fields (Fields, 2001). An increase in charged residues might increase the stability of a protein via the increased formation of ion pairs and networks. Furthermore, the increased residue hydrophobicity makes the hydrophobic effect stronger. This is a relevant feature, as the major driving force in protein folding and stability is considered to be the hydrophobic effect (Dill, 1990). Additionally, the increased residue volume may enhance the stability via more efficient packing (Fields, 2001).

Reflecting the residue content of a protein are the bonds and interactions they can form. Interactions contributing to stability are hydrogen bonds, electrostatic interactions (ion pairs), aromatic interactions, hydrophobic interactions, disulfide bonds, intersubunit interactions and oligomerization, and metal binding (Li, Zhou, and Lu, 2005; Vieille and Zeikus, 2001). Vogt et al. (Vogt, Woell, and Argos, 1997) have observed that the number of internal hydrogen bonds increase by 11.7 per chain per 10°C (0.036 per residue) and the number of ion pairs (salt bridges) increase by 1.8 per chain per 10°C (0.007 per residue). In line with this, the α-amylase structure has been reported to be stabilized by an increased number of salt-bridges making it less prone to thermal inactivation (Tomazic and Klibanov, 1988). Disulfide bridges are believed to stabilize a protein by decreasing the entropy of the protein’s unfolded state (Matsumura, Signor, and Matthews, 1989). It
has been observed that the unfolding temperature of the T4 phage lysozyme increased by 23.4°C by the addition of three engineered disulfide bridges. Furthermore, important factors for protein stabilization are the docking of the N- and C-termini to the core of the protein or to other thermini, anchoring loops, and post-translational modifications, with glycosylation and lysine methylation being the most prevalent ones identified so far (Vieille and Zeikus, 2001). Lastly, extrinsic factors are known to stabilize proteins. These include stabilization by salt, by the substrate or by pressure (Vieille and Zeikus, 2001).

1.9 Viruses from extreme environments

As the theme of this thesis is the characterization of extremophilic, icosahedral viruses, a selection of these will be described below. When this thesis was initiated, only two extremophilic icosahedral viruses, SH1 and STIV, were described in detail. However, as the first publication included in this thesis is on *Thermus* infecting viruses, I will begin with a short introduction of those.

1.9.1 Viruses infecting *Thermus* species

Several bacteriophages infecting *Thermus* species have been isolated and characterized. These viruses consist of the filamentous PH75 (Pederson et al., 2001), the siphoviruses P23-45, P74-26 (Berdygulova et al., 2011; Minakhin et al., 2008) and TSP4 (Berdygulova et al., 2011; Lin et al., 2010; Minakhin et al., 2008), the myoviruses φTMA (Tamakoshi et al., 2011) and φYS40 (Naryshkina et al., 2006; Sakaki and Oshima, 1975; Sevostyanova et al., 2007), and the unclassified TS2126 (Blondal et al., 2005). The most extensive study thus far, conducted by the Promega Corporation (Yu, Slater, and Ackermann, 2006), characterizes 115 of the bacteriophages infecting *Thermus* species. These viruses belong to the *Myoviridae, Siphoviridae, Tectiviridae*, and *Inoviridae* families (Figure 2). Based on these studies, over 50% of the phage isolates were non-tailed. This is in strong contrast to what has been described for the phages characterized so far – of the previously characterized phages, merely 4% are non-tailed while the majority are tailed (Ackermann, 2007). This observed bias is in part due to the fact that certain *Thermus* strains are selectively prone to infections by tectiviruses and inoviruses (Yu, Slater, and Ackermann, 2006).

Of the 11 siphoviruses identified in the study conducted by Promega (Yu, Slater, and Ackermann, 2006), P23-45 and P74-26 have been studied in more detail (Berdygulova et al., 2011; Minakhin et al., 2008). P23-45 infects *Thermus* species of the genus *aquaticus* and *thermophilus*, whereas P74-26 is specific to *Thermus thermophilus* (Yu, Slater, and Ackermann, 2006). The genomes of P23-45 and P74-26 are dsDNA of 84,202 bp and 83,319 bp, respectively. P23-45 encodes 117 and P74-26 encodes 116 predicted ORFs, of which 113 are common to both. Three of these predicted ORFs are shared by hypothetical proteins from φIN93 (Matsushita and Yanase, 2009a) and φYS40 (Naryshkina et al., 2006), indicating that *Thermus* phages have access to a common gene pool, as has been observed for other phages (Minakhin et al., 2008; Pedulla et al., 2003). Furthermore, the gene order in P23-45 and φIN93 appears conserved, consistent with modular phage evolution (Minakhin et al., 2008; Pedulla et al., 2003). Since predicted ORFs in P23-45 are similar to proteins in mesophilic phages, it suggests that the gene pool accessible by these thermophilic phages is not limited to genes from thermophiles (Minakhin et al., 2008). Mass-spectrometric analysis of highly purified P23-45 virions revealed 16 structural proteins, seven of which show homology to hypothetical proteins or
conserved domains in phages and bacteria (Minakhin et al., 2008). Importantly, one of these proteins (gp86) has a putative function as a portal protein (Minakhin et al., 2008).

The Thermus thermophilus myovirus φYS40 was isolated and characterized already in 1974 (Sakaki and Oshima, 1975), and was rediscovered by another group 32 years later (Naryshkina et al., 2006). Initial studies on the virus focused on characterizing its thermostability. Unsurprisingly, φYS40 has been shown to be most stable in the water of the host spring it has been isolated from, as opposed to various buffer solutions or bacterial growth media (Sakaki and Oshima, 1975). The genome of φYS40 is 152,372 bp long dsDNA with 170 predicted ORFs (Naryshkina et al., 2006). The predicted proteins of φYS40 are between 43 and 1744 amino acids long. Based on mass-spectrometric analyses, 33 of these are structural proteins (Naryshkina et al., 2006).

The other Thermus thermophilus myovirus described in detail, φTMA, has a similar morphology and genome length to that of φYS40 (Naryshkina et al., 2006; Tamakoshi et al., 2011). The genome of φTMA is 151,483-bp long circular dsDNA with 168 predicted ORFs, of which 159 are shared with φYS40 (Tamakoshi et al., 2011). The amino acid identity between orthologs of φTMA and φYS40 ranges from 43 % to 100 % with an average of 94 %. A striking difference between the φTMA and φYS40 is the presence of a transposase and a resolvase gene in φTMA (Tamakoshi et al., 2011). In addition to the transposase gene found on φTMA, the unclassified Thermus virus φIN93 has been shown to contain a transposase encoded on an insertion element that has transposed from Thermus thermophilus T22 into φIN93 (Matsushita and Yanase, 2009b). It is currently unknown whether the transposase and resolvase proteins encoded by these genes are active. Both φTMA and φYS40 are lytic phages, even though no holin or lysozyme genes have so far been identified in either genome (Naryshkina et al., 2006; Tamakoshi et al., 2011).

φTMA has a broader host-cell specificity than φYS40 and is able to infect Thermus thermophilus, T. flavus, and T. aquaticus (Tamakoshi et al., 2011). The infectivity of φTMA is dependent on the presence of type IV pili on the host-cell surface, suggesting that this could be the phage receptor. In T-even related phages (such as T4) it has been showed that the host specificity resides in the C-terminal portion of the gp37 tail fibers that bind to the receptor on the host-cell surface (Tétart et al., 1996). The C-terminal portion of the tail fibers is strikingly different in φTMA and φYS40. This may explain the difference in host-cell specificity of the two phages, if the tail fibers of thermophilic phages have the same function as those of mesophilic phages (Tamakoshi et al., 2011).

1.9.2 Sulfolobus Turreted Icosahedral Virus (STIV)

STIV was the first icosahedral archaeal virus to be described (Rice et al., 2004). STIV was observed in the culture supernatant of Sulfolobus solfataricus isolate YNPRC179 (Yellowstone National Park Rabbit Creek (thermal area) 179), an aerobic facultative chemolithotroph (Rice et al., 2004). The optimal growth conditions of STIV were determined to be pH 3.3 and 80°C in DSM88 minimal media (Rice et al., 2004). The buoyant density of STIV in Cs₂SO₄ equilibrium centrifugation was 1.236 g cm⁻³ (Rice et al., 2004), which is close to that of PM2 and PRD1 in CsCl equilibrium centrifugation – 1.28 and 1.265 g cm⁻³, respectively – indicating the presence of lipids (Bradley and Rutherford, 1975; Espejo and Canelo, 1968). The STIV genome consists of circular dsDNA of 17,663-bp with 37 predicted ORFs (Maaty et al., 2006; Rice et al., 2004).
1.9.2.1 Three-dimensional STIV architecture and the structural STIV proteins

The initial three-dimensional reconstruction of STIV to 27 Å resolution revealed an icosahedral virus with an internal membrane (Rice et al., 2004). The diameter of the capsid is 74.7 nm vertex base-to-vertex base, 73 nm edge-to-edge, and 69.2 nm facet-to-facet (Khayat et al., 2005). The T-number of STIV is $T = 31d$ (Khayat et al., 2005; Rice et al., 2004). Characteristic for the virion, giving STIV its name, are the large, turreted attachments with decorating petal-like appendages at the viral five-fold vertices (Rice et al., 2004). These turrets are suggested to be involved in host-cell recognition and attachment (Rice et al., 2004). After the initial three-dimensional STIV virion structure, three new viral morphologies have been described (Khayat et al., 2010). The initial STIV structure published by Rice et al. (Rice et al., 2004) will hereafter be referred to as the decorated STIV virion (Khayat et al., 2010) (Figure 8a, left hand side). The three other virion morphologies described are the: i) undecorated particle, missing the petal-like appendages (Figure 8a, right hand side); ii) the nodular, disassembled particle, missing the turrets altogether; and iii) the spherical lipid core (Khayat et al., 2010).

![Figure 8. Structure of STIV. (a) Structure of the decorated STIV particle, petal-like appendages attached to the five-fold symmetric turrets (left) and the undecorated particle (right). (b) The turret of the undecorated particle is rendered in yellow and the petals of the decorated particle are in purple. A postulated cementing protein located at the base of the turret in the decorated particle (discussed below) is shown in blue. (c) Densities corresponding to the petals and cementing protein colored as in (b). Reprinted from (Khayat et al., 2010) with permission from the publisher.](image)

The structure of the major capsid protein B345 of STIV was solved in 2005 (Khayat et al., 2005). B345 is a double β-barrel protein similar to those of PRD1, adenovirus, and PBCV-1 (Athappilly et al., 1994; Benson et al., 2002; Nandhagopal et al., 2002); the superposition of all the structures yields a RMSD
< 2.2 Å (Khayat et al., 2005). Furthermore, the average center-to-center distance between the B345 capsomers in the STIV capsid shell is 73.8±1.4 Å, virtually identical to the distance described for the PRD1 P3 capsomers (73.5±0.3 Å) and PBCV-1 (74 Å) (Abrescia et al., 2004; Khayat et al., 2005; Nandhagopal et al., 2002). In the decorated STIV virion the 26 C-terminal amino acids of B345 (2.7 kDa) are suggested to form a helix that contacts the outer membrane leaflet, much like the N-terminal helix of PRD1 P3 (Abrescia et al., 2004; Khayat et al., 2010; Khayat et al., 2005; San Martin et al., 2002). In the undecorated particles the C-terminus of B345 seems to take a different conformation. Reaching towards the viral membrane, the modeled helix is located between the two β-barrels of each B345 subunit (Khayat et al., 2010). The main difference between the modeled C-terminal helices in the decorated and undecorated structures is the location where the helix reaches down towards the membrane (Khayat et al., 2010; Rice et al., 2004).

The B345 protein is remarkably stable. It is stable in the acidic growth media at high-temperatures for weeks and 7M GuHCl is required to fully denature the protein (Khayat et al., 2005). There are several features about B345 that might contribute to the stability of the protein. Firstly, B345 seems to be more tightly packed than the major capsid proteins of the aforementioned viruses based on cavity volume calculations. PRD1 P3, for instance, has twice the calculated cavity volume of B345 (Khayat et al., 2005). Tight packing of structures and filling of cavities is known to increase the thermal stability of proteins (Ishikawa et al., 1993). Secondly, the proline content of B345 is 1.3 % higher than that of an average protein sequence (Khayat et al., 2005; McCaldon and Argos, 1988). Proline residues have been suggested to increase the thermal stability of a protein by reducing its flexibility and decreasing the entropy of the unfolded state (Watanabe et al., 1991). Thirdly, B345 has shorter loops than the other aforementioned major capsid proteins, which might increase its stability (Khayat et al., 2005). Lastly, B345 is glycosylated most likely at the β-barrel tower and glycosylation has been shown to increase the thermal stability of some proteins (Khayat et al., 2010; Maaty et al., 2006; Wang et al., 1996). The protein most likely glycosylating B345 is STIV A197 (Larson et al., 2006).

The postulated host-attachment structures located at the viral five-fold vertices are striking turreted protrusions “oddly resembling World War II sea mines” (Ackermann, 2007), extending 135 Å above the outer edge of the capsid shell (Khayat et al., 2005; Rice et al., 2004). The stalk of the turret is decorated by petal-like appendages that extend 82 Å laterally (Figure 8b and c, purple densities). The width of the entire complex is 244 Å (Khayat et al., 2005). The base of the turret extends 50 Å into the capsid from the bottom of the B345 shell spanning the gap between the capsid and the membrane (Khayat et al., 2005). The centre of the turret contains a 3 nm channel (in diameter) that could provide access between the exterior and the interior of the virion. It has thus been suggested that each turret might act as a portal for the translocation of the viral genome (Rice et al., 2004). The calculated mass for the turret complex is 637 kDa (Khayat et al., 2005) with the density of the petal-like appendages being 63 kDa (Khayat et al., 2010). Initially, fold recognition algorithms predicted C381 and A223 to be similar to the PRD1 spike protein P5 (Maaty et al., 2006; Merckel et al., 2005). Furthermore, C557 was suggested to form the petal-like appendages of the turret based on the fact that the N-terminus of C557 has a sequence similar to that of anklyn repeat and that the anklyn domains resemble a curved cylinder, much like the petal-like appendages attached to the turret (Maaty et al., 2006). As the calculated mass for the turret complex is 637 kDa, each monomer of the pentameric turret has been suggested to be formed of one copy of C381, A223, and C557 (Khayat et
al., 2010; Maaty et al., 2006). This would give an expected mass of 125 kDa for the monomeric subunit, which is close to the estimated 127 kDa subunit of the pentamer (Maaty et al., 2006). The studies on the undecorated STIV particles by Khayat et al. (Khayat et al., 2010) confirmed that the petals are formed by C557, as the respective band is missing in SDS-PAGE gels when analyzing this particle type. However, at low contour levels of the undecorated virions, traces of the petals can be seen, indicating partial occupancy (Khayat et al., 2010). Furthermore, as the predicted structure of C381 has some resemblance to cadherin molecules responsible for cell-cell adhesion in animal tissues, Khayat et al. (Khayat et al., 2010) have suggested that C381 is located at the tip of the turret where it could be responsible for host-cell interaction. This arrangement would place A223 at the base of the turret (Khayat et al., 2010). Surrounding the base of the turret complex in the decorated particles is density corresponding to a 19.8 kDa protein (Khayat et al., 2005). This density is, however, missing in the undecorated particles and in the undecorated particles the turrets do not contact the viral membrane (Khayat et al., 2010).

Nine virion proteins have been identified by mass-spectrometry; namely, C557, C381, B345, A223, B164, B130, B109, A78, and A55 (Maaty et al., 2006). Furthermore, four additional viral proteins, B129, B264, C92, and A61, have been shown to be abundantly expressed during infection (Maaty et al., 2012). Nine of these proteins have been identified to be associated with membranes, namely A55, A61, A223, A510, B130, B164, B345, C92, and C381 (Maaty et al., 2012). Of these proteins, B345 is known to be glycosylated (Khayat et al., 2010; Maaty et al., 2006) and B109, B130, and C92 are suggested to have post-translational modifications based on the presence of several different size spots in 2D protein gels (Maaty et al., 2012). Of the structural STIV proteins, A510, A55, A78, B109, and B130 currently have no assigned function (Maaty et al., 2012). Based on sequence and structural analysis B164 is the postulated 19 kDa genome packaging P-loop ATPase (Maaty et al., 2006). The structure of the DNA binding protein B116 has been solved and the suggested functions include regulation of gene expression or the modification, synthesis or packaging of nucleic acids (Larson et al., 2007b).

1.9.2.2 STIV assembly and life cycle

A model for the assembly of STIV has been proposed (Brumfield et al., 2009) (Figure 9), attempting to answer some of the open questions relating to the assembly of icosahedral capsids with an internal membrane – more precisely, on how the membrane core assembles and whether it is formed independently or co-assembled with the capsid. For STIV it appears that the membrane that will be internalized first forms a curved crystalline area (Brumfield et al., 2009) and is co-assembled with the protein capsid (Fu et al., 2010). Currently no proteins assisting the assembly process of STIV have been identified. The membrane to be internalized and the membrane of the host do not seem to be connected (Brumfield et al., 2009), supporting the fact that the lipids forming the STIV virion membrane are selectively acquired from the host (Maaty et al., 2006). This is similar to what has been described for SH1 (Bamford et al., 2005) and the lipid-containing bacteriophages PRD1, Bam35, and PM2 (Laurinavicius, Bamford, and Somerharju, 2007; Laurinavicius et al., 2004b). In the lipid containing dsRNA virus φ6, the lipids are selectively acquired from the host cytoplasmic membrane.
The rationale for this selectivity is that the φ6 virion proteins might preferentially bind certain lipids thus affecting the lipid composition in the virion (Mindich, Sinclair, and Cohen, 1976). Likewise, it has been suggested for PRD1 and Bam35 that positively charged side chains in the amino acids forming the capsid or membrane proteins might preferentially bind negatively charged phosphatidylglycerol molecules which therefore become enriched in the viral membrane (Laurinavicius et al., 2004b) accounting for the preferential lipid composition in these bacteriophages. Alternatively, the viral infection might modulate the host’s lipid synthesis resulting in the accumulation of a certain type of host lipids in the viral membrane (Laurinavicius et al., 2004b). In addition to STIV, the lipids in the archaeal, icosahedral virus SH1 seem to be selectively acquired from the host (Bamford et al., 2005).

The C-terminus of the STIV major capsid protein B345 has been shown to interact with the viral membrane (Khayat et al., 2005). Thus, it has been proposed that the lipids to be incorporated into the viral membrane are derived de novo and associate through their hydrophobic nature. As STIV is known to encode for at least nine membrane proteins (Maaty et al., 2012), these are suggested to be embedded in the nascent lipid creating the observed curvature and providing a docking place for the major capsid protein (Fu et al., 2010). After formation of this lipid and protein vesicle, the viral turrets are assembled and inserted into the lipid membrane (Brumfield et al., 2009). This early assembly product is thick-walled, spherical, and devoid of the viral genome. Upon genome encapsidation, the capsid becomes more angular and its protein capsid thinner (Brumfield et al., 2009). Comparison of empty (DNA-lacking) and full STIV particle reconstructions from electron cryo-tomograms of infected cells demonstrate that despite the angularization of the STIV virion upon genome packaging, no large-scale reorganizations of the capsid or membrane occur upon genome packaging (Fu et al., 2010). Furthermore, confirming the observation by Brumfield et al. (Brumfield et al., 2009), it has been shown that the turret structures were assembled onto the particle prior to genome packaging and these turret structures were similar in both the procapsid and the virion (Fu et al., 2010). What is currently unknown is whether STIV has a specialized packaging vertex similar to that of PRD1 (Gowen et al., 2003; Strömsten, Bamford, and Bamford, 2003).
Figure 9. A model for STIV assembly and maturation. STIV-encoded membrane proteins (not shown) are embedded in the nascent lipid-bilayer and induce the curvature of the membrane and provide an attachment site for the major capsid protein. The membrane proteins are thought to function as tape-measure scaffolding proteins and facilitate the correct assembly of the capsid proteins. As in tailed phages, the assembly is thought to initiate at a dedicated genome packaging vertex. The capsid proteins and the membrane co-assemble forming a procapsid. The genome is subsequently packaged either via a turret vertex or a specialized portal vertex. The procapsid matures into a virion without major conformational changes, excluding the angularization of the capsid. The figure is redrawn based on (Fu et al., 2010) with permission from the publisher.

STIV is suggested to infect the host in a manner similar to that of PM2, during which the capsid shell is dismantled and the lipid core is exposed allowing fusion with the host’s membrane in order to deliver the genome (Khayat et al., 2010; Kivelä et al., 2004). The comparison is based on the fact that a lipid core, much like the one observed in PM2 and SH1 (Huiskonen et al., 2004; Jääniloja et al., 2008; Kivelä, Kalkkinen, and Bamford, 2002), can be isolated from STIV (Khayat et al., 2010). Even though the STIV capsid could dissociate during infection, there are currently no suggestions available on how the S-layer of the STIV host S. solfataricus would be penetrated prior to membrane fusion (Khayat et al., 2010).

Initial STIV infection studies of S. solfataricus P2 indicated that only 10% of the host population became infected, suggesting that the isolate used was a mixed population with respect to susceptibility to STIV infection (Ortmann et al., 2008). By performing several rounds of single-colony isolation, a new S. solfataricus P2 host 2-2-12 highly susceptible to STIV infection was isolated (Ortmann et al., 2008). This strain was used in microarray studies on the STIV genes transcribed during infection (Ortmann et al., 2008). Transcription of virus genes was first detected at 8 h post-infection (p.i.), peaking at 24 h p.i. with subsequent lysis of cells at 32 h p.i.. The early genes transcribed at 8 h p.i. correspond to the individual genes c92, c121, and a137 as well as two gene clusters: i) b116, a53, c118, and a56; and ii) c57 to a106 (Ortmann et al., 2008), suggesting the latter
ones are operons (Maaty et al., 2006; Ortmann et al., 2008). STIV seems to take an approach similar to that of PRD1 to virus production, since none of the identified structural proteins encode a DNA or RNA polymerase (Ortmann et al., 2008). This means that STIV has to utilize the protein synthesis machinery present in the cell (Ortmann et al., 2008). Furthermore, as the expression of only 177 host genes (merely 6% of the total number of host encoded genes) was affected by STIV infection, it indicates that only a small number of host genes are reprogrammed during infection in order to avoid major stress response (Ortmann et al., 2008). Subsequently, thin-sections showed that between time points 24 h p.i. and 32 h p.i., virus-like particles begin to appear within the host-cells with up to 50 virus-like particles being visible inside an individual section (Brumfield et al., 2009).

STIV was the first known lytic virus infecting Sulfolobus spp. (Maaty et al., 2006; Rice et al., 2004). The lysis mechanism remained unknown until 2009, when Brumfield et al. discovered that STIV exits the host-cells via striking pyramid-like structures (Brumfield et al., 2009). These pyramid-like protrusions do not seem to be covered by the Sulfolobus cell wall S-layer and they appear thicker in cross-section than the cytoplasmic membrane, suggesting that the pyramids have a composition differing from that of the surrounding cellular membrane (Fu et al., 2010). The pyramids start to form before the viral particles become visible (Fu et al., 2010). It has been later shown that the STIV gene product C92 is responsible for pyramid formation and that it has been essential for virus production (Snyder et al., 2011b). C92 is suggested to have a function similar to the proteins involved in the holin/endolysin system of bacteriophages (Snyder et al., 2011b). It does not, however, cause the cells to lyse – it merely opens up a gate for viral exit – indicating that there is another viral-mediated event causing host-cell lysis (Snyder et al., 2011b). Such a role has been implicated for the endosomal sorting complex required for transport-like proteins possibly causing cell lysis upon pyramid formation (Snyder and Young, 2011c). Intriguingly, simultaneously to the report of the STIV pyramids, Bize et al. reported similar pyramid-like structures formed during Sulfolobus rod-shaped virus 2 (SIRV2) driven host cell exit (Bize et al., 2009), with the pyramids being formed by a protein homologous to the C92 (Quax et al., 2010; Quax et al., 2011), which seems to be conserved in a handful of crenarchaeal viruses (Quax et al., 2010; Snyder et al., 2011a). Both the STIV and SIRV2 pyramids are heptagonal and are 56-107 nm high and 90-135 nm wide at the heptagon base (Quax et al., 2011; Snyder et al., 2011b).

Our understanding of extremophilic icosahedral viruses has been greatly advanced with the establishment of a genetic system for STIV (Wirth et al., 2011). The STIV genome was amplified and cloned by PCR. Furthermore, transfection of the genome into S. solfataricus resulted in the production of virions. Interestingly, transfection of S. solfataricus with a linear genome did not result in a productive infection, indicating that viral DNA replication either relies upon a circular DNA replication mechanism or on a circular template for transcription (Wirth et al., 2011). Based on gene disruption and frame-shift mutations, genes a197, b345, c381, and c557 produced non-infectious particles (Fulton et al., 2009; Wirth et al., 2011). Furthermore, a frame-shift mutation in b116 delayed the infection cycle by 24 hours (Wirth et al., 2011).
1.9.3 The haloarchaeal virus SH1

The haloarchaeal virus (halovirus) SH1 infecting the euryarchaea *Haloarcula hispanica* was first described in 2005 (Porter *et al.*, 2005). SH1 was named based on its isolation site (Serpentine Lake, Australia) and its host (*hispanica*) (Porter *et al.*, 2005). The natural habitat of SH1 and its host is a hypersaline (> 3 M NaCl) salt lake.

SH1 is an icosahedral, membrane-containing archaeal virus (Jäälinoja *et al.*, 2008). It is stable between pH 6 and 9, and up to 50°C (Porter *et al.*, 2005). The virion was shown to be very sensitive to ionic concentration, with divalent cations (Mg²⁺) being an absolute requirement (Porter *et al.*, 2005). The genome of SH1 is linear, 30,898-bp long dsDNA with 309-bp inverted terminal repeats and it encodes for 56 potential ORFs longer than 40 amino acids (Bamford *et al.*, 2005; Porter *et al.*, 2005). SH1 has 15 structural proteins with apparent molecular masses ranging from 4 to 185 kDa, of which 11 have been confirmed by mass-spectrometry (Bamford *et al.*, 2005; Porter *et al.*, 2005). The most abundant virion proteins are VP3 (37.5 kDa), VP4 (25.7 kDa), VP7 (20.0 kDa), and VP12 (9.8 kDa). VP4 and VP7 have been suggested to form the capsid (Bamford *et al.*, 2005) and have, furthermore, been shown to form homo- and heteromultimeric complexes in the size range of 45 to 75 kDa (Bamford *et al.*, 2005; Porter *et al.*, 2005). Dissociation studies of the SH1 virion showed that VP2, VP3, VP4, VP6, VP7, and VP9 were released as soluble proteins in low ionic strength in 3M urea. Furthermore, in these conditions 40-50 % of VP5 and 15 % of VP1 were retained with the virion (Kivelä *et al.*, 2006). Lipid cores, similar to those observed in PM2, were observed in SH1 (Kivelä, Kalkkinen, and Bamford, 2002; Kivelä *et al.*, 2006). When treated with urea, VP1 and the genomic DNA were retained with these lipid cores. Other membrane-associated proteins are VP10, VP12, VP5, VP4, and VP13 (Kivelä *et al.*, 2006).

There are some details available about the SH1 life cycle. Viruses have been observed to attach to the host-cells 2 h p.i. (Porter *et al.*, 2005). After 12 h p.i., empty, DNA-lacking procapsids together with DNA-filled virions were detected inside the host cells (Porter *et al.*, 2005). Cell lysis begins at 5 to 6 h p.i. with an average burst size of 230 pfu ml⁻¹. However, SH1 virus production has been shown to be attenuated, with approximately 20 % of the host cells remaining uninfected (Porter *et al.*, 2005). This is most likely due to the fact that there are two different host cell morphologies observed, with only the thin-celled ones becoming infected (Porter *et al.*, 2005). Thin-section EM images suggest a release mechanism based on cell disruption (Porter *et al.*, 2005).

A detailed structure of SH1 to 9.6 Å resolution revealed a virus with an icosahedral capsid with an underlying lipid membrane and massive symmetry-mismatched host-attachment structures (Jäälinoja *et al.*, 2008). The size of the virus is 115 nm from spike-to-spike, 78.5 nm edge-to-edge, 78.0 nm facet-to-facet, and 79.5 nm vertex-to-vertex (Jäälinoja *et al.*, 2008). The membrane of SH1 is 2.4 nm thick on average and follows the shape of the capsid. Furthermore, the membrane contains icosahedrally-ordered membrane proteins and clear transmembrane structures can be seen underneath the capsid at the viral five-folds (Jäälinoja *et al.*, 2008). The triangulation number of the SH1 capsid is $T = 28d$ (Jäälinoja *et al.*, 2008). The viral capsid shell is thus built of 270 capsomers with hexameric bases and 12 pentameric capsomers, with 4.5 independent capsomers in the asymmetric unit. The capsid is built of two distinct capsomer types – the type II and type III capsomer. The type II capsomer has two additional decorating towers on top of the hexameric base, whereas the type III
capsomer has three (Jäälinoja et al., 2008). The average calculated mass of the type II capsomer is 178 kDa and the type III capsomer 195 kDa. It has been suggested that VP4 (25.7 kDa) forms the hexameric base and VP7 (20.0 kDa) the decorating towers (Jäälinoja et al., 2008). Based on the skewed conformation of the hexameric bases and the fact that there is a two-fold symmetric capsomer sitting on the two-fold axis of symmetry, it has been proposed that the capsomers are built of vertical single β-barrels as opposed to the vertical double β-barrels observed in viruses such as PRD1 (Benson et al., 2002), PM2 (Abrescia et al., 2008), STIV (Khayat et al., 2005), and PBCV-1 (Nandhagopal et al., 2002). Furthermore, each of the six capsomer subunits in the hexameric base makes a bridge-like connection to an adjacent capsomer and the capsomer base is connected to the underlying membrane (Jäälinoja et al., 2008).

The large, bifurcated host-attachment structures attached to the five-folds are unique, although they share some similarity with the smaller bifurcated structures on Acidianus filamentous virus 1 (AFV-1) shown to be involved in virus attachment to host pili (Bettstetter et al., 2003) (Figure 10). A 34 Å resolution reconstruction of the horns indicates that they are built of 10 parallel tubular domains with a calculated mass of 1.3 MDa for the whole structure (Jäälinoja et al., 2008). Based on dissociation and infection studies, it has been shown that the spikes are formed of the viral proteins VP3 (37.5 kDa) and VP6 (Jäälinoja et al., 2008) and it has been suggested that the horns are built of heteromultimers of these proteins. Furthermore, the release of VP3 and VP6, as well as VP3, VP6, and VP2 led to a decrease in infectivity, confirming that the spikes are used in host-cell adsorption and infection (Jäälinoja et al., 2008).

The SH1 horns are attached to the pentameric capsomer at the viral five-folds. The protein forming the pentamer seems to be a single β-barrel the size of the P31 homopentamer (68.7 kDa) forming the PRD1 pentameric capsomer (Jäälinoja et al., 2008; Rydman et al., 1999). The calculated mass for the SH1 pentamer is 85 kDa (Jäälinoja et al., 2008). Thus, the likely virion protein candidates for the pentamer are VP7 (20.0 kDa) in a different conformation or VP9 (16.5 kDa) (Jäälinoja et al., 2008). There are five arm-like structures in the pentamer attaching it to the adjacent capsomers with hexameric bases. Furthermore, a shaft that is located under the pentamer traverses the membrane and this shaft is surrounded by five additional groups of transmembrane proteins (Jäälinoja et al., 2008). Removal of VP3 and VP6 does not change the transmembrane structure at the vertex, but removal of VP2, VP3, and VP6 disorders the membrane, the DNA, and the membrane proteins (Jäälinoja et al., 2008). The empty procapsids (Porter et al., 2005) together with the postulated genome packaging ATPase ORF17 (Bamford et al., 2005) indicate that SH1 packages its genome into preformed procapsids. The shaft located underneath the viral five-fold vertices spanning the capsid and the membrane could be a potential prefusion complex and it has been suggested that the shaft and the surrounding peripheral proteins are likely involved in genome translocation into the host cell (Jäälinoja et al., 2008).
Figure 10. Electron micrographs showing the bifurcated host-attachment structures of SH1 and AFV1. (a) An electron cryo-micrograph of the vitrified SH1 virions (V) and lipid cores (LC). The horn-like structures are indicated with black arrows. The inset shows a schematic representation of the adjacent SH1 particle with three pronounced structures attached. Courtesy of Pasi Laurinmäki. (b) An electron micrograph of negatively-stained AFV1 particles with the horn-like structures clearly visible. Bars, 100 nm. (b) reprinted from (Bettstetter et al., 2003) with permission from the publisher.
2. AIMS OF THE PRESENT STUDY

The central theme of this thesis is icosahedral viruses living in extreme environments where temperatures rise above 70°C. The questions I have addressed in this study include how macromolecules, such as proteins and lipids, assemble, and how these assemblies stay functional even in extreme environments. Furthermore, I have addressed the fundamental issue of virus evolution. As three-dimensional structure is more conserved than sequence, the structure of viruses and their component proteins allow us to find related viruses, the similarity of which escape merely sequence-based analyses. Finally, with biotechnological applications in mind, I looked for the opportunity to find and characterize thermostable viral enzymes.

The specific aims of this thesis were:

- To determine the three-dimensional structure of the bacteriophage P23-77 infecting *Thermus thermophilus* using 3DEM
- To determine the three-dimensional structure of the archaeal virus STIV2 infecting *Sulfolobus islandicus* using 3DEM
- To determine the genome sequence of STIV2
- To identify structural STIV2 proteins
- To homology model P23-77 and STIV2 gene products
- To biochemically characterize the genome packaging NTPase B204 of STIV2
- To solve the structure the B204 NTPase using X-ray crystallography
- To generate a hypothesis for the assembly and genome encapsidation of STIV2 and related viruses
3. MATERIALS AND METHODS

3.1.1 Virus propagation and purification for cryoEM studies (I, II)

The icosahedral virus P23-77 was obtained from the Promega Corporation collection (Yu, Slater, and Ackermann, 2006) and propagated on *Thermus thermophilus* (ATCC 33923). *T. thermophilus* cells were grown in *Thermus* virus (TV)-medium containing 4 g yeast extract, 8 g peptone, 2 g NaCl per L at pH 7.5, with aeration at 70°C. For virus propagation, liquid cultures of *T. thermophilus* cells (density $7 \times 10^8$ cfu ml$^{-1}$) were infected with a multiplicity of infection of 10. After lysis, cell debris was removed by centrifugation (Sorvall SLA3000 rotor, 7,000 rpm, 20 min, and 25°C) and virus particles were precipitated from the supernatant with 12 % w/v polyethylene glycol (PEG) and 0.5 M NaCl (final concentrations) while stirring for 35 min at 28°C. Precipitated viruses were collected by centrifugation (Sorvall SLA3000 rotor, 7,000 rpm, 20 min, and 25°C). The virus pellet was rinsed with TV-buffer (20 mM Tris-HCl pH 7.5, 5 mM MgCl$_2$, 150 mM NaCl) and resuspended in 1:20 volume of the original lysate volume. Viruses were purified by rate zonal centrifugation (linear 5-20 % (w/v) sucrose gradient in TV-buffer; Sorvall AH629 rotor, 23,000 rpm, 45 min, and 25°C). The light scattering zone was collected, diluted with TV-buffer and the viruses concentrated by differential centrifugation (Beckman SW50.1 rotor, 40,000 rpm, 40 min, and 25°C). The pelleted viruses were resuspended in TV-buffer and immediately used for the preparation of vitrified specimens for cryoEM.

The icosahedral STIV2 virus was isolated from the liquid growth medium of *Sulfolobus islandicus* G4ST-2. G4ST-2 cells were grown in T-medium containing 25 mM (NH$_4$)$_2$SO$_4$, 3 mM K$_2$SO$_4$, 1.5 mM KCl, 20 mM glycine, 40 µM MnCl$_2$, 10.4 µM Na$_2$B$_4$O$_7$, 0.38 µM ZnSO$_4$, 0.13 µM CuSO$_4$, 62 nM Na$_2$MoO$_4$, 59 nM VOSO$_4$, 18 nM CoSO$_4$, 19 nM NiSO$_4$, 0.1 mM HCl, 1 mM MgCl$_2$, and 0.3 mM Ca(NO$_3$)$_2$ adjusted to pH 3.5. Alternatively, ST-medium was used, which corresponds to the T-medium with added elemental sulfur. G4ST-2 was grown for 3 to 5 days (until the optical density (OD) at 600 nm did not increase anymore) and the medium was cleared by low-speed centrifugation (as for P23-77 above). The resulting supernatant was concentrated 12,000x by ultrafiltration (Sartorius AG Vivaflow 200 cassettes, 100 kDa cutoff), filtered through 0.45 and 0.22 µm filters (Whatman), and further concentrated by ultrafiltration (Millipore Amicon Ultra, 100 kDa cutoff). The concentrated supernatant was purified by rate zonal centrifugation (linear 15-35 % (w/v) sucrose gradient in 25 or 50 mM sodium citrate; Beckman SW50.1 rotor, 40,000 rpm, 25 min, and 18°C). The light scattering zone was collected, diluted 1:5 with sodium citrate buffer, and the viruses concentrated by differential centrifugation (Beckman SW50.1 rotor, 40,000 rpm, 40 min, and 18°C). The pelleted viruses were resuspended in 25 or 50 mM sodium citrate pH 3.5 or 4.5 and immediately used for the preparation of vitrified specimens for cryoEM.

3.1.2 CryoEM (I, II)

Aliquots of virus (3 µl) were vitrified on holey carbon film coated grids (Quantifoil R 2/2) in liquid ethane as described (Adrian et al., 1984). Both P23-77 and STIV2 were imaged at -180°C in a FEI Tecnai F20 field emission gun transmission electron microscope operating at 200 kV and using a Gatan 626 cryoholder. P23-77 virus images were recorded on Kodak SO163 film or on a Gatan UltraScan 4000 CCD camera under low dose conditions, at nominal magnification of 50,000× and...
68,000x, respectively. Alternatively, for the unpublished 1.0 nm resolution reconstruction of P23-77, images were recorded on Kodak SO163 film at a nominal magnification of 62,000x. The micrographs recorded on film were developed in full-strength Kodak D19 film developer for 12 min. The STIV2 virus images were collected on a Gatan UltraScan 4000 CCD camera under low dose conditions as described for P23-77.

3.1.3 CryoEM image processing (I, II, unpublished)

Films were digitized at 7-µm intervals on a Zeiss Photoscan TD scanner, resulting in a nominal sampling of 0.14 nm pixel$^{-1}$ for images collected at 50,000x and 0.113 nm pixel$^{-1}$ for images collected at 62,000x. Both sets of images were binned by 2 to speed up calculations; the resulting samplings were 0.28 nm pixel$^{-1}$ and 0.2258 nm pixel$^{-1}$, respectively. The sampling of the images collected on CCD was 0.22 nm pixel$^{-1}$ and the STIV2 images were further binned by 2 for a final sampling of 0.442 nm pixel$^{-1}$. CTFFIND3 (Mindell and Grigorieff, 2003) was used to estimate the contrast transfer function. Drifted and astigmatic images were discarded. ETHAN (Kivioja et al., 2000) was used to locate the virus particles in the micrographs and particles were extracted in the EMAN program BOXER (Ludtke, Baldwin, and Chiu, 1999). BSOFT (Heymann, 2001) was used for further image processing.

For the processing of P23-77, the three-dimensional structure of SH1 (EMDB:1353) (Jäälinoja et al., 2008) scaled to the appropriate size was used as a starting model to determine the orientations and origins of the icosahedral particles in a model-based approach (Baker and Cheng, 1996). Likewise, for the processing of STIV2, a three-dimensional model of STIV (kind gift of J. Johnson), scaled to the appropriate size, was used as the starting model. PFT2 and EM3DR2 (Baker and Cheng, 1996) were used in the initial rounds of refinement, and PO$^2$R and P3DR (Ji et al., 2006; Marinescu and Ji, 2003) for subsequent rounds. The contrast transfer function was fully corrected in P3DR using a Wiener filter (Marinescu and Ji, 2003). The resolution of the three-dimensional reconstructions was assessed using a Fourier shell correlation (FSC) cutoff of 0.5 (Harauz and van Heel, 1986). The structures have been deposited in the EMDB database with the accession numbers 1525 (P23-77) and 1679 (STIV2).

The unpublished 1.0 nm resolution reconstruction of P23-77 was generated using FREALIGN (Grigorieff, 2007). The reconstruction was calculated using 1,260 virus particles from micrographs collected at 62,000x magnification. The particle images extracted using BOXER (Ludtke, Baldwin, and Chiu, 1999) were pre-processed in IMAGIC (van Heel et al., 1996). By using the search function of FREALIGN, the particles were initially aligned against the 1.4 nm resolution reconstruction of P23-77 (EMDB: 1525). This step was followed by subsequent refinement cycles converging into the 1.0 nm resolution reconstruction. The resolution of the three-dimensional reconstruction was assessed using a FSC cutoff of 0.143 (Rosenthal and Henderson, 2003). Assessed at the FSC cutoff 0.5, as the published reconstruction, the resolution is 1.32 nm. A negative temperature factor of 500 Å$^2$ was applied to the final reconstruction.

3.1.4 Structural docking, map segmentation and difference imaging (I, II, unpublished)

A trimeric model of the homology modeled STIV2 major capsid protein A345 monomer (see 3.1.10) was generated by superimposing the monomer onto the trimeric PRD1 P3 (PDB: 1hx6) in CHIMERA.
(Pettersen et al., 2004). Furthermore, in order to generate the full icosahedral asymmetric unit, the trimeric A345 model was fitted into each capsomer in the asymmetric unit in the 2.0 nm resolution map of STIV2 (EMDB:1679) using CHIMERA and the resulting model was exported using COOT (Emsley et al., 2010). To generate a pseudoatomic model of the whole capsid, symmetry operations were applied in BSOFT. This pseudoatomic model was filtered to 2.0 nm resolution prior to density subtraction using BSOFT. This same procedure was used in order to generate a trimeric model for the STIV major capsid protein B345 (PDB: 2bbd) for the segmentation of the STIV turret. Additionally, in order to create a single β-barrel with a hexameric base for docking into the 1.0 nm resolution P23-77 reconstruction, the PM2 major capsid protein P2 V2 barrel was selected in PYMOL and fitted into each individual barrel in the hexameric P2 structure (PDB: 2vff).

The turrets of STIV, STIV2 and the capsomers of P23-77 were manually segmented using QSEGMENT (Ludtke, Baldwin, and Chiu, 1999). The two types of the P23-77 capsomers were further averaged as follows: In order to determine the rotations and translations between the capsomers, a pseudoatomic model of each type of capsomer was generated in SITUS (Wriggers, Milligan, and McCammon, 1999). These models were further fitted into the reconstruction using COLORES (Chacon and Wriggers, 2002) and subsequently re-segmented in QSEGMENT. The molecular mass of the STIV and STIV2 turrets, and the P23-77 capsomers were estimated in EMAN from the segmented densities using a density threshold of 1 or 2 σ above the mean and a protein density of 1.35 g ml⁻¹.

The difference imaging of the STIV and STIV2 turrets was done using BSOFT. To visualize the part of the turrets buried in the viral capsid, the pseudoatomic models filtered to the respective resolutions of the three-dimensional reconstructions were subtracted from the virion densities. Furthermore, in order to analyze the difference between the STIV and STIV2 turrets, the generated difference maps were scaled to the same size and resolution and the STIV2 turret was subtracted from that of STIV decorated particles.

### 3.1.5 DNA cloning (II, III)

The cloning of the STIV2 genome was done at the Danish Archaea Center in Copenhagen, Denmark, and is described in detail in publication II.

The ORF coding for the STIV2 genome packaging NTPase B204 was amplified from the *Sulfolobus islandicus* G4ST-2 strain harboring the STIV2 virus using the forward primer 5’ CGCCCGCATATGAACTCGGATGATATAGT and the reverse primer 5’ GTGGTGCTCGAGAATCGGCTTGTGTATCTT, and the Phusion DNA-polymerase (Finnzymes, Finland) in GC-reaction buffer. The PCR-product was purified using the Roche HighPure PCR purification kit and digested using *NdeI* and *XhoI* for cloning into pET22b. Restriction enzymes were purchased from New England Biolabs. Plasmid DNA was extracted using the GeneJET Plasmid Miniprep Kit and DNA was sequenced at Eurofins MWG Operon (Germany).

### 3.1.6 STIV2 genome annotation (II)

The gene recognition, annotation, and comparison to other known sequences were done using MUTAGEN (Brügger, Redder, and Skovgaard, 2003) and NCBI BLAST (Altschul et al., 1990).
3.1.7 Expression and purification of recombinant B204 (III)

B204 was expressed in *Escherichia coli* ER2566/pTF16 (TaKaRa Bio Inc.) in Luria-Bertani medium supplemented with ampicillin (100 µg ml⁻¹), chloramphenicol (34 µg ml⁻¹), L-arabinose (5 µg ml⁻¹), and 1 drop of antifoam 204 (Sigma). Cells were grown at 37°C and 220 rpm to an optical density of 0.5-0.6, and induced with 0.5 mM isopropyl-β-D-thiogalactopyranoside (IPTG) for 1 hour. The cells were collected by centrifugation (Sorvall SLA3000 rotor, 8,000 rpm, 15 min, 4°C). Pelleted cells were stored frozen at -20°C.

To produce L-seleno-methionine labeled B204 for multi-wavelength anomalous diffraction (MAD), B834/pTF16 cells transformed with the plasmid expressing B204 were grown in Luria-Bertani medium supplemented with ampicillin (100 µg ml⁻¹), chloramphenicol (34 µg ml⁻¹), L-arabinose (5 µg ml⁻¹), and 1 drop of antifoam 204. Cells were grown at 37°C and 220 rpm to an optical density of 0.5-0.6, and collected by centrifugation (Sorvall SLA3000 rotor, 8,000 rpm, 10 min, 25°C). The cell pellets were resuspended and transferred into 400 ml of minimal MOPS-medium supplemented with 17 amino acids (Neidhardt, Bloch, and Smith, 1974), L-seleno-methionine (50 mg l⁻¹), ampicillin (100 µg ml⁻¹), chloramphenicol (34 µg ml⁻¹), L-arabinose (2.5 µg ml⁻¹), and 1 drop of antifoam 204 and induced with 0.5 mM IPTG for 2 hours. The cells were collected by centrifugation (Sorvall SLA3000 rotor, 8,000 rpm, 15 min, 4°C). Pelleted cells were stored frozen at -20 °C.

Frozen cell pellets were thawed, and resuspended in 100 mM MES pH 6.5, 100 mM NaCl, 10 mM sodium phosphate, 5 mM MgCl₂. Resuspended cells were supplemented with deoxyribonuclease I (4 µg ml⁻¹), lysozyme (40 µg ml⁻¹), and 1 mM Pefabloc protease inhibitor (final concentration) (Sigma). Cells were lysed with the French Press cell disrupter (Thermo Electron Corporation) in the 20K cell at 1200 psi at room temperature (22°C). Cell debris was removed by centrifugation (Sorvall SS34 rotor, 18,000 rpm, 30 min, 20°C). The supernatant was applied to a HiTrap Heparin HP-column on an ÄKTA™ Prime chromatography pump system at room temperature. Bound B204 was eluted with a 0.1-2 M NaCl gradient in 100 mM MES pH 6.5, 10 mM sodium phosphate, 5 mM MgCl₂. The fractions containing B204 were diluted 1:2.5 with 100 mM MES pH 6.5, 100 mM NaCl, 20 mM imidazole, 10 mM sodium phosphate, 5 mM MgCl₂, and loaded on a HisTrap HP-column. Bound protein was eluted with 0.8 M imidazole in 100 mM MES pH 6.5, 100 mM NaCl, 10 mM sodium phosphate, 5 mM MgCl₂. The fractions containing B204 were diluted 1:5 with 100 mM MES pH 6.5, 100 mM NaCl, 10 mM sodium phosphate, 5 mM MgCl₂, and loaded onto a HiTrap SP HP-column. Bound protein was eluted with 2 M NaCl in 100 mM MES pH 6.5, 10 mM sodium phosphate, 5 mM MgCl₂. The fractions containing B204 were concentrated using ultrafiltration (Millipore Amicon Ultra 10 kDa cutoff) to a final concentration of 10 mg ml⁻¹. The concentrated protein was subsequently loaded onto a Superdex 200 HR 10/300 gel filtration column on a Waters chromatography pump system, equilibrated with 100 mM citric acid pH 5.0, 50 mM NaCl, 5 mM MgCl₂ (flow rate 0.5 ml min⁻¹). The fractions containing B204 were concentrated using ultrafiltration (Millipore Amicon Ultra 10 kDa cutoff) to a final concentration of 12.5 mg ml⁻¹, and pure, concentrated protein was snap frozen in liquid nitrogen and stored at -80°C.
3.1.8 Protein analysis (I, II, III)

Proteins in the virus preparations and the purified B204 protein were resolved on SDS-polyacrylamide gels (Laemmli, 1970) for standard analysis. A Bio-rad Precision Plus unstained protein standard (size range 10 to 250 kDa) or a Fermentas PageRuler Plus prestained protein ladder (size range 11 to 250 kDa) were used as standards. For mass-spectrometric analyses of the STIV2 structural proteins, the proteins were resolved on a 20 % SDS-polyacrylamide gel, stained with Coomassie blue or silver nitrate, and the lanes containing the virus sample cut into pieces. Each gel piece was analyzed separately by liquid chromatography mass-spectrometry (LC-MS/MS) (see section 3.1.9).

Transmembrane helices were predicted using TMpred (Hofmann and Stoffel, 1993) and TMHMM (Krogh et al., 2001). The number of salt-bridges was analyzed using ESBRI (Costantini, Colonna, and Facchiano, 2008) and the number of hydrogen bonds using PDBePISA.

3.1.9 Mass-spectrometry (II, III)

The STIV2 proteins were identified using mass-peptide fingerprinting at the Protein Chemistry Core Facility, Institute of Biotechnology, University of Helsinki. The purified B204 protein was similarly identified prior to crystallization trials. The LC-MS/MS analysis was performed using an Ultimate 3000 nano-LC and a QStar Elite hybrid quadrupole time-of-flight (TOF) mass-spectrometer with nano-electrospray (nano-ES) ionization. The LC-MS/MS samples were first loaded on a ProteCol C18 trap column (3 µm, 10 mm × 150 µm, 120 Å), followed by peptide separation on a PepMap100 C18 analytical column (5 µm, 15 cm × 75 µm, 100 Å) at 200 nl min⁻¹. The separation gradient consisted of 0-50 % B in 20 min, 50 % B for 3 min, 50-100 % B in 2 min and 100 % B for 3 min (buffer A: 0.1 % formic acid; buffer B: 0.08 % formic acid in 80 % acetonitrile). MS data were acquired using Analyst QS 2.0 software. The LC-MS/MS data was searched with the in-house Mascot version 2.2 against a database of all STIV2 ORFs.

Produced L-selenomethionine-labeled B204 was compared against the total mass of the native B204 in order to confirm the desired labeling. Protein masses were determined using a MALDI-TOF mass-spectrometric analysis in an Ultraflex TOF/TOF instrument equipped with a nitrogen laser operating at 337 nm. The mass spectra were acquired in positive ion linear mode using α-Cyano-4-hydroxycinnamic acid as the matrix. The samples were either diluted with 0.1 % TFA (trifluoro acetic acid) or desalted using Millipore® C4 Zip Tip.

3.1.10 Protein homology modeling (II, IV, unpublished)

Proteins in this study were homology modeled using the I-TASSER server (Roy, Kucukural, and Zhang, 2010; Zhang, 2008), as its threading algorithm proved to generate more reliable models than the Phyre or SWISS-MODEL servers (Kelley and Sternberg, 2009; Schwede et al., 2003). Only I-TASSER models with a TM-score > 0.5 were considered to have the correct topology (Zhang, 2008; Zhang and Skolnick, 2004). The TM-score is a structural similarity measurement, and gets values in the range [0,1] (Zhang and Skolnick, 2004). Furthermore, a TM-score ≤ 0.17 corresponds to the similarity between two random models selected from the PDB library and, as stated above, a TM-score ≥ 0.5
corresponds to two models with similar topology (Zhang, 2008; Zhang and Skolnick, 2004). When creating homology models for the NTPases of the PRD1-like viruses, the structure of B204 was used as an additional template in I-TASSER.

3.1.11 Differential scanning fluorimetry (unpublished)

The thermal stability of B204 in different buffers (Table 2) was assessed by differential scanning fluorimetry (ThermoFluor) in a 48-well format using a MiniOpticon Real-Time PCR System (Biorad). Briefly, 20 µl of buffer was mixed with 5 µl of B204 at 1 mg ml⁻¹ and 25 µl SYPRO Orange dye solution. The reactions were carried out in a Multiplate Low-Profile Unskirted PCR plate covered with Bio-Seal 7. The plate was heated from 25 to 99°C and a melting curve of B204 for each well was recorded in 0.5°C increments by monitoring the fluorescence at 540-700 nm.

3.1.12 Enzyme activity assay (III)

NTPase activity was measured using the malachite green assay, as described previously (Ghosh et al., 2011; Lanzetta et al., 1979). Briefly, purified protein (7 µg) was preheated for 2 minutes in 50 µl of assay buffer. Reactions were initiated by adding ATP and incubated for a further 10 min. The reaction was stopped by snap freezing in liquid nitrogen. The amount of released inorganic phosphate was determined by adding 800 µl of malachite green molybdate reagent (0.034 % (w/v) malachite green, 1.05 % (w/v) ammonium molybdate, 1 N HCl, 1 % Triton X-100) to the reactions. The reactions were incubated for 6 min and stopped by adding 100 µl 34 % (w/v) citric acid. The reactions were further incubated for 1 hour at room temperature prior to measuring the absorbance at 660 nm using a Hitachi U-2001 spectrophotometer. The data was corrected for non-enzymatic ATP hydrolysis. KH₂PO₄ was used in all experiments to generate standard curves.

To determine the activity of B204, the assay was performed at pH-values between 3.0-7.0 (100 mM citric acid pH 3.0-5.0, 100 mM MES pH 5.5-6.0, and 100 mM HEPES pH 7.0). All buffers contained 100 mM NaCl and 5 mM MgCl₂. The temperature range analyzed was from 50°C to 90°C. In order to analyze the reaction requirements of B204, ATP was substituted by GTP, CTP, or UTP, and likewise MgCl₂ with CaCl₂ or MnCl₂. To determine whether the activity of B204 is stimulated by nucleic acids, 700 ng of linear dsDNA (PCR-product of STIV2 a345, 1.8 kbp) or φX174 RF I DNA (dsDNA, covalently closed, circular, 5.386 kbp; NEB) was used per reaction.

3.1.13 Electrophoretic mobility shift assay (III)

Nucleic acid binding to B204 was analyzed using an electrophoretic mobility shift assay (gel-shift). In order to study the kind of DNA molecules B204 binds, 0.1, 0.5, and 1.0 µg of B204 was incubated with 100 ng of DNA-molecules in 10 µl reaction mixtures. The DNA-molecules used were linear dsDNA (PCR-product of the STIV2 a345 major capsid protein gene, 1.8 kb), φX174 RF I DNA (dsDNA, covalently closed, circular, 5,386 bp, 3.50 × 10⁶ Da), φX174 RF II DNA (dsDNA, nicked (relaxed), circular, 5,386 bp, 3.50 × 10⁶ Da), and φX174 virion DNA (ssDNA, circular, 5,386 bp, 1.70 × 10⁶ Da). φX174 DNA was purchased from New England Biolabs (NEB). The reactions were performed at room temperature in a reaction buffer of 100 mM citric acid pH 4.5, 100 mM NaCl, 5 mM MgCl₂ for 30 min. Thereafter, 2 µl of 6 × DNA sample buffer (Fermentas) was added and the samples were analyzed in
0.8 % SeaKem LE-agarose gels in 1 × Tris-acetic acid-EDTA buffer with 0.5 µg ml⁻¹ ethidium bromide (EtBr) (6 V/cm, 90 min) and recorded using a Biorad Molecular Imager ChemiDoc XRS system.

In order to study the length of dsDNA that B204 binds, 0.5 and 1.0 µg of B204 was incubated with 100 ng of DNA molecules in 5 µl reaction mixtures. DNA molecules used were 15, 20, 25, and 35 bp No Limits DNA fragments (Fermentas). The reactions were performed at room temperature in a reaction buffer of 100 mM citric acid pH 5.0, 50 mM NaCl, 5 mM MgCl₂ for 10 min. Thereafter, 1 µl of 6 × DNA sample buffer (Fermentas) was added, and the samples were analyzed in 20 % polyacrylamide (PAGE) gels in 1 × Tris-borate-EDTA buffer (5 V/cm, 200 min), stained with EtBr, and recorded as above.

3.1.14 Protein crystallization and data collection (III)

Native and Se-Met labeled B204 crystals were grown using the hanging-drop vapor-diffusion method at room temperature by mixing 2 µl of protein solution (12.5 mg ml⁻¹) with 2 µl of reservoir solution (0.1 M sodium cacodylate pH 6.0, 0.21 M ammonium sulphate, 0.05 M MgCl₂, 35 % PEG-6000). B204 was co-crystallized with AMP in sitting-drops at room temperature by mixing 200 nl of protein solution (12.5 mg ml⁻¹) with 200 nl reservoir solution (0.1 M Tris-HCl pH 8.5, 0.2 M MgCl₂, 30 % PEG-4000). B204 was co-crystallized with AMPPCP or ATPγS in sitting-drops at room temperature by mixing 1 or 2 µl of protein solution (12.5 mg ml⁻¹) with 1 or 2 µl of reservoir solution (0.1 M Tris-HCl pH 8.3, 0.2 M MgCl₂, 30 % PEG-8000) respectively. Crystals were cryo-protected in Paratone-N and flash-frozen in liquid nitrogen.

Diffraction data was collected at the ESRF (European Synchrotron Radiation Facility) beamlines ID14-4 and ID23-2. Statistics for the data collection and refinement are reported in publication III.

3.1.15 Data processing, structure solution and refinement (III)

The diffraction images were processed using XDS (Kabsch, 2010a; Kabsch, 2010b). The Se-sites were located using SHELXD (Sheldrick, 2008) and the two wavelength MAD data, together with the native data, were used for phasing with SHARP (Bricogne et al., 2003) and SOLOMON (Abrahams and Leslie, 1996). The initial model was built using ARP/wARP. Molecular replacement was done using PHASER (McCoy et al., 2007; Winn et al., 2011), with manual model building in COOT (Emsley et al., 2010). Refinement was performed in REFMACS (Murshudov et al., 2011) and PHENIX (Adams et al., 2002; Afonine, Grosse-Kunstleve, and Adams, 2005).
4. RESULTS AND DISCUSSION

The aim of this study was to investigate life in extreme environments using two extremophilic, icosahedrally-ordered, membrane-containing viruses as research systems. One of these viruses, P23-77, originates from the Promega Corporation collection and infects the thermoalkaliphilic *Thermus thermophilus* bacterium (Yu, Slater, and Ackermann, 2006). The other virus, STIV2, was isolated in this study and infects the hyperthermoacidophilic archaean *Sulfolobus islandicus*. When this study was initiated, only two icosahedrally-ordered archaeal viruses, SH1 and STIV, had been studied in detail. Furthermore, none of the studied viruses infecting *Thermus* species were membrane-containing and those that had an icosahedrally-ordered head were tailed bacteriophages. Thus, this study yields valuable novel information on the structure of icosahedrally-ordered, membrane-containing extremophilic viruses. Additionally, the sequence sampling of these viruses is heavily underrepresented in public databases, making comparisons to other viruses difficult and in some cases impossible. This study provided the genome sequence of the hyperthermoacidophilic virus STIV2, with immediate implications on the analysis of a close relative, STIV. Furthermore, while this study was ongoing, the genome of P23-77 was sequenced and released and used to identify related viruses and genetic elements. In addition to studying the overall virion architecture using cryoEM, the structure of the genome packaging NTPase of STIV2 was solved using X-ray crystallography. This structure further allowed me to model the corresponding protein in P23-77. Since the structure of the genome packaging NTPase of STIV2 is the first one to be described for a membrane-containing virus, I propose a model of action for the genome packaging of these viruses using STIV2 as an example.

4.1 Extremophilic viruses within a single β-barrel lineage

One of the aims of this thesis was to characterize the icosahedrally-ordered, membrane containing bacteriophage P23-77 using 3DEM. As described in detailed in section 4.1.3, the structure of P23-77 was solved to 1.4 nm resolution and its architecture was surprisingly identified as similar to that of the haloarchaeal virus SH1. Further research allowed me to generate a higher resolution reconstruction to 1.0 nm of the P23-77 virion which will be presented below as unpublished results; the quest for a further improved structure is ongoing. At the time of publication of the P23-77 structure, the genome sequence of P23-77 was not available. Upon the release of the P23-77 genome sequence, I used homology modeling to study the possible functions of the P23-77 proteins. One such protein, encoded by ORF34, yielded a model of a putative excisionase missing in the deposited sequence annotation. These findings will be presented below as unpublished results. Furthermore, as I solved the structure of the STIV2 genome packaging NTPase B204 in publication III, the structure of B204 was used as a model in order to generate a homology model of the P23-77 genome packaging ATPase ORF13.

4.1.1 Comparison of *Thermus* phages (I)

We initially obtained seven icosahedral, non-tailed, dsDNA phages from the Promega collection originating from hot, alkaline springs on the North Island of New Zealand (Yu, Slater, and Ackermann, 2006). These seven viruses were P23-65H, P23-72, P23-77, P37-14, P78-65, P78-83, and P78-86. Three of these viruses – P23-65H, P23-72, and P23-77 – grew to high titers (> $1 \times 10^{11}$ pfu
and were chosen for more detailed analysis. The host for these viruses is *Thermus thermophilus* (ATCC 33923) with optimal growth conditions of 70°C and pH 7.5. All three viruses are lytic with an average burst size of 200 pfu. Based on thin section samples, virus particles are associated with the cell surface 15 min p.i. and after 45 min p.i. virions are visible inside the cells. Host-cell lysis was detected 70 min p.i. (I, Figure 1B). Additionally, all three viruses appeared indistinguishable in negative stain and based on SDS-PAGE analysis of highly purified virions, the protein profiles of these three viruses were similar (I, Figure 2). Despite their similar appearance in our hands, the viruses are not identical. According to Yu et al. (Yu, Slater, and Ackermann, 2006) the main difference between these three viruses can be summarized as follows: i) P23-65H infects *Thermus thermophilus* (ATCC 33923), whereas P23-72 and P23-77 additionally infect *Thermus* sp. ATCC 27978; and ii) the genomes of P23-65H and P23-72 produced nine bands upon Acc651 restriction enzyme digestion, whereas P23-77 produced only four.

Quite surprisingly, P23-77 and P23-72 were most stable at 28°C and P23-65H at 22°C despite their host being thermophilic. Whereas P23-65H and P23-72 dropped in titer upon storage at temperatures between 4 to 37°C, P23-77 was not affected. However, when incubated at 70°C, P23-77 lost 70% of its infectivity within 24 h. Despite poor stability at 70°C, the maximum adsorption rate of P23-77 to *T. thermophilus* was reached within 10 min, when 70 % of the phages were surface-bound, compared to the adsorption rate at room temperature which was less than half of this (I, Figure 1C). The observed inconsistency in virion stability and host-cell adsorption rates in the analyzed temperatures might be a reflection of the viruses’ adaption to the fluctuating temperatures in the hot springs – the viruses stay viable even in the case the temperature drops, but are more likely to initiate infection in the physiological conditions of the host. Alternatively, and likely, the buffers used in the study lack certain ions present in the native environment that stabilize the virions at higher temperatures, as has been described for φYS40 (Sakaki and Oshima, 1975).

### 4.1.2 Biochemical characterization of bacteriophage P23-77 (I)

Due to its superior stability during storage, P23-77 was chosen for further structural and biochemical studies. The low buoyant density of the virions (1.22 g ml⁻¹ in sucrose and 1.21 g ml⁻¹ in CsCl) indicated the presence of lipids. The lipid composition of P23-77 was thus analyzed using thin-layer chromatography (TLC) and was found to consist mainly of neutral lipids acquired selectively from the host (I, Figure 3B). Later on, the lipids of P23-77 were analyzed more in-depth using a combination of TLC and mass-spectrometry (Jalasvuori et al., 2009). Based on these studies, the lipids of P23-77 consist of glycolipids, and other unidentified lipids.

As revealed by the analysis of the protein pattern of P23-77 (I, Figure 3C), it was evident that P23-77 has at least 10 protein bands with apparent molecular masses of 8 to 35 kDa, two of which were abundant (20 and 35 kDa), indicating the existence of two major capsid proteins. A similar protein pattern has been described for the haloarchaeal viruses SH1 and HHIV-2 (Jaakkola et al., 2012; Porter et al., 2005). Upon subsequent P23-77 genome publication it has been shown that the P23-77 genome is a 17,036-bp long circular dsDNA molecule encoding 37 putative ORFs. As anticipated from our study, ten of these ORFs have been confirmed to be true genes. Furthermore, the two major
capsid protein bands were identified as VP16 (19.1 kDa) and VP17 (31.8 kDa) based on mass-peptide fingerprinting (Jalasvuori et al., 2009).

4.1.3 Structure of P23-77 to 1.4 nm (I)

I used highly purified virions from 1× and 2× sucrose gradients for electron cryo-microscopy. Using 880 images from 25 micrographs I calculated an icosahedral reconstruction to 1.4 nm resolution (FSC 0.5) (I, Figures 4 and 5). Approximately 2% of the particles observed were empty, lacking DNA. I thus used 49 images from 44 CCD-micrographs in order to calculate an icosahedral reconstruction of the empty viruses to 3.0 nm resolution (FSC 0.5). The reconstructions revealed a multilayered virus structure. Both types of particles are roughly spherical and have a diameter of 78 nm. By comparing the radial intensity profiles of these two reconstructions, I could confirm the presence of an internal lipid bilayer by deducing which of the peaks of density belong to DNA and which belong to the membrane (I, Figure 4C). In the multilayered P23-77 structure, the thickness of the outer protein shell is 6 nm and the average spacing between the membrane leaflets is 3.0 nm. This is close to what has been described for other viruses with internal membranes, such as SH1 with an average spacing of 2.4 nm (Jäälinoja et al., 2008) and Bam35 with an average spacing of 2.5 nm (Laurinnäkö et al., 2005) between the membrane leaflets. Furthermore, the average spacing between the concentric P23-77 DNA rings is 3.1 nm. Assuming that the genome fully occupies the space enclosed by the membrane, the average packaging density of the P23-77 genome is 0.38 bp nm\(^{-3}\), once again close to what has been described for other viruses with internal membranes, such as SH1 (0.45 bp nm\(^{-3}\)) (Jäälinoja et al., 2008) and PM2 (0.35 bp nm\(^{-3}\)) (Huiskonen et al., 2004).

The triangulation (T) number of a virus defines the number of pseudo-identical subunits with quasi-similar environments in an icosahedral asymmetric unit (Caspar and Klug, 1962). The T-number of P23-77 is 28d and the capsid lattice in this case consists of 270 capsomers with hexameric bases (4.5 per asymmetric unit) and 12 pentameric capsomers (I, Figure 5A). Currently, SH1 is the only other virus for which such a T-number has been described (Jäälinoja et al., 2008). The P23-77 surface looks like it is covered by small towers. These towers are attached to the base of the capsomers. The position of these towers depends on the position of the capsomer in the asymmetric unit (I, Figures 5 and 6). Coincidentally, similar towers have been observed on the SH1 capsomers with hexameric bases (Jäälinoja et al., 2008). In SH1 these capsomers are referred to as type II or type III capsomers depending on the number of attached towers. In contrast, the number of towers per capsomer in P23-77 is always two, but the location on the hexameric base varies between two towers on the same side (type A capsomer) or on opposite corners (type B capsomer) (I, Figure 6). As SH1 has three towers in the type III capsomer, it was impossible at the time of publication to identify the rotational relationship between the capsomers (Jäälinoja et al., 2008). However, as the P23-77 type A capsomer has two asymmetrically ordered towers (I, Figure 6A), this analysis suddenly became possible. In the nucleo-cytoplasmic large DNA viruses (NCLDV) CIV and PBCV-1, combined cryoEM and X-ray analysis show that a capsomer in the pentavalent position at the vertices is rotated 72° relative to its neighbouring pentavalent capsomer and 60° to the adjacent hexavalent capsomer (Simpson et al., 2003). Furthermore in these viruses, including PpV01, the hexavalent capsomers in each icosahedral facet are oriented in the same direction (Simpson et al., 2003; Yan et al., 2005; Yan et al., 2009; Zhang et al., 2011b). In contrast to this, all capsomers in the P23-77 facet do not face
the same direction, but are rotated 120° around the three-fold symmetry axis and only face the same direction within the asymmetric unit (I). Assuming that the SH1 type III capsomer is equivalent to the P23-77 type A capsomer, then this capsomer rotation pattern could be true in SH1 as well.

The unusual even T-number of P23-77 and SH1 required a closer analysis of the capsid structure. It has been shown for large lipid-containing viruses, such as PBCV-1, PpV01, and CIV (Cherrier et al., 2009; Simpson et al., 2003; Yan et al., 2005; Yan et al., 2009; Zhang et al., 2011b), that their double β-barrel, pseudo-hexameric capsomers are arranged as triagonal arrays forming the facets of the virus. It has been observed by Wrigley that when such viruses are dissociated, triangular trisymmetrons and pentagonal pentasymmetrons are released (Wrigley, 1969). These symmetrons are considerable in size in these large viruses with big T-numbers; PpV01 (T = 219), PBCV-1 (T = 169d), and CIV (T = 147) (Yan et al., 2005; Yan et al., 2009; Zhang et al., 2011b). In each of these viruses the pentasymmetron consists of 31 capsomers (1 pentamer and 30 trimers). In the much smaller P23-77 and SH1, the pentasymmetrons are formed of six capsomers; one pentamer and five hexamers (I, Figures 5 and 6, yellow and red capsomers). Furthermore, the trisymmetrons of PpV01 have been confirmed to consist of 91 trimers and those of PBCV-1 and CIV of 66 and 55, respectively (Yan et al., 2005; Yan et al., 2009; Zhang et al., 2011b). In P23-77 and SH1, the trisymmetrons consist of six hexamers each and the disymmetrons of three hexamers each.

It has been observed that for T-numbers where h is odd (T = h^2 + hk + k^2), adjacent trisymmetrons make contact across the two-fold axis of symmetry, as is the case in PpV01, PBCV-1, and CIV (Yan et al., 2005; Yan et al., 2009; Zhang et al., 2011b). However, when h is even, Wrigley predicted that linear disymmetrons would be required to form the contacts between the trisymmetrons (Simpson et al., 2003; Wrigley, 1969). In the case where both h and k are even, as is in the case of P23-77, the central capsomer of the disymmetrons would sit on the icosahedral two-fold axis of symmetry. In such a case, a virus with trimeric capsomers would either have to abandon icosahedral symmetry or use a special two- or six-fold symmetric capsomer (Simpson et al., 2003). In P23-77 and SH1 we see the first examples of viruses containing the predicted linear disymmetrons. As predicted, the central capsomer sitting on the two-fold axis of symmetry is two-fold symmetric, but interestingly, so are the other capsomers in the disymmetron (I, Figures 5 and 6, blue capsomers). This means that at least the capsomer sitting of the two-fold axis has to be formed of six single β-barrels in both P23-77 and SH1, as opposed to three double β-barrels. Due to the skewed appearance of the SH1 capsomers (Jäälinnoja et al., 2008) (I, Figure 5D), it has been suggested that all the capsomers would be formed from single β-barrels. I proposed that the two-fold capsomer in P23-77 is also formed of single β-barrels (together with the other capsomers in the disymmetron). Furthermore, I proposed that the type A capsomers are formed of single β-barrels due to the non-symmetric location of the towers. If the type A capsomers would be formed of double β-barrels as in PRD1, a tower is missing. These observations suggest that we have a new lineage of viruses with vertical single β-barrel capsomers. Interestingly, there is a recent publication describing the biochemical characterization of the haloarchael virus HHIV-2 (Jaakkola et al., 2012). As HHIV-2 shares the same protein profile as P23-77 and SH1, it has been suggested that HHIV-2 has similar capsid architecture to that of P23-77 and SH1 with capsomers with hexameric bases formed by single β-barrel proteins. Furthermore, according to Jaakkola et al. (Jaakkola et al., 2012), a fourth virus with a similar protein pattern, SSIP-1, has been isolated and a paper describing it has been submitted for publication. While these
results are highly intriguing and build on the hypothesis of a new viral lineage consisting of viruses with vertical single β-barrel major capsid proteins deviating from those of the picornaviruses (Krupovic and Bamford, 2008a; Rossmann and Johnson, 1989), it will be interesting to see if the studies are followed up with structural description of the viruses and what these structures will reveal.

The average mass of both type A and B capsomers calculated from segmented volumes in P23-77 is 150 kDa. In SH1, the respective masses are 178 and 195 kDa for the type II and III capsomers, respectively (Jäälinoja et al., 2008). In both P23-77 and SH1, the individual capsomers appear hollow, as is commonly seen in β-barrel structures at these resolutions. In both P23-77 and SH1, the presence of two major protein bands in the SDS-PAGE gel indicates the presence of two major capsid proteins. In SH1, these proteins are VP4 and VP7, with VP7 (20 kDa) suggested to be forming the hexameric base and VP4 (25.7 kDa) the decorating towers (Jäälinoja et al., 2008). As described previously, in P23-77 these proteins have been identified as VP16 and VP17 (Jalasvuori et al., 2009). If we assume that the arrangement is similar to that of SH1, with VP16 forming the single β-barrels composing the hexameric base, and VP17 the decorating towers – the mass of such a complex would be 178 kDa. Alternatively, VP17 could form the single β-barrels of the hexameric base and VP16 the decorating towers – the mass of this kind of an arrangement would be 229 kDa; clearly too large to correspond to the observed density. With the sequence data in hand, other types of mixtures are possible, such as four copies of VP16 and two copies of VP17, which would give a mass of 140 kDa (closest to the mass estimated from the reconstruction). In order to accurately decide on the correct threshold to use in the reconstruction and to decide which of these models is correct, one can use the fit of an X-ray crystallographic model in the cryoEM density (Baker, Olson, and Fuller, 1999). The lack of such a model (as was the case with the interpretation of the P23-77 and SH1 capsomer masses) makes the threshold chosen for mass calculation more difficult. This might contribute to the difference in mass observed between the calculated estimate and the proposed proteins forming the assemblies. Furthermore, the threshold on which to estimate the mass for segmented densities is also affected by structure flexibility, contrast transfer correction, partial occupancy, and varying resolution within the reconstruction (S. J. Butcher, personal communication).

Wrigley’s scheme was originally introduced in the context of the assembly of very large viruses and there the natural interpretation for the occurrence of di-, tri- and pentasymmetrons was as intermediates in virus assembly (Wrigley, 1969). In the case of small(er) capsids, this interpretation might not be true, but these interaction sites might rather be important in interactions with minor capsid proteins involved in regulation of assembly, much like the minor protein observed between the facets in PRD1 (Abrescia et al., 2004). In order to analyse the possible interaction sites between di-, tri- and pentasymmetrons in P23-77, a higher resolution cryoEM reconstruction, preferably advanced with atomic models of the major capsid proteins, is required. Even a modest boost in resolution (to around 7 Å) may allow the identification of individual protein boundaries in the capsid. Additionally, as the dissociation products identified by Jalasvuori et al. (Jalasvuori et al., 2009) were done based on SDS-PAGE gel, an interesting follow-up question would be to address whether the treatments that dissociated VP16, VP17 and VP11 would produce the di-, tri-, and pentasymmetrons we predicted (I).
Despite the fact that the overall capsid architectures of P23-77 and SH1 are very similar, the host-attachment structures residing at the icosahedral five-fold vertices are clearly different. In P23-77, the host-attachment structures are 15 nm long, thin spikes as measured from the micrographs (I, Figure 4A). They appear shorter in the icosahedral reconstruction, which could be due to low occupancy, flexibility and/or inappropriate averaging in the icosahedral reconstruction process (Briggs et al., 2005; Huiskonen et al., 2007a) (I, Figure 7A). In SH1, the host-attachment structures are huge, bifurcated, and two-fold symmetric (Jäälinoja et al., 2008) (Figure 10a). The estimated mass for the P23-77 pentameric capsomer with spike attached is 90 kDa and for that of SH1, excluding the horns, 85 kDa (Jäälinoja et al., 2008). Thus, the pentamers of these viruses seem to be formed of a protein in the size range of 17-18 kDa, and is suggested to be a single β-barrel as seen in the PRD1 pentamers formed of protein P31 (Abrescia et al., 2004; Jäälinoja et al., 2008; Rydman et al., 1999). In P23-77 we observed three bands in the SDS-PAGE gel of apparent molecular weights between 15-22 kDa, any of which could form the pentameric bases. Both P23-77 and SH1 have strong densities below the icosahedral five-fold vertices, indicating the presence of membrane proteins (I, Figure 7). In P23-77 these proteins seem to be peripheral membrane proteins, compared to transmembrane ones in SH1. Further dissociation studies of P23-77 showed that VP11, VP16, and VP17 are capsid-associated, whereas VP15, VP19, VP20, VP22, and VP23 are membrane-associated proteins (Jalasvuori et al., 2009). Since the protein forming the pentamer is currently unidentified, it will be interesting to see whether or not further analyses can reveal if VP11 (22 kDa) forms the pentamer, or whether it is the membrane-associated VP15 (14.7 kDa), VP20 (24.3 kDa), or putative protein VP6 (19.4 kDa) with two predicted transmembrane helices (Jalasvuori et al., 2009). The other membrane-associated proteins are too small to fit the predicted size range. In addition to identifying the pentamer, dissociation studies combined with infectivity assays would be valuable in identifying the protein forming the host-attachment spike.

4.1.4 Structure of P23-77 to 1.0 nm (unpublished)

After the P23-77 structure was published, more data from P23-77 micrographs were collected. I also had the privilege of visiting Professor Nikolaus Grigorieff at Brandeis University (MA, USA) to learn new processing methods using the FREALIGN software developed by his research group (Grigorieff, 2007; Li, Grigorieff, and Cheng, 2010).

In order to get a higher resolution reconstruction of the full, DNA-containing P23-77 virus particle, I collected cryoEM data on Kodak D19 film at 62,000× magnification and processed 1,260 particles using the FREALIGN software to 1.0 nm resolution (FSC 0.143). Importantly, as opposed to the FSC cutoff of 0.5 usually used as a standard in assessing the resolution of cryoEM reconstructions, the cutoff 0.143 is used in FREALIGN (Rosenthal and Henderson, 2003; Zhang et al., 2008). Furthermore, the resolution of the unpublished map at FSC 0.5 is 1.32 nm; however, when comparing panels (c) and (d) in Figure 11, it is clear that more details are discerned in the unpublished reconstruction. For instance, the details in the protein capsid are much more pronounced, with the β-barrels in the capsomer base better refined (Figure 11).
Figure 11. Capsid architecture of P23-77. (a) A radially depth-cued isosurface representation of the 1.4 nm resolution reconstruction of P23-77 (EMBD:1525) drawn at 2σ above the mean and (b) of the 1.0 nm resolution reconstruction drawn correspondingly. The triangulation number that describes the $T = 28d$ arrangement of the capsomers is indicated in (a) with white dots. Symmetry axes are indicated in (b) with a white ellipse (two-fold), white triangle (three-fold), and white pentagon (five-fold). (c) and (d) represent close-ups of the respective isosurface representations. The representations were rendered in CHIMERA (Pettersen et al., 2004).

At the time of publication of the P23-77 structure I tried to fit in the smaller β-barrel of the PRD1 double β-barrel major capsid protein P3 (PDB: 1hx6) into the P23-77 capsid density, much as has been done for SH1 (Jääniloja et al., 2008); the barrel was, however, too large to fit properly. Since the publication of the P23-77 the structure of the PM2 double β-barrel (PDD: 2vvf) became available (Abrescia et al., 2008). The major capsid protein P2 of PM2 is the smallest one in the lineage of PRD1-like viruses lacking the terminal extensions and loops present in other members (Abrescia et al., 2008). Due to the fact that the PM2 major capsid protein is a double β-barrel, I selected the smaller one of these barrels (V2), created a hexameric structure of it, and fitted it into the P23-77 density. Interestingly, the 1.0 nm resolution P23-77 reconstruction has several features that stand out compared to the published structure: i) when I fitted the hexameric PM2 P2 V2 into the capsid density, I noticed that there is density unaccounted for between the capsomers, much like the bridge-like connections seen between the capsomers in SH1 (Jääniloja et al., 2008); and ii) the hole in the pentameric capsomer seen in the cross-section of the virion seems more pronounced and could be a channel for viral DNA entry and exit (Figure 12). The calculated size of the V2 barrel of
PM2 P2 is 14.8 kDa, corresponding more closely to that of P23-77 VP16 (19.1 kDa), indicating that this protein could form the hexameric base of the P23-77 capsomers. However, since the sequence similarity of PM2 P2 V2 to VP16 is only 27 % and to VP17 32 %, atomic models of these proteins are required to dissect the fine architecture of P23-77 in more detail.

In other viruses producing tri- and pentasymmetrons there are finger proteins observed linking the hexameric capsomers together within the trisymmetrons (Yan et al., 2009; Zhang et al., 2011b). Whether these proteins are unique to the NCLVDs or whether they are presented by minor capsid proteins in P23-77 and SH1 remains to be determined. A similar finger protein could explain the bridge-like densities seen between the P23-77 and SH1 capsomers. Alternatively, as crosslinking has been observed between the SH1 capsid proteins, the bridge-like densities could be formed by crosslinked networks of capsid proteins (Jäälinoja et al., 2008; Porter et al., 2005). As an example, the lambdoid bacteriophage HK97 shows extensive autocatalytic crosslinking during capsid maturation (Gan et al., 2004). This crosslinking is, however, linked with major conformational changes in the capsid. As there is no evidence of major conformational changes during capsid maturation in either P23-77 (I) or SH1 (Porter et al., 2005), it is highly unlike that autocatalytic crosslinking combined with capsid expansion similar to that of HK97 takes place.
Based on the I-TASSER scoring system (see section 3.1.10), seven of these proteins yielded reliable models (products of ORFs 1, 3, 9, 10, 21, 32, and 34). The fold prediction for the products of ORFs 1, 3, 9, and 10 confirmed the functions assigned by Jalasvuori et al. (Jalasvuori et al., 2009). The modeled folds of products from ORFs 1 and 32 suggest α-helical structures with functions as a DNA primase (PDB: 2au3) (Corn et al., 2005) and as a transcription factor (PDB: 2zhg) (Watanabe et al., 2008), respectively. The fold of the translation product of ORF34 indicates that it is a DNA-binding protein similar to the excisionase (Xis) protein of the conjugative transposon Tn916 (RMSD: 4.1) (PDB: 1y6u) (Abbani, Iwahara, and Clubb, 2005) (Figure 13). Bioinformatic results, however, have always to be experimentally validated.

The structure of P23-77 was the first Thermus virus structure to be solved by 3DEM and together with the results from SH1 founded the basis for proposing a new vertical single β-barrel lineage of viruses, differing from the tangentially oriented β-barrels seen on picornaviruses (Rossmann and Johnson, 1989). The presence of a postulated excisionase protein of P23-77 might indicate that its whole genome, or parts of it, have been involved in horizontal gene transfer that was missed in the...
original annotation (Jalasvuori et al., 2009). Additionally, the Thermus virus φTMA has both a transposase and a resolvase gene in its genome, and φIN93 harbors an insertion sequence (Matsushita and Yanase, 2009a; Tamakoshi et al., 2011). Since the exchange between genetic material of Thermus phages has been described, I suggest that P23-77 also has access to the large gene pool shared by thermophilic and mesophilic phages and their hosts. This statement is supported by the excisionase gene identified in this study, which is similar to that encoded by the Tn916 transposon. Furthermore, Jalasvuori et al. (Jalasvuori et al., 2009; Jalasvuori, Pawlowski, and Bamford, 2010) have shown that the closest relatives of the Thermus phages P23-77 and φIN93 are virus-related, genome integrating elements, further strengthening the evidence of horizontal gene transfer.

**Figure 13. Stereoview of the homology model of P23-77 gp34.** The homology modeled gp34 of P23-77 is shown in rainbow coloring and its closest structural homologue, the Xis protein of the transposon Tn916 (PDB: 1y6u) in grey. The N- and C-termini of gp34 are indicated. The figure was rendered in PyMOL.
4.2 Structural studies on the hyperthermo-acidophilic virus STIV2

The other virus studied in this thesis is the crenarchaeal *Sulfolobus* turreted icosahedral virus 2 (STIV2), observed in a culture supernatant of *Sulfolobus islandicus* G4ST-2. The work on STIV2 is more complete than that of P23-77, including the genome sequencing, the structural protein identification, and the structure determination. Furthermore, with the genome sequence available, I was able to identify the genome packaging NTPase B204. This identification led to the expression, purification, and biochemical characterization of this protein, as well as the determination of its three-dimensional structure using X-ray crystallography.

4.2.1 The STIV2 genome (II)

The STIV2 genome is a 16,622-bp long circular dsDNA molecule with 64 predicted ORFs > 50 amino acids long, if the use of GTG and TTG start codons common in Archaea are considered as well. Based on the presence of Shine-Dalgarno and archaeal TATA box sequences, 34 of these are proposed to be genes (II). These genes encode for proteins in the size range of 6 to 67 kDa.

Of these 34 suggested genes, 2 ORFs (A103 and E132) are unique to STIV2 without any homologues. Overall, the genome organization is similar to that of STIV with some regions exhibiting over 70 % sequence identity (II, Figure 2). This not only allowed me to annotate the STIV2 genome, but to reannotate that of STIV, adding 7 predicted ORFs to STIV, namely, *e93, f60, c55, c88, d161, f65a*, and *b111* (II, Supplemental table S2). Furthermore, comparison of ORF lengths in STIV2 and STIV suggested that two STIV genes could be extended using a TTG initiation codon. In addition to sharing similarities with STIV, three ORFs are common to archaeal viruses, namely B72 (confirmed structural protein in STIV2; see 4.2.2), *b60, and b116*. The naming convention used for STIV2 genes and gene products follows that of many other archaeal viruses, with the letter (a-f) corresponding to the reading frame and the number stating the number of amino acids in the gene product.

There are some differences in the STIV and STIV2 protein encoding genes. I proposed that there is a large insert in STIV2 in the gene *b631*, the corresponding gene in STIV being *a223*. Furthermore, I proposed that gene *c381* suggested to be involved in turret formation in STIV is missing in STIV2. Khayat et al. astutely pointed out that STIV2 *b631* could actually be a fusion of STIV *a223* and *c381* (II, Figure 2) (Khayat et al., 2010). Additionally, gene *c510* of STIV2 has a 5’ deletion compared to *c557* of STIV, and there is a 5’ deletion in the postulated STIV genome packaging ATPase (*b164*) compared to that of STIV2 (*b204*).

4.2.2 Identification of structural STIV2 proteins (II, III)

Based on SDS-PAGE analysis together with mass-peptide fingerprinting, I identified 12 structural virion proteins (*Figure 14a*). These proteins are *A259, C141, A55, B72, E51, A103, E132, B631, E69, E76b, A345*, and *C510*. Based on sequence similarity, STIV2 A345 was identified as the major capsid protein. Furthermore, our homology model of A345 (see 4.2.3) is almost identical to the structure of STIV B345 (II, Supplemental figure S2). Likewise, based on similarity to STIV proteins, STIV2 proteins B631 and C510 were suggested to form the turrets (*Figure 14b*). B72 was concluded to be a DNA binding protein based on similarity to VP2 of *Sulfolobus* spindle shaped virus 1 (SSV1) (Reiter et al.,
1987). A homologous protein is found in STIV (A78) (Maaty et al., 2006), SSV6 (Redder et al., 2009), Acidianus spindle-shaped virus 1 (Redder et al., 2009), and the Thermus phage φIN93 (Matsushita and Yanase, 2008). Proteins A259, C141, A55, A103, E132, C510, E51, and E76b are predicted to have transmembrane regions.

4.2.3 Homology modeling of structural STIV2 proteins (II)

Having identified the STIV2 structural proteins, I was able to generate homology models for five of them (II, Supplemental figure S2). Four of these were modeled based on their STIV counterparts for which structures have been solved by X-ray crystallography (Khayat et al., 2005; Larson et al., 2007a; Larson et al., 2007b; Larson et al., 2006). These proteins include the major capsid protein B345, the DNA-binding proteins B116 and F98, and the glycosyltransferase B197. Furthermore, a reliable model was generated for the postulated genome packaging P-loop NTPase based on its similarity to the Pseudomonas aeruginosa FtsK DNA translocase (Massey et al., 2006), which later inspired me to examine its structure and function in more detail (III).

4.2.4 Structure of the archaeal virus STIV2 and suggested virion arrangement (II, III)

I used enriched virus preparations for electron cryo-microscopy. I calculated an icosahedral reconstruction to 2.0 nm resolution (FSC 0.5) using 713 particles from 358 CCD-micrographs (II, Figure 3). Very rarely, empty, DNA-lacking virions were observed. The low number of particles used in the reconstruction is due to the low titers of virus obtained (see section 4.2.5). The reconstruction of STIV2 revealed a multilayered virus structure, with an outer protein capsid and an internal lipid membrane following the shape of the icosahedral capsid. The triangulation number of STIV2 is \( T = 31d \), with 300 copies of the major capsid protein A345. Thus, the overall architecture of STIV2 is similar to that of STIV (Rice et al., 2004). In STIV, the position where the C-terminal helical domain of the major capsid protein is suggested to interact with the underlying membrane has been pinpointed (Khayat et al., 2010; Rice et al., 2004). As the C-terminal helix in our homology model is disordered (II, Supplemental figure S2) and as the STIV2 virion difference map does not have the same kind of density protrusions as STIV (II, Figure 3), it is currently unclear whether STIV2 has the same kind of an interaction pattern between the C-terminus of A345 and the virion membrane as that observed in STIV.

STIV2 has a striking resemblance to STIV, excluding the turreted structures at the viral five-fold vertices (II, Figure 3). The turrets of STIV2 extend 11 nm outside the capsid shell and the total length of the turret is 24 nm. The turrets are 10 nm across at the top and 7.5 nm at the capsid level. Much like in the STIV turret, a clear channel can be seen in the middle (II, Figure 3d, open arrow). The calculated size of the STIV2 turret is 520 kDa, assuming a protein concentration of 1.35 g ml\(^{-1}\) and a contour level of 1σ above the mean. The most striking difference between STIV2 and the decorated STIV particle (Khayat et al., 2010; Rice et al., 2004) is the lack of petal-like densities attached to the turrets. Furthermore, the undecorated STIV particle resembling STIV2 had these petals at low occupancy (Khayat et al., 2010). Khayat et al. (Khayat et al., 2010) suggested that STIV2 could have these petal-like appendages at low occupancy; however, I have never observed turrets looking like those of decorated STIV in either G4ST-2 culture supernatants or purified STIV2 virions. Further work needs to be carried out on the STIV2 virions in order to resolve this issue. This requires a much more efficient production system for STIV2, or structure solution of the individual proteins concerned.
Concluding the discussion on the petals, they do not seem to confer host-cell specificity in the narrow sampling assayed, as both STIV and STIV2 infect the same *Sulfolobus solfataricus* P2 strain 2-2-12 ([II, and [Ortmann et al., 2008]). Ultimately, the analysis of both STIV and STIV2 would benefit from a subnanometer resolution cryoEM reconstruction combined with X-ray crystallographic data of the major capsid components.

Figure 14. STIV2 structural genes and virion assembly revisited. (a) The STIV2 is depicted with a solid black bar. Predicted ORFs are depicted as light arrows. Confirmed genes encoding structural proteins are depicted as black arrows and named. The dark gray arrow (a105) indicates a new STIV2 annotation by Maaty *et al.* ([Maaty et al., 2012]). Light grey arrows with names indicate genes for which the encoded protein structures have been solved in STIV. (b) Suggested structural organization of the STIV2 virion.
4.2.5 Life cycle of P23-77 and STIV2 (I, II)

This study deals with two extremophilic viruses with very different life cycles. P23-77 is evidently a lytic virus with a rapid (70 min) life cycle, yielding high-titers of purified virions (> 1 \times 10^{11} \text{ pfu ml}^{-1}) when propagated on a commercial host strain (I). STIV2, on the other hand, does not seem to lyse the host cells, is grown for 3-5 days until the host-cells are harvested and the virions collected from the supernatant, and the maximum titer observed was > 1 \times 10^{2} \text{ pfu ml}^{-1} (II).

The life cycle of P23-77 is similar to that described for other viruses in the lineage of PRD1-like viruses (Butcher, Manole, and Karhu, 2012). To me, the most interesting currently unanswered questions relate to whether or not P23-77 produces procapsids, or whether the rare, empty particles observed are viruses broken during purification (I, Figure 4). Furthermore, it is currently not clear how P23-77 enters the host cells for infection and exits them upon lysis. According to Jalasvuori et al. (Jalasvuori et al., 2009), P23-77 encodes four proteins potentially used in these events – two endolysins and two lysozymes. Comparing to PRD1, the endolysins (gp9 and gp10) might correspond to PRD1 P7 and be used in host-cell entry, whereas the lysozymes (VP29 and gp31) might correspond to PRD1 P35, P36, and P37, and be used in host-cell exit (Krupovic, Cvirkaitė-Krupovic, and Bamford, 2008c; Rydman and Bamford, 2000; Rydman and Bamford, 2002; Rydman and Bamford, 2003).

The life cycle of STIV2 is more puzzling. Based on colony-PCR, STIV2 was determined to persistently infect all analyzed G4ST-2 colonies. However, I did not observe any virions inside G4ST-2 or S. solfataricus P2 2-2-12 cells in thin section samples. In addition, when using S. solfataricus P2 2-2-12 as a host, the highest titer of enriched virions achieved was merely 10^7 \text{ pfu ml}^{-1}. These observations oblige us to ask questions on the STIV2 life cycle. First of all, STIV2 does not contain any known integration genes, ruling out a lysogenic life style, forgetting for the moment the possibility of one of the unassigned genes in the STIV2 genome encoding for a yet unknown integration gene. Secondly, the cell lysis mechanism of STIV2 strongly deviates from that of STIV and SIRV2, viruses that exit the host cells via striking pyramid-like structures (Brumfield et al., 2009; Quax et al., 2010; Quax et al., 2011). A gene corresponding to STIV c92 and SIRV2 p98, shown to be responsible for the pyramid formation, is missing from the STIV2 genome. Thus it is highly likely that STIV2 exits the host cell in a different manner. One possibility is that the currently unassigned proteins A103 or E132 participate in cell lysis. Another possibility is that STIV2 infects both G4ST-2 and S. solfataricus P2 2-2-12 chronically. Hence there are many unanswered questions about the STIV2 life cycle. Considering the high similarity of STIV2 and STIV, I would be tempted to speculate that the c92 might have ended up on the STIV gene via a recombination event and stayed there due to the likely added fitness to the virus.

4.2.6 Isolation and stability of B204 (III, unpublished)

I identified the postulated genome packaging NTPase B204 of STIV2 based on sequence similarity to other P-loop ATPases (II). Furthermore, I managed to generate a reliable homology model of B204 (II) based on its similarity to the Pseudomonas aeruginosa FtsK DNA translocase (Massey et al., 2006). Inspired by these results, I decided to clone, express, and purify B204 for structural studies. B204 was cloned by PCR from G4ST-2 cells into the pET22B expression vector, yielding a construct with a C-terminal 6 \times \text{His-tag}. Initially, the protein was rather insoluble and precipitated rapidly,
much as described by Žiedaite et al. while working with the genome packaging ATPase P9 of PRD1 (Žiedaite et al., 2009). I thus carried out differential scanning fluorimetry in order to identify buffers in which the unfolding temperature of B204 would be as high as possible, indicating stabilization of the protein (Table 2). The results can be summarized as follows: i) buffers with a pH between 5 and 6 tend to stabilize B204 as measured by an increase in the melting temperature; ii) ammonium sulfate in the buffer stabilizes B204 more efficiently than sodium chloride; iii) higher salt concentrations (0.5 M) stabilizes B204 more than lower ones (0.125-0.15 M).

As it has been reported that extrinsic factors, such as metals and substrates might increase the thermal stability of a protein (Vieille and Zeikus, 2001), I analyzed the melting temperature of B204 in the presence of MgCl₂ and nucleotides in a subset of the buffers assayed in Table 2. When analyzed in the presence of MgCl₂, I observed an increase in the melting temperature of B204. The melting temperature in 0.1 M MES pH 6.0 and 0.125 M NaCl increased by an average of 3.6°C when 5 mM MgCl₂ (final concentration) was added. Furthermore, the melting temperature of B204 in 0.1 M MES pH 6.5 and 0.125 M (NH₄)₂SO₄ increased by 12°C when 5 mM MgCl₂ (final concentration) was added. Moreover, the addition of nucleotides to a selection of the assayed buffers increased the melting temperature of B204. When 10 mM ADP (final concentration) was added to a buffer containing 0.1 M citric acid pH 5.0 and 0.1 M NaCl, the melting temperature increased by 5°C and when the same amount of ATP was added the melting temperature increased 3°C. Furthermore, the same additions to a buffer containing 0.1 M MES pH 5.0 and 0.1 M NaCl increased the melting temperature of B204 by 10.5°C and 10°C, respectively.

Based on the thermofluor results, I chose a 0.1 M MES (pH 6.5) based buffer for purification and a 0.1 M citric acid (pH 5.0) based buffer for gel filtration and concentration.

### Table 2. Buffers used and results from the differential scanning fluorimetry experiments

<table>
<thead>
<tr>
<th>pH</th>
<th>Buffer</th>
<th>Tm (°C)</th>
<th>0.125 M NaCl</th>
<th>0.150 M NaCl</th>
<th>0.5 M NaCl</th>
<th>0.125 M (NH₄)₂SO₄</th>
<th>0.5 M (NH₄)₂SO₄</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5</td>
<td>0.1 M sodium acetate</td>
<td>ND</td>
<td>ND</td>
<td>48.0±0</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>4.0</td>
<td>0.1 M citric acid</td>
<td>63.7±0.7</td>
<td>62.9±0.5</td>
<td>63.5±0.5</td>
<td>63.0±0</td>
<td>67.1±0.9</td>
<td>67.8±1.1</td>
</tr>
<tr>
<td></td>
<td>[n=6]</td>
<td>[n=6]</td>
<td>[n=2]</td>
<td>[n=2]</td>
<td>[n=2]</td>
<td>[n=4]</td>
<td>[n=2]</td>
</tr>
<tr>
<td>4.0</td>
<td>0.1 M sodium acetate</td>
<td>59.5±1.1</td>
<td>60.3±0.4</td>
<td>59.25±0.25</td>
<td>61.0±0</td>
<td>66.5±0</td>
<td>67.8±1.8</td>
</tr>
<tr>
<td></td>
<td>[n=4]</td>
<td>[n=2]</td>
<td>[n=2]</td>
<td>[n=2]</td>
<td>[n=2]</td>
<td>[n=2]</td>
<td>[n=2]</td>
</tr>
<tr>
<td>4.5</td>
<td>0.1 M ammonium acetate</td>
<td>ND</td>
<td>ND</td>
<td>64.25±0.25</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td></td>
<td>[n=2]</td>
<td>[n=2]</td>
<td>[n=2]</td>
<td>[n=2]</td>
<td>[n=2]</td>
<td>[n=2]</td>
<td>[n=2]</td>
</tr>
<tr>
<td>5.0</td>
<td>0.1 M MES</td>
<td>64.3±0.4</td>
<td>66.0±0.7</td>
<td>ND</td>
<td>69.0±0</td>
<td>70.8±0.4</td>
<td>70.3±0</td>
</tr>
<tr>
<td></td>
<td>[n=4]</td>
<td>[n=2]</td>
<td>[n=2]</td>
<td>[n=2]</td>
<td>[n=2]</td>
<td>[n=2]</td>
<td>[n=2]</td>
</tr>
<tr>
<td>5.0</td>
<td>0.1 M citric acid</td>
<td>68.8±0.3</td>
<td>69.5±0</td>
<td>70.0±0</td>
<td>72.5±0</td>
<td>70.8±0.4</td>
<td>73.5±0.7</td>
</tr>
<tr>
<td></td>
<td>[n=4]</td>
<td>[n=2]</td>
<td>[n=2]</td>
<td>[n=2]</td>
<td>[n=2]</td>
<td>[n=2]</td>
<td>[n=2]</td>
</tr>
<tr>
<td>5.0</td>
<td>0.1 M sodium</td>
<td>65.3±1.1</td>
<td>67.0±0.4</td>
<td>67.5±0</td>
<td>ND</td>
<td>72.3±1.8</td>
<td>ND</td>
</tr>
</tbody>
</table>

Tm = melting temperature  
[n] = number of repeats  
ND = not determined
4.2.7 B204 enzymatic activity (III)

I had difficulties finding a good assay for the studies of the enzymatic activity of B204. I began by using the Enzcheck assay (Molecular Probes), which has been used in our research program before and for which there was local expertise available. My results using this assay were unreliable, most likely due to the fact that B204 is not very active at the room temperature in which the assay was performed (25°C). I thus turned to a malachite green assay described initially by Lanzetta et al. (Lanzetta et al., 1979) and used by, among others, Ghosh et al. (Ghosh et al., 2011) for the study of an archaeal flagellar ATPase. The benefit of this assay is that the temperature and pH of the reaction can be determined by the user.

Initial screening of the pH optimum for B204 activity was done using buffers ranging from pH 3 to 7 at 80°C (III, Figure 2a) with the rational that the host for STIV2 is grown at 80°C (II). I also explored the activity of B204 in the optimal pH of 4.5 at temperatures ranging from 50°C to 90°C (III, Figure
In this study I solved four different structures of B204: i) together with a sulphate ion; ii) with AMP; iii) with ATPγS (apparently hydrolyzed to ADP); and iv) with AMPPCP. The root-mean-square deviation (RMSD) value between the main chains between the four structures is less than 0.5 Å. The structures are all well ordered, except for two loop regions between amino acids 41-47 (A-loop) and 74-78 (B-loop) that are disordered in most cases.

B204 is a 24.8 kDa monomer with nine β-strands and eight α-helices. The topology of B204 follows the FtsK-HerA superfamily fold belonging to the family of A32-like packaging ATPases (Iyer et al., 2004a; Iyer et al., 2004b) as I have suggested earlier based on homology modeling (II).

In this study I solved four different structures of B204: i) together with a sulphate ion; ii) with AMP; iii) with ATPγS (apparently hydrolyzed to ADP); and iv) with AMPPCP. The root-mean-square deviation (RMSD) value between the main chains between the four structures is less than 0.5 Å. The structures are all well ordered, except for two loop regions between amino acids 41-47 (A-loop) and 74-78 (B-loop) that are disordered in most cases.

The residues interacting with the nucleotide are K13, K14, G16, K17, S18, Y19, Y186, and I204 (III, Figure 3). The adenine moiety stacks between Y19 and Y186, whereas the sugar moiety does not interact with B204. While most of the interactions involve main chain atoms, the side chain of K17 binds the β-phosphate of ADP and AMPPCP, and the K13 side chain binds the γ-phosphate of AMPPCP (III, Figure 3b and c). The sulphate ion binding site maps to that of the β-phosphate. In the structure with ADP, the catalytic Mg²⁺-ion is coordinated by S18 in the Walker A motif, the β-phosphate, and four water molecules. In the AMPPCP structure one of the water molecules is replaced by a γ-phosphate oxygen. The structure in complex with AMP lacks the Mg²⁺-ion.

Two active site conformations were detected depending on the presence or absence of nucleotide in the P-loop (Table 3). The A chains in all structures are occupied by their respective ligands. Furthermore, in the structures with sulphate and ADP, all monomers in the AU are occupied by ligands. In the AMPPCP structure, chains A and B are occupied by the nucleotide and an Mg²⁺-ion, whilst the nucleotide in chains C and D is apparently hydrolysed to AMP. The P-loop in the AMP structure is occupied by a nucleotide in chain A, whereas the P-loop in chain B is empty. I thus compared chain B in the AMP structure and chains C and D in the AMPPCP structure to their respective A chains. The empty P-loops have a closed conformation compared to the open conformation in the presence of a nucleotide (III, Figure 3f). The maximum difference between these two P-loop conformations is 1.8 Å. On comparison of the sulphate ion structure with the AMPPCP structure (III, Figure 3c and d), the following changes in the active site in the absence of a nucleotide were noticed: i) the gap between Y19 and Y186 opens up; ii) K13 and K14 bend away from the nucleotide binding site; iii) E49 moves further away from the active site; and iv) S18 moves away from the cation-binding site (III, Figure 3f).

I also observed an additional metal-binding site formed by D39, H41, D73, and H108 in molecule A of the nucleotide-containing structures. The coordination geometry and ligands exclude Mg²⁺, but no other metal was present in the crystallization conditions. When the metal ion is absent, H41 flips
approximately 90° (III, Figure 4b). Based on anomalous difference Fourier maps, I observed peak heights of 4-8 σ at the suggested secondary metal ion binding site; precluding the density from being a water atom and confirming the presence of a transition metal such as Fe³⁺, Zn²⁺ or Ni²⁺ with tetrahedral coordination (Hsin et al., 2008). Thus, it was modeled as a Fe³⁺ ion (III, Figure 4a and b) (Table 3). Experiments are ongoing to determine the nature of this metal.

Table 3. B204 active site conformations.

<table>
<thead>
<tr>
<th>Structure</th>
<th>ChainA</th>
<th>ChainB</th>
<th>ChainC</th>
<th>ChainD</th>
<th>Ligands</th>
</tr>
</thead>
<tbody>
<tr>
<td>SO₄²⁻</td>
<td>SO₄²⁻</td>
<td>SO₄²⁻</td>
<td>N/A</td>
<td>ligand</td>
<td></td>
</tr>
<tr>
<td>empty</td>
<td>empty</td>
<td>empty</td>
<td>open</td>
<td>metal</td>
<td></td>
</tr>
<tr>
<td>empty</td>
<td>empty</td>
<td>empty</td>
<td>open</td>
<td>conformation</td>
<td></td>
</tr>
<tr>
<td>AMP</td>
<td>AMP</td>
<td>empty</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>empty</td>
<td>empty</td>
<td>empty</td>
<td>open</td>
<td>nucleotide</td>
<td></td>
</tr>
<tr>
<td>Fe³⁺</td>
<td>open</td>
<td>closed</td>
<td>closed</td>
<td>metal</td>
<td></td>
</tr>
<tr>
<td>AMP</td>
<td>ADP</td>
<td>ADP</td>
<td>ADP</td>
<td>nucleotide</td>
<td></td>
</tr>
<tr>
<td>Mg²⁺</td>
<td>Mg²⁺</td>
<td>Mg²⁺</td>
<td>Mg²⁺</td>
<td>cofactor</td>
<td></td>
</tr>
<tr>
<td>empty</td>
<td>Fe³⁺</td>
<td>open</td>
<td>open</td>
<td>metal</td>
<td></td>
</tr>
<tr>
<td>open</td>
<td>open</td>
<td>open</td>
<td>open</td>
<td>conformation</td>
<td></td>
</tr>
<tr>
<td>ATPγS</td>
<td>AMP</td>
<td>AMP</td>
<td>AMP</td>
<td>AMP</td>
<td></td>
</tr>
<tr>
<td>empty</td>
<td>Mg²⁺</td>
<td>empty</td>
<td>empty</td>
<td>nucleotide</td>
<td></td>
</tr>
<tr>
<td>Fe³⁺</td>
<td>open</td>
<td>empty</td>
<td>empty</td>
<td>metal</td>
<td></td>
</tr>
<tr>
<td>open</td>
<td>open</td>
<td>empty</td>
<td>empty</td>
<td>conformation</td>
<td></td>
</tr>
<tr>
<td>AMPPCP</td>
<td>AMPPCP</td>
<td>AMPPCP</td>
<td>AMPPCP</td>
<td>nucleotide</td>
<td></td>
</tr>
<tr>
<td>Mg²⁺</td>
<td>Mg²⁺</td>
<td>empty</td>
<td>empty</td>
<td>cofactor</td>
<td></td>
</tr>
<tr>
<td>empty</td>
<td>Fe⁺⁺</td>
<td>empty</td>
<td>empty</td>
<td>metal</td>
<td></td>
</tr>
<tr>
<td>open</td>
<td>open</td>
<td>closed</td>
<td>closed</td>
<td>conformation</td>
<td></td>
</tr>
</tbody>
</table>

4.2.9 Proposed catalytic cycle of B204 (III)

I would like to propose a model for the catalytic activity of B204, similar to that described for T4 gp17 (Sun et al., 2008) and φ12 P4 (Mancini et al., 2004). In φ12 it has been proposed that the nucleotide binding locks the active site in the “up” configuration, which in B204 would correspond to the “open” conformation. Upon ATP hydrolysis, the P-loop in φ12 P4 has been shown to swivel down, simultaneously pushing the RNA translocation. This “down” conformation would, in B204, correspond to the “closed” conformation. In T4 gp17 it is the C-terminal domain responsible for DNA-binding that translocates the genome, whereas the N-terminal domain provides the ATPase activity (Sun et al., 2008). In T4, the gp17 structure flexes between the “tensed” DNA-bound with the hydrolysis reaction occurred and the “relaxed” structures (Sun et al., 2008). Likewise, the “tensed” structure could be compared to the B204 “closed” structure. With the current knowledge at hand, I cannot propose a more detailed description of the possible catalytic cycle of B204 or confirm which amino acids are responsible for STIV2 genome translocation. This awaits mutagenesis studies and further structural work.
4.2.10  B204 DNA-binding and proposed DNA-binding site (III)

I performed electrophoretic mobility shift assays to analyze B204 binding to DNA. B204 binds both linear and circular DNA, as well as both dsDNA and ssDNA (III, Figure 2e). Interestingly, B204 did not show any sequence specificity in DNA binding, as the linear dsDNA used was a PCR product of the STIV2 major capsid protein A345, and the other DNA molecules were φX174 DNA. I also ordered short dsDNA fragments in order to study the minimal length of DNA molecule bound by B204. Using the electrophoretic mobility shift assay I showed that B204 binds dsDNA-fragments from 20 bp upwards, whereas 15 bp seems to be too short for B204 to bind (III, Figure 2f).

Many viral genome packaging ATPases exist as higher multimers and many of the structural B204 homologues identified by DALI (Holm and Rosenström, 2010), such as TrwB (Gomis-Rueth et al., 2001), FtsK, (Massey et al., 2006) and the DnaB helicase (Bailey, Eliason, and Steitz, 2007), exist as hexamers. However, our structures are monomeric, perhaps unsurprisingly, as many viral DNA-translocating proteins form the functional, ring-like multimer only upon binding to the virus capsid (Guo and Lee, 2007). I thus created a model of a hexameric B204 based on the FtsK (Massey et al., 2006) hexamer (III, Figure 4d) (RMSD 1.8 Å). In this model the nucleotide binding site of B204 is at the interface between two monomers, as in the hexameric φ12 P4 (Mancini et al., 2004) and FtsK (Massey et al., 2006). Furthermore, FtsK is known to process dsDNA and in the hexameric model of B204 the channel is of a suitable size for dsDNA genome translocation. The DNA-binding domain of FtsK has been shown to reside in the channel of the hexamer (Sivanathan et al., 2006). This region has 30 % sequence similarity to the corresponding region in B204 and orients towards the interior of the hexameric channel in our model (III, Figure 4d, red color).

4.2.11  Proposed model for STIV2 genome packaging using B204 (III)

Based on our electrophoretic mobility shift assays, I can conclude that B204 binds DNA in a non-sequence-specific manner since it binds DNA both from STIV2 and φX174, as well as short, commercial dsDNA oligonucleotides. This indicates that B204 does not bind to a special genome packaging sequence on the STIV2 genome. This further leads me to hypothesize that STIV2 encodes for a DNA-binding protein that binds the genome and recruits it to the biological, multimeric B204 complex for DNA translocation. So far, I have identified one DNA-binding protein in STIV2 (B72) that could perform this task (II). The genome packaging would initialize with B72 binding to the postulated linear dsDNA genome of STIV2 (III, Figure 5). Simultaneously, procapsid assembly would begin at the postulated packaging vertex of STIV2. It is assumed that B204 takes it proper quaternary form when assembling into the capsid. Small membrane proteins would form the pore in the membrane for genome translocation, much like the membrane proteins P20 and P22 in PRD1 (Gowen et al., 2003; Strömsten, Bamford, and Bamford, 2003). STIV2 has 7 suggested membrane proteins (A259, C141, E132, A55, A103, E51, and E76b) that could carry out this task (II, III; Supplemental material). The most likely candidates are the small membrane proteins E51, A55, and E76b. Furthermore, in STIV, the transmembrane proteins have been suggested to serve as tape-measure scaffolding proteins and assist in correct assembly of capsid proteins (Fu et al., 2010). Whether this is true for STIV2 is not yet known.
The recruitment of the STIV2 genome to a multimeric form of B204 assembled on the capsid by B72 would form the motor complex that, upon ATP hydrolysis, would translocate the genome inside the procapsid. The STIV2 genome was shown to exist as a circular dsDNA molecule when isolated from virions (II). I believe that the genome is packaged in linear form and circularized once inside the capsid. In STIV it has been shown that the procapsid does not undergo major conformational changes upon maturation (Fu et al., 2010). The same is believed to be true for STIV2.

4.2.12 Implications for the PRD1-like viruses (III)

B204 of STIV2 is the first genome packaging NTPase from a membrane-containing virus for which the structure has been solved. Furthermore, the structure of B204 is the first structure of an NTPase from a non-tailed dsDNA virus to be solved, as well as the first structure from an archaeal virus to be described. The work on B204 thus sets the foundation for future studies on other genome packaging NTPases from membrane-containing viruses.

According to the theory of the viral self, there are structures and functions in the viral life cycle that form the innate self of a virus, such as virion assembly and genome packaging (Bamford, 2003). This is based on observations that the genes coding for the major capsid protein and the genome packaging ATPase are next to each other in the genome of many viruses and implies that they are inherited as one genetic entity in the course of evolution (Bamford, 2003; Bamford, Burnett, and Stuart, 2002). Since some viruses have been divided into lineages based on the fold of their major capsid protein (Bamford, 2003; Benson et al., 2004), and since I postulated that STIV2 would belong to the PRD1-adenolinessage of viruses based on the fold in its homology modeled major capsid protein (II), I took a closer look at the ATPases of the PRD1-adenolinessage. I used B204 as a structural constraint and generated homology models of the ATPases of the viruses suggested to belong to the PRD1-adenolinessage (III, Supplemental figure 3). The viruses for which I managed to generate reliable models belong both to the double β-barrel (Bam35, PM2, CIV, PRD1, Mimivirus, and STIV) and the single β-barrel lineage of viruses (P23-77 and SH1) (Krupovic and Bamford, 2008a). This further strengthens the argument for the viral self hypothesis.

Notably, the closest relative to STIV2 is STIV. The sequence identity between the STIV B164 and STIV2 B204 is 70 % (III, Supplemental table 1). The structure of STIV2 B204 and the homology modeled STIV B164 are similar (Figure 15), except that STIV B164 is missing the two C-terminal α-helices. Due to the high sequence identity of B204 and B164, it is likely that soluble B164 might be produced and successfully purified using the expression and purification protocol that I developed (III). Wirth et al. (Wirth et al., 2011) have recently described the development of a genetic system for STIV studies. If a B164 mutant STIV virus could be produced, this would open up new avenues for the study of the function of the PRD1-like genome packaging NTPases in an extremophilic environment to compare with studies on PRD1 P9 (Strömsten, Bamford, and Bamford, 2005; Žiedaitė et al., 2009).

4.2.13 Proposed factors contributing to the thermostability of B204 (III)

In an attempt to understand the factors contributing to the thermostability of B204, I took advantage of our homology models and compared the structures to each other. The main interactions I looked at were the number of hydrogen bonds and salt bridges. The analysis on the
number of disulfide bridges were omitted, as B204 does not contain any cysteine residues. Furthermore, based on the sequences, I analyzed whether the number of isoleucines, arginines, glutamic acids, lysines or prolines were increased, or whether the number of serines, asparagines, glutamines, threonines or metionines were reduced; Haney et al. (Haney et al., 1999) have observed that these are trends found in hyperthermophilic proteins. As I have three thermophilic models available (Figure 15) – the NTPases of STIV2, STIV and P23-77 – a trend was considered reliable if observed in all of these compared to five mesophilic models (the NTPases of PRD1, Bam35, PM2, CIV, and Mimivirus).

Figure 15. Structures of thermophilic NTPases. (a) STIV2 B204 in complex with AMPPCP and Mg$^{2+}$. (b) Homology model of STIV B164. (c) Homology model of P23-77 ORF13. B164 lacks the two C-terminal α-helices compared to B204 (highlighted with black open circles). Furthermore, the core β-strands in the model of the P23-77 ORF13 are missing. The figure was rendered in PyMOL.

Summarizing (III, Supplemental table 2), the thermostability of B204 seems mainly to be due to the amount of hydrogen bonds and salt-bridges present. These properties were, on the other hand, not shared by STIV B164 or P23-77 ORF13, which seemed to have the same number of hydrogen bonds and salt bridges as their mesophilic homologues. In contrast, the features shared by all three thermophilic viruses were the reduced number of asparagine, glutamine and threonine residues as compared to the mesophiles. Furthermore, in STIV2 B204 and STIV B164 the number of isoleucine residues is increased compared to the mesophiles. In addition to being stabilized by the intrinsic factors discussed above, extrinsic factors such as nucleotides stabilize B204 as well (see section 4.2.6).
4.3 The origin of viruses and viral lineages

What is a virus? Nucleic acid surrounded by a protein capsid? Are viruses alive? How have viruses emerged? There are several unresolved questions regarding the nature of viruses. Currently there are three conflicting theories trying to explain the emergence of these parasites, as summarized by Forterre, and Hendrix and colleagues (Forterre, 2006; Hendrix et al., 2000). Firstly, viruses can be seen as relics of a pre-cellular life form. Secondly, viruses might be derived by reduction from unicellular organisms. Lastly, viruses might originate from fragments of genetic material that escaped from the control of the host cell and became parasitic. The first theory has been dismissed for some time, since all currently known viruses are obligate parasites. We are, however, faced with the fact that we might soon have to revisit our view on the nature of viruses, as the giant mimivirus encodes many genes involved in protein translation during its infection cycle (Abergel et al., 2007; Claverie and Abergel, 2010; Legendre et al., 2010). It has been suggested that a more proper division than living (cells) and non-living (viruses) creatures could be made into ribosome-encoding (cells) and capsid encoding (viruses) organisms (Raoult and Forterre, 2008). The question of the definition of life and whether viruses are living or not, still causes heated debate (Claverie and Ogata, 2009; Koonin, Senkevich, and Dolja, 2009; Ludmir and Enquist, 2009; Moreira and Lopez-Garcia, 2009; Raoult, 2009) more than a century after the discovery of bacteriophages (Beijerinck, 1898; Iwanowski, 1892). The second theory has also been rejected for some time, since until the discovery of mimivirus there was no known intermediate between a virus and a cell. Once again, mimivirus might challenge this theory (Claverie and Abergel, 2010; Raoult and Forterre, 2008). Lastly, the escape theory has acquired support from the observation that contemporary viruses can integrate host genes into their genome. Furthermore, believers in this theory regard plasmids and mobile genetic elements as viral precursors (Forterre, 2006). It has further been suggested that certain genes might occasionally mutate to encode proteins that can self-assemble into (icosahedral) capsids and that these capsids occasionally might trap genetic material, forming a proto-virus contributing to genetic exchange and enhanced fitness of its host by horizontal gene transfer (Hendrix et al., 2000). Such proteins self-assembling into icosahedral particles have been isolated, such as the virus-like particle of Pyrococcus furiosus with an HK97-fold (Akita et al., 2010). Despite controversy regarding the origin of viruses they are suggested to have a pivotal role in the early evolution of cellular life (Forterre, 2006; Hendrix et al., 2000; Koonin, Senkevich, and Dolja, 2006) and still are one of the main driving evolutionary forces (Brockhurst et al., 2007; Paterson et al., 2010).

Leaving the topic of the origin of viruses, an equally challenging and stimulating debate is going on about the relationship between different viruses (Bamford, 2003; Bamford, Burnett, and Stuart, 2002; Bamford, Grimes, and Stuart, 2005; Benson et al., 2004; Hendrix et al., 1999; Iyer, Aravind, and Koonin, 2001; Iyer et al., 2006). These taxonomical classifications are blurred to some extent by the high level of mosaicity in the tailed bacteriophages (Hendrix, Hatfull, and Smith, 2003; Pedulla et al., 2003). The variability in virus genomes poses a significant technical challenge for comparative evolutionary genetics (Shakhnovich and Shakhnovich, 2008). For instance, as discussed above, it has been shown that no single gene is shared by all bacteriophages (Rohwer and Edwards, 2002). While the genomes of tailed dsDNA phages have been shown to be highly mosaic (Hendrix et al., 1999; Pedulla et al., 2003), those of the Tectiviridae family seem to have escaped horizontal gene transfer
(Bamford et al., 2005). The fact that we do not observe horizontal gene transfer events in tectiviruses could, on the other hand, be attributed to the small number of tectiviruses classified. According to the International Committee on the Taxonomy of Viruses, the *Tectiviridae* family has 5 species assigned, namely Bacillus phage AP50, Bacillus phage Bam35, Enterobacteria phage PRD1, and Thermus phage P37-14 as well as the unassigned Bacillus phage φNS11. This is in strong contrast with the number of isolated tailed phages (*Caudovirales*) that constitute up to 95% of the approximately 5,500 phages isolated to date (Ackermann, 2007).

Due to horizontal gene transfer and genome mosaicity, the classifications of viruses relies mainly on their structure and on their nucleic acid type (see section 1.2). Furthermore, the structure of the viral major capsid proteins (see section 1.4) and the presence of a RNA-dependent RNA polymerase in the viral genome are equally used basis for classifying viruses and for attempting to explain their evolutionarily relationships (Bamford, 2003; Bamford, Grimes, and Stuart, 2005; Koonin et al., 2008). Additionally, as the major capsid proteins and the genome packaging ATPases seem to be linked during the course of evolution, the proposed virus classification into lineages utilizes the presence of these so-called viral “self” genes (Bamford, 2003; Bamford, Burnett, and Stuart, 2002; Bamford, Grimes, and Stuart, 2005). Such lineages identified to date are the PRD1-like, the HK97-like, and the bluetongue virus-like viruses (Bamford, Grimes, and Stuart, 2005) as well as the monophyletic NCLDV (Iyer, Aravind, and Koonin, 2001; Iyer et al., 2006).

### 4.3.1 Evolution of viruses

As introduced, viruses are proposed to form families and lineages that precede the division of the three domains of life. This idea seems to be more established for certain viral families, such as the PRD1-like, HK97-like, and the BTV-like viruses. The similarities of the major capsid proteins, the arrangement of the viral “self” genes and the proposed viral lineages and families have raised questions about virus evolution. There are three main explanations for this similarity between viruses infecting different domains of life. Firstly, there could have been massively convergent evolution, due to the fact that there are many different virus morphologies that can infect Archaea, Eukarya, and Bacteria. A second explanation is that there has been horizontal gene transfer among viruses infecting the different domains of life. In this case, it is likely that the horizontal gene transfer has happened a long time ago in order for the sequence resemblance to have disappeared. If this is the likely scenario, some mechanisms that maintain the sequence similarities of the major capsid protein and the genome packaging ATPase – the viral “self” (Bamford, 2003; Bamford, Burnett, and Stuart, 2002) – must exist. Interestingly, in all the prokaryotic viruses proposed to belong to the PRD1-like viruses, the genes for the major capsid protein and the genome packaging ATPase are always found close to each other, with the ATPase residing upstream of the major capsid protein encoding gene (Krupovic and Bamford, 2008b). The reason for this has been suggested to be to keep these genes together despite possible recombination (Krupovic and Bamford, 2008b). To me, an equally likely explanation is that these genes are moving together via horizontal gene transfer due to the fact that they have proven to provide an efficient toolbox for gene encapsidation. This order of the ATPase and major capsid protein encoding genes is conserved in STIV2 and P23-77 as well. Furthermore, as proposed in this study, P23-77 encodes a postulated excisionase gene. This indicates that horizontal gene transfer might occur in the *Tectiviridae* lineage as well, as has been shown for tailed phages (Pedulla et al., 2003). Finally, there might have existed a common ancestor
to the PRD1-like viruses that would have preceded the divergence of the three domains of life some 3 billion years ago (Bamford, 2003; Bamford, Burnett, and Stuart, 2002; Rice et al., 2004).

So are STIV and STIV2 related? Despite the fact that one is isolated from a hot spring in Yellowstone National Park (USA) and the other from a host spring on Iceland, the genomic organization and sequence similarity between STIV and STIV2 are highly conserved, as are the identity of the virion associated proteins and the overall three-dimensional organization. Even though they were initially described to infect different Sulfolobus species, I later showed that they are capable of infecting the same host. To me, these viruses are clearly related. The question remains on how their ancestor has crossed the ocean and ended up in two globally very isolated hot springs. Results by Snyder et al. (Snyder et al., 2007) show that geographically isolated hot springs do exchange viruses, and that the viral diversity in these ponds are foremost affected by migration as opposed to mutation. Furthermore, these migrating viruses have been shown to be airborne and viral migration between metapopulations is suggested to possibly be a global phenomenon. This would be a very convenient explanation for the detection of two highly similar viruses – STIV and STIV2 – in such distant locations.

How does this relate to the similarity between P23-77 and SH1? The gene order over the viral “self” genes is conserved, as is the overall three-dimensional structure, omitting the host-attachment structures. Both viruses are suggested to encode two major capsid proteins that would form single β-barrel proteins. Each capsomer with a hexameric base would thus be composed of six copies of this single β-barrel, as opposed to the PRD1-like viruses in which the hexameric capsomers are composed of three copies of the double β-barrel major capsid proteins. This double β-barrel protein seems to have arisen via gene duplication; a common mechanism in protein evolution. In P23-77 and SH1 we see the only identified viruses thus far with single β-barrels as their major capsid proteins that could have been the predecessor of the double β-barrel protein viruses. However, I would like to argue that the similarity seen in P23-77 and SH1 does not necessarily mean that their ancestor existed before the divergence of Bacteria and Archaea. The closest relatives to P23-77 and SH1 are the plasmid pH205 of Halobacterium salinarum and the chromosome integrated provirus IHP of Haloarcula marismortui (Jalasvuori et al., 2009). Later, five additional integrated genomic elements similar to P23-77 and SH1 were identified (Jalasvuori, Pawlowski, and Bamford, 2010). These elements have been predicted to contain genes encoding integrases and proteins required for homologous recombination. The presence of such genes indicates the mobility of these elements and suggests that horizontal gene transfer might have been prevalent at some stage. Thus, nothing precludes that P23-77 and SH1 were actually created after the divergence of the domains Archaea and Bacteria, especially as viruses have been proposed to have access to a global genetic pool. Even though the host-attachment structures of P23-77 and SH1 differ significantly, the genes encoding these proteins are suggested to belong to the viral “non-self” genes (Bamford, 2003) and can thus be exchanged between unrelated viruses or even between the virus and the host chromosome through horizontal gene transfer (Hendrix et al., 1999). Thus, I would like to conclude that P23-77 and SH1 most likely form a single β-barrel lineage predating the double β-barrel viruses and are related despite infecting hosts from different domains of life. I furthermore would like to argue that genetic exchange is still prevalent in this lineage of viruses and has contributed to both the spreading of these single β-barrel genome integrated proviruses and to the acquisition of genes encoding for the host-attachment structures of P23-77 and SH1. However, as P23-77 infects a
bacterium living in a hot, alkaline spring and SH1 is an archaeon living in a salt lake, it is almost certain that they are evolutionarily more distant than STIV and STIV2.
5. CONCLUSIONS AND FUTURE PROSPECTS

In this thesis I have studied two extremophilic icosahedrally-ordered viruses with an internal membrane – P23-77 and STIV2. The publication on P23-77 included in this thesis (I) describes the biochemical characterization of the virus, as well as its overall three-dimensional structure analyzed by cryoEM. Furthermore, a detailed structural comparison to the SH1 virus with similar capsid architecture highlights the common themes in assembly of viruses with the proposed single β-barrel major capsid proteins. One of the main obstacles in analyzing the P23-77 virus was that I lacked its genome sequence. Thus, when the studies on STIV2 were undertaken, the genome sequence of this virus was deduced and provided me with the possibility to analyze and compare it with other viruses. Furthermore, with the genome sequence available, it was possible to make a more in-depth analysis of the viral proteins using mass-peptide fingerprinting to identify the structural STIV2 proteins. With the genome sequence at hand and the structural proteins identified, the analysis of the three-dimensional STIV2 structure solved by cryoEM became much more feasible and informative. As I also managed to homology model a handful of the STIV2 proteins, the major capsid protein being one of them, I was able to do difference imaging on the viral capsid, highlighting the details of the host-attachment structures, and to assign STIV2 to the PRD1-like virus lineage, which paved the way for my hypothesis on the genome packaging of these viruses.

Even though the genome packaging machineries on many icosahedrally-ordered viruses have been studied in detail, the structure of a genome packaging NTPase from a membrane-containing virus has remained elusive. In publication III included in this thesis, I described the structure of the genome packaging NTPase B204 of STIV2 solved by X-ray crystallography. I additionally studied the nucleotide hydrolysis of B204 and showed that it binds DNA – a necessity for dsDNA genome packaging. One of the main difficulties in the studies on P23-77 and STIV2 has been the lack of homologous sequences in public databases, reflecting the under sampling of viruses from extreme environments. This has made homology based modeling of proteins impossible in many cases. With the structure of B204 at hand, I managed to generate homology models for three additional extremophilic viral genome packaging NTPases – P23-77 being one of them – as well as for five genome packaging NTPases of viruses from mesophilic hosts. The aforementioned are NTPases from the PRD1-like viruses, previously recalcitrant to homology modeling or structural determination. As B204 is derived from a hyperthermoacidophilic virus, and because I had eight other generated structures to compare it to, I could deduce some general trends for the reasons of the thermostability of B204. Thermostable enzymes are of high interest due to their industrial and biotechnological applications, and being able to engineer proteins for higher thermostability is of interest. If the factors contributing to thermostability of proteins can be deduced by comparing thermostable proteins to mesophilic ones, as opposed to trial-and-error mutagenesis, it would have added value for industry.

The questions addressed in this thesis raise many new ones. A deeper analysis of both STIV2 and P23-77 would benefit from subnanometer cryoEM reconstructions. Currently, this path will not be followed for STIV2 due to the immense difficulty of virus production. For P23-77 I am currently working on a higher resolution reconstruction. Furthermore, the major capsid proteins VP16 and VP17 of P23-77 have been crystallized (Rissanen et al., 2012) and the impending structures should
provide added benefit to the analysis of the P23-77 capsid architecture. Continued dissociation studies or controlled proteolytic cleavage of the P23-77 virion should provide insight into the nature of the pentamer and the host-attachment spike. In case the spike protein can be identified and isolated, this would allow infection studies of the host. In combination with studies on host entry and exit, it would provide important knowledge on the infection process of extremophilic viruses. Equally puzzling is the structure and composition of the STIV and STIV2 host attachment structures, which would require a more in-depth study. As for P23-77, the infection process of these viruses is largely unknown.

I would furthermore like to dare to address the issue on the monophyletic origin of various viral families and lineages. As more and more sequences have been analyzed (Jalasvuori et al., 2009; Jalasvuori, Pawlowski, and Bamford, 2010; Krupovic and Bamford, 2008a; Krupovic and Bamford, 2008b), it has become evident that many of the members proposed to belong to the vertical β-barrel lineage are what seems to be, or have been mobile genetic elements, such as plasmids, integrated proviruses with identified genes encoding for transposases, and Mavericks. The high number of these elements – eight all together - compared to the number of contemporary viruses – nine in the double β-barrel and two in the single β-barrel lineage according to Krupovic and Bamford (Krupovic and Bamford, 2008a) – indicates to me that horizontal gene transfer has contributed to the spread of this lineage.

The key findings in this thesis highlight the conservation of the major capsid protein and genome packaging NTPase structures, as well as add to our sampling on viruses proposed to belong to the vertical β-barrel lineage of viruses. Furthermore, evident both in P23-77 and STIV2, is the conserved arrangement of the major capsid protein and the genome packaging NTPase in the genomic DNA; a feature shared by viruses belonging to the vertical β-barrel lineage of viruses. Even though previous studies have shown that the NTPase P9 of the type virus PRD1 is not able to package genomes other than its own (most likely due to its requirement of the genome terminal P8 protein), an interesting aspect of the viral self genes would be to analyze to what extent these key components – the major capsid protein and the NTPase – are exchangeable between viruses, such as STIV and STIV2 or, on the other hand P23-77, SH1, and HHIV-2.
ACKNOWLEDGEMENTS

This work was carried out between 2007 and 2011 at the Finnish Centre of Excellence in Virus Research (2006-2011), at the Institute of Biotechnology and Department of Biosciences at the University of Helsinki. I would like to thank Professors Mart Saarma and Tomi Mäkelä for providing the excellent research facilities and the stimulating working environment at the Institute of Biotechnology. This work was financed by the Viikki Doctoral Programme in Molecular Biosciences (VGSB). I would like to thank Professor Marja Makarow for accepting me as a student into VGSB, and Dennis Bamford who took over the directorship of VGSB and additionally acted as the head of the Finnish Center of Excellence in Virus Research, for providing inspiring student get-togethers and valuable infrastructure in our CoE.

My greatest thanks and biggest hugs go to my supervisor Professor Sarah J. Butcher. I wish to express my deepest gratitude to Sarah for all her support during these hectic years. Sarah has been an immense source of inspiration, energy, ideas, and evil laughter that I will miss greatly.

I would like to thank my follow-up group members Ari Huovila, Alexander Plyusnin, and Roman Tuma for their time and interest in the wonder of extremophilic viruses. Our meetings, although rare, have been a source of new ideas and inspiration, as well as guidance. Roman Tuma is especially thanked for introducing me to mass-spectrometry, for his willingness to travel with me to Alabama – not once, but twice – and foremost for his valuable comments on the manuscript on the STIV2 NTPase.

Nicola G. A. Abrescia and Pirkko Heikinheimo are thanked for carefully reviewing this thesis and for their helpful comments.

Kirsty Williamson is warmly thanked for carefully correcting the English language in this thesis.

I would like to thank all my co-authors (in chronological order) – Silja Jaatinen, Pasi Laurinmäki, Dennis Bamford, Peter Redder, Xu Peng, Laila Reigstad, David Prangishvili, Tommi Kajander, Lassi Liljeroos, and Adrian Goldman – for good collaboration and intellectual discussion on the extremely interesting topics at hand.

Professor Nikolaus Grigorieff is warmly thanked for letting me visit his research group and for teaching me a vast amount on 3DEM, in addition to providing stimulating discussions on life as a scientist. The work carried out in his research group is presented in this thesis as unpublished results.

Heidi Repo and Robert Kolodziejczyk are warmly thanked for their help with the Thermofluor assays. Sonja-Verena Albers is acknowledged for sharing her expertise on enzyme assays. The Sicilian ERASMUS-students Annunziata Pennino and Salvatore di Girolamo are thanked for their contribution to the STIV2-project.
The Electron Microscopy Unit, the Protein Chemistry Core Facility, and the Protein Crystallization Facility at the Institute of Biotechnology are warmly acknowledged for excellent collaboration and technical expertise.

Professors Tapio Palva and Juha Partanen at the Department of Genetics are thanked for their help during my PhD studies. Jussi Alho is acknowledged for his help with all the bureaucracy related to graduation, and Anita Tienhaara for help and advice while trying to put this thesis into its final format. A special thanks goes to Pekka Heino at the Department of Genetics for dealing with all study related issues during these years, for always having his door open, and for always having time for a chat.

I especially want to thank all the past and present members of the colorful Butcher group for providing a stimulating, fun and crazy research environment filled with laughter. I also wish to thank all the wonderful people in the Goldman, Heikinheimo, Iwai and Bamford groups for shared research interest, friendship, laughter and an occasional beer or two.

Very special thanks go to all my fantastic friends. Thank you for all the shared joy, drinks, copious amounts of food, excellent discussions, exercise, sewing clubs, and for having lovely children I have been allowed to play with to ease the frustration of failed experiments and approaching deadlines.

My mother Yrsa and aunt Siv are thanked for their endless love and support during these hectic years; for providing me with food and money, for accompanying me on cross-country skis, on tennis courts and golf courses, and for providing refreshing travel to balance the workload. My aunt Siv is especially thanked for providing access to heaven on Earth – Engudden.

Kalle and Kaisa, thank you for your endless interest (flaggar ni redan?) in my Supertheory of Supereverything – flax!

Finally, I would like to express my deepest love and gratitude to my best friend and husband, my collaborator and my own walking wikipedia, Esko.

Malmö, Sweden April 2012
REFERENCES


Koonin, E. V., Senkevich, T. G., and Dolja, V. V. (2009). Compelling reasons why viruses are relevant for the origin of cells. *Nat Rev Microbiol* 7(8), 615; author reply 615.


Raoult, D. (2009). There is no such thing as a tree of life (and of course viruses are out!). *Nat Rev Microbiol* **7**(8), 615; author reply 615.


