Structural investigation of two viral proteins involved in DNA-packaging

Juan Luis Loredo Varela

PhD

University of York
Chemistry

September 2014

Abstract

Viral DNA-packaging motors are molecular machines that enable viruses to replicate by providing the means of storing the viral genome into empty procapsids. In double stranded DNA bacteriophages the molecular motor is comprised by three elements: the portal protein and the small and large terminases. The portal protein nucleates polymerisation of the capsid and scaffold proteins, initiating procapsid assembly, besides allowing passage of DNA. The small terminase recognises the viral genome and presents it to the large terminase which possesses both nuclease activity to cleave concatemeric DNA at the initiation of the packaging, and ATPase activity to drive translocation.

Although currently several X-ray structures for the different components of the motor from different bacteriophages are available fundamental questions regarding the DNA recognition mechanism, stoichiometry and orientation of the motor components *in vivo* and the mechanism of ATP-driven DNA-translocation remain. This project focused on elucidating the X-ray crystal structures of (i) the major capsid protein, from *Bacillus subtilis* bacteriophage SPP1 and (ii) the small terminase protein from *Thermus thermophilus* bacteriophage G20C.

Several constructs of the SPP1 capsid protein, including truncations at the N- and C-termini, single and double mutants and engineered proteins were generated in order to produce a protein suitable for crystallisation. Mutant uG*13*P;T104Y;A261W was the only construct that produced native and Se-Met crystals. The X-ray structure of the SPP1 capsid protein was determined at 3.0 Å resolution by single wavelength anomalous diffraction. The structure exhibited the HK97-fold consisting of the axial and peripheral domains and the extended E-loop.

The X-ray structure of the small terminase from phage G20C was solved at 2.5 Å resolution by single wavelength anomalous diffraction. The structure consisted of circular nine-mers where the N- terminal domains of each subunit reside at the periphery of the assembly and the C-terminal oligomerisation domains form a central channel. The conserved structural features between small terminases suggest that the DNA-recognition mechanism might be conserved. DNA-binding experiments demonstrated that the G20C small terminase binds to viral and non-viral DNA with both the N- and C- terminal domains playing an important role.

The structural information generated from these two elements of SPP1 and G20C bacteriophages provides insight into two different aspects of the assembly process: (i) how the capsid protein's conformational plasticity may assist the assembly and (ii) DNA-recognition during virus particle construction. This information is critical for understanding similar processes in other viruses, in particular, in the evolutionarily related herpes viruses.

List of contents

List of figures

List of tables

Author’s preface

This thesis describes the experimental work performed on producing, purifying, crystallising and solving the X-ray structure of two phage proteins, one from the SPP1 and one from the G20C bacteriophages. Structural data are analysed in order to gain insight into the phage assembly and in particular, into DNA packaging.

The introduction provides the reader with the necessary information, ordered to follow the content to acquire a general perspective of the knowledge available for their counterparts in other dsDNA phages. Moreover, the introduction highlights questions regarding DNA-translocation that remain unclear to date and that several research groups around the world strive to answer. Due to the Antson Laboratory’s long history of working with the SPP1 phage as a model regarding the single components of the molecular motor, and the amount of biochemical data available for this system from other laboratories, the SPP1 molecular motor is used as a reference point in the introduction in the same way as the HK97 capsid protein serves as a reference for the common fold in capsid proteins in dsDNA phages.

The outcome of this four-year project is organised in eight chapters that describe the efforts to produce crystals of the SPP1 capsid protein and G20C small terminase, the progress made in elucidating the X-ray structure of each protein and the biochemical characterisation of the G20C small terminase. A wealth of data was produced, however, only the most relevant information for the project will be presented. Where necessary, additional data will be presented in figures, tables or appendices. Not-shown data will also be referred to where similar results were found in repeated experiments.

Most of the protocols used to develop this project are common practice in protein production laboratories or X-ray crystallography facilities, therefore just a brief description of each protocol will be provided, with the understanding that the specific settings were varied for each protein or construct in turn.

Finally, it is important to mention the significance of studying each protein. For SPP1, there are X-ray structures for the large terminase nuclease domain, portal protein and small terminase from the SPP1-like phage SF6, all from the Antson Laboratory. NMR structures are available for proteins connecting the portal with the tail and 3D cryo-EM reconstructions are available for the whole capsid and tail (as detailed in the Introduction section). The crystal structure of the capsid protein will account for one of the key structural components of the phage. Structural plasticity of this protein plays a key role during capsid maturation and capsid assembly allowing (i) construction of 5-fold and 6-fold symmetrical circular assemblies that are the building blocks of the capsid and (ii) controlling changes in capsid's shape and volume during different steps in its assembly.

Investigation of the G20C small terminase adds one more structure to the available collection of small terminase structures. How exactly this protein interacts with DNA remains a matter of debate in the scientific community. Thus biochemical and structural characterisation of the G20C small terminase is valuable for understanding how this protein functions. Moreover, to date little is known about proteins from thermophilic phages, and no accurate structural information was available about their components.

Acknowledgements

I would like to thank my supervisor and mentor, Professor Alfred Antson, for his support and guidance throughout my PhD. I would also like to thank him for the opportunity to come to England and be part of a project that has fulfilled all my professional and personal expectations. Thanks to his leadership and support I can say that my projects were successful both with regards to the experimental results obtained and the knowledge and skills I gained during the time I was part of his most capable and welcoming group.

Special thanks are owed to my funding council, the Mexican National Council of Science and Technology (Consejo Nacional de Ciencia y Tecnología de México, CONACyT) for the sponsorship to cover my tuition fees and living expenses during the four years of my PhD.

Thanks are extended to Mrs. Maria Chechik, for her extensive teaching and guidance in the laboratory and help in data analysis and experiment design in every stage of my projects since my first year at the YSBL.

I would also like to thank to Dr. Callum Smits for his help to shape the SPP1 project, to Dr. Sandra Greive for assisting during the experiment design in the G20C small terminase project and discussion during presentations and writing of this thesis, and to Dr. Huw Jenkins for his help during solving the structures of the SPP1 capsid protein and G20C small terminase. The three persons mentioned above additionally provided help at the laboratory, teaching, help in data analysis and experiment design.

Thanks to Dr Carina Bütner, Dr Elena Blagova, Weisha Luan and Dan Peters for helping during either experiment design or general advice at the lab.

Thanks to Oliver Bayfield for proofreading this thesis and offered advice to complete the literature review. Thanks to Christian Roth for the advice to write chapter 4.

Thanks to the Antson's group and the rest of the YSBL people for making my time at the laboratory enjoyable and for the general support during these years.

Thanks to my family and friends for the moral support and motivation to continue in this journey that has been so far my most ambitious project.

I want to dedicate this work to my mother for the continuous support to my professional career and to my sister, brother, father, niece and nephew for inspiring me to challenge myself.

Author’s declaration

I declare that I have written this thesis, and that the work described in it is my own. The work described here has not been published before the start of the project. The work in this thesis has not been used to obtain previous degrees. Work that has been performed by someone else is detailed below:

At the beginning of my project I was provided with G*13*P_WT and ΔC recombinant DNAs that were designed and cloned by Dr. Callum Smits and Mrs. Maria Chechik, members of the Antson Laboratory.

The G20C small terminase full length was identified and cloned by Dr. Callum Smits and Mrs. Maria Chechik. Additionally, Mrs. Maria Chechik purified the native protein used in the DNA-binding experiments and crystallised the native protein, providing the optimal conditions for the crystallisation of the Se-Met small terminase by myself.

Some of the output research in this thesis has been published in the following papers:

I. Loredo-Varela, J., Chechik, M., Levdikov, V.M., Abd-El-Aziz, A., Minakhin, L., Severinov, K., Smits, C. and Antson, A.A. (2013) The putative small terminase from the thermophilic dsDNA bacteriophage G20C is a nine-subunit oligomer. *Acta Crystallographica Section F*, **69**, 876-879. Included in Appendix 11. Derived Publications.

II. Crystal structure and DNA-binding activity of thermophilic phage G20C small terminase. Juan Loredo-Varela, Maria Chechik, Huw T. Jenkins, Sandra J. Greive, and Alfred A. Antson. (In preparation)

III. Structural studies on the SPP1 capsid protein. Juan Loredo-Varela, Huw T. Jenkins, Maria Chechik, Callum Smits, and Alfred A. Antson. (In preparation)

Chapter 1

1. Introduction

1.1 dsDNA bacteriophages SPP1 and G20C

Bacteriophages, or phages, are viruses whose hosts are bacteria, and are the most abundant organisms on earth (1). They can be classified according to the type of infection they produce (lytic or lysogenic), the genome nucleic acid type they contain and the morphology and symmetry of their capsid (2).

Double stranded (ds) DNA tailed bacteriophages belong to the order *Caudovirales*, that is further divided into three main families: *Myoviridae* that includes phages with contractile tails; *Podoviridae*, with phages that have short non-contractile tails, and *Siphoviridae,* that groups phages with long non-contractile tails (2).

The mesophilic *Siphovirus* SPP1 that infects the Gram-positive rod bacterium *Bacillus subtilis* was isolated from soil and initially characterised by Riva *et al* (3). Subsequently, other research groups sequenced the complete genome (4) and characterised genomic sections comprising the DNA-packaging operon (5). The 44,007 bp long SPP1 genome contains 81 *orf*s organised in five early and four late operons (GenBank accession code: NC_004166.2) (4). Examples of the identified proteins include the products of genes *1* and *2* (G*1*P and G*2*P), which correspond to the small and large terminase subunits, respectively. G*6*P, G*7*P, G*11*P, G*12*P, and G*13*P were identified respectively as the structural proteins: portal protein, minor head protein, scaffolding, decoration and capsid (also known as head protein or major coat protein) proteins (4).

Thermus thermophilus phage G20C, a thermophilic *Siphovirus*, was isolated from hot springs in the Kamchatka region in Siberia, Russia (6). The G20C genome, which is 81,159 bp long, shares high DNA sequence homology with the thermophilic phages P23-45 and P74-26 (GenBank codes EU100883.1 and EU100884.1) (7). From the 117 putative *orf*s predicted for phage P23-45, a number of encoded proteins were annotated on the basis of their homology to proteins or domains with known function and on the basis of genomic context. Examples of putative proteins are the large terminase, portal, scaffolding and capsid proteins, encoded by genes 85, 86, 87, and 89 respectively (7).

1.2 Assembly of dsDNA viruses

dsDNA bacteriophages exhibit a general assembly process (Figure 1), although particular stages are present for distinct phages. Initially, multiple copies of the capsid protein are polymerised around the portal protein by the scaffolding protein to form an empty procapsid (prohead). Once the procapsid is complete the complex of concatemeric DNA-small and large terminases is docked to the portal protein to start the ATP-fuelled translocation of DNA, beginning with the priming of the lead end of the DNA by cleavage at the genome *pac* site. The hetero oligomer formed by the portal protein and terminase subunits is called the molecular motor. After a one-genome length fragment has been packaged, the concatemer bound-small and large terminase complex leaves to find another empty procapsid and repeat the cycle. The final step to produce the infective phage is the attachment of completion proteins to the portal vertex to avoid the leakage of DNA from the capsid and mediate in the docking of the tail. Minor completion proteins, decoration and/or glue proteins may also be involved at certain stages of the assembly (8-10).

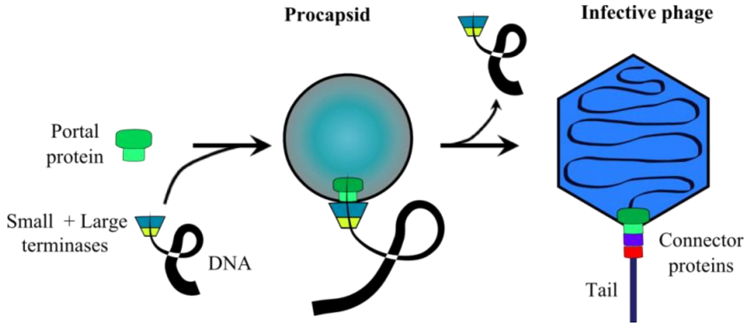


Figure 1. General assembly process in dsDNA bacteriophages

Figure adapted from Lebedev *et al* (11).

In the following sections, the SPP1 molecular motor will be described in detail as an example of molecular motors found in dsDNA phages. Although this project involved X-ray structure determination of the small terminase from another *Siphovirus*, G20C bacteriophage, close evolutionary relationship suggests that any data on this protein will be relevant also to other phages.

1.3 DNA-packaging molecular motor

The components of the molecular motor from phage SPP1 are the portal protein (G*6*P), the small (G*1*P) and large (G*2*P) terminase proteins (Figure 2) (12).

1.3.1 Small terminase

The translocation of DNA starts when the small terminase (ST) protein is believed to specifically recognise the viral genomic DNA synthesised by the host machinery from an environment that also contains host DNA. More information about the nature of ST-DNA interaction is presented in sections 1.9 and 1.10. Once the ST has recognised DNA it presents it to the large terminase (LT) which makes a cut that generates a free end in the DNA concatemer (see section 1.3.2) (9,10). In some phages it has been shown that the ST regulates the large terminase enzymatic activities (13-15).

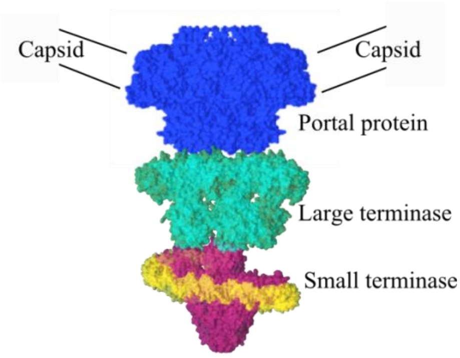


Figure 2. Model of SPP1 molecular motor

Portal protein (blue), large terminase (cyan) and small terminase (magenta) are docked to reflect biochemical data, but no structure of a complex is available. The portal and ST models are derived from PDB entries 2JES and 3ZQQ (11,16). Large terminase PDB 3CPE (17). DNA is in yellow. All figures, unless otherwise indicated, were made using CCP4mg (18)

The SPP1 ST has been the focus of extensive work, leading to the accumulation of a vast amount of biochemical and structural data that has allowed partial understanding of how the ST interacts with a specific region of DNA (5,19,20) and how it is related to the ST from other phages (21).

In 2012, the X-ray structure of the ST from phage SF6, another *Siphovirus* closely related to SPP1 (16,21) was solved. Besides sharing 82 % sequence identity and exhibiting similar domain arrangement (21), SF6 ST is able to supply the small terminase function to SPP1 phages defective in gene 1 (21).

The SF6 full-length ST forms ring-like structures of nine subunits in solution and in the crystal (Figure 3a). The nine subunits define a central channel that consists of two structural parts: a nine *β*-stranded barrel on top, and the base at the bottom of the assembly, formed by the N-terminal domains and the central oligomerisation domain (OD) of the monomer. The OD is composed by two short antiparallel α-helices and one *β*-hairpin. The N- terminal domain (NTD) of every subunit consists of the DNA-binding domain, which hangs at the periphery of the ring (16). More details and figures about the SF6 ST and that from other dsDNA phages are given in section 1.8.

1.3.2 Large terminase

One LT subunit is a dual-activity enzyme, one activity of which is to cleave the concatemeric DNA twice to produce single genome-length DNA fragments, after the ST has recognised its binding site. The second enzymatic activity, ATPase, provides the energy required for the molecular motor to translocate DNA into the procapsid (9,22). The endonuclease and ATPase activities are associated with the C- and N-termini, respectively. Additionally, both the C- and N-terminal domains comprise interaction sites with the portal protein and ST (Figure 2) (23-25).

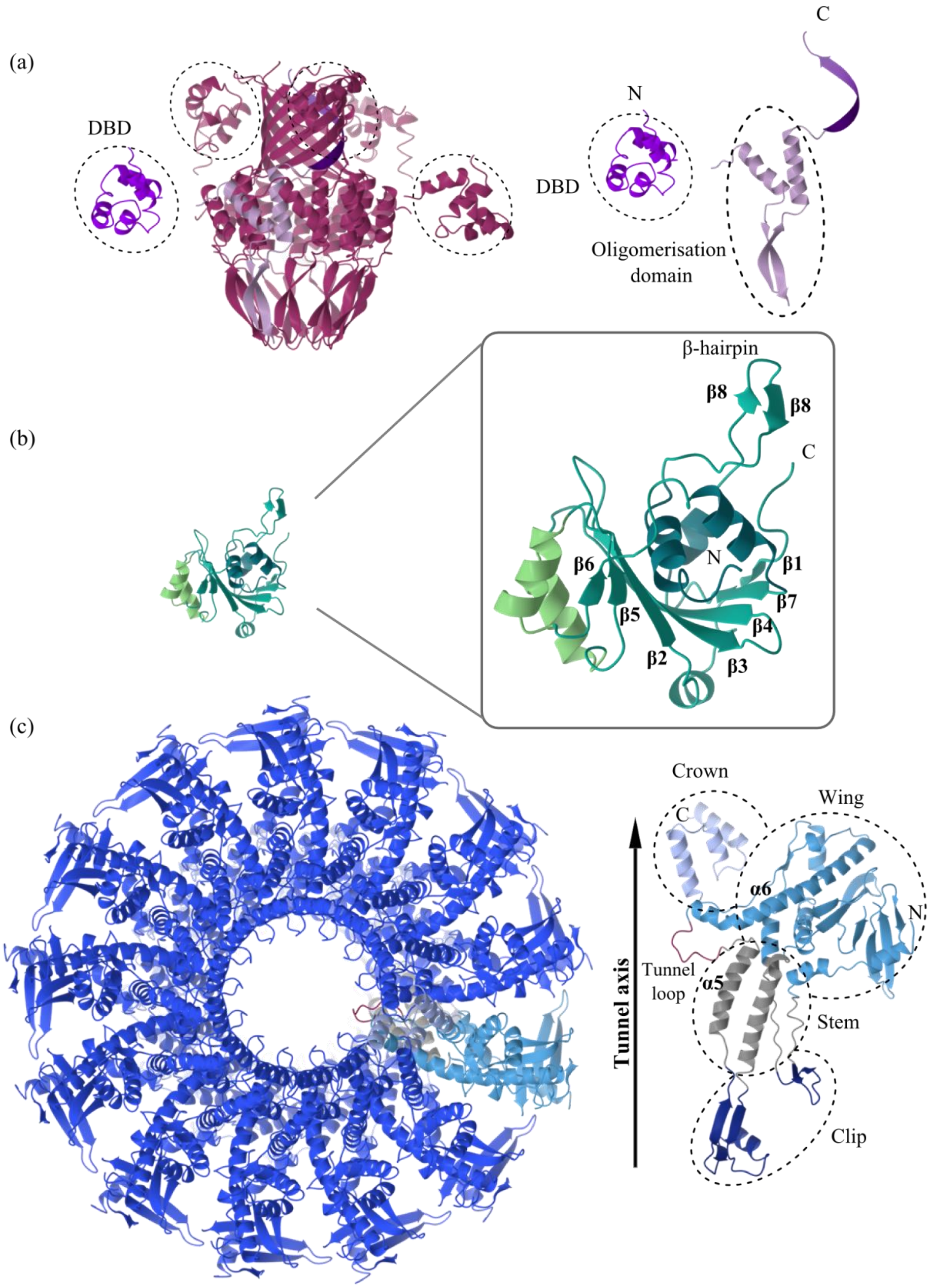


Figure 3. X-ray structures of the individual components of the SPP1 molecular motor

(a) Full-length SF6 ST oligomer (left) and monomer (right) (PDB 3ZQQ). The DBDs and OD are highlighted with dashed lines. (b) SPP1 LT nuclease domain (PDB 2WBN). The β-strands in the core and β-hairpin are numbered. (c) SPP1 portal protein (13-mer, left) and monomer (right) (PDB 2JES). Domains are indicated by dashed ovals. All the structures, except the inset are to scale.

It has been showed that G*2*P FL is a monomer in solution (13). The X-ray structure of the C-terminal nuclease domain of SPP1 LT was determined at 1.9 Å resolution, showing that the G*2*P_ΔN construct comprises seven β-strands with a set of α-helices at each side of the β-sheet and a short β-hairpin at the C-terminus (Figure 3b) (22). The fold of the RNase H endonucleases family (26) was identified suggesting that SPP1 LT might possess a similar catalytic mechanism.

Furthermore, several acidic residues and His400 were shown to be important for the cleavage of DNA and the C-terminal β-hairpin movements were implicated in switching the nuclease activity on and off (22).

1.3.3 Portal protein

The portal protein is a 12-subunit homo-oligomer embedded in the capsid, replacing one of the 12 pentamers of the capsid icosahedron (27,28). Besides providing the passage of DNA from the protoplasm to the interior of the capsid during the assembly and in the reverse direction during host infection, the portal protein is believed to provide the site for the scaffolding protein, SPP1 G*11*P, to associate and thus to begin polymerising the assembly of the major capsid protein, G*13*P (29,30). The cryo-EM structure showed that the functional oligomerisation state of the portal protein is a 12-subunit ring-like structure that forms a central channel through which the DNA is translocated. However, subsequent determination of the X-ray crystal structure of recombinantly expressed G*6*P in *E. coli* showed that a 13-subunit assembly is formed which is also shown to exist in solution (11,28,31). The G*6*P X-ray structure provided information about the interaction of the portal protein with DNA, highlighting the potentially key role of the portal protein's tunnel loops in DNA translocation (11).

G*6*P is a mixed α/β structure composed of four domains, named from top to bottom, crown, wing, stem and clip (Figure 3). The stem and wing make contact with the capsid protein, while the clip interacts with the terminase complex. The tunnel loop, mentioned above, comprises 21 residues connecting the twisted helices α6 with α5. Its movement is likely to accompany DNA translocation through the portal (11).

The 13-mer has an inner diameter at its most constricted portion, defined by tunnel loops, of 27.7 Ǻ (vdW diameter). In the functional 12-mer assembly, however, the internal diameter of this ring is reduced to 18.1 Ǻ. The position of the loop in the 13- and 12-mer structures differs by ~ 3 Å shift along the tunnel axis, demonstrating potential movement of these loops in the protein. (11)

The translocation model proposed by Lebedev *et al* overcomes the reduced diameter of the tunnel that would imply clashes between DNA sugar phosphates and the tunnel loops by introducing two structural rearrangements. First, it was proposed that the DNA is shifted so that it is no longer coaxial with the portal protein, in this way the major groove can be accommodated by approximately three loops at one time. The second rearrangement occurs when loops slide relative to each other in a sequential mode, as a “Mexican wave”, allowing two base pairs to be translocated while one ATP molecule is hydrolysed by the viral ATPase (11).

1.4 Icosahedral capsid architecture

Capsids, also known as heads or shells, are protein assemblies designed not only to contain and protect the viral genome, but also to transport it from one host cell to another (32).

1.4.1 Quasiequivalence

In many viruses, the capsid is built by the aggregation of a high number of copies of a single protein, but in more complex viruses two or more proteins can be involved. Such behaviour is a perfect example of viral efficiency; for a virus it is more economical to store the genetic information using only a single gene rather than two or more. In addition, when within the host cell, it is easier to direct the production of multiple copies of a single (or few) protein(s) rather than many different ones (32).

Based on Watson and Crick’s observations about the structure of small plant viruses (33), Caspar and Klug postulated that only a limited number of designs for a biological container made up of a high number of the same protein was possible, the helical tubes and icosahedral particles (32).

In helical tubes, the protein subunits are packed in helical array and every subunit lies in identical structural and chemical environments, except by the subunits at the top and bottom of the rod. As the same contacts are used by the subunits over and over through the helical array, it is reasonable to say that they are all in the same environment and are equivalent (32).

The architecture of flexible filamentous viruses can be explained if some of the bonds that hold the helical array together are slightly distorted. Since all the bonds are no longer the same but the overall bonding pattern has not changed considerably, the subunits remain quasi-equivalently related (32).

Caspar and Klug first introduced the concept of quasi-equivalence in 1962 to argue that the same principles that governed the construction of helical viruses applied for virus particles with icosahedral symmetry (32).

Icosahedral symmetry permits the highest number, that is 60, of identical proteins to cluster in spherical arrangement where they are identically accommodated, a situation that is only valid for small viruses (Figure 4a). To construct larger viruses, which have a higher number of subunits, it is necessary to deviate from the equivalence and adopt quasi-equivalence; Individual subunits can be packed with slightly distorted bonds so they will lie in similar, but not exactly identical environments. In this way the capsid is essentially still held together by the same type of inter-subunit interactions, but they differ slightly for subunits in different environments (32).

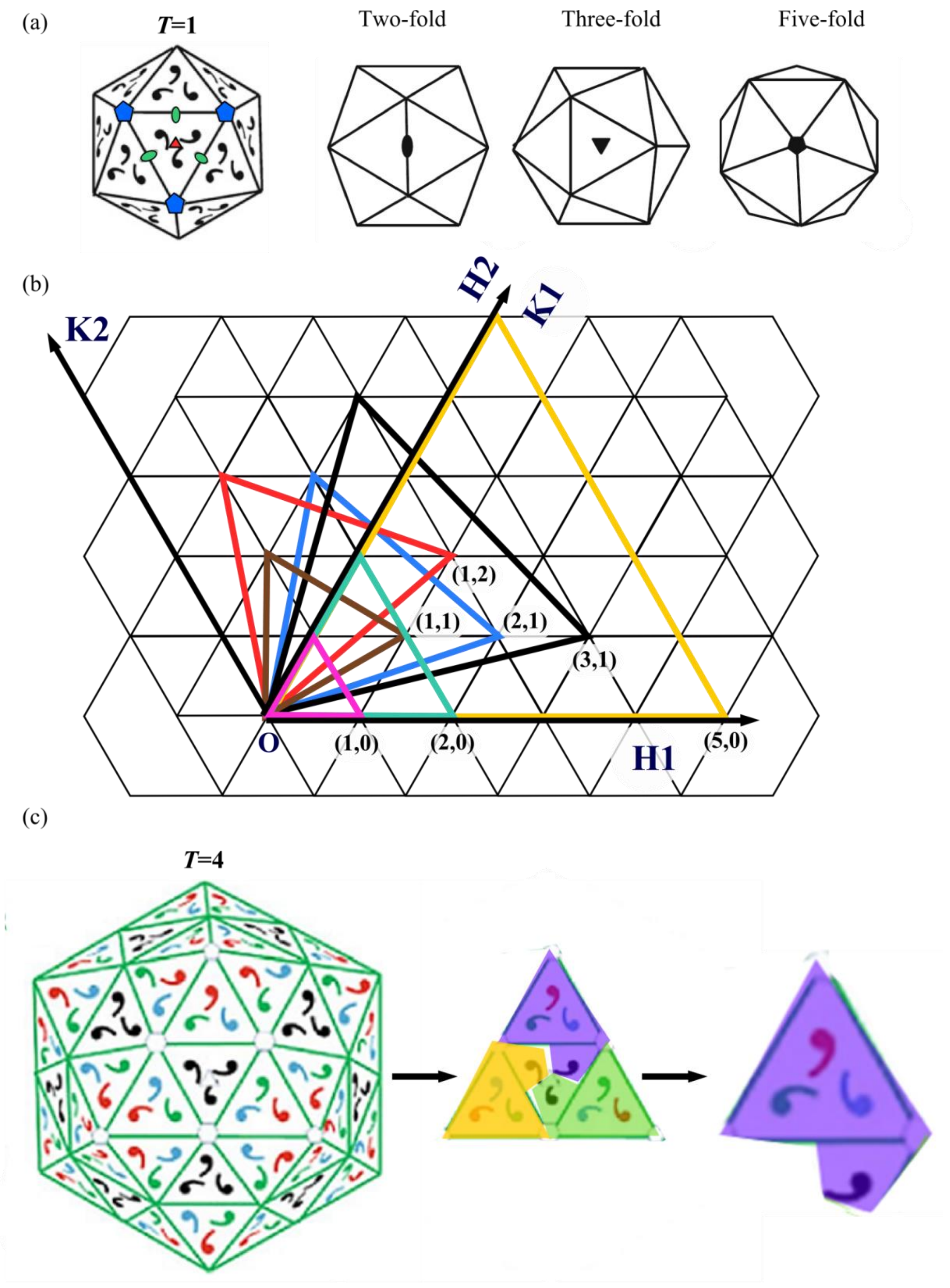


Figure 4. Icosahedral symmetry in capsids

(a) Symmetry axes in a *T*=1 icosahedral capsid. The 2-, 3- and 5-fold axes are indicated by ovals, triangles and pentagons. At the most left figure the subunits on each triangular facet are represented by black commas. The smallest icosahedron fits three individual subunits on each triangular facet, designated as one unit triangle. (b) Drawing of triangular facets on a hexagonal lattice. Calculation of the *T* number is done by selecting an origin, *O*, for the initial pentamer and then finding the closest pentamer centre. The coordinates (*H*,*K*) describe, in unit triangles, how far the pentamers are to each other in the capsid. One icosahedron with *T*=7, has either (*2,1*) (blue) or (*1*,*2*) (red) values. For the facet with (*1*,*2*) values, the closest pentamer to the origin is at position one unit triangle along the H axis and two unit triangles along the K axis. Facet triangles belonging to icosahedra with numbers *T*=1 (pink), 3 (brown), 4 (cyan), 13 (black) and 25 (yellow) are also plotted. (c) Icosahedral asymmetric unit in a *T*=4 capsid. One triangular facet extracted from the icosahedron shows the four unit triangles and twelve subunits that form it (centre). One third of the triangular facet is shown at the right to highlight only one asymmetric unit (shadowed in purple), formed by four copies of the capsid protein. Figure adapted from Prasad & Schmid, 2012 (34).

The quasi-equivalence is clearer to understand if it is emphasised that the same subunit can form pentamers and hexamers (Section1.4.2). The subunits and contacts at the pentamers are different from those at the hexamers, however not significantly, as the environment at the five- and three-fold axes (centre of hexamer or six-fold axis) cannot be too different because of the intrinsic properties of the individual subunits that forbid extreme variances.

1.4.2 Icosahedral symmetry

Two general shapes are commonly observed in viral capsids, helical tubes or isometric particles with cubic symmetry. Three types of cubic symmetry are found in isometric capsids: tetrahedral, octahedral and the one that is preferred by viruses, the icosahedral symmetry (32,34).

One icosahedron has 20 identical facets, each defined by an equilateral triangle formed by three copies of the capsid protein, making it possible for the icosahedral container to accommodate 60 copies of capsid protein, the maximum number of copies among particles with cubic symmetry, as there are only 12 for the tetrahedron and 28 for the octahedron (Figure 4a) (32,34).

The icosahedrons are characterised by exhibiting 2-, 3- and 5-fold rotational symmetry axes. Application of these rotational symmetries to the icosahedral asymmetric unit (Section1.4.4) is enough to determine the complete capsid structure (Figure 4.a,c).

Every icosahedral capsid is composed of 10*T*+2 morphological units (pentamers and hexamers) built by the same type of capsid protein (See section 1.4.3 to read more about the triangulation number *T*). The smallest viruses are formed by pentamers only, distributed symmetrically at the vertices of the icosahedron, making a total of twelve pentamers. Larger virus design incorporate hexamers, distributed on the triangular facets, that allow the architecture of the capsid to enlarge when required. Even though pentamers and hexamers form the shells, the assembly into pentamers and hexamers is not necessary before capsid formation. The number of morphological units can be calculated by the formula 10(*T*-1) hexamers + 12 pentamers (32,34).

1.4.3 Triangulation number

It was mentioned earlier that the smallest icosahedral capsid can accommodate 60 subunits in 20 triangular facets, that is 3 subunits per facet, all in identical environments (Figure 4.a). If more subunits are incorporated, for example in larger viruses, the triangular facet needs to be enlarged by the equation $T=H^2 + HK + K^2$, where T is the triangulation number and H and K are 0 or positive integers that represent coordinates along the H and K axes that cross at a 60° angle in the hexagonal lattice (Figure 4.b) (32,34-36).

Icosahedron with T=1 has (H,K) values of (1,0) and its facet is defined as unit triangle. To find the T value for larger icosahedra, one origin, O, is chosen at any pentamer centre. Next, a vector is drawn from the origin to the closest neighbouring pentamer centre, that has certain (H,K) values. The triangular facet is enclosed by drawing a second vector from the point (H,K) to the centre of a third pentamer (Figure 4.b). Thus, the T number indicates the distance between pentamers and the number of unit triangles on each facet. Two capsids with (*2,0*) and (2,1) (H,K) values have numbers T=4 and T=7 with 4 and 7 unit triangles per facet, respectively (Figure 4.c). The number of subunits is calculated by 60T, derived from the condition that all icosahedra have 20 triangular facets, with T unit triangles formed by 3 subunits each: 20 *x* 4(or 7) *x* 3 = 240 or 420 respectively (32,34-36).

1.4.4 Icosahedral asymmetric unit

The icosahedral asymmetric unit is defined as one third of the icosahedral triangular facet, which results from dividing the triangular facet into three identical parts at the three-fold axis (Figure 4.a,c). Producing the 60T subunits of the capsid is possible by applying 5-3-2 rotational symmetry to the asymmetric unit. A capsid with T=1 has one copy of the capsid protein in the asymmetric unit, while a capsid with T=4, has four. Thus, the triangulation number also describes how many copies of the capsid protein are fitted in the asymmetric unit (34).

1.4.5 Prolate icosahedra

Icosahedral morphologies different to the isometric version are observed in phages T4 and φ29 (37,38), where capsids are assembled as prolate icosahedra. Prolate is the name given to icosahedra that present with the main cylindrical body elongated (Figure 5). Thus, these icosahedra are formed by two types of triangular facets, equilateral triangles at the top and bottom, defined by the T number, and larger triangles in the cylindrical body. The larger facets are defined by two different vectors, the base vector of the triangle is still determined by $T= H^2 + HK + K^2$, whereas the elongated sides are determined by the elongation number, Q, calculated by Q=H1H2 + $H$1$K$2+ $K$1$K$2.

Figure 5 shows the pair of triangular facets for prolate capsids of both φ29 (T=3, Q=5) and T4 (T=13*l*, Q=20) phages. The equilateral facets have coordinates φ29 (*1,1*) and T4 (3,*1*) while the elongated sides' coordinates are φ29 (*1*,*2*) and T4 (*4*,*2*). To draw the elongated facet, the K1 axis becomes the H2 axis and a new K2 axis, at 60° with respect to the H2/K1 is sketched.

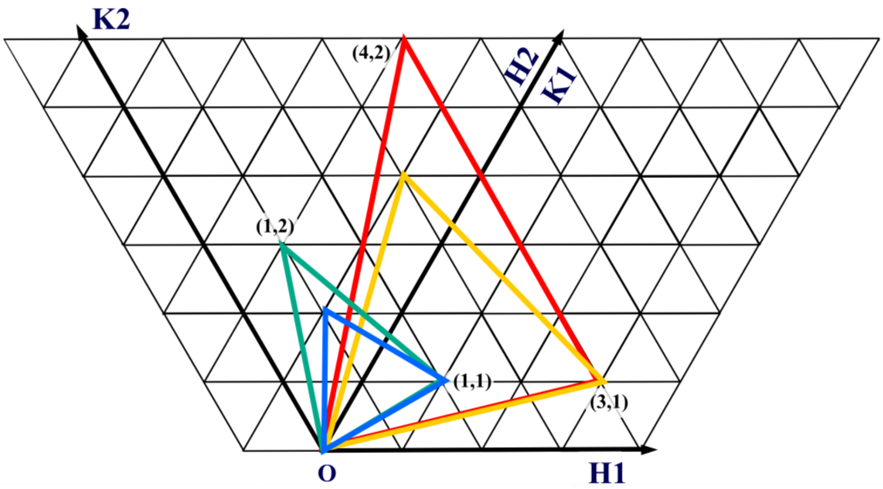


Figure 5. Prolate Icosahedron

Equilateral and elongated facets of phage φ29 are shown in blue and cyan, and those of T4 in yellow and red. Coordinates for each phage, the origin and axes are indicated. Figure adapted from Prasad & Schmid, 2012 (34).

1.5 HK97 capsid protein

In 2000, the X-ray structure of the complete capsid from the dsDNA phage HK97 was determined at 3.6 Å resolution (Figure 6. (39). The HK97 capsid, with *T*=7, exhibits icosahedral symmetry with 12 pentamers at the vertexes and hexamers at the rest of the triangular facets. Through the capsid, three different levels of organisation are observed. First, residues K169 and N356 (shown as black balls in the asymmetric unit in Figure 6.b and in ball and stick representation in Figure 6.c) create isopeptide bonds between neighbouring subunits. The second level corresponds to the pentamers and hexamers formed by the capsid protein and the third one to the catenated links that create the "chainmail" all over the capsid (39). As of 2014, the isopeptide bond has only been observed in the capsid of HK97 but not in other capsids.

The asymmetric unit from HK97 has seven slightly differently folded copies of gp5, six from a hexamer and the seventh from a pentamer (Figure 6.b). Every monomer is an α/β structure organised into peripheral (P-) and axial (A-) domains. The P-domain spans the quasi-three and two-fold axes of the capsid (Figure 6.). The main features of the peripheral domain are the extended loop (E-loop), helix α3', and residues K169 and N356, responsible of holding the capsid together. Whereas the P-domains interact with P-domains both from the same or different morphological units, the A-domains are only in close contact with A-domains from the same morphological unit, at the 5-fold and 6-fold symmetry axes (39).

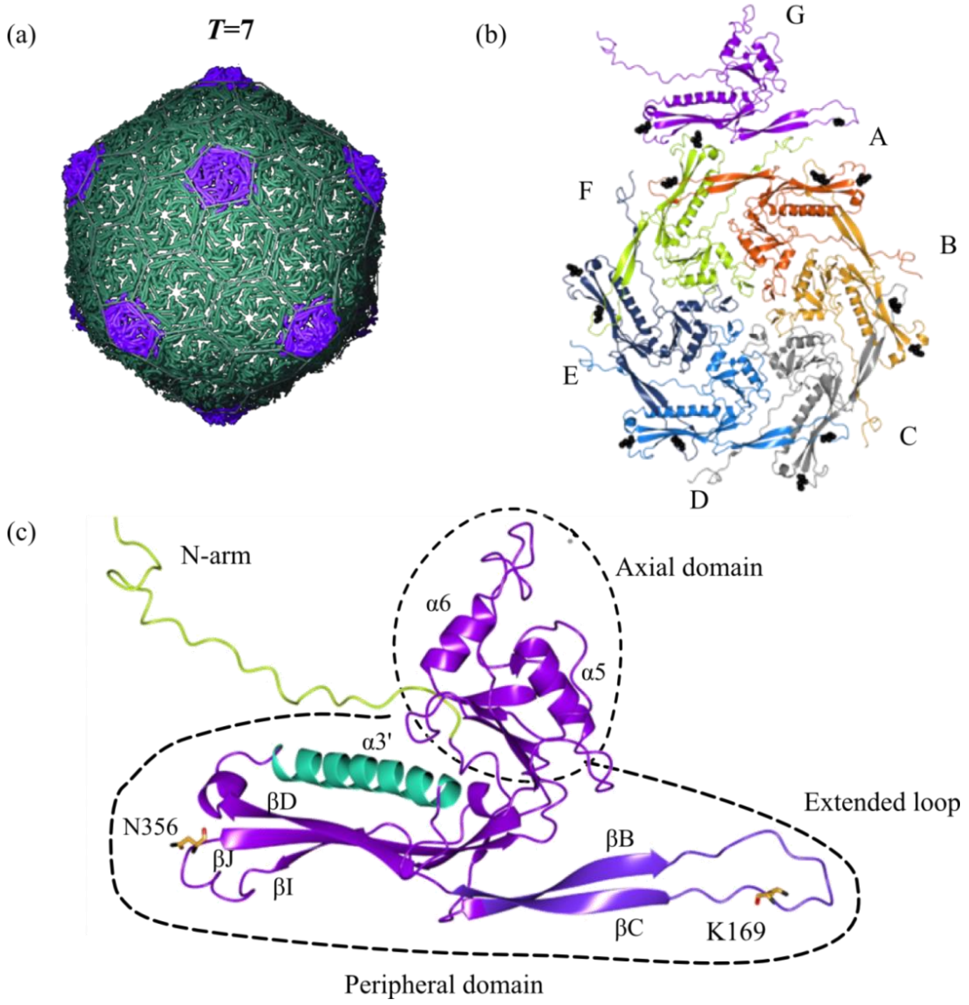


Figure 6. Structure of the HK97 capsid

(a) In the HK97 phage capsid, pentamers are shown in purple and hexamers in cyan. (b) Asymmetric unit extracted from the capsid. Residues are shown as black balls to illustrate their distribution in the asymmetric unit. (c) Individual subunit organisation. A- and P-domains are circled by dashed lines. The N-arm is shown in green and helix α3' in cyan. Residues K169 and N356 are shown in ball and stick representation. Part a was extracted from Wikoff, *et al*, 2000 and used with permission of the editor (39). Parts b and c were made using the PDB code 1OGH.

Initially, residues 1-103 facilitate and guide the assembly of the immature form of the viral capsid, called prohead I (procapsid I). At a later stage, residues 1-103 are cleaved off by the serine protease encoded by HK97 (39,40). The 29 N-terminal residues of the 104-385 protein construct comprise the N-arm, which within the capsid, lies beneath adjacent monomers. The flexibility of the N-arm and the E-loop is thought to be critical for allowing gp5 to oligomerise into pentamers and hexamers (39).

1.5.1 Structural conservation of capsid proteins

Tailed dsDNA bacteriophages are at first glance very distant from each other since they adopt different morphologies and contain tails of varying length. Phages T4 and φ29 present prolate icosahedra (37,38), while HK97 and SPP1 adopt isometric forms (39,41). They can have different *T*/*Q* numbers, for example, φ29 has *T*=3, *Q*=5, T4, *T*=13*l*, *Q*=20 and P22, HK97, SPP1 have *T*=7*l* numbers. In addition dsDNA bacteriophages can have somewhat different assembly pathways, belong to different families and infect different hosts, which can be both Gram-positive and Gram-negative bacteria (Table 1)

Despite all these significant differences, it has been proposed that all tailed dsDNA bacteriophages have evolved from a common ancestor (42). This hypothesis is supported by the common HK97-fold of the capsid protein, named so because the first X-ray structural data were obtained for the capsid of this phage (Figure 7) (39).

Table 1. Capsid proteins in dsDNA bacteriophages

Phage	**HK97**	**SPP1**	**T4**	**P22**	**φ29**	**PfV**	**Bb_PhRP**	**Ec_Pcp**
Family	*Siphoviridae*	*Siphoviridae*	*Myoviridae*	*Podoviridae*	*Podoviridae*	Domain *Archaea*	---	---
Host	*E.coli*	*B.subtilis*	*E coli*	*Salmonella* s. *typhimurium*	*B.subtilis*	*Pyrococcus furiosus*	*Bordetella bronchiseptica*	*E. coli*
Diameter (Å)	660	610	850 width 1150 height	628	Prolate: 450 Isometric: 425 width x 540 height	---	---	---
T	7*l*	7*l*	*T*=13*l* *Q*=20	7*l*	*T*=3 *Q*=5	*T*=3	---	---
Protein	gp5	G*13*P	gp24 Pentamers gp23 Hexamers	gp5	gp8	---	---	---
Size (kDa)	30.7	35	56 48.7	~47	49.8	38.8	34.1	39.6
Aa's	282cleaved	324	521 422	430	488	345	316	351
Technique	X-ray	Cryo-EM	X-ray Cryo-EM	Cryo-EM	Cryo -EM	X-ray	X-ray	X-ray
Resolution (Å)	3.6	8.8 (Full particle)	gp24 2.9 Capsid 22	ExH-8.2 Shell-9.1	Isometric fibreless 7.9 Isometric fibered 8.7 Prolate fibreless 12.7	3.6	2.05	2.2
Special features	Protease: gp4 Inter-subunit crosslinking between K169 and N356	Scaffolding: G*11*P Decoration protein: G*12*P	Insertion domain Decoration proteins: gp hoc,gp soc Scaffolding: gp22- Assembly protease: gp21	Telokin-like domain Scaffolding: gp8-	BIG2 domain 174b pRNA Scaffolding: gp7 Head fibres gp8.5	---	---	---
PDB	1OHG	4AN5	1YUE 1Z1U	3IYH 3IYI	1YXN	2E0Z	3BJQ	3BQW
Reference	(39)	(41)	(37)	(43)	(38)	(44)	(45)	(46)

--- Information not available.
For particles with isometric icosahedral shape only *T* is shown.
1Z1U is a theoretical model for T4 hexameric protein, gp23.

Likewise, protein folds observed in capsid proteins of other groups of viruses are also suggested to be shared, like the PRD1 phage-like fold and the blue-tongue virus-fold (BTV), that are found in viruses that infect mainly Eukarya (42).

To date there are X-ray structures available for the already described HK97 gp5, for the vertex protein gp24 from T4 (38), for the putative capsid protein of prophage of *Escherichia coli* CFT073 (Ec_Pcp) (46), for a phage-related protein BB3626 from *Bordetella bronchiseptica* (Bb_PhRP) (45), for a viral capsid-like particle found in *Pyrococcus furiosus* (PfV) *(44)* and the shell-forming encapsulins, from bacteria *Thermotoga maritima* (47) and *Myxococcus xanthus* (48) (Table 1, Figure 7*)*. Cryo-EM structures, deposited as poly-Ala models, have been solved for the capsid proteins from P22 (gp5) (43,49), φ29 (gp8) (38) and for SPP1 (G*13*P) phages (41). In T4, proteins gp23 forms hexamers and gp24, pentamers. The structure of gp23 was determined by building a homology model using the gp24's X-ray structure (37). The HK97-fold has also been identified in the cryo-EM reconstructions of Sf6 (50), T5 (51), T7 (52), λ (53), P-SSP7 (54) and ε15 phage capsids (55).

Structures of all these capsid proteins are homologous to the HK97 capsid protein, gp5, to some extent (Figure 7) and are hence described as HK97-like. All the head proteins appear to be organised into A- and P-domains. For bacteriophage P22, poly-Ala models for the expanded head (ExH) and empty shells were determined. While in the ExH particle that includes no pentamers, the A-domain is easy to identify, in the shell particle the A-domain lies in a "hidden" location, suggesting that it can adjust its position in the different maturation stages (43).

A-domains establish contacts with A-domains from neighbouring subunits at the five- or six-fold axis. Helices α5 and α6 and the apical loop that connects them are partly responsible of keeping the subunits of one morphological unit in place. Helices α5 and α6 from the same subunit are packed next to helices α6 and α5, respectively, from different subunits at each side.

The P-domains from all the capsid proteins make extensive contacts with their counterpart in the same or adjacent capsomers. Specifically the long β strands at the bottom of the P-domain are the source of such interactions at the two-fold symmetry axes. In T4 it was observed that the distance between the closest C*α* atoms between hexamers is around 16 Å in particles without decoration proteins. Such a long distance was attributed to variation in the structure of the loops at the P-domain in gp24 and gp23 or calibration errors in the amplification of the electron microscope (38).

More specific features are observed in the analysed set of capsid proteins, for example, the long helix spanning most of the P-domain has proved to be one of the most remarkable structural characteristics of the HK97-fold. Helices α3, α5 and α6 (shaded) have served a reference in cryo-EM 3D reconstructions because they are easily recognisable and have been used as a starting point for building models (41,43).

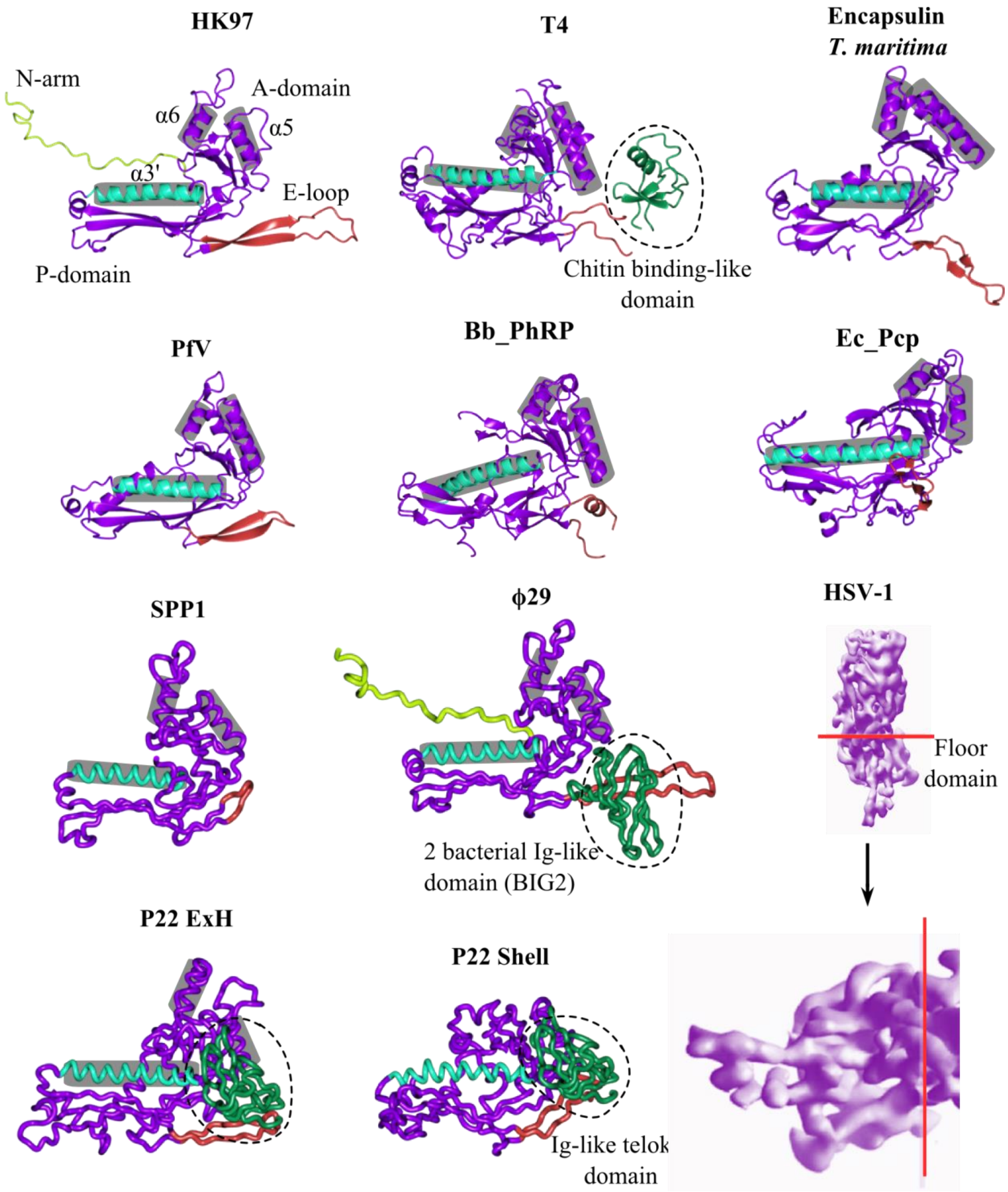


Figure 7. Capsid proteins with HK97 fold

Side views of the structures mentioned in the text and Table 1 are shown as ribbon diagrams (X-ray structures) or worm representations (Poly-Ala models from cryo-EM reconstructions). The N-arm from HK97 is shown in lime colour and so is a segment from φ29. Insertion domains are shown in green (delimited by dashed ovals). Helix α3 (following numbering in HK97) is shown in cyan and the E-loop is shown in salmon. Conserved helices (α3, α5 and α6) are highlighted with grey background. Corresponding PDB entries and references are found in Table 1.

The E-loop has been identified in other capsid proteins, however its conformation is less conserved than the long helix spanning the P-domain. In T4 and Bb_PhRP, it was not possible to trace it as it is highly disordered, probably due to its intrinsic flexibility (38,41). In P22, a putative capsid protein from an *E. coli* prophage, and in the viral capsid-like particle from *P. furiosus*, the E-loop is considerably shorter than that in HK97 capsid protein (43,44).

One more difference between HK97 gp5 and other capsid proteins is the presence of one additional globular domain (shown in green) at the distant end of the E-loop in T4 (37), P22 (43,56) and φ29 (38). This domain is called the insertion domain in T4 (it presents topological similarity to the chitin-binding domain of chitiquinase), the telokin-like domain (or immunoglobulin-like) in P22 and group 2 bacterial immunoglobulin-like domain (BIG2) in φ29. The additional insertion domain is thought to play an analogous role to the crosslinking and glue proteins in HK97 and T4 since it is involved in intra- and inter-capsomer contacts.

The finding of a virus-like particle in *P. furiosus* was intriguing because it proved that viruses with HK97-fold are also present in the *Archaea* domain (44). The purified particle does not carry DNA nor is it able to infect *P. furiosus*. Little is known about the origin of the corresponding gene though it is speculated that it is an immature or aberrant form of another capsid. (The evidence is that mutants of φ29 capsid proteins fail to assemble into prolate icosahedra but form isometric particles instead). A similar change could have occurred to this capsid-like individual subunit since its shape is almost spherical, which is supported by the absence of encapsidated DNA, known to contribute to shape the final icosahedral morphology in other phages. Other essential proteins involved in the assembly or DNA-packaging process are also likely to be missing (44).

As a result of the accumulation of structural information about the capsid protein in viruses, it is conceivable that viruses use a limited range of capsid protein folds to preserve their DNA since the HK97-fold is conserved across bacteria and in the evolutionarily related archaeal and eukaryotic viruses (42).

1.6 SPP1 phage capsid protein

1.6.1 Morphogenesis

The assembly of the SPP1 capsid is initiated when the scaffolding protein, G*11*P (23.4 kDa) promotes polymerisation of 415 protomers of the major capsid protein, G*13*P (35.3 kDa), around the portal protein (12 copies of G*6*P, 63kDa each) into icosahedral lattices with T=7*l* (Figure 8) (29). To form a viable procapsid, all the listed components were shown to be necessary, for instance removal of G*6*P leads to populations of capsids with *T*=7 and *T*=4 geometries (30). Removal of G*11*P produces immature proheads and aberrant heads tubes. As expected, removal of G*13*P produces no capsids, while its overexpression in *E.coli* leads to formation of big aggregates that elute in the void volume of gel filtration columns, evidence of its intrinsic capability to polymerise. Individually purified G*11*P and G*13*P polymerise into filament like-structures and empty or open coiled capsid-like structures, as revealed by negative-stain EM. Co-expression of both G*11*P and G*13*P leads to the immature/aberrant structures described earlier (29,30).

The scaffolding protein's secondary structure was predicted to consist of long α-helices and two heptad ((I/L)X(E/D)LXXX) sequences that might confer amphipatic character to G*11*P which then would be able to form dimers. It has been demonstrated that G*11*P forms tetramers although dimers of refolded protein were found to bind more efficiently to G*13*P than the native His-tagged tetramers (57).

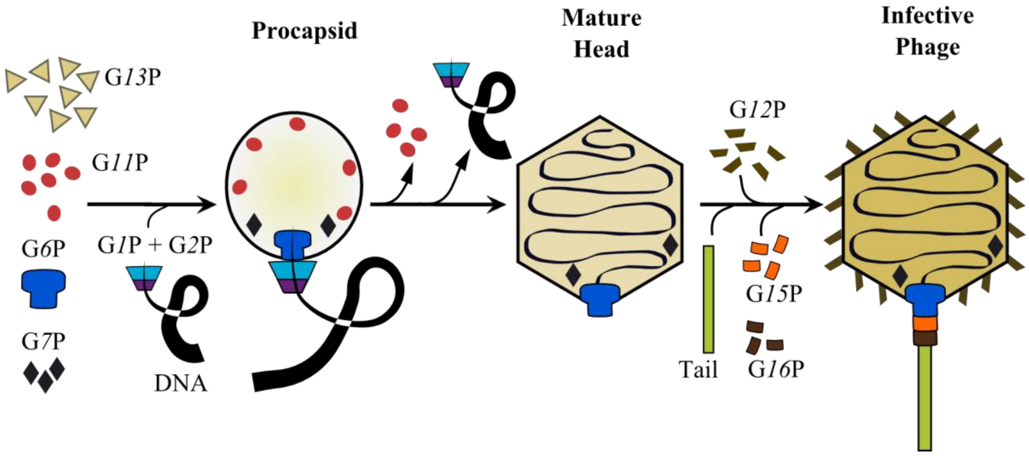


Figure 8. Assembly of SPP1 phage

All elements that are known to be involved in the process are shown. G*6*P, G*7*P, G*11*P and G*13*P form the procapsid. Once completed, the procapsid serves to dock the G*1*P-G*2*P-DNA complex. As the DNA fills the procapsid, G*11*P exits through gaps between G*13*P protomers. Full procapsids change from the initial spherical to the characteristic icosahedral shape. After a complete genome is translocated the G*1*P-G*2*P complex leaves and the portal is free to attach the completion proteins and tail. G*12*P is then also added to the final capsid. The number of shown components does not reflect the real stoichiometry at any step. The location of G*7*P during assembly is not clear. Figure adapted from Lebedev *et al*, 2008 (11) and Becker *et al*, 1997 (29).

Early research found that the minor protein, G*7*P (34K kDa), is present during the complete assembly process, however it was not possible to identify its function (29). Further experiments demonstrated clear interactions with G*6*P. One or two copies of G*7*P are found in the SPP1 capsid and it is possible that these are associated with the portal (30). The most recent findings about G*7*P indicated that mutants lacking gene 7 produced less infectious progeny than when G*7*P was supplied. However, the less infectious phages showed the same efficiency of DNA packaging and processivity of translocation cycles as the wild-type (WT) phages. If G7P does not participate in the capsid assembly or in the DNA-translocation process, then it is likely to play a role in the efficient delivery of DNA to the host as phages missing G*7*P detach from the ejected DNA but the ones with G*7*P remain attached to one end of the viral DNA. G*7*P might act as a clamp that slows down the ejection rate to a level that allows efficient passage to the protoplasm, which is compatible with the fact that removal of this gene produces faster ejection of DNA than in the WT, although the G*7*P is not co-ejected to the cell host (58). Two copies of G*7*P per phage strongly interact with G*6*P via the C-terminus of the latter protein (59).

The function of G*8*P and G*9*P remains unclear, but it is known that removal of these genes produced no heads or tailless heads, respectively (29).

After DNA packaging, the decoration protein, G*12*P (6.6 kDa) is attached to the outer surface of the capsid, whilst completion or “stopper” proteins G*15*P and G*16*P are docked to the portal to avoid leakage of DNA and at the same time provide the site for attachment of the tail (29,60).

NMR structures have been solved for G*15*P (11.6 kDa) and G*16*P (12.5 kDa) monomers (Figure 9). G*15*P consists of four helices and a short β-strand. G*16*P is organized into seven β-strands that form a β-barrel with four sheets at every side. β1-β2 and β2’-β3 pairs are connected by long loops. Docking of G*15*P structure onto the Cryo-EM density maps (11) was only possible by dissecting the structure. Twelve subunits were fitted in different ways in a ring-like conformation, suggesting that G*15*P reorganises itself during the assembly with G*6*P and G*16*P (hence called the adaptor protein).

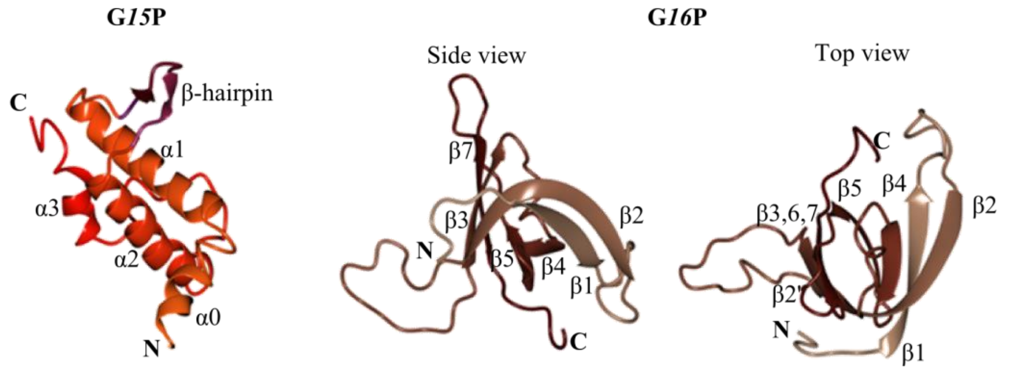


Figure 9. Completion proteins of SPP1 phage

Ribbon diagrams of the main models from the NMR structures of the adaptor (G*15*P) and stopper (G*16*P) proteins. The numbering is the same as in the original reference. The long β2-strand is present at both sides of the barrel, thus is indicated as β2 and β2’. G*15*P PDB 2KBZ, G*16*P PDB 2KCA.

On the other hand, Twelve G*16*P subunits were fitted as rigid bodies in a way that one long loop from every subunit projects toward the centre of the cylinder to prevent DNA leakage. G*16*P does not undergo major structural rearrangements to interact with other phage proteins nor to oligomerise as G*15*P does. G*16*P provides the interface to attach the tail (61).

1.6.2 Cryo-EM structure of the SPP1 capsid

Pseudo-atomic structures of SPP1 capsid, determined by Cryo-EM, at late stages of assembly showed icosahedral arrangements with *T*=7*l* geometry. Interestingly, the shape and thickness of the protein layer (~ 27 Å) of tailless, DNA-filled, decorated and non-decorated capsids, and tailed, empty and DNA-filled capsids were essentially the same, making it apparent that no large changes occur to the capsid when the tail or decoration proteins are attached or when the DNA is ejected (41).

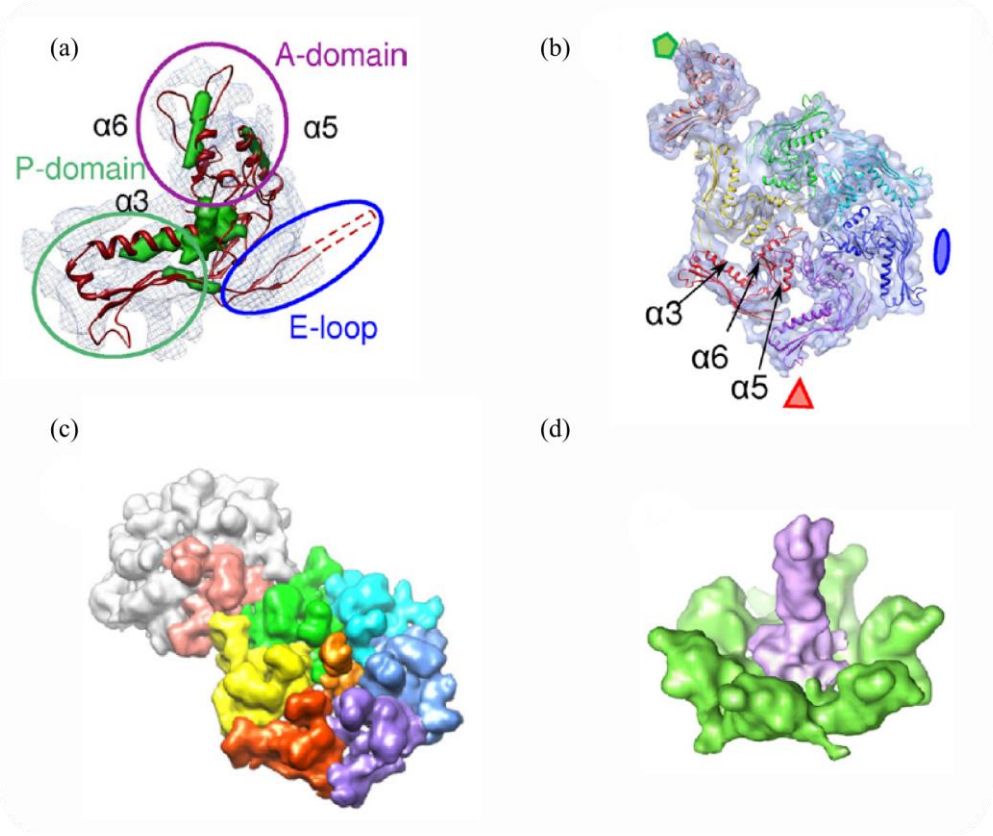


Figure 10. Cryo-EM structure of the SPP1 capsid protein.

(a) Pseudoatomic model of G*13*P. HK97 gp5 is shown in red while identified α-helices in SPP1 are shown in green. (b)(c) Asymmetric unit that features seven copies of G*13*P, six forming the hexamer and one forming a pentamer. The density of G*12*P is shown in orange in (c). (d) Superposition of hexamers from decorated and non-decorated capsid. Density corresponding to G*12*P is coloured in purple. Figure extracted from White *et al*, 2012 and used with permission of the editor.

Helices α3, α5 and α6 were identified in the EM density maps at the same position they are in HK97 (Figure 10a), therefore a model for SPP1 G*13*P using alignments of secondary structure predictions of proteins which exhibit a similar fold could be constructed. G*13*P displays the HK97-fold, being organised in A- and P-domains that are involved in inter-subunit contacts. The E-loop could not be traced in SPP1 but the G*13*P model fitted into the extracted asymmetric unit shows some unassigned density that might indicate either an interaction between the E-loop of one subunit with the adjacent P-domain or a different conformation of G*13*P in the region where the E-loop is in HK97 (Figure 10b,c) (41).

Superposition of hexamers extracted from decorated and non-decorated capsids exposed the location of G*12*P (Figure 10d). Three copies of G*12*P bind to each of the sixty flat hexameric centres. However the spike density observed is not long enough to account for three G*12*P subunits implying a degree of flexibility in the most distal subunit. The function of G*12*P remains elusive since decorated and non-decorated capsids exhibited the same morphology, released packed DNA and underwent protein denaturation at the same temperature when thermostability was tested, ruling out a potential role in defining the capsid shape or conferring stability (41,62).

1.7 Example of protein-engineering to facilitate crystallisation

A clear example of how altering some characteristics in wild type (WT) proteins allows the study of key functional details in protein structure is presented by the human β_2-adrenergic receptor (β_2-AR), which was engineered to solve its X-ray structure (63,64). The β_2-adrenergic receptor, a membrane protein that belongs to the G protein-coupled receptors, is located principally in the smooth muscle, and changes at the genetic level lead to diseases like asthma, hypertension and heart failure.

The inherent flexibility and complexity of the WT receptor’s flexible intracellular loop (ICL3) and its C-terminus hampered the growing of diffraction quality crystals. This obstacle was overcome by replacing the ICL3 that links transmembrane α-helices V and VI with T4-lysozyme (T4L) (65) and by truncating the C-terminal portion (residues 366-413) that was shown to present an extended conformation (66). T4L was chosen due to its solubility, propensity to crystallise and the proximity between the N- and C-termini (10.7 Å), which is similar to the distance predicted between α-helices V and VI. Residues 2-164 in T4L replaced region 234-259 in the β_2-AR. C54T and C97A mutations were introduced in the T4L protein (63,64).

The X-ray crystal structure of the β_2-AR-T4L engineered protein in complex with the agonist carazolol was solved at 2.4 Å resolution (Figure 11). The β_2-AR comprises seven transmembrane long α-helices, one short intracellular α-helix and one short extracellular loop (ECL2) that is folded in α-helical arrangement. The N- and C-termini from T4L are linked to α-helices V and VI respectively (63,64). The fold of T4L in complex with β_2-AR remains unchanged in comparison to the crystallised T4L monomer while the overall fold of the α-helices from the β_2-AR are likely to be very similar to the WT one as it has been shown that the two regions connected by the ICL3 (helices I-V and VI-VII) can be expressed separately and still assemble into functional receptors (67).

Investigation of the functional properties of the β_2-AR-T4L protein revealed an active phenotype that exhibits higher affinity for the agonists used in the study than the WT. The affinity for the antagonist was normal. β_2-AR-T4L did not bind to G proteins as could be predicted because of the removal of the ICL3.

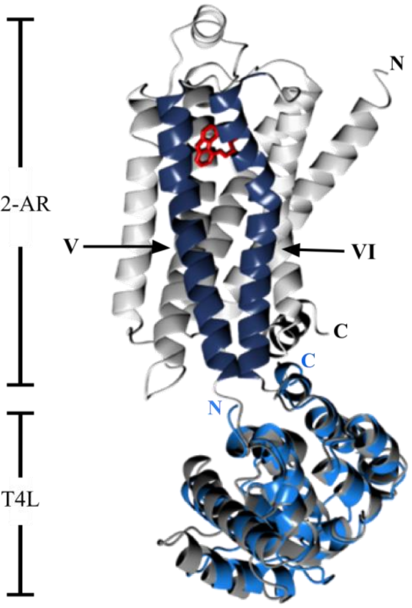


Figure 11. Human β_2-AT-T4 lysozyme engineered protein

The β_2-AR is shown in white and T4L in grey. α-helices V and VI are highlighted in dark blue. Carazolol is shown in red. WT T4L is shown in light blue. The N- and C-termini from the engineered protein and WT T4L are shown in black and blue respectively. PDB entries: β_2-AR-T4L: 2RH1, T4L: 2LZM

1.8 Structural diversity of small terminases

Several high resolution crystal structures of STs from mesophilic phages including SPP1-like phage SF6 (16), P22-like Sf6 (68), P22 (14) and T4-like phage 44RR (69), are available (

Table 2, Figure 12). A fifth X-ray structure is available for the putative ST from *Bacillus cereus* ATCC 14579 prophage phBC6A51 (70). The ST has also been identified in the *Pseudomonas aeuroginosa* phage PaP3 (15) for which functional studies are available and in the *Lactococcus lactis* bacteriophage ASCC454 (71) whose ST has been crystallised but no X-ray structure has been published.

All the elucidated structures exhibit a ring-like organisation enclosing a central channel of varying diameter. The number of subunits varies from eight to twelve, nevertheless, the general architecture and the overall outer diameter are similar. P22 (14,72), Sf6 (68) and prophage phBC6A51 oligomerise with a fixed number of subunits while 44RR (69) and SF6 (16) can crystallise in two different forms: 11-mers or 12-mers (44RR), and 9-mer or 10-mers (SF6).

Table 2. Small terminases in dsDNA phages

	SF6 (SPP1-like)	**Sf6 (P22-like)**	**P22**	**44RR (T4-like)**	**Prophage phBC6A51**
Family	*Siphoviridae*	*Podoviridae*	*Podoviridae*	*Myoviridae*	---
Host	*Bacillus subtilis*	*Shigella flexneri*	*Salmonella s. typhimurium*	*Aeromonas salmonicida*	*Bacillus cereus* ATCC 14579
Technique	X-ray	X-ray	X-ray	X-ray	X-ray
Resolution (Å)	1.85/2.19 1.68 3.0 4	1.65	1.75 3.35	2.8 1.8	1.9
Oligomeric state	10-mers 9-mer 9-mer 9-mer	8-mer	9-mer	11-mer 12-mer	9-mer
Protein and truncations	**G*1*P**-53-120 **G*1*P**-53-120 **G*1*P**-65-141 **G*1*P**- FL (1-164) (only 1-141 in crystal)	**gp1**-FL (140 residues), crystallised from 10-139 or 10-132	**gp3** **c**leaved: (1-140) **gp3**-FL(1-162), last 23 aa's are disordered.	**gp16**-FL 1-154 26-112, 11-mer 25-114, 12-mer	----
Overall diameter (reported) Å	65 (53-120) DBDs flexible	99	95	75.5 80.6	~100
Height (reported)	~88	66	70 (in both crystals) 100 (plus modelled α-helix barrel	~40	~45
Channel diameter (most constricted part)[a]	14.3, 9-mer	~15.5	19.4	24.2 27.8	~9
PDB	3ZQM/3ZQN 3ZQO 3ZQP 3ZQQ	3HEF	3P9A	3TXQ 3TQS	2AO9
Reference	(16)	(68)	(14)	(69)	To be published (70)

--- Information not available

[a] The presented diameters correspond to the most constricted section of the channel, which might be not the top part of the assembly.

Interestingly, in the case of 44RR, the subunits that constitute both oligomeric forms are identical (69), whereas in SF6, constructs missing the C-terminus assemble into both 10-mer and 9-mers, but the full length protein only forms 9-mers, suggesting that the C-terminus is important in defining the oligomerisation state (16).

For P22, Sf6 and prophage phBC6A51 the overall diameter is comparable, however, for SF6 and 44RR, that have flexible or disordered NTDs, it is not possible to accurately measure the outer diameter since it is likely that the NTDs occupy different positions (Figure 12). The overall heights vary to a greater extent as the assemblies have distinctive C-terminal domains:

β-stranded barrels in SF6, Sf6 and prophage phBC6A51, or an α-helical barrel in P22. Moreover, in most cases portions of the C-terminus are disordered in the crystal.

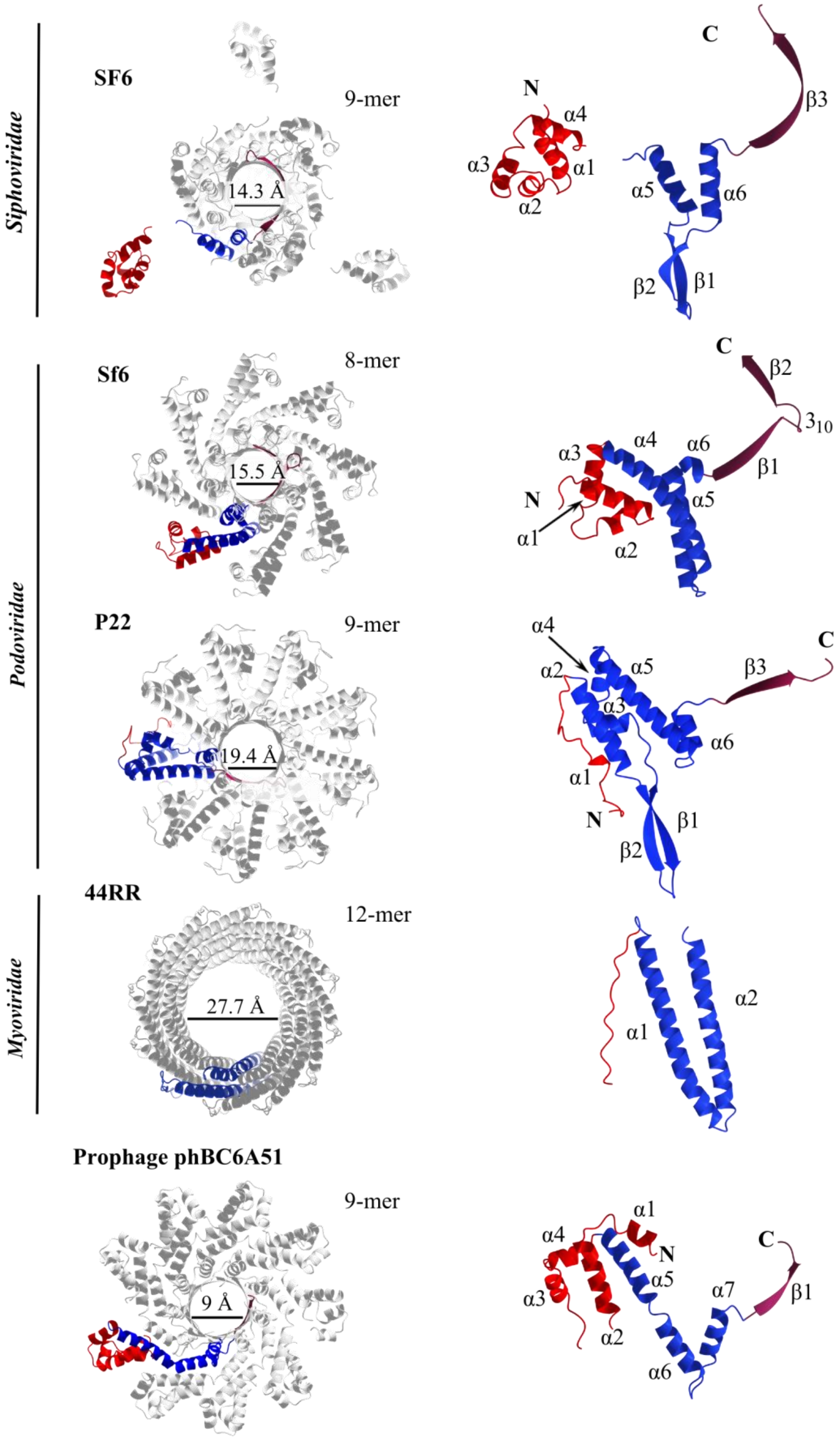


Figure 12. Small terminases in dsDNA phages

Ribbon diagrams for the oligomer (top view, left) and monomer (side view, right) of every ST with published X-ray structure. Diameter for the most constricted part of the channel is indicated (atom-to-tom distance). The highlighted subunit is shown in the side view with the NTDs in red, ODs in blue and C-terminal β-strand in purple. The numbering of the secondary structure elements is the same as in the original publications. Table 2 lists corresponding PDB entries and references.

With the exception of P22, the terminases present defined ODs that pack closely in a circular array and establish contacts with adjacent subunits to form a central channel. The helical N-terminal domains (NTD), identified as the DBD in SF6 (16) and Sf6 (68), hang at the periphery of the structure surrounding the oligomer. Helix-turn-helix in the DBDs motifs of SF6 and Sf6 are formed by α-helices 2 and 3 and their connecting loop. Superposition of SF6 and Sf6 DBDs show an almost identical fold (shown in section 8.3). Other shared features between some STs are a β-hairpin barrel at the bottom of the SF6 and P22 structures and a β-stranded barrel on top of SF6, Sf6, P22 and phBC6A51 STs. A structural comparison will be presented in chapter 8.

1.9 Proposed models for the small terminase-DNA interaction

The currently available structures provide insights about the oligomerisation state of the ST, the general organisation of the assembly and domain organisation of individual subunits, however, the structures in isolation are still a limited source of information when addressing how the ST interacts with DNA. Two major models that take into account distinct properties of the ST have been proposed to describe such interaction (Figure 13) (16,69,73). The first model posits that the dsDNA wraps around the circular assembly to establish contacts with the HTH motifs present in the DBD from every subunit (Figure 13a). This model is supported by the finding that multiple DNase-protected sites were found in the *pac* recognition site of SPP1 (19) separated by ~10 bp stretches (full turn of B-form DNA), in good agreement with the predicted distance of ~35 Å between adjacent DBDs in SF6 (16). The spacing of DBDs in SF6 is similar to that observed experimentally between DBD in Sf6 (68) or between the traces of NTD (including a predicted HTH motif) in 44RR (17) even though the number of subunits per oligomer differs.

In contrast, the second model considers the central channel as the binding site for DNA (Figure 13b), proposing that DNA could pass through the channel, on its way to the procapsid, during mechanical translocation by the LT. Although this model assigns function to the channel, it has several vulnerable points. For instance, only the channel from phage 44RR (69) is wide enough to accommodate DNA (~23 Å van der Waals diameter) since the rest of the channels have portions that would lead to clashes with DNA (Table 2). Moreover, threading DNA through the channel would imply one free end before interaction with the ST, which is not likely to occur as the ST needs to present the DNA to the LT to generate the free end. The possibility of the disassembly of the ST to embrace DNA is unlikely due to the high stability of the oligomer (Discussed on reference (69)).

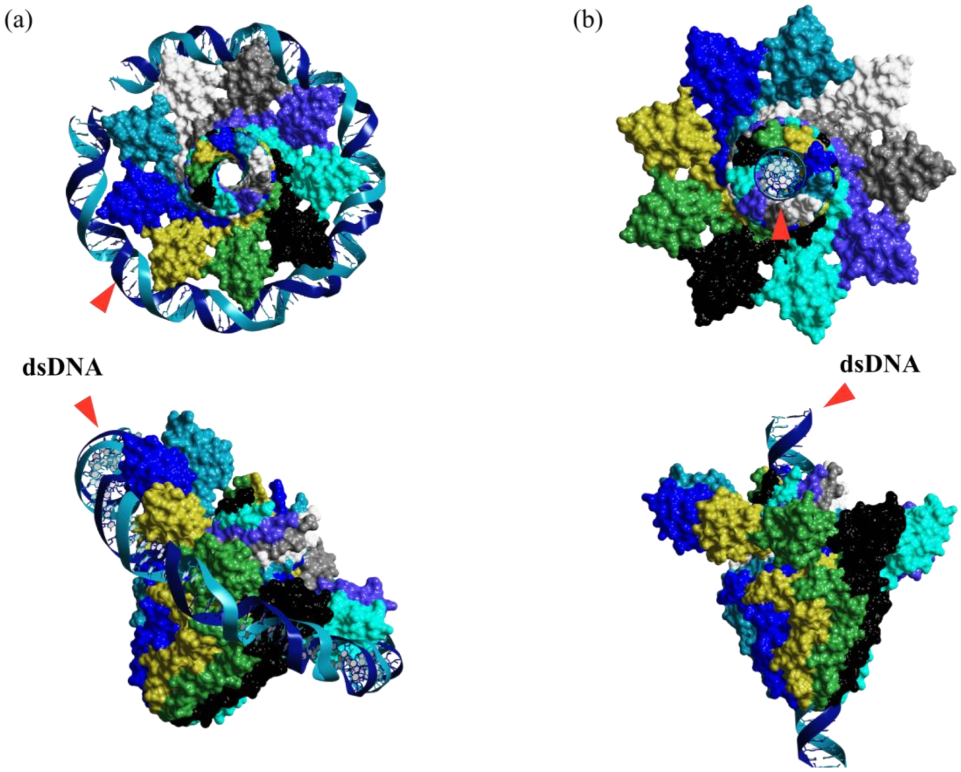


Figure 13. Small terminase:dsDNA models

Molecular surface representations of SF6 ST with dsDNA bound around the outside (a) and passing through the assembly (b). Top and side views are presented for clarity. DNA is indicated by orange arrowheads. The models for this figure (including nine DBD and circular DNA) were derived from the X-ray structure of SF6 ST (16) (PDB 3ZQQ). Figure adapted from Buttner *et al*, 2012 (16).

To date, little evidence exists to support the ST channel-dsDNA model. Little or no interaction was observed between the DNA and SF6 ST missing the DBD by SPR and EPR experiments, strongly suggesting the DNA binding around the outside. The truncated versions of SF6 ST (53-120 and 53-145) exhibited only residual binding to the tested DNA fragments in SPR experiments, in comparison with the full-length (FL) protein. The labelled version of the second construct, with only one spin label (facing the channel) introduced in one subunit to avoid coupling between adjacent subunits, failed to show any interaction with DNA in EPR experiments, when a 22-bp dsDNA extracted from the SPP1 *pac* sequence was added to the reaction (16).

1.10 Small terminase interaction with DNA: recognition sites and bent DNA

1.10.1 Bent DNA

Bending of DNA is a recurrent theme in the discussion of protein-DNA interactions as it is often involved in increasing the specificity of the protein towards DNA. The selectivity for DNA introduced by the bending is achieved because this allows the nucleic acid to adopt a conformation required for protein binding. In some cases, bending is induced by the bound protein, however most of the time, it is an inherent characteristic of the DNA sequence (74).

Two hypotheses to explain the mechanism of bending were initially put forward based on anomalously migrating DNA fragments in acrylamide gels (reviewed in (75)): the junction and wedge models (Figure 14).

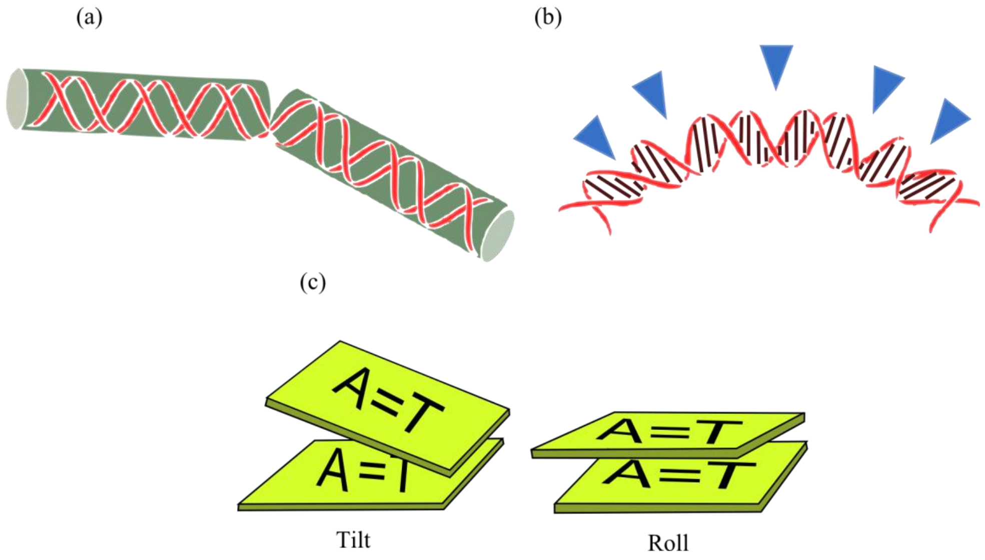


Figure 14. Models for bent DNA

(a) The junction model. Segments that represent B-form and non-B-form DNA are indicated by cylinders. (b) The wedge model. Consecutive bends caused by an AA dinucleotide are indicated by a blue triangle. (c) Tilt and roll wedges. Figure adapted from Sindens, 1994 (75).

The junction model assumes that bending occurs at a B- and A-form DNA junction associated with A-tracts. The position of bending could be either at the 3’, 5’-end or both ends of the A-tract (76).

The wedge model suggests that there is a “wedge” angle at an AA dinucleotide that causes deviation from the axis of the double helix. Bending is observed when consecutive wedge angles add on. Two different wedge angles are observed according to their location: tilt angle, opening in the direction of the phosphate backbone and the roll angle, which opens in the direction of either the minor or major groove (77,78).

In both models the phasing of A-tracts or AA dinucleotides is stated as the optimal condition to lead to bending, as for the sequence, tracts of $A_{4\text{-}9}$ were shown to be bent, however sequences without A-clusters also can be bent (75). Likewise, AT-rich segments, exposing the minor groove to a protein, have been pointed out as a source of bending. Removal of the water molecules from this site decreases the rigidity associated with the minor groove, making the DNA capable of adopting the conformation imposed by the binding protein (79). According to the literature 8.7° or 18° bends can be generated by each AA dinucleotide (77) or A-tracts, respectively (74).

1.10.2 Recognition sites: SPP1 pac site

The ST from SPP1 phage recognises a specific site on the viral genome to start packaging, the *pac* site (Figure 15a). The *pac* site is further divided into three segments with no apparent sequence similarity, *pac*L (left), *pac*C (centre) and *pac*R (right). G*1*P binds only to *pac*L and *pac*R as was evident by the presence of DNase protected segments. One of the protected regions, at *pac*L, is surrounded by poly-A sequences separated by approximately 10 bp that may induce bending upon binding to G*1*P (19).

Based on the observations that G*1*P is an oligomer in solution with several HTH motifs at the NTDs and that G*1*P binds to both the encapsidated and non-encapsidated ends, *pac*R and *pac*L, respectively, a model was proposed to explain how the SPP1 ST interacts with DNA and how it makes DNA accessible to the LT (G*2*P) (Figure 15b). Briefly, two different G*1*P oligomers bind separately to the *pac*L and *pac*R sites. Protein-protein interactions between the oligomers bring the two sites together and at the same time the *pac*C site is exposed so G*2*P can bind to G*1*P and make a cut in the DNA producing a free end. In Figure 15b, G*2*P monomers are represented by small spheres that interact with DNA and with G*1*P, but it is unknown whether the G*2*P contacts the same or different G*1*P molecule (19).

This model, proposed in 1995, was the first wrap-around model published and to date some of its key points are largely accepted by the community.

Besides SPP1, other packaging initiation regions have been identified in phages λ (site *cos*B, 120 bp long) (80), P1 (*pac* site, 161 bp long) (81) and P22 (*pac* site, 120bp long) (82).

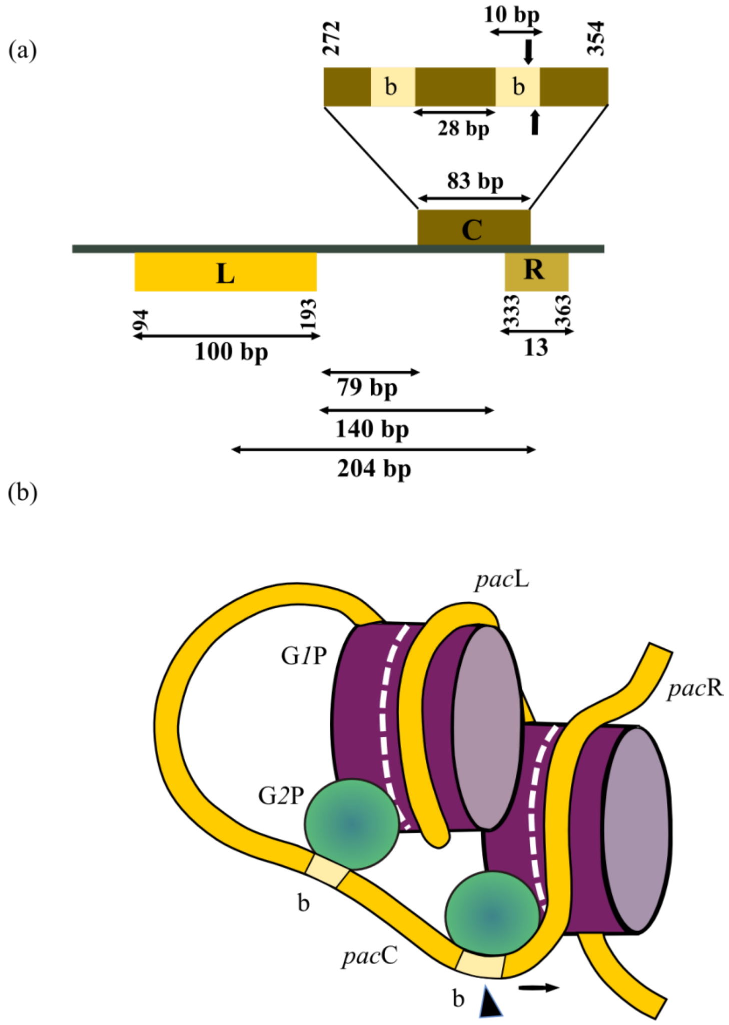


Figure 15. Model for ST complex with LT and *pac* DNA

(a) Scheme of the *pac* site DNA of the SPP1 phage (not to scale). The length and position of each fragment in the genome is indicated. Box b, where G2P is expected to bind and cut is shown. Distances between the end and start of *pac*L-*pac*C and *pac*L-*pac*R are indicated by double-headed arrows. The centre-to-centre distance of *pac*L-*pac*R is also shown. (b) Model of the G*1*P nucleoprotein complex with the *pac* site DNA. Two different G*1*P assemblies recognise the *pac*L and *pac*R sites, while G2P establishes contact with G*1*P and recognises box B. The site of cut is pointed out with a black arrowhead. The direction of packaging is signalled by the horizontal arrow. Figure adapted from Chai *et al*, 1995 (19).

1.11 Objectives

There is enough evidence to conclude that dsDNA phages make a huge impact on the biosphere of our planet. Evolutionary related animal viruses, such as members of the herpes family of viruses, lead to several health problems resulting in huge economic costs (42,44). The dsDNA bacteriophages are the focus of extensive research for several reasons. First, they serve as excellent model systems for understanding how similar animal viruses function (9). Secondly, their hosts include bacteria of industrial or clinical relevance, such as *Lactococcus lactis (71)*, a fermentative organism in the dairy industry or *Escherichia coli* and *Salmonella*, both foodborne pathogenic bacteria (83,84). Therefore we can benefit from learning how bacterial viruses assemble and how they function.

In dsDNA phages, the icosahedral capsid is the container that viruses use to store and protect their nucleic acid molecule. At one of the twelve vertexes, the portal protein is embedded to provide the gate of entry for DNA and the surface for the DNA-terminase complex docking (9). The aim of this thesis is to elucidate X-ray structures of the capsid protein G*13*P from mesophilic phage SPP1, and that of the ST from thermophilic phage G20C.

The X-ray structure of the SPP1 capsid protein could then be fitted into the available Cryo-EM maps (41) to reconstruct the complete capsid. Fitting the X-ray structure of the portal protein will make possible to study the nature of the capsid-portal interaction. One long term goal of this project is to introduce mutations in G*13*P to disrupt interactions between other G*13*P protomers to leave the portal surrounded by a single layer of capsid. These complexes could potentially be attached to any surface to deliver DNA after the assembly of the molecular motor. The possible uses of these devices include gene therapy or other areas of biotechnology. The capsid structure will allow the understanding of the mechanism SPP1 uses to stabilise its capsid during maturation. A comparison with other capsid proteins would be useful for elucidating segments that remain most conserved across evolution.

Solving the structure for the G20C ST will provide further knowledge about the oligomerisation state, domain organisation and characteristic features. One additional aim is to address the ST-DNA interaction by investigating binding activity towards viral and non-viral DNA. Several constructs will be tested to determine which domain(s) are responsible for DNA-binding. A comparison with available ST X-ray structures will be carried out to identify common features. Finally, the X-ray structure and experimental data will be used to propose a model for a functional complex of the G20C ST with DNA.

Chapter 2

2. Materials and methods

The chapter two of this thesis concisely describes how the constructs were handled during cloning, purification, crystallisation and biochemical investigation. Each subsection offers a general description of the protocol, and the additional details about every construct are presented in the appendices or results sections. Those approaches that were developed specifically for this project are described in the results section 3.4 (production of engineered proteins) and chapter 9 (cloning of DNA for probing DNA-binding activity).

2.1 Cloning

The full-length genes G*13*P, coding for the SPP1 capsid protein, and the one coding for the putative ST in G20C, were amplified from the genomic DNA of SPP1 and G20C phages by PCR. The pairs of primers for each construct used in this project are listed in appendix **1**. Inserts were then cloned into vector pET28a (Novagen) by ligation with T4 ligase according to manufacturer instructions (New England BioLabs) (Figure 17, section 2.3). The cells were transformed (Section 2.2) with the recombinant DNA and incubated overnight at 37°C. The following morning, the presence of the insert in the vector was tested by colony PCR using the pair of primers specific to every construct but using the colony as source of DNA. From a single positive colony, overnight cultures were prepared. Plasmid DNA was extracted according to manufacturer's instructions from the cell pellet. Purified recombinant DNA was sent for sequencing (GATC-BIOTECH, Germany) to confirm the sequence of the vector insert. The positive plasmids were used to transform cells for overexpression of the protein of interest (POI).

Details about the types of modifications that were made to the wild type proteins, in addition to the affinity purification tag and vector used are listed on appendix **1**.

In general, all the inserts were produced by one amplification reaction, with the exception of the G*13*P-Anti-TRAP constructs, for which two reactions were set up in order to clone the engineered proteins. A detailed explanation about the design and production of both constructs is given in the results section 3.4.

Two types of cloning were used in subsequent constructs. The first one consisted in using T4 ligase to create phosphodiester bonds between the insert and vector, both with complementary sticky ends generated by endonucleases in specific restriction sites (Figure 16). The second approach takes advantage of the DNA-recombination activity of the In-Fusion® enzyme (85). Briefly, both insert and vector are amplified so they have 15-bp complementary regions at the 5' and 3’ ends. During the incubation period, the poxvirus DNA polymerase uses its 3’-5’ exonuclease activity to generate sticky ends at the provided substrates by removing nucleotides at the 3’ end, producing complementary ends that can then anneal. Nicks present within the recombinant DNA are repaired by bacterial enzymes following transfection (86).

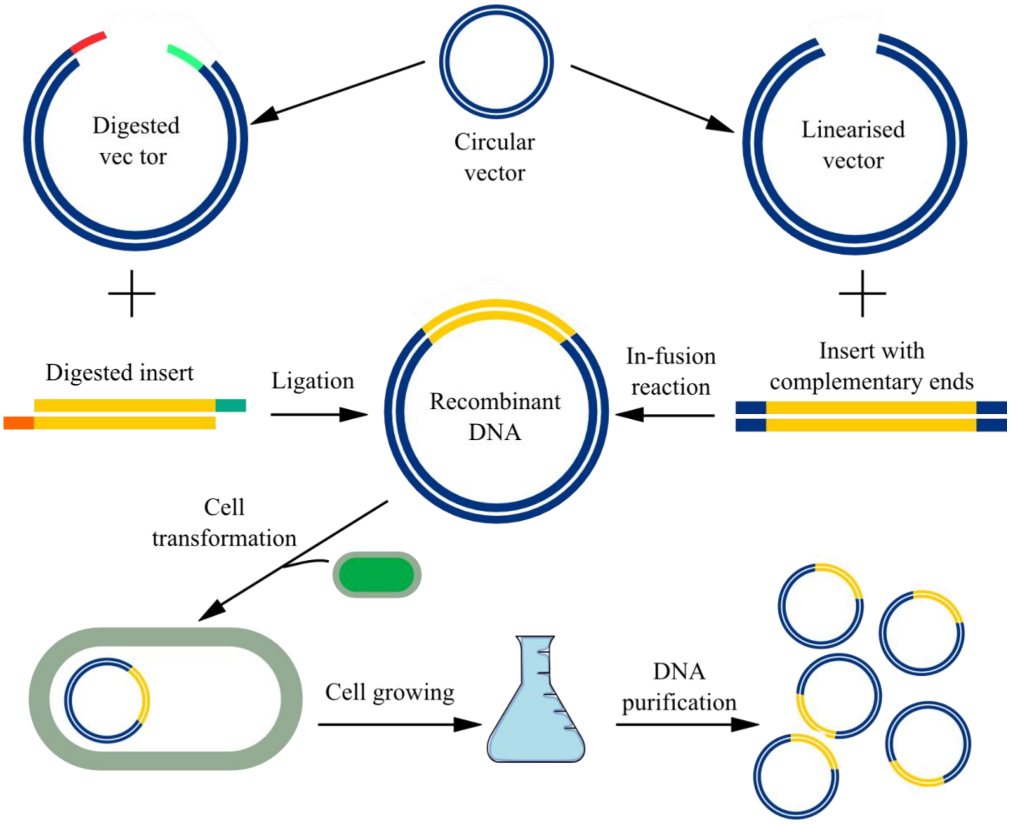


Figure 16. Production of recombinant DNA

Schematic representation of approaches used for production of large amounts of recombinant DNA.

Table 3. Settings of PCR, ligation and In-fusion reactions

PCR reaction		**Ligation**		**In-fusion**	
Component		Component		Component	
0.2-1 µM	Primers A /B	50 ng	Vector DNA (4 kb)	50-100 ng	Vector DNA (5.5 kb)
0.4 µL	10 mM dNTPs	37.5 ng	Insert DNA (1 kb)	50-100 ng	Insert DNA (1 kb)
x ng	DNA template	2 µL	10X T4 DNA Ligase Buffer	2 µL 5X	In-Fusion HD Enzyme Premix
4 µL	5X Phusion HF buffer		H_2O	________	H_2O
0.4 µL	Phusion DNA polymerase	1 µL	T4 DNA Ligase		
______	H_2O				
20 µL		20 µL		10 µL	
		- Incubation at RT during 2h		- Incubation at 50°C during 15 min	

PCR program for G*13*P_FL_D194W

98 °C	2 min
98 °C	15 s
48 °C	20 s
72 °C	3 min
72 °C	5 min

2.2 Cell transformation

Several *E. coli* modified strains were used to produce multiple copies of plasmids or recombinant proteins. The type of competent cell was selected according to the experimental requirements; one particular example will be described below. The manufacturer’s instructions for every cell strain were followed. Most of the cells used in this study required 30 minutes of incubation on ice with the DNA (5-10 ng) after cell thawing. Heat shock at 42 °C during 45 s-1 min followed. The cells were again incubated on ice for 2 min followed by addition of ~ 200 µL of Luria-Bertani medium (LB). After 1 h incubation at 37 °C in the shaker at 180 rpm, the samples were plated on LB agar plates with antibiotics suitable for selecting for the cells harbouring the desired plasmid, and incubated overnight at 37 °C. B834 cells were used to produce native protein when the protein was expressed in a higher amount than when using other cells. The type of cell used in the solubility screen is mentioned in the legend of the figures.

BL21 Star^{TM} DE3 cells were used to express G*13*P-AT full length constructs because these cells carry a truncated version of the RNase E enzyme gen (*rne*) of which the C-terminal portion is involved in mRNA degradation. The truncated gene only lacks the C-terminal portion of the RNase E enzyme but keeps the N-terminal portion that is involved in rRNA processing and cell growth, making the expressed mRNAs more stable (87).

The pLysS plasmid was incorporated in Rossetta 2 or BL21 cells used for this study as is compatible with the pET system. The pLysS plasmid encodes the T7 lysozyme to lower the basal expression of the gene of interest under the control of the T7 promoter since it inhibits the T7 RNA polymerase (88,89).

2.3 Cell growth and protein overexpression: pET system

Cultures (Luria-Bertani medium) started from single colonies were grown to optical density of 0.6 (O.D $_{600\ \mathrm{nm}}$) at 37 °C. When cultures were ready, protein expression was induced with 1 mM Isopropyl β-D-1-thiogalactopyranosid (IPTG).

All of the recombinant DNAs used the pET expression system to produce large quantities of POI.

The pET system is a bacterial plasmid that contains the gene for the *lac* repressor protein (*lacI*), the T7 promoter, specific to T7 RNA polymerase, a *lac* operator to block transcription, origins of replication f1 and genes for antibiotic resistance (Figure 17). The gene that codes for the POI is situated at the poly-linker region that includes several restriction sites, the region coding for the tag is situated at the 5' end of the poly-linker region. Moreover, the host's genome is modified to incorporate the T7 RNA polymerase and the *lac* promoter and operator.

The addition of IPTG, an analogue of lactose, displaces the *lac* repressor from the *lac* operators both on the pET plasmid and host's chromosome so the T7 RNA polymerase is expressed to then express the gene that encodes the POI.

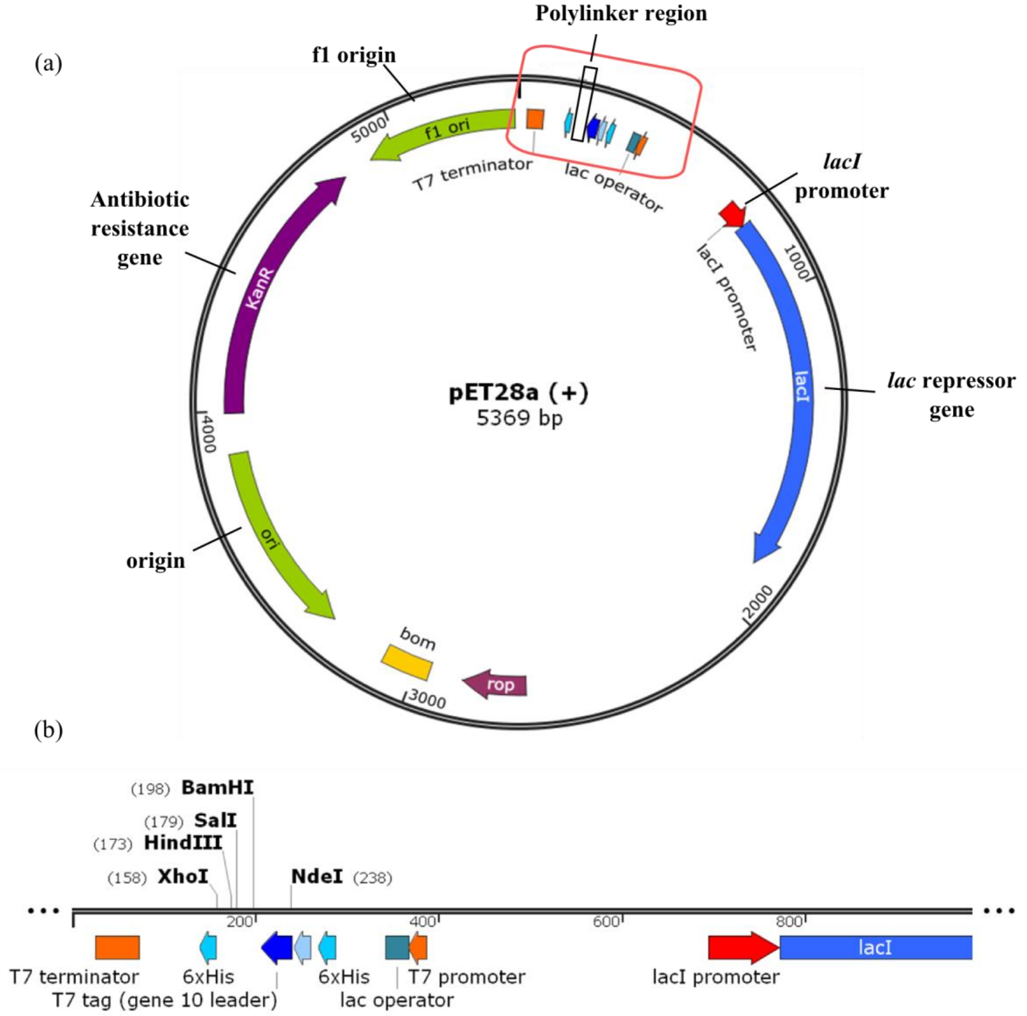


Figure 17. pET28a (+) vector

(a) Features of the pET system. (b) The squared area, part of the poly-linker is magnified for detail. Only some of the restriction sites used in this project from the polylinker area are shown.

For solubility screens, the induction time was for 3 h at 37 °C to an optical density at 600 nm (OD_{600}) of ~0.6 or overnight at 16 °C. 1 mL culture aliquots were prepared and centrifuged to keep only the pellet.

For large scale production, 1:100 dilutions were prepared with 7.5 mL of overnight culture (37 °C) and 750 mL of autoclaved LB medium. The baffled flasks were grown up to 0.6-0.8 $O.D_{600}$ at 37 °C, 180 rpm. Protein overexpression was induced with addition of 1 mM IPTG at the temperature where the higher solubility was observed in the solubility screens.

2.4 Production of Seleno-Methionine (Se-Met) proteins

An overnight culture of B834 cells expressing the POI in LB medium was used to inoculate 500 mL of 40 μg mL^{-1} Se-Met medium [including 1μg mL^{-1} vitamins, 40 μg mL^{-1} amino acids and 0.4% (w/v) glucose, 2 mM $MgS0_4$, 25 μg mL^{-1} and 2X M9 salts] and incubated at 37°C until $O.D0_{600\ nm}$ ~ 0.8 1 mM IPTG was added to induce protein overexpression at 16 °C overnight.

B834 cells are methionine auxotrophs, which means they cannot synthesise their own methionine, being forced to incorporate the Se-Met supplied from the medium. Tryptone is absent in the medium as it contains sulphur-based Met.

2.5 Solubility tests in SDS-PAGE

1 mL aliquots of cultured bacteria were centrifuged after protein induction to produce pellets that were resuspended in 200 μL of buffer (Figure 18). The resuspended bacteria were sonicated to obtain the total lysate sample (soluble and insoluble). A sample of 12 μL mixed with 4X loading dye was loaded on the SDS-PAGE gel. Lysates were clarified by centrifugation at 13,300 rpm for 1 min at room temperature (RT). A further 12 μL of the supernatant was loaded in the gel to estimate the fraction of protein remaining in solution, which is the soluble protein. The same procedure was followed with aliquots taken immediately before induction.

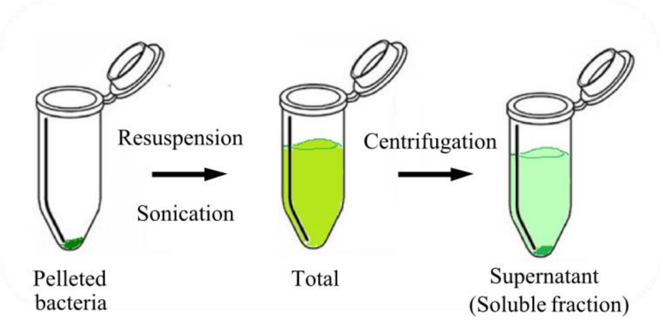


Figure 18. Solubility screen

12 or 15 % polyacrylamide gels (stock: 30 % (w/v) polyacrylamide, 0.8 % (w/v) Bis-acrylamide, ratio 37.5:1) were prepared and run in 25 mM Tris-HCl, 192 mM glycine, 0.1 % SDS for 50 min at 200 V. The gel was stained with Coomassie blue.

2.6 His-Tag stain

The InVision™ His-tag in-gel stain (Life technologies) was used to specifically stain His-tagged proteins. The protocol described by the manufacturer consists in fixing the gel (carefully washed with ultrapure water immediately after the run,) during 1 hour in 100 mL of 40 % (v/v) ethanol and 10 % (v/v) acetic acid, followed by 10 min washes with ultrapure water. The gel was incubated with the InVision™ His-tag In-gel stain for 1 hour to later be washed 2 times with 20 mM NaH_2PO_4 pH 7.8. The gel was visualised at 560-590 nm immediately after.

2.7 Nickel-affinity chromatography

Nickel-affinity chromatography (NAC) was used to purify most of the protein constructs in this study. By adding a poly-Histidine tag (poly-$(His)_6$ or His-tag) to the POI, it was possible to separate it from the bacterial cell debris and endogenous *E.coli* proteins as the poly-His segment binds to nickel ions, which are chelated by the immobilised group from the cross-linked agarose beads (Sepharose^{TM}) of the His-Trap^{TM} column (GE healthcare) (Figure 19a). Subsection 2.7 is based on protocols described in references (90,91).

Prior to the chromatography, the crude lysate was prepared by resuspending the pelleted bacteria in buffer containing 25 mM Tris pH 7.5, 200 mM NaCl, 20 mM Imidazole, 100 μg mL^{-1} lysozyme, 0.7 μg mL^{-1} pepstatin A, 0.5 μg mL^{-1} leupeptin and 100 mM 4-(2-Aminoethyl)-benzenesulfonyl fluoride hydrochloride (AEBSF). To see the buffers used for every construct see appendix **2**. The sonicated solution was spun down at 15,000-17,500 rpm to produce a clear supernatant.

The clear supernatant was loaded onto a His-trap high performance column (His-trap HP, bead size: 30 μm) attached to the ÄKTA purifier (GE Healthcare). The His-tagged protein was eluted over a gradient mixture of loading buffer with buffer containing 0.5 M free imidazole that has higher affinity towards the Ni^{+2} in the matrix than the His residues and hence disrupts the His-Ni^{+2} coordination bonds (Figure 19a).

Samples from the lysate, supernatant and fractions were loaded onto 12 % or 15 % (w/v) polyacrylamide gels (37.5:1) to estimate the presence and purity of the POI. Those fractions that contained the POI were pooled together to undergo thrombin digestion followed by size-exclusion chromatography (SEC)

When the His-tag was cleaved by thrombin protease (BD Biosciences), digestion was carried out at the same time that the sample was dialysed against the binding buffer (no imidazole) using 1 unit of thrombin to digest 1 mg of protein overnight at 4 ºC. A second NAC was performed to separate the cleaved protein from non-cleaved. The cleaved protein was further purified by SEC.

The His-tag was replaced by the glutathione S-Transferase-tag (GST) in three constructs because of its bigger size than the His-tag and its propensity to form dimers that in theory could help stop G*13*P from aggregating. Beads with immobilised glutathione were used to bind the POI. Elution was achieved by adding reduced glutathione to the buffer. GST fusion proteins featured a 3C protease cleavage site between GST and the G*13*P construct to facilitate further removal.

One of the G*13*P-Anti-TRAP (G*13*P) constructs was cloned into the H-MBP vector that encodes both the His-tag and the maltose-binding protein (MBP) that is usually included because it can increase the solubility of overexpressed proteins in bacteria. It is then possible to purify MBP fusion proteins by chromatography on cross-linked amylose and then elute with 10 mM amylose, however it was decided to make use of the His-tag and purify the construct by NAC

There was also some variation regarding the type of columns used to purify the constructs. His-trap fast flow crude columns (His-trap FF) (GE Healthcare) were chosen when constructs precipitated after centrifugation at high speed. The bead size in His-trap FF crude columns is ~ 90µm, making it possible to load clear lysate at relatively fast flow rates without blocking the column. The principles of binding and elution from the Ni-beads are the same as for His-trap HP columns.

The Hi-trap Q fast flow columns (or only FF Q, bead size: 45-165 µm) are anion exchange columns that contain quaternary ammonium ($-N^+(CH_3)_3$) as the anion exchanger (GE Healthcare) (Figure 19c). This column was used to isolate the untagged G*13*P constructs. The bacterial pellet was resuspended in binding buffer (low NaCl concentration), sonicated, centrifuged and loaded onto the column. The POI was then eluted with buffer containing high concentration of anion (Cl^{-1}).

2.8 Size-exclusion chromatography

Samples from NAC were concentrated and filtered through a 0.45 µm filter to remove big aggregates and rubbish from the sample that could block the SEC column filter. 500 µL or 5 mL were loaded manually in appropriate loops onto Superdex^{TM} 200 10/30 or Hi load 16/60 columns (GE Healthcare), respectively. The buffers used for all the constructs are listed in appendix **2**. When no S200 column was available, S75 16/60 columns were used.

The three type of columns listed contain a matrix of cross-linked agarose and dextran with particles of mean size of 34 µm that contain pores in their surface. As the sample goes through the column, smaller components will be delayed by passage through the pores while bigger particles will elute faster as they are unable to insert into the pores and instead pass through the void volume between the particles (Figure 19b).

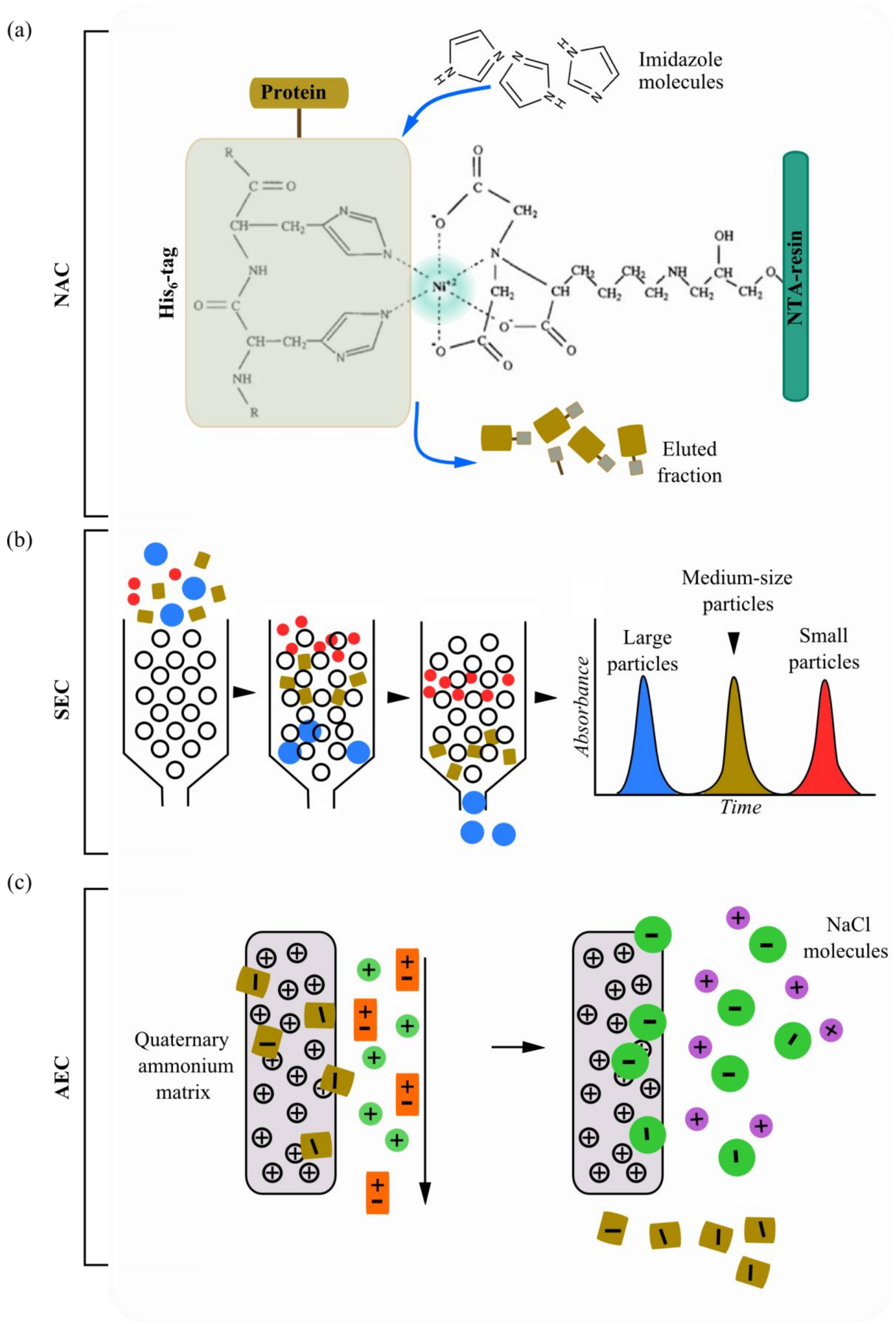


Figure 19. Protein purification techniques

(a) Nickel–affinity chromatography. One coordination bond between a His-tagged protein and the nitriloacetic acid (NTA) resin is shown. The Ni^{+2} atom is highlighted in cyan. Adapted from (90) (b) Size-exclusion chromatography. The separation of three protein molecules in a SEC matrix is illustrated. (c) Anion exchange chromatography. The matrix from a Q column binds to negatively charged molecules which are eluted with Cl^{-}. Protein or matrix components are not to scale. Figures b,c adapted from (92).

Those particles that are too big to fit into the bead pores will elute at the early stages of the run in the void volume. S200 and S75 columns can resolve particles that range in size from 3-10^3-7x10^4 and 1x10^4-6x10^5 (M_r), respectively. The manufacturer's instructions where followed to complete the protein separation.

2.9 Anion exchange Chromatography

Anion exchange chromatography (AEC) was included during the optimisation of the purification process of untagged G*13*P double mutants and during the purification of G20C ST. Buffer with low NaCl concentration was used to resuspend bacterial pellets and a clear supernatant was obtained as described above. The supernatant was loaded onto a Mono Q column 5/50 (GE Healthcare) that is packed with polystyrene/divinyl benzene beads that have immobilised quaternary ammonium as the anion exchanger. The -N+$(CH_3)_3$ in the beads adsorb molecules with opposite charge. As the counter-anion Cl^{-1} is added to higher concentrations, the POI is desorbed from the matrix according to its charge (Figure 19c).

2.10 On-column refolding

Fast expression of recombinant protein often leads to the formation of insoluble aggregates called inclusion bodies. Even though the inclusion bodies are formed by insoluble protein they have the advantage of producing the protein in high yield, this protein is protected from proteolysis and can be solubilized into folded active species (93).

The on-column refolding method used to purify one His-tagged protein from this study consisted of refolding the denatured protein while it was bound to the Ni-beads. One important aspect of this protocol is that it allows lowering of the denaturing agent concentration at a rate that prevents aggregation but allows folding. It also avoids exposure of the protein to intermediate denaturant agent concentrations for a long time period, like in dialysis, in which the POI is prone to aggregation since it is neither entirely folded nor denatured (93).

The pelleted bacteria were resuspended in cold denaturing solution (Table 4) and incubated on ice for 45 min. Quick sonication of the sample to disrupt DNA was followed by manual loading onto a His-Trap FF (GE Healthcare) column previously equilibrated with denaturing solution. Once the column was reconnected to the ÄKTA purifier, denaturing solution 2 was flowed over to remove any non-specifically bound protein. The denaturing agent was washed off with solution III that contained the detergent triton X-100 to prevent aggregation and 0.5 M NaCl to remove contaminating protein molecules bound to the target protein or column. β-cyclodextrin was included in solution IV to remove the triton X-100 from the protein-detergent complex and to promote folding. Before detaching the protein with elution buffer, solution V was flowed through the column to wash β-cyclodextrin off.

Table 4. On-column refolding solutions

Solution	Components:
I. Denaturing solution	6 M Guanidine hydrochloride 50 mM Tris pH 7.5 10 mM β-Mercaptoethanol
II. Denaturing solution 2 (2 CV)	Solution I + 20 mM Imidazole
III. Detergent buffer (4 CV)	50 mM Tris pH 7.5, 0.5 M NaCl, 0.1 % Triton X-100
IV. Buffer for detergent removal (15 CV)	50 mM Tris pH 7.5, 0.1 M NaCl, 5 mM β-Cyclodextrin
V. Basic Buffer (5 CV)	50 mM Tris pH 7.5, 0.1 M NaCl
VI. Elution Buffer (20-25 CV)	50 mM Tris pH 7.5, 0.1 M NaCl, 0.5 M Imidazole

2.11 Pull-down assays

Pull-down assays were used to determine where the spontaneous truncation in G*13*P-AT engineered proteins was taking place. 30 μL of nickel-beads (from His-TRAP columns) were washed two times with 25 mM Tris pH 7.5, 250 mM NaCl to remove the storage solution. 100 μL of soluble fraction from the BL21 Star cultures (section 2.2) were added to the beads. The mix was incubated at room temperature with continuous shaking for 15 min. 12 μL were kept to load later on SDS-PAGE as the total sample. The mix was centrifuged at 13,300 rpm for 1 min, and the supernatant was discarded. The beads were washed two times with buffer. 12 μL of supernatant (SND) and beads were loaded onto SDS-PAGE. Finally, the bound protein was eluted with the binding buffer plus 0.5 M imidazole. Two dilutions of the eluted protein were loaded.

2.12 Size Exclusion Chromatography coupled with Multi-Angle Laser Light Scattering (SEC-MALLS)

SEC-MALLS is a technique used to measure molecular weight based on how molecules scatter laser light. SEC-MALLS was used to estimate the molecular weight of G*13*P or the oligomeric state of G20C ST. Scattering of light is dependent of the POI concentration, refractive index, molecular weight and scattering angle.

His-tagged or non-tagged samples were diluted to a fixed concentration in 20 mM Tris pH 7.5, 250 mM NaCl, and loaded onto a BioSep SEC-s3000 gel filtration column (Phenomenex) which was equilibrated with 20 mM Tris pH 7.5, 250 mM NaCl. Size-exclusion chromatography was carried out on a Shimadzu HPLC system with flow rate of 0.5 mL/min. The elution was monitored at 280 nm by a SPD20A UV/Vis detector. Light scattering data were recorded by a Dawn HELEOS-II 18-angle light scattering detector and the concentration of the eluting protein was measured by an in-line Optilab rEX refractive index monitor (Wyatt Technology). Data were analysed using the ASTRA V software package (Wyatt Technology).

2.13 Protein crystallisation experiments

Only when protein molecules precipitate in an ordered manner can crystals form. Variables influencing crystallisation include the nature of the protein, the nature of the precipitating agent, the protein or crystallising agent concentrations, pH, temperature, time, etc. The dynamic representation of the interaction of several variables and resulting solubility with respect to the crystallisation outcome is known as phase diagram (Figure 20a).

The phase diagram is organised into several zones that describe how the phase of the protein behaves as the variables change, for example, protein and crystallising agent concentrations (Figure 20a). In the undersaturated zone the protein will remain in liquid solution. When a crystallisation experiment is set up it is expected that the protein concentration will increase until the labile zone is reached, which is where the formation of the nuclei occurs. Protein molecules will aggregate around the nuclei so the protein concentration will decrease until nucleation will no longer occur and a new zone, the metastable zone, will be reached. In this zone, the crystal growth is supported only while the protein concentration is high enough. Liquid-liquid phase separation, visible as protein rich droplets under the microscope, happens at the metastable zone, where the protein concentration is high. At some point, the crystal will get to the saturation zone, indicated by the solubility curve, where the crystals, if present, will be in equilibrium with the solution. The crystal would dissolve if the solution changes to the undersaturated zone. In cases where the protein concentration is too high to sustain ordered aggregation, precipitates will form. The supersaturated precipitation zone is characterised by the formation of microcrystalline or amorphous precipitates, with the former being able to redissolve, keep their native conformation and sometimes act as good seeds for crystallisation.

2.13.1 Sitting drop vapour diffusion

The POIs were utilised to set up crystallisation trays using commercial screens: INDEX (Hampton Research), PACT (QIAGEN), MPD (QIAGEN), ammonium sulphate screen (QIAGEN), MORPHEUS (Molecular Dimensions) and Clear Strategy Screens I and II (CSS 1&2). All were selected as initial screens because of the variety of salts, precipitants, pH value and additives they incorporate in order to cover an extremely diverse variety of chemical conditions. Sitting drop vapour diffusion experiments were set up in 96-well MRC crystallisation plates with a Mosquito nanolitre pipetting robot (TTP lab-tech). The program carried the instructions to aliquot 54 µL of each screen solution into each well and then to aliquot 150 and 300 nL of sample to the top and bottom drop, respectively plus 150 nL of reservoir solution to each drop.

In this kind of experiment, the protein-reservoir solution drop is placed in vapour equilibrium with the reservoir. Since the concentration of reagents is lower in the drop, water will diffuse from the drop to the reservoir, producing an increase in the protein and reagent concentrations, a fact that could take the drop solution into the labile zone for crystal nucleation (Figure 20b). The trays were stored in a temperature-controlled room at 18°C unless otherwise stated.

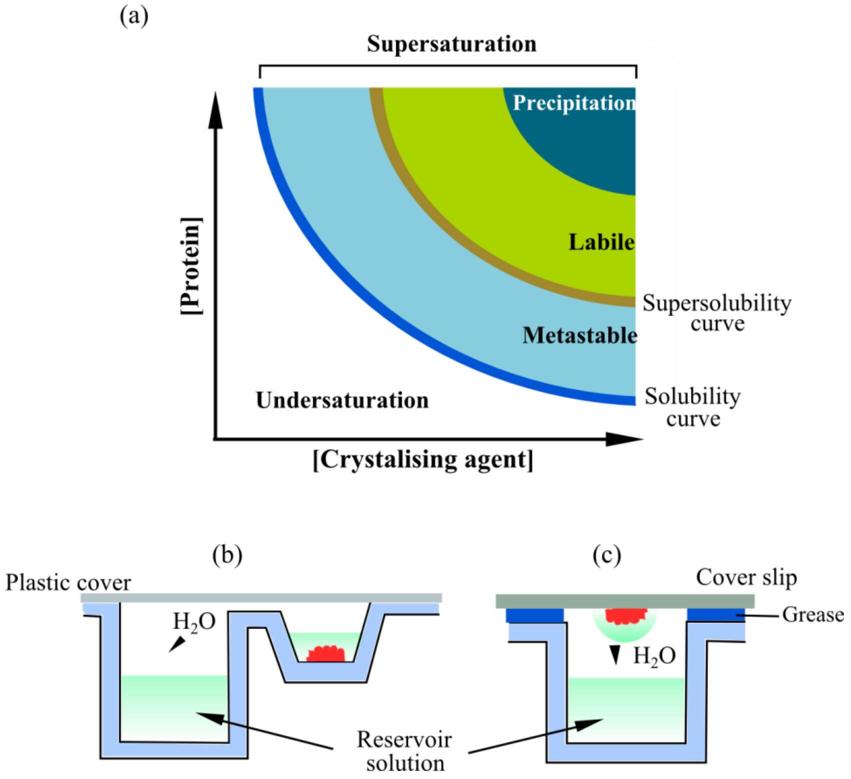


Figure 20. Protein crystallisation experiments

(a) Phase diagram describing the crystal input as the concentrations of protein and crystallising agent increase. Adapted from Luft, *et al*, 2011 (94) .(b), (c) Sitting and hanging drop vapour diffusion settings. Adapted from reference (95).

2.13.2 Manual optimisation: hanging drop vapour diffusion experiments

When hits (formation of crystalline material) were observed in the 96-well trays, the original conditions were occasionally manually reproduced and optimised, varying the protein and reagent concentrations or additives. Optimisation was carried out with hanging drops equilibrated with 500 µL of reservoir in previously grease-sealed 24-well trays. Wells were

covered with siliconised glass coverslips and incubated at room temperature. Usually 1 μL of protein was mixed with 1 μL of reservoir, when indicated, 0.5 μL of seed stock were added.
The principle of the hanging drop and sitting drop are the same, with differences lying in the speed of setup or easiness to fish the crystals (Figure 20c), and theoretical maximum dimensions of crystals grown due to the size of the drop and equilibration times.

2.13.3 Seeding

Seeding is a technique where already formed crystals are used as nuclei that are introduced into new reservoir-drop condition in a lower level of supersaturation to nucleate crystal formation (96).

Two types of seeding were used: macroseeding and microseeding. Macroseeding implicated the transfer of macro seeds or crystal (5-50 µm in diameter) to the new drop. For this type of seeding, the seed stock preparation consisted of mixing drops that contained crystals (under similar conditions) and then streaking drops prepared as described for manual optimisation with a needle previously soaked in the seed stock.

The preparation of the seed stock was more elaborate for the microseeding. Crystals from single drops were collected separately and then resuspended in 50 - 100 µL of ice-cold stabilising solution placed in a seed bead (Hampton research). The conditions on the stabilising solution were essentially the same as in the original reservoir solution but either the concentration of the main precipitant was slightly increased or glycerol was added to prevent the crystals from dissolving. The seed bead was vortexed for 1 min and then centrifuged at 14, 000 rpm at 4°C for 1 min. The vortexing and centrifugation were repeated once. To use the seed stock, a final vortexing was implemented. The seed stock was added directly to the optimisation drop (0.5 µL) or aliquot by the Mosquito robot into each of the 96 conditions from the commercial screens (50 nL).

2.13.4 Crystal testing

The fished crystals were flash-cooled in liquid nitrogen to be tested in-house using a MSC-RUH3R X-ray generator (Rigaku) with rotating anode and MAR345 detector (Marresearch). When the diffraction was about 3.5 Å resolution, the crystals were kept in liquid nitrogen and sent to the Diamond Light Source, UK, facility to collect data at the I04 beamline. When some of the crystals were fished out, cryo-protectant solution was used to keep the crystal in optimal condition.

2.14 Bioinformatics, molecular graphics and modelling software

The following online software was utilised to investigate SPP1's G*13*P or G20C ST's sequences or to create figures of proteins relevant to this thesis from the published PDB files published in the RCSB PDB (97).

2.14.1 Bioinformatics

- The Position-Specific iterated Basic Local Alignment Search Tool (PSI-BLAST) was used to align sequences of capsid proteins with published X-RAY structure to G*13*P (98).
- The secondary structure of G*13*P was predicted with the JPred server (99).
- Multiple sequence alignment and multiple secondary structure prediction were carried out using PRALINE using DSSP (100) and PSIPRED (101).
- Nucleotide sequences were aligned using ClustalW (102).
- Disorder prediction for G*13*P was done with DisEMBL^{TM} (103).

2.14.2 Molecular graphics

- All molecular representations like ribbon diagrams or electrostatic surface representation were created using CCP4mg (18).

2.14.3 Analysis of X-ray crystal structures

- The protein topology was sketched using the Pro-origami server (104).
- PDBePISA was used to identify H-bonds and buried/exposed surface (105).

2.14.4 Modelling

- PDBefold was used to identify crystallised protein-DNA complexes that could be superposed to G20C ST. Scores will be explained in section 8.4 (106).

2.15 Elucidation of X-ray crystal structures

The following subsection provides the reader with a list of software used to solve X-ray structures. All the references are included for more details about the calculations and details of each program.

2.15.1 Data processing

The set of diffraction patterns collected for crystals of native or Se-Met SPP1 G*13*P or G20C ST were processed to deduce crystal symmetry and cell dimensions, and to produce integrated intensities of the reflections present in every diffraction image using the program XDS (X-Ray Detector Software) (107) .

The previously integrated reflections were merged by the program Aimless (108,109). Merging of reflections consists in averaging reflections equivalent by symmetry and of similar intensity.

2.15.2 Phasing by single anomalous diffraction (SAD)

The SHELX*C* and SHELX*D* (110) programs provided the location of the Selenium atoms present in SPP1 G*13*P and G20C ST.

The phases of the data set were calculated using the software Phaser SAD (111,112) providing the integrated intensities, the substructure from SHELXD and the amino acid sequence. To investigate the structure of the crystallised G20C ST and SPP1 G*13*P double mutant in the lower-resolution native data sets, molecular replacement (MR) was carried out. The Phaser MR option was used to determine the phases inputting the refined Se-Met protein models of both proteins and the processed native data sets.

2.15.3 Density modification and automated model building

The program Parrot (113) was run to perform density modification on the phases determined by Phaser SAD in order to obtain an electron-density map to submit for automated model-building.

The modified electron-density map, the sequence of G*13*P and the coordinates of the heavy atoms were fed to the program Buccaneer (114) to perform automatic model building.

2.15.4 Refinement and validation

The refinement, which is the process to improve the agreement of the model and the X-ray data was performed using the REFMAC software (115).

Coot (116) was used for validation of the model.

Additional miscellaneous data processing was carried out using diverse programs clustered in the CCP4 interface (117).

2.16 DNA-binding tests using electrophoretic mobility shift assays (EMSA)

To test the capability of the G20C ST WT to bind DNA, the DNA concentration, either G20C_F1 or pUC18 (Chapter 9), was adjusted to 0.015 µM in a 20 µL total volume containing 25 mM Tris pH 7.5, 100 mM NaCl, 20 mM Mg^{+2} and 1 mM ATP (a master mix was prepared to transfer the same volume to all the reaction tubes). The ST WT concentration was adjusted to 3.75, 7, 15 and 30 µM to have molar ratios of 9-mers to 1 kbp DNA fragment of 250:1, 500:1, 1000:1 and 2000:1, respectively. DNA controls included no protein to test the migration of DNA alone. Reactions were incubated at 37°C during 45 min and then loaded on 0.75 % (w/v) agarose gel. The gel was run at 60 V, 45 mA at room temperature for 1 h 20 min. SYBR safe® was added during the preparation of the agarose gels.

To test the capability of the ΔC, ΔN and DBD G20C constructs to bind dsDNA, the concentration of ΔC, ΔN constructs was adjusted to 30 µM while the concentration of DBD was 270 µM. DNA concentration was kept at 0.015 µM for all the constructs so the final construct to 1 kbp DNA fragment molar ratio was 2,000:1 for ΔC and ΔN, and 18,000:1 for the DBD. Solution conditions were the same as stated above, as was the procedure followed to run the agarose gel.

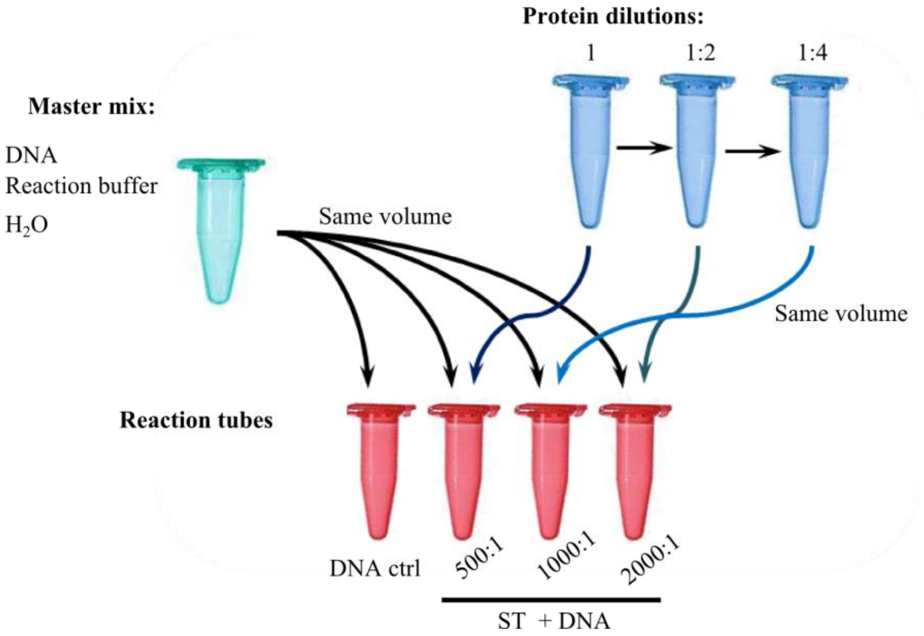


Figure 21: DNA-binding experiment settings

Chapter 3

3. Production of SPP1 capsid protein crystals

The current chapter summarises approaches carried out to produce homogeneous SPP1 capsid protein to use in crystallisation experiments. The outcome of this investigation is arranged in subsections that describe the main results for the different types of modification that G*13*P underwent. Such modifications include truncations at the N- or C-termini, single and double point mutagenesis, and protein engineering by fusion. The information presented flows in a logical and chronological ordering. The design and purification of every modified construct are presented in the same subsection while the crystallisation work performed on every construct found to be suitable for crystallisation trials is summarised in section 3.5. Most of the experimental details are described in Chapter 2. Whenever appropriate, protocols and/or protocol modification are briefly described.

3.1 Purification of G*13*P wild type

The full length gene coding for G*13*P_WT was cloned into pET28a vector between the NdeI and XhoI restriction sites. The pET28a vector contains a Hexa-Histidine tag (His-tag) at the 5' end of the NdeI restriction site, separated by a flexible linker that includes one thrombin-cleavage site. Information about the cloning design for the WT protein and subsequent constructs can be found in appendix **1**.

The recombinant DNA was used to transform BL21 and Rossetta 2 pLysS cells. Overnight cultures of 5 mL medium were set up to perform a solubility screen from which it was observed that G*13*P_WT is expressed in a higher amount and in soluble form when induced in BL21 cells at 37 °C than in Rossetta 2 pLysS or at 16 ºC. Expression was scaled up to cultures of 750 mL medium to purify G*13*P_WT by NAC.

3.1.1 Nickel Affinity Chromatography and Size exclusion chromatography

The protocol for purification commenced with NAC and continued as described next. It was observed that the lysate had a cloudy aspect due to the presence of small granular white precipitate, indicating the formation of protein aggregates. Cell lysate containing G*13*P_WT, purified in 25 mM Tris pH 7.5, 150 mM NaCl using an imidazole gradient (Figure 22), eluted in four peaks which indicated four species with different affinity towards nickel ions. The peaks were analysed separately. SDS-PAGE confirmed that the SPP1 capsid protein was present in all peaks and that the migrating size corresponded to the expected size, ~35 kDa.

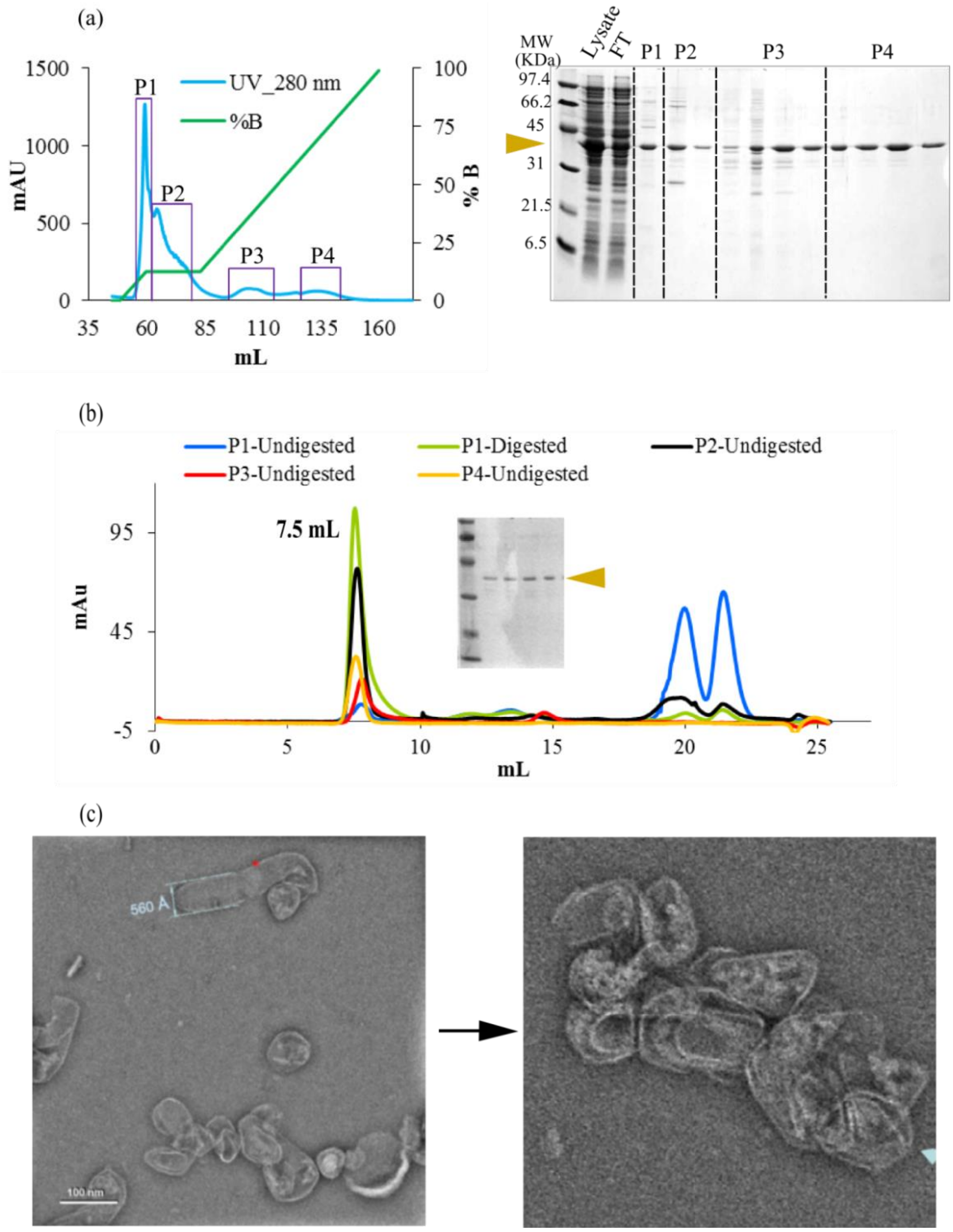


Figure 22. Purification of G*13*P_WT

(a) NAC profile. Peaks and fractions are numbered 1-4 according to the peak numbering. The yellow arrowhead points to G*13*P. The lysate, flow through (FT) and eluted fractions were loaded onto the SDS-PAGE. (b) SEC for the four peaks from NAC. The elution volume of the main species is indicated. The gel in the inset represents fractions from SEC runs. (c) 2% Uranyl acetate stain for cryo-EM. The diameter of one structure is indicated by light blue numbers and a single capsid wall is indicated by a red star.

The same volume from each peak was run through a S200 10/30 SEC column to further separate species in the samples by size. Even though protein concentration is different in the four samples, it is observed that G*13*P_WT elutes in the void volume of the column, which is the volume confined to species that are too big to migrate through the gel pores due to its aggregated form. Peaks 1, 2, 3 and 4 eluted at ~7.5 mL. By running commercial SEC markers, it was estimated that monomeric species would be expected to have an elution volume of around 14.4 mL, as an example, ovalbumin (44 kDa), eluted at 14.4 mL under the same conditions G*13*P_WT was run.

To rule out the His-tag playing a role in the aggregation process, one thrombin-digested sample was analysed by SEC. The digested sample showed the same elution volume as the undigested samples, confirming that both samples form aggregates, and thus indicated that aggregation is probably mediated by G*13*P-G*13*P contacts.

SEC fractions of non-digested sample were analysed by cryo-EM by collaborators at the Birkbeck College, London (Figure 22c). Deformed capsid-like structures of ~ 560 Å in diameter and ~ 35 Å-thick walls were visible. The capsid-like structures were shown to pack against each other, as had been observed with the overexpression of G*13*P_WT in the absence of scaffold protein G*11*P which also led to the formation of aberrant capsids (29).

The discovery that the size of the capsid-like structure is smaller than that of the mature capsid (~610 Å (41)) determined by cryo-EM but the wall is thicker (~27 Å thick in the mature capsid) can be explained by the absence of G*11*P or G*6*P that when present, lead G*13*P to organise into a proper icosahedral shell. Also, the absence of the molecular motor to pack DNA leads to the immature versions of the capsid that do not undergo expansion when full with DNA. Thickness of the presented capsid-like structure is also bigger than that of the mature HK97 capsid (~18 Å) (39), supporting the idea that reorganisation of the capsid protein is needed to produce thinner but more stable walls.

3.1.2 Solubility test

To overcome the G*13*P_WT polymerisation property in pursuit of obtaining stable and soluble monomer, a more thorough solubility screen was carried out (Figure 23a). New overnight cultures were resuspended in buffers containing 50-500 mM NaCl and 50 mM of buffer (Na acetate pH 4, Tris pH 7.5 or Ches pH 10). Different additives like glycerol, sucrose or metals were also added. After sonication, total samples were taken to use as reference and the remaining solution was centrifuged at the start time (T_o) and incubated at RT during thirty minutes or two hours. After this time, samples that would make a new total sample were taken (2h) and the rest of the solution was spun down at 13.3 krpm to obtain the soluble fraction (S2h). Where indicated by a star (*), samples were spun down at 17.5 krpm, the same speed that is used to obtain a clear lysate for NAC.

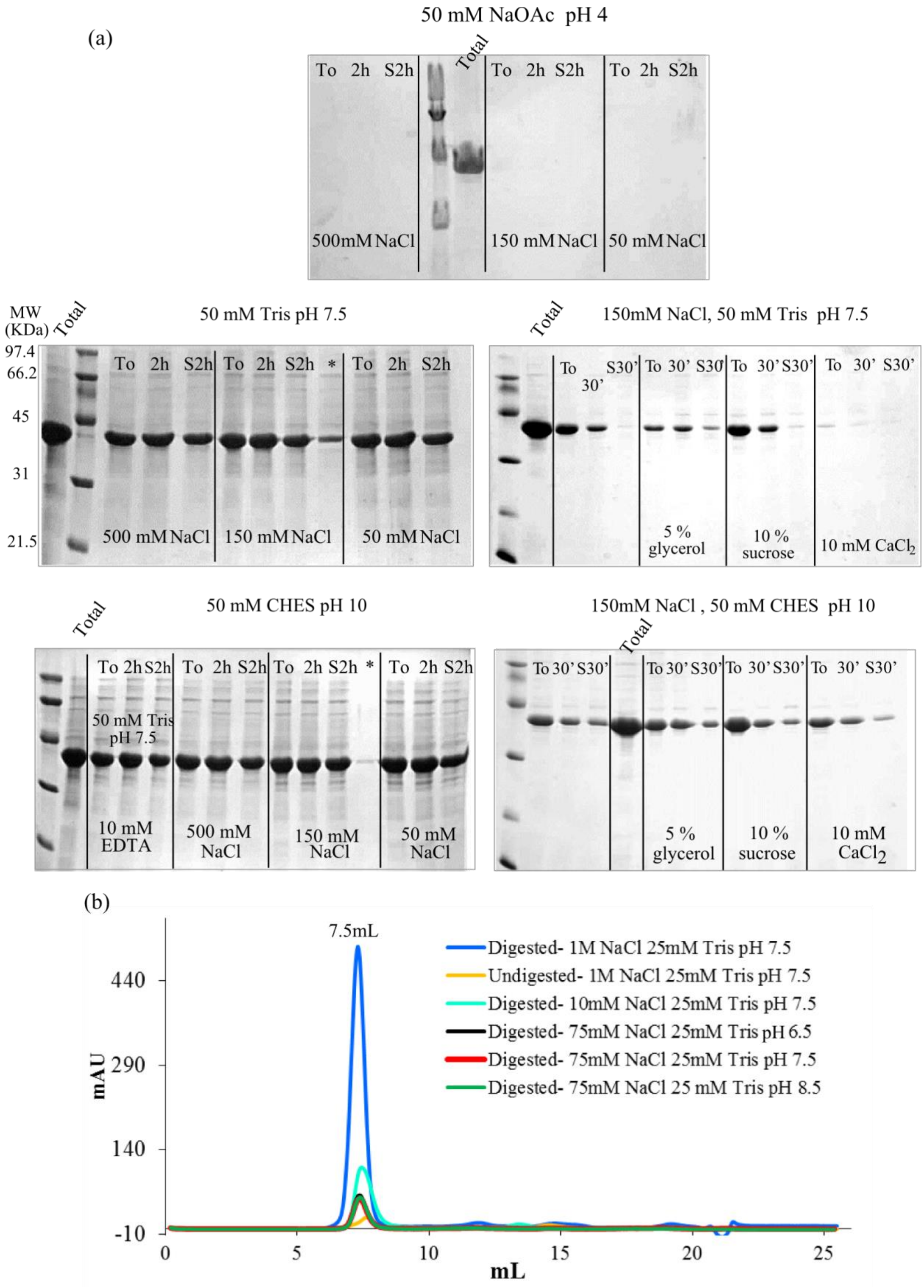


Figure 23. Screening of conditions to disrupt aggregation

(a) Solubility screen at pH 4 (top), 7.5 (centre) and 10 (bottom). The time (T_0 or 2 h) at which total samples were taken is indicated at the top part of the lanes and the NaCl concentration or included additives are indicated at the bottom part. “S2h” and a star (*) denote the final soluble fractions after the 2 h of incubation and after centrifugation at high speed (17.5 krpm), respectively. (b) SEC of thrombin-digested and digested G*13*P _WT at varying pH and NaCl concentration. The elution volume of the main specie is indicated.

None of the conditions tried had a significant effect in stopping aggregation as can be seen in samples after centrifugation. For example, when pH 7.5 was used the amount of protein at T_o was reduced to about 50%, of the original amount in the three tested NaCl concentrations.

After two hours, the same amount of protein remained in solution (2h), however, subsequent centrifugation decreased the amount of protein (S2h). When the sample in 150 mM NaCl was centrifuged at fast speed (17.5 krpm), most of the protein precipitated (Figure 23).

In experiments including additives, samples were only incubated during 30 min but the behaviour observed is the same as explained above for samples with no additives.

In the case of acidic pH, samples precipitated when SDS was added to the mix, forming solid aggregates that did not migrate through the polyacrylamide gel.

3.1.3 Size exclusion chromatography with variable NaCl concentration

To confirm whether changes in the pH or NaCl concentration had any effect in leading G*13*P_WT towards the formation of smaller or monomeric species more suitable for crystallisation, thrombin-digested samples from NAC were run onto a SEC column using conditions different to the originally tested (Figure 23b). The logic behind this screen is to find those conditions where solvent-protein contacts are preferred over protein-protein contacts. This approach demonstrated that the size of the aggregates does not vary upon change in pH or NaCl concentration as all the samples eluted at 7.5mL, corresponding to the void volume.

3.2 Purification of truncated G*13*P

3.2.1 Secondary structure prediction

Considering that the WT protein was not suitable for crystallographic studies, truncations at the N- and C-termini were introduced based on the secondary structure and disorder predictions (Figure 24) in an attempt to produce more ordered monomers by removal of disordered regions. Often, short regions at both ends do not have a defined secondary structure and their flexibility prevents the whole protein from establishing regular contacts with other subunits to form a crystal. It is not unusual that removal of a few residues at either end results in more soluble or non-aggregating proteins. Ten truncated versions of G*13*P were designed (appendix **1**) and cloned: only one construct lacked a C-terminal portion while the other nine lacked several lengths of the N-terminus. Moreover, three of these constructs were GST-tagged and the rest had an N-terminal His-tag, removable with thrombin protease.

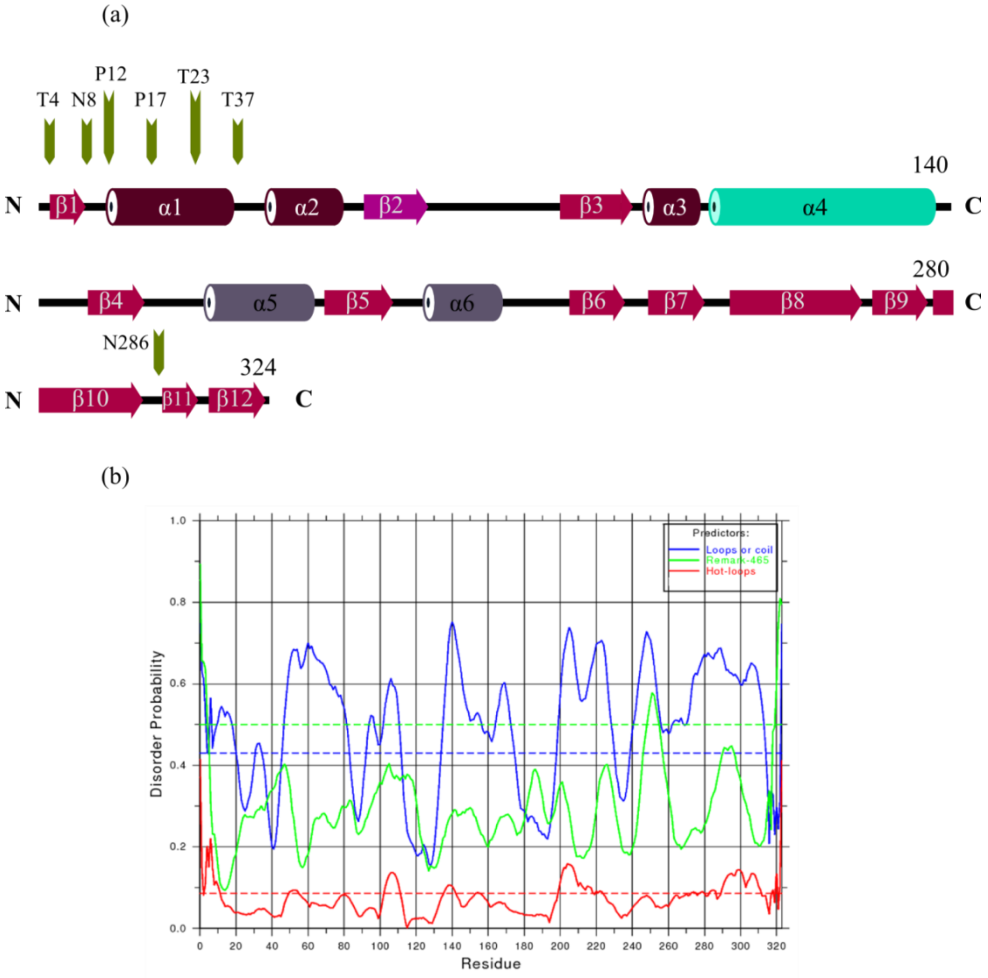


Figure 24. Design of truncations in G*13*P_WT

(a) Scheme of the secondary structure prediction by JPred (not to scale). α-helices, β-sheets and loops are represented by cylinders, thick arrows and a black line, respectively. Numbering is included. Positions where truncations where introduced are indicated by green arrowheads. The residue showed on top of each arrowhead is the first residue after the start methionine. (b) Disorder prediction was carried out using the DisEMBL server (103). The disorder probability is plotted versus the amino acid number. The blue line represents the propensity to form loops, the green line the disordered regions and the red line the floppy loops that have high temperature factors.

3.2.2 Solubility test

The designed constructs were cloned and expressed (see appendices 1-3) in two bacterial strains at 16 and 37 °C. The solubility tests showed that constructs expressed at 16 °C and 37 °C G*13*P_ΔC, ΔN15 His-tagged, ΔN15 GST-tagged, ΔN21 His-tagged, ΔN21 GST-tagged and ΔN27 were insoluble in both strains in spite of being expressed in high amount (Figure 25a). Only results from one strain, one temperature or a small number of buffers are shown. Unless otherwise mentioned, similar findings were observed in all conditions.

When the protein was soluble, variations in pH or NaCl concentration did not seem to make a difference since the amount of protein found in solution was similar between the samples. For construct G*13*P_ΔN10 it was observed that expression at 16 °C produced soluble protein in contrast to expression at 37 °C, where even though the POI was expressed, it was not soluble.

The introduction of the GST-tag in constructs G*13*P_ΔN15 and ΔN21 did not improve solubility. The GST-tag was chosen due to its high solubility and large size that would impede G*13*P aggregation. One version of the G*13*P_WT with GST tag was cloned, expressed and purified, but it did not produce a high yield (not shown). G*13*P_ΔN6 exhibited high solubility in all the tested conditions.

The truncated constructs were designed and expressed at different times, therefore the conditions tried varied based on experience with previously characterised constructs.

3.2.3 Nickel Affinity Chromatography and Size exclusion chromatography

The two truncated versions that were soluble were expressed on a large scale followed by NAC. It was observed in repeated experiments that centrifugation at high speed of the lysate of cells containing G*13*P_ΔN10 led to precipitation. New crude lysate was loaded onto a His-Trap FF crude column that allows loading crude lysate without further centrifugation. Using this approach made it possible to obtain enough G*13*P_ΔN10 to perform SEC to characterise its oligomeric state (Figure 25b). Both thrombin-digested and non-digested samples were analysed, demonstrating that G*13*P_ΔN10 still forms big aggregates that elute in the void volume.

G*13*P_ΔN6 was purified by standard NAC, however its behaviour in SEC was the same as G*13*P_ΔN10 as it also eluted in the void volume.

Only G*13*P_ΔC included truncation at the C-terminus and it was the first construct characterised after the WT. A thorough solubility screen that involved testing the conditions tried with the ΔN constructs was carried out. In addition, protein auto-induction was also included. After several efforts to make it soluble it was abandoned in favour of the ΔN constructs (Figure 25a).

Truncating the G*13*P_WT protein proved not to be an effective approach to stop aggregation. Nevertheless, it established the extent to which the protein could be truncated. This information was useful when designing new modifications that will be described later in this chapter.

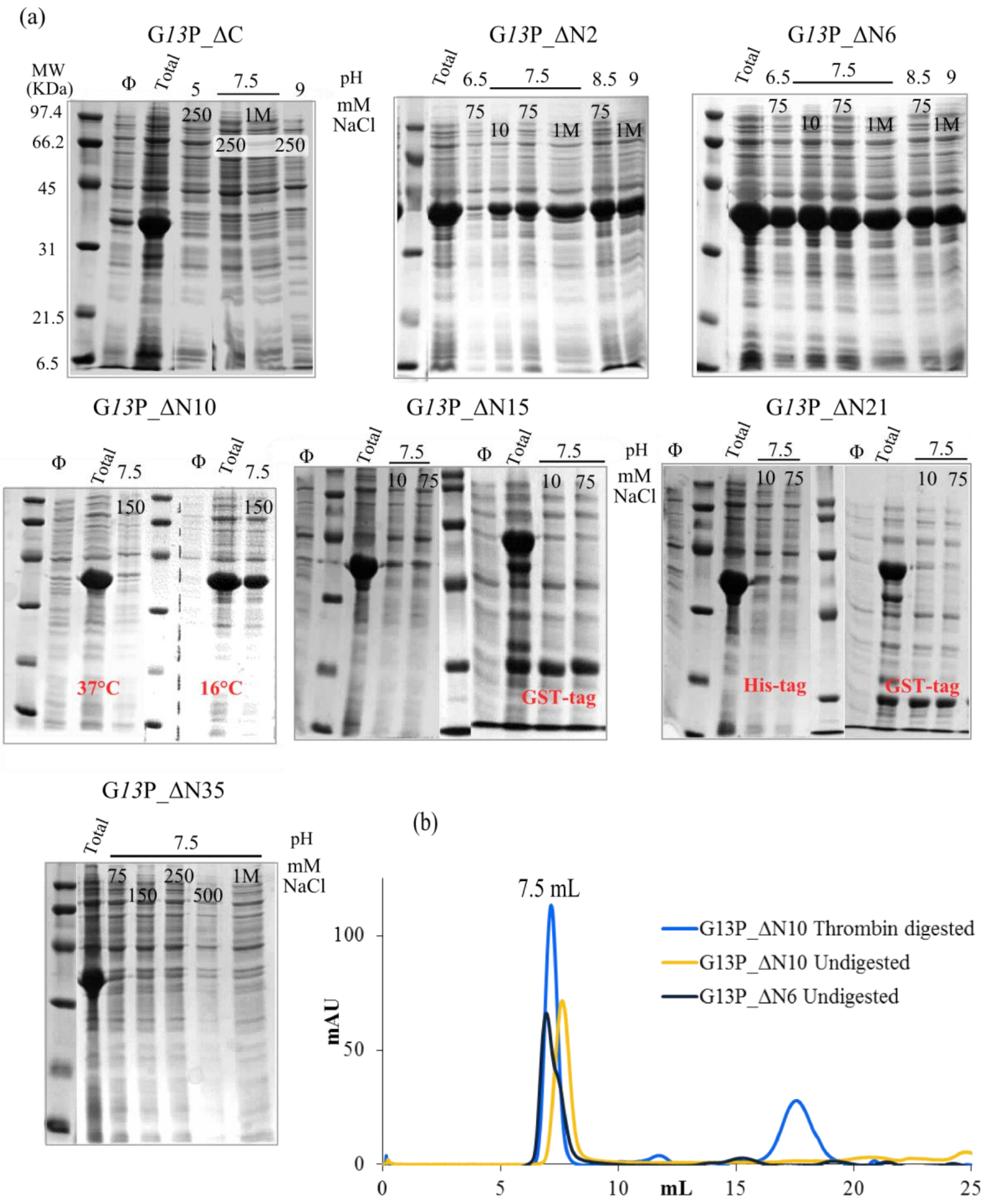


Figure 25. Production and purification of the G*13*P truncated constructs

(a) Solubility screen of truncated constructs. Only results from BL21 cells induced at 16°C are shown. The pH and NaCl concentration are indicated at the top part of the gel. The gels are not to scale to each other, for that reason the protein markers are included in every gel. Φ corresponds to samples before protein induction. Total samples correspond to the sonicated sample after induction and the rest of the lanes to the soluble fractions. (b) SEC profile of G*13*P_ΔN6 and G*13*P_ΔN10. The elution volume is indicated at the top of the peaks.

3.3 Design and purification of single and double mutants

3.3.1 Design of single mutants

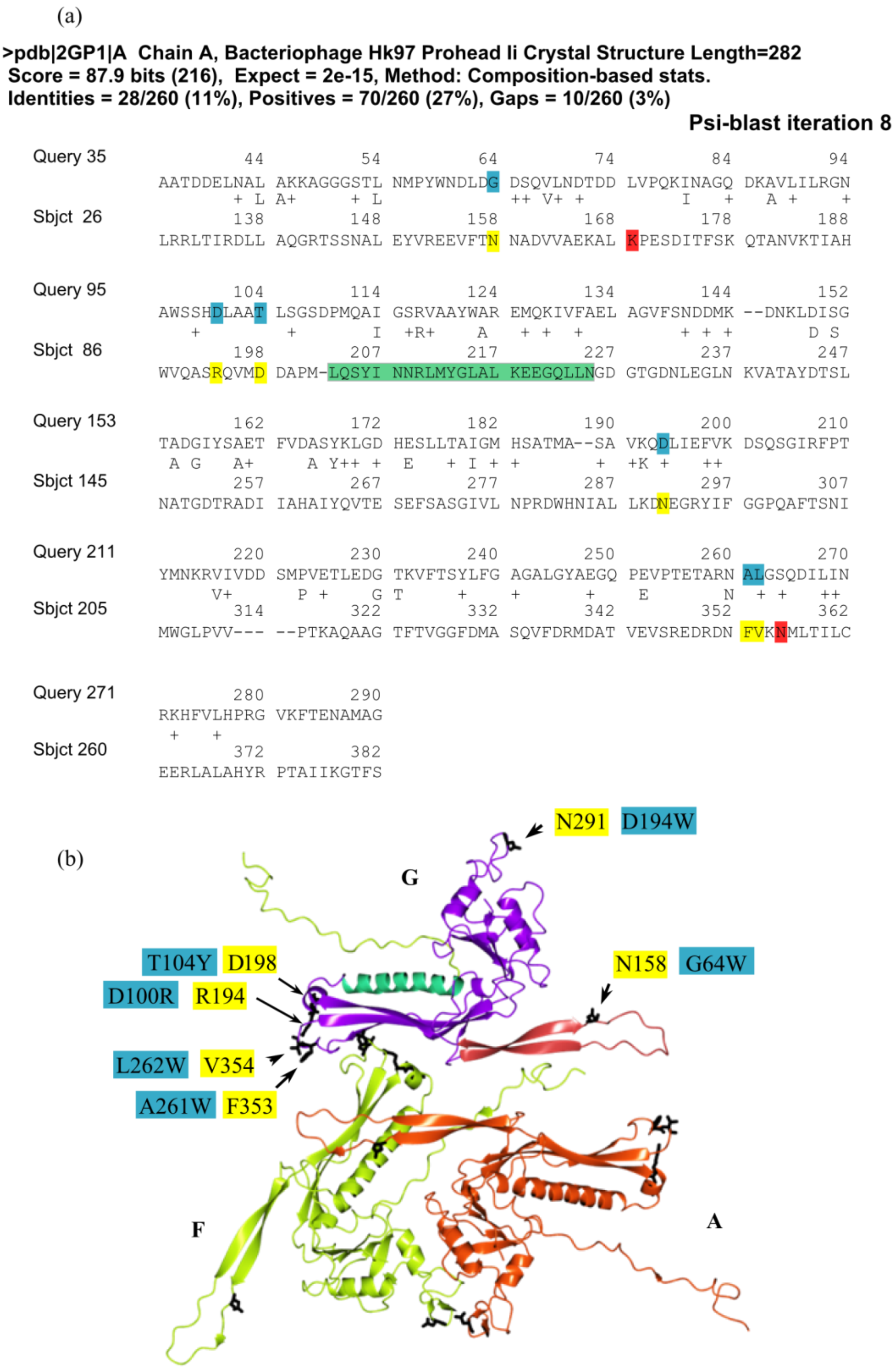
From section 1.5.1 it is evident that in spite of the low sequence homology between the known capsid proteins from dsDNA bacteriophages, they all have the HK97-fold. Position-specific iterated Basic Local Alignment Search Tool (PSI-BLAST) was used to find proteins with published structure that would guide us in the design of mutations that stop interactions between G*13*P_WT (No cryo-EM model of G*13*P was available during the early years of this project). The mutations were expected to disrupt the formation of the capsid-like structures promoting the formation of hexamers, pentamers or monomers in isolation. The PSI-BLAST was used to produce a multiple protein sequence alignment with HK97 gp5 that showed 28 identical residues (11%) and 70 conservative substitutions (27%) introducing 10 gaps (Figure 26a). Once HK97 gp5 was predicted to have structural homology to G*13*P, the analysis of the PDB file was carried out using the PDBePISA server.

3.3.2 PDBePISA Analysis

Residues participating in inter-subunit contacts in the HK97 icosahedral asymmetric unit were identified with PDBePISA, which lists the H-bonds created at the different interfaces between the seven subunits (Figure 26). Table 5 and Table 6 list some of the residues that establish H-bonds between subunits F & A, both from the hexamer, and between subunits G & F, with G being part of the pentamer (complete tables can be found in appendices **4** and **5)**.

Six residues were selected in HK97 gp5 based on their likely significance in forming positive interactions with neighbouring subunits because of their positioning or H-bonding ability. The aligned residues in SPP1 G*13*P, based on PSI-BLAST were mutated to facilitate crystallographic studies (Figure 26).

Residue N291 from HK97 gp5 is situated at the top of the A-domain and it creates a circular network of H-bonds with the same residue in the adjacent subunits in both pentamers and hexamers. In SPP1, the same position identified by the iterated BLAST sequence alignment is occupied by residue D194. This residue was mutated to tryptophan by single point mutagenesis, due to the tryptophan's size and steric bulk that would be expected to disrupt the H-bonding network at the subunit-subunit interface, and at the same time keep A-domains of other subunits further from each other (Figure 26 and Figure 27).



```
>pdb|2GP1|A  Chain A, Bacteriophage Hk97 Prohead Ii Crystal Structure Length=282
 Score = 87.9 bits (216),  Expect = 2e-15, Method: Composition-based stats.
 Identities = 28/260 (11%), Positives = 70/260 (27%), Gaps = 10/260 (3%)
                                                              Psi-blast iteration 8

Query 35          44         54         64         74         84         94
          AATDDELNAL AKKAGGGSTL NMPYWNDLDG DSQVLNDTDD LVPQKINAGQ DKAVLILRGN
                 + L A+     + L            ++ V+ +         I   +   A +    +
Sbjct 26         138        148        158        168        178        188
          LRRLTIRDLL AQGRTSSNAL EYVREEVFTN NADVVAEKAL KPESDITFSK QTANVKTIAH

Query 95         104        114        124        134        144        152
          AWSSHDLAAT LSGSDPMQAI GSRVAAYWAR EMQKIVFAEL AGVFSNDDMK --DNKLDISG
             +        +       I  +R+    A   + +  +         + + +       D S
Sbjct 86         198        207        217        227        237        247
          WVQASRQVMD DAPM-LQSYI NNRLMYGLAL KEEGQLLNGD GTGDNLEGLN KVATAYDTSL

Query 153        162        172        182        190        200        210
          TADGIYSAET FVDASYKLGD HESLLTAIGM HSATMA--SA VKQDLIEFVK DSQSGIRFPT
           A G   A+     A Y++ +  E   + I + +       +  +K +   ++
Sbjct 145        257        267        277        287        297        307
          NATGDTRADI IAHAIYQVTE SEFSASGIVL NPRDWHNIAL LKDNEGRYIF GGPQAFTSNI

Query 211        220        230        240        250        260        270
          YMNKRVIVDD SMPVETLEDG TKVFTSYLFG AGALGYAEGQ PEVPTETARN ALGSQDILIN
               V+      P +    G T     +    +      +    E       N  + +   ++
Sbjct 205        314        322        332        342        352        362
          MWGLPVV--- --PTKAQAAG TFTVGGFDMA SQVFDRMDAT VEVSREDRDN FVKNMLTILC

Query 271        280        290
          RKHFVLHPRG VKFTENAMAG
           +   +
Sbjct 260        372        382
          EERLALAHYR PTAIIKGTFS
```

Figure 26. Design of mutations in SPP1 G*13*P_WT

(a) Alignment of SPP1 G*13*P and HK97 gp5 sequences obtained by iterated PSI-BLAST procedure. Key residues from HK97 gp5 that were identified to be important for inter-subunit contacts are shown in yellow boxes, while the corresponding residues, that were mutated in SPP1, are shown in blue boxes. Residues that covalently link HK97 capsid are shown in red boxes. Helix α3 sequence is squared in a cyan box. Identical residues and conservative substitutions are indicated by the one-letter code or plus signs (+), respectively. (b) Location of residues in HK97, expected to be important in the SPP1 capsid (thick bond representation). Only subunit G was labelled.

Table 5. H-bonds at subunits F-A subunits interface

##	Structure 1	Dist. [Å]	Structure 2
20	F:MET 119[SD]	3.3	A:THR 143[OG1]
21	F:ILE 125[O]	3.5	A:GLU 153[N]
22	F:MET 126[O]	3.1	A:ARG 372[NH2]
23	F:THR 185[O]	2.7	A:VAL 163[N]
24	F:ALA 187[O]	3.3	A:ASP 161[N]
25	F:TRP 189[O]	2.9	A:SER 172[N]
26	F:GLY 214[O]	3.5	A:ASN 158[ND2]
27	F:ASN 291[O]	3.2	A:ASN 291[ND2]
28	F:GLU 292[O]	3.1	A:ASN 291[N]
29	F:PRO 300[O]	2.7	A:TRP 309[N]
30	F:GLN 301[O]	2.7	A:GLY 310[N]

The complete table is shown in appendix **4**. Residues of interest are presented with grey background.

Table 6. H-bonds at the F-G subunits interface

##	Structure 1	Dist. [Å]	Structure 2
3	G:ARG 338[NH2]	3.0	F:ASP 198[O]
4	G:ARG 338[NH2]	3.1	F:ASP 199[OD1]
5	G:ARG 365[NH1]	3.0	F:ASP 199[OD1]
7	G:ARG 350[N]	3.8	F:GLU 348[O]
8	G:SER 346[OG]	2.5	F:GLU 348[OE1]
9	G:ARG 347[N]	3.5	F:GLU 348[OE1]
10	G:GLU 348[N]	3.3	F:GLU 348[OE1]
11	G:ARG 350[N]	3.2	F:ARG 350[O]
12	G:ASP 173[OD2]	3.1	F:SER 104[N]
13	G:GLU 171[OE2]	3.5	F:SER 104[N]
14	G:GLU 344[OE2]	2.5	F:ARG 194[NH2]

The complete table is shown in appendix **5**. Residues of interest are presented with grey background.

Residue N158 from gp5 is located in the E-loop and it is H-bonded to residue G214 from helix α3 in the right-hand adjacent subunit (Figure 26 and Figure 27). Residue G64 from SPP1 aligns with N158 and it was substituted by another tryptophan, following the same logic that with D194W mutation.

In HK97 gp5 pairs R194-D198 and F353-V354 are located at the P-loop and helix α2 (Figure 27). R194 and D198 form H-bonds with E344 and R338 from the adjacent subunit of a neighbouring morphological unit. Mutations D100R, T104Y, A261W and L262W were introduced in G*13*P since these residues align with residues R194, D198, F353 and V354 in gp5, respectively. All mutations consist of bulky amino acids that would keep subunits from other morphological units distant. While mutation D194W is aimed to disrupt interactions at the centre of the morphological units and theoretically produce monomers, the rest of the mutations aimed to keep hexamers or pentamers in solution.

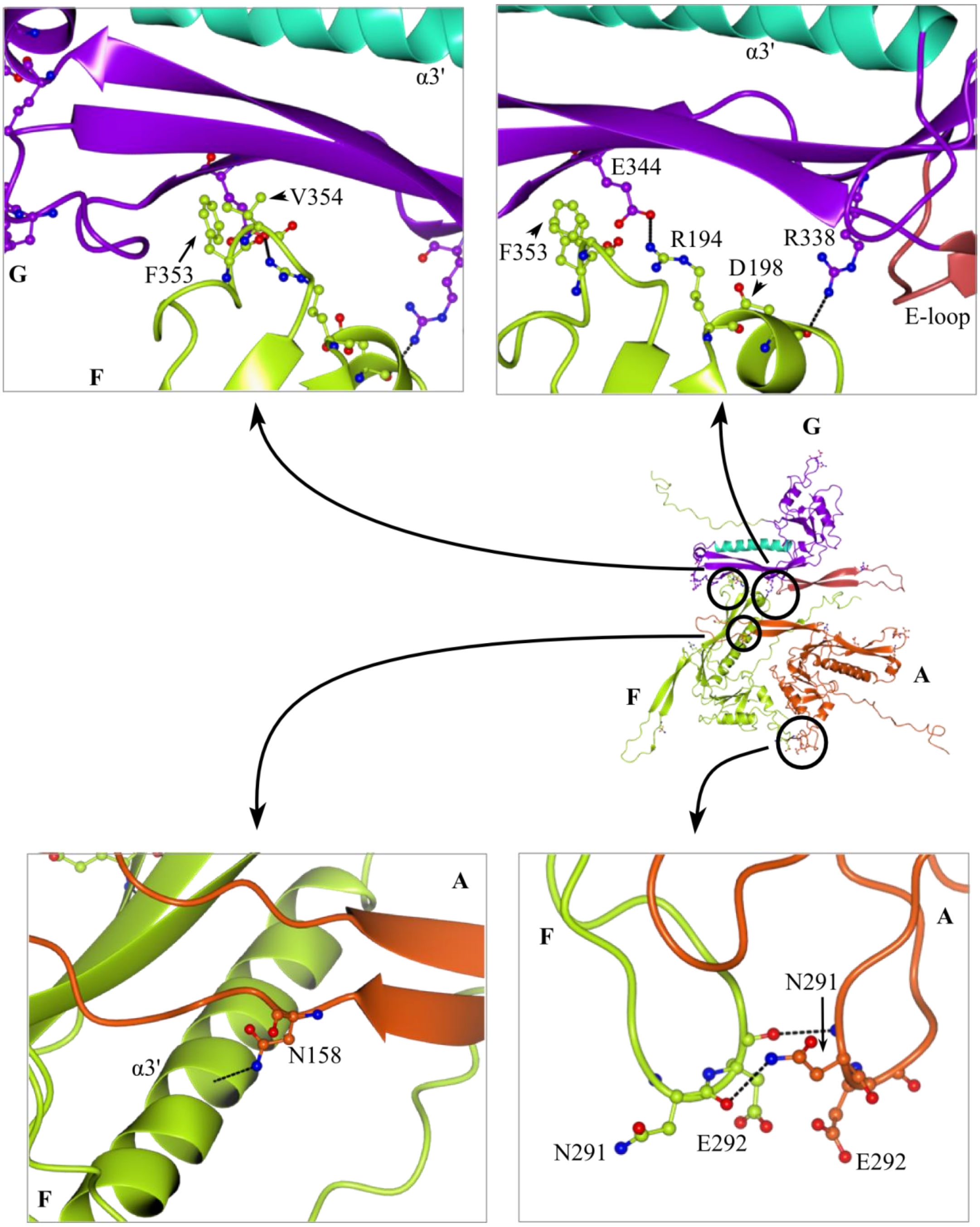


Figure 27. Scheme of H-bonding of key residues in HK97 gp5

Ball and stick representation of several residues that establish H-bonds in HK97 gp5. Counterparts of these residues in SPP1 were targeted for mutagenesis. Carbons atoms are coloured according to the chain they belong to. Oxygen and nitrogen atoms are shown in red and blue. The corresponding chain is labelled and so are some secondary structure elements (representations not to scale).

3.3.3 *Solubility test*

All mutants were cloned by site directed mutagenesis using the G*13*P_WT DNA as template. Briefly, forward and reverse primers included the desired mutation at the centre of the primer permitting the complementary regions at each end to anneal to the template whilst the modified codon created a bubble that in subsequent cycles allows the annealing of the complete primer. The recombinant DNA was then used to transform two types of bacterial strains and induce expression at two different temperatures. Samples were resuspended in several buffers to screen their solubility (Figure 28) following the described protocol. All mutants showed solubility at some extent in at least one of the expression conditions tested. The buffer or NaCl concentration did not make a significant difference as approximately the same amount of protein was observed in SDS-PAGE.

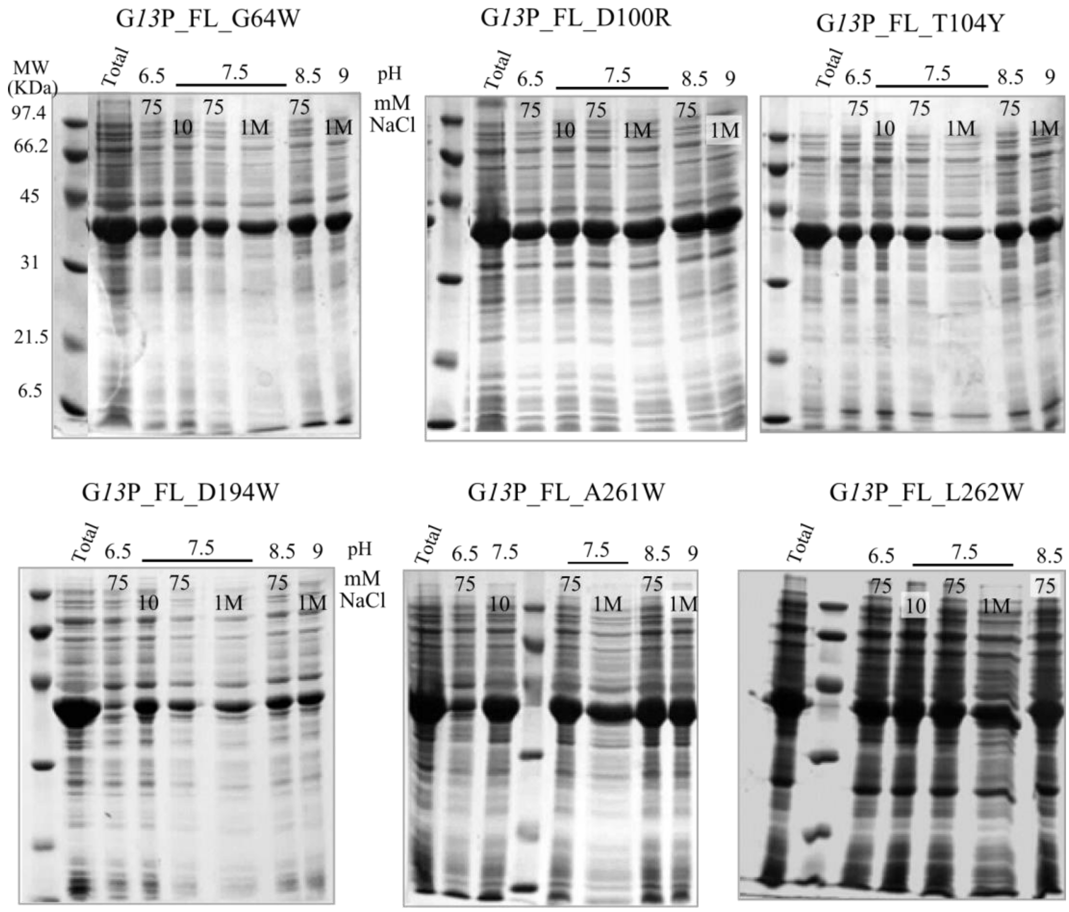


Figure 28. Solubility screen for G*13*P mutants

Only results from BL21 cells induced at 16°C are shown. The pH and NaCl concentration are indicated. The gels are not to scale with each other. For G*13*P_FL_L262W results from BL21 pLysS cells cultured at 37°C are shown as these conditions showed the highest solubility. FL refers to full length constructs.

3.3.4 Purification of single mutants

Production of mutants was scaled up to cultures of 750 mL of LB medium and NAC and SEC were performed as a way of testing which of the mutations prevented aggregation. Because of previous experiences with the truncations and other modifications that were made on G*13*P, it was decided to keep on characterising different constructs in the case a particular one did not show optimal solubility or yield. Even though in theory all mutations were expected to affect aggregation, the low percent of homology with HK97 gp5 gave a degree of unpredictability to the mutations in G*13*P.

G*13*P_FL_D100R and G*13*P_FL_D194W did not give a high yield from NAC, which could be explained by the difference in the amount of protein between the lysate and the loaded supernatant (Figure 29a). Even when the lanes with the lysate are overloaded with protein it is possible to see that the mutants were expressed in high amount, but centrifugation at high speed led to the precipitation of most of it. In the case of G*13*P_FL_D100R, the lane containing the lysate shows a broad protein band of ~40 kDa that likely contains G*13*P but with retarded migration. A dilution of the lysate included at the left side of the lysate confirms that G*13*P was in solution before centrifugation. No further work was made with these mutants.

The rest of the mutants were soluble and subsequently purified by NAC and SEC. The amount of G*13*P_FL_A261W, T104Y and L262W after SEC was too low to setup crystal trays, nevertheless, it was observed that also G*13*P_FL_A261W and T104Y aggregated but a fraction of the sample eluted where a monomer or a similarly sized impurity would elute. G*13*P_FL_L262W did not exhibit aggregation but the yield from SEC was low (Figure 29b).

G*13*P_FL_G64W was produced in high amount after SEC. Aggregation was observed, although most of the protein remained as low MW species. Fractions from this peak were pooled together and concentrated using a 10 kDa cut-off centrifugal concentrator, where the formation of proteinaceous white threads was observed when the protein concentration started to rise. Non-concentrated samples were run onto a SEC-MALLS instrument and the estimated molecular size was $1x10^6$ Da (Not shown). Such observation is attributed to aggregation. The G64W mutation, located at the E-loop, was expected to weaken interactions between the capsid proteins, but other residues for example at the P-loop or A-domain still promoted the formation of aggregates. From a protein that establishes multiple contacts using diverse surfaces, it is not surprising that a single mutation is not sufficient to stop its aggregation.

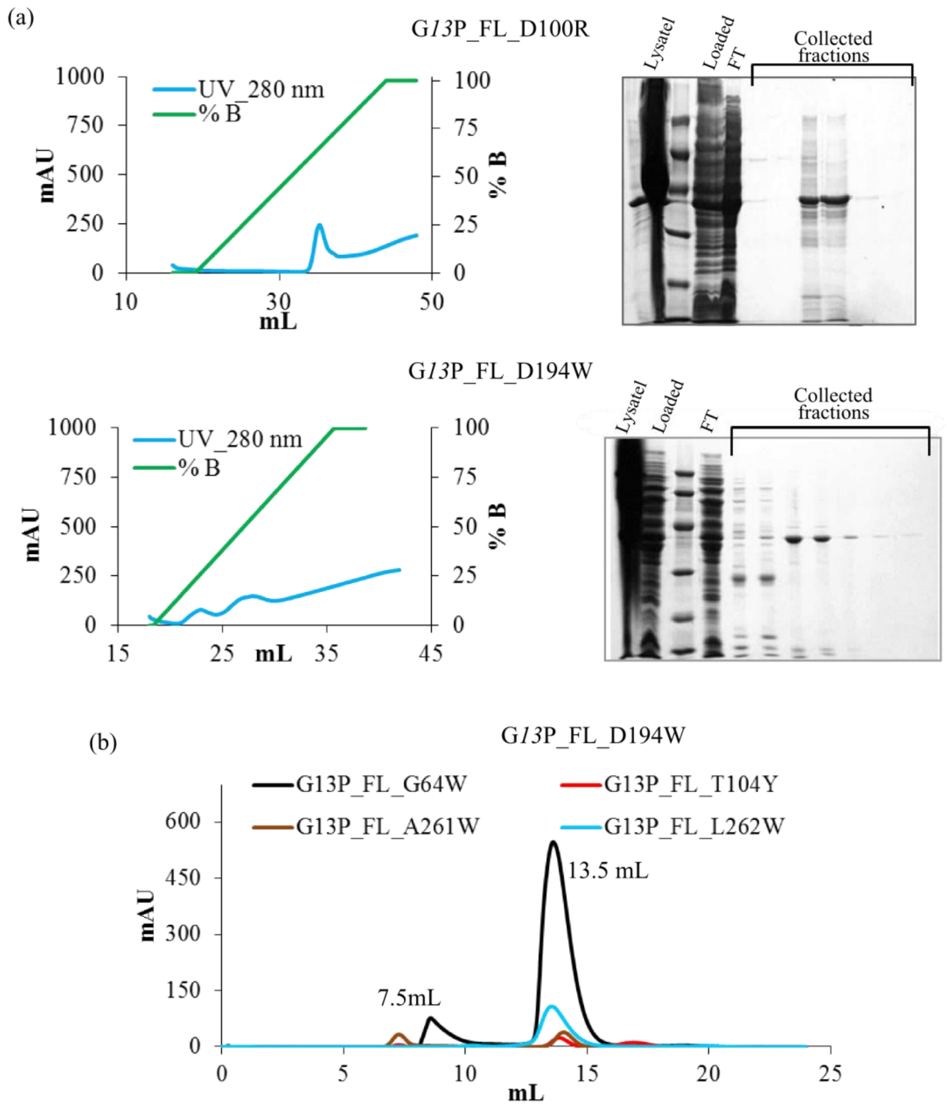


Figure 29. Purification of G*13*P single mutants

(a) NAC for G*13*P_FL_D100R and G*13*P_FL_D194W. The corresponding gels are included at the right side and every lane is labelled to show the most important samples. (b) SEC profile for G*13*P_FL_ G64W, T104Y, A261W and L262W. The elution volume of both species is indicated.

3.3.5 Tagged and untagged double mutants

As none of the single mutations were suitable for isolating stable monomer, the three most successful single-point mutations were combined in pairs to generate three double mutants to investigate whether their joint effects were sufficient to produce protein that could exist stably in solution without aggregation.

The selected double mutants were: G*13*P_FL_G64W;A261W, G*13*P_FL_T104Y;A261W and G*13*P_FL_G64W;T104Y. The second mutation was introduced by site directed mutagenesis using either single mutant DNA as template and primers for the desired second mutation. As done with the truncated and mutant constructs, bacterial cells were transformed with recombinant DNA and protein expression was induced at 16 and 37 °C (Figure 30a). G*13*P_FL_G64W;T104Y was produced in a higher amount when cultured at 37 °C in comparison with G*13*P_FL_T104Y;A261W and G*13*P_FL_G64W;A261W where the major amount of expressed protein was obtained at 16 °C.

Figure 30b shows the NAC profile of G*13*P_FL_G64W;A261W using a His-trap FF crude column. During the standard NAC it was observed that most of the POI precipitated after centrifugation at high speed (not shown). A new crude lysate was loaded onto the FF crude column however no binding to the Ni-beads was observed. Characterisation of the rest of the mutants followed and no further work was performed with this particular double mutant.

G*13*P_FL_G64W;T104Y and G*13*P_FL_T104Y;A261W were soluble and purified successfully by NAC and SEC (Figure 30c). The SEC profile shows that both double mutants eluted like homogenous species where a monomer would be expected. A reasonable amount of protein was produced out of subsequent SEC runs with the rest of the purified protein by NAC. SEC fractions were concentrated and no aggregates were visible. Crystallisation work for these constructs will be described in section 3.5.

The molecular weights of G*13*P_FL_G64W;T104Y and G*13*P_FL_T104Y;A261W, determined by SEC-MALLS were 35.0 and 34.9 kDa, respectively, both corresponding to monomers of G*13*P within experimental error of the technique.

Interestingly, the double mutant which did not include the T104Y mutation was the only construct that precipitated, suggesting that T104 is establishing contacts with adjacent subunits. It is possible that the replacement by tyrosine kept apart residues proximal to T104 that make inter-subunit contacts.

The individual effects of G64W or A261W mutations did not lead to the production of monomers, but both did when combined with the T104Y mutation. The mechanism for this disruption by T104Y mutation might be due to the relatively bigger gap that the new tyrosine creates between the area above the P-loop and the P-domain (Figure 27). Introducing such a mutation would disrupt H-bonds between R338-D198 and R194-D100 (in HK97 gp5) at the same time. Although the single T104Y mutant protein was monomeric to some extent, the presence of a second mutation is likely to further separate subunits from each other (Figure 29, individual SEC for T104Y).

Untagged versions of the double mutants that were monomers in solution were the last constructs to be explored during this project. They are included in this section to keep constructs classified by the type of alteration that was introduced. It was decided to continue working with the double mutants after the lack of successful results with the engineered protein (To be discussed in section 3.4).

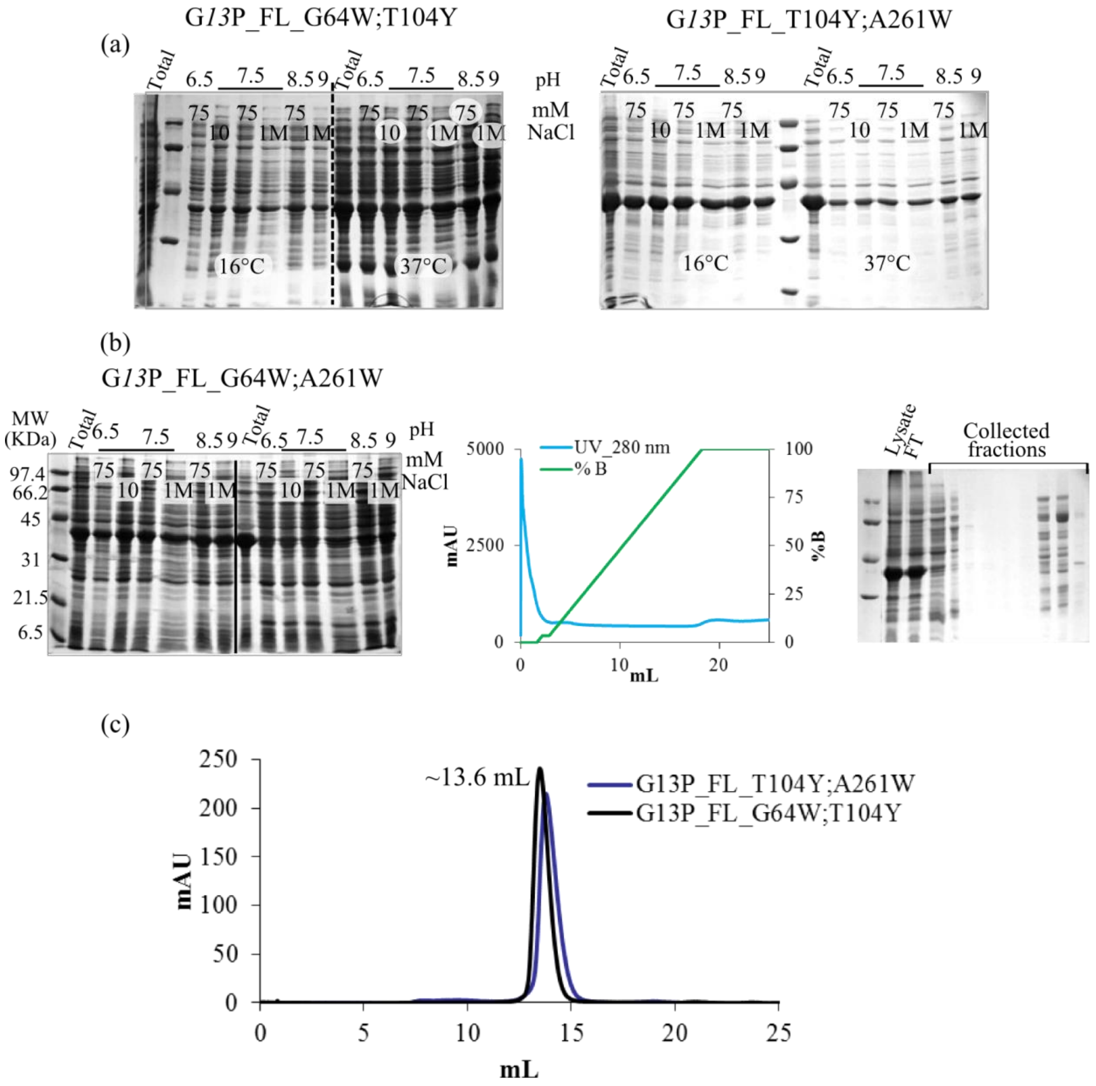


Figure 30. Production of G*13*P double mutants

(a) Solubility screens for G*13*P_FL_G64W;A261W (in B834 cells) and G*13*P_FL_T104Y;A261W (in BL21 PLysS cells).(b) Solubility screen and NAC profile (FF crude column) for G*13*P_FL_G64W;A261W. (c) SEC profile of G*13*P_FL_G64W;T104Y and G*13*P_FL_T104Y;A261W.

The cloning protocol consisted in amplifying the full length double mutants with primers that had fifteen bases from the insert and fifteen bases from pET22b expression vector, specifically, between the restriction sites NdeI and XhoI. The start methionine encoded by one of the codons of the NdeI restriction site was used as the start codon and one stop codon was introduced after G*13*P's last residue codon to avoid including the His-tag at the 3' end that is part of pET22b vector.

Only one strain of bacterial cells were transformed and protein expression was carried out at 16°C as it was known that these double mutants are generally well expressed and soluble. Both untagged-G*13*P_FL_G64W;T104Y and untagged- G*13*P_FL_T104Y;A261W (abbreviated as u-construct from here on) were soluble under the conditions used that include pHs 6, 7.5 and 9 and NaCl concentration from 10 mM to 1 M (Figure 31a).

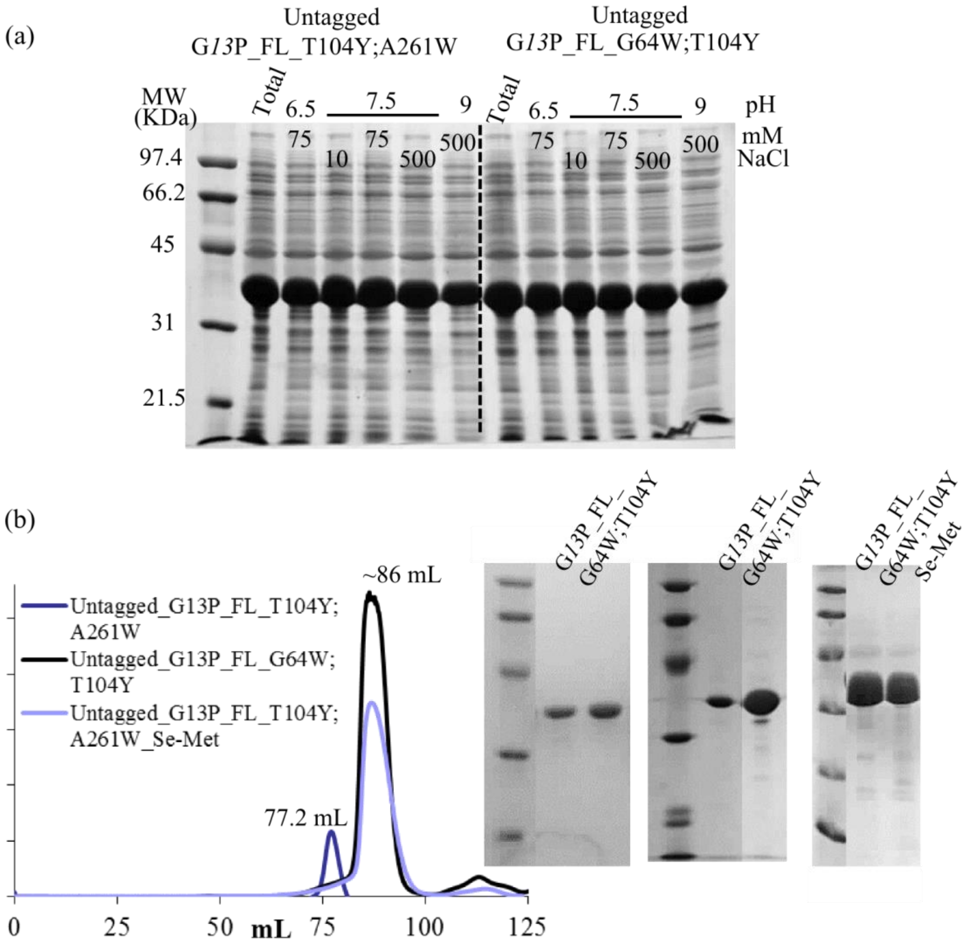


Figure 31. Production of G*13*P untagged double mutants

(a) Solubility screen for u-G*13*P_FL_G64W;T104Y and u-G*13*P_FL_T104Y;A261W (in B843 cells). (b) SEC profile in a S200 16 60 column (absorbance at 280 nm). Fractions from each run are shown in the gels at the right side.

Production was scaled up and the u-G*13*P_FL_G64W;T104Y was purified from the bacterial pellet by $(NH_4)_2SO_4$ precipitation and a variety of chromatographic techniques (not shown). Briefly, the final optimised protocol used to purify both constructs included getting a clear supernatant by spinning down the cell lysate. The supernatant was then loaded onto a FF Q column without previous $(NH_4)_2SO_4$ as it was observed that after centrifugation the supernatant was relatively pure. Protein was then eluted using a NaCl gradient. SEC followed the anion exchange chromatography.

Both constructs, plus Se-Met u- G*13*P_FL_T104Y;A261W were produced in high amount and concentrated for crystallographic studies (Sections 3.5). Constructs eluted as monomers in the

S200 16/60 column used for the large scale production (86 mL). Difference in the elution volume of the Se-Met construct might be due to flow rate, AKTA purifier used, or degree of cleanness of the column as this particular purification was performed weeks after the native constructs. The gels at the right of the graph show that the purified protein was pure enough to set up crystal trays.

One truncated version of G*13*P_FL_T104Y;A261W was cloned to test if the removal of the N-terminal portion in addition to the double mutation led to crystallisable protein. G*13*P_ΔN6_T104Y;A261W was soluble when expressed in B834 cells at 16°C. SEC analysis showed that it existed mainly as monomer in solution; however a fraction of the sample also aggregated. When the monomeric sample was run again in SEC the aggregate was generated again, suggesting that existed equilibrium between the aggregate form and the monomer (not shown). No further work was performed on this construct.

3.4 Design and purification of G*13*P-Anti-TRAP constructs

Following the example of protein-engineering that was presented in section 1.7 where a flexible ICL3 loop was replaced by T4 lysozyme to regions contributing to the movement of transmembrane α-helices in the β_2-AR, a similar approach was followed with G*13*P by introducing the Zinc-binding domain from anti-TRAP into the region in G*13*P that aligns with the P-loop from HK97 gp5 (Figure 32).

Anti-TRAP (AT) is an antagonist of TRAP, a common protein in Gram-positive bacteria involved in tryptophan metabolism regulation. The AT crystal structure was solved at 2.8 Å resolution by the Antson Laboratory. In the crystal, AT exists as a dodecamer formed by four AT trimers. Each L-shaped monomer displays two wings separated by a 100º angle (Figure 32e). One of the wings contains two pairs of small antiparallel β-strands and one zinc-binding domain that consists of four Cys residues coordinating one ion of Zn^{+2}. The second wing consists of one short α-helix (118).

AT was selected as fusion partner to G*13*P for several practical reasons. The first one concerns the presence of the coordinated Zn^{+2} atom that once integrated into the engineered protein would facilitate structure solving by anomalous scattering. Secondly, the ordered fold of the Zn-binding domain correctly integrated into the G*13*P monomer would be expected to disrupt inter-P-domain contacts as well as replacing a loop that might be flexible and be contributing to disorder in the G*13*P structure. Finally, the distance between the residues at the ends of the AT portion that was inserted in G*13*P sequence is similar to that predicted in G*13*P based on homology modelling against the HK97 gp5, thus it is more likely to yield a structure resembling the native fold.

Two segments from AT, both comprising the Zn-binding domain, were included in the fusion construct: V10-I35 and A11-V34 (Figure 32a,b). In HK95 gp5 such segment from AT would feasibly replace residues E347-M357 (Figure 32c), however in the ten aligned residues in G*13*P this would likely disrupt some secondary structure elements, predicted from homology modelling, and thus the range in G*13*P that was substituted for AT segments was A257-Q265, which was predicted to connect strands β9-β10 in G*13*P.

The distance between AT residue pairs V10-I35 and A11-V34 (which are allocated at the ends of the two AT segments) are 3.9 and 5.4 Å respectively, comparable to the distance between pair E347-M357 in HK97 gp5, which is 7.8 Å.

A model using AT and gp5 PDB entries was created with Coot by merging the desired fragments from each protein (Figure 32e), using gp5 as a model for G*13*P. In theory, all G*13*P domains would be expected to exhibit their native folds in the designed fusion constructs since the β-hairpin from the Zn-binding domain would bring the C-terminal portion of G*13*P to proximity with the rest of G*13*P to complete the P- and A-domains as is predicted in the wild type structure.

3.4.1 Design of primers

The recombinant DNA for constructs G*13*P-V10-I35AT and G*13*P-A11-V34AT (two residues shorter) was produced in two rounds of PCR using G*13*P_WT and AT as templates. Both G*13*P and AT DNA sequences were combined to sketch the engineered proteins' sequences and design primers that included both sequences (Figure 33, Table 7). The logic to design primers for both constructs was the same and hence it will be explained in detail only for G*13*P-V10-I35AT, more information about G*13*P-A11-V34AT can be found in appendices **6** and **7**.

Three pairs of primers were used to generate three DNA fragments that contained overlapping G*13*P or AT sequences for construction of the complete construct coding sequence (Figure 34). For instance, single DNA chains of fragment A generated during the first PCR had a short segment of AT at the 3' end in addition to the large segment of G*13*P at the 5' end encoding the N terminal portion. Chains from fragment B consisted of the complete introduced AT sequence plus additional G*13*P sequences at each end. Fragment B included the 3' segment of G*13*P plus a short sequence of AT at the 5' end. More information about the sequence of the primers and the templates they amplify is given in Table 7.

Separate PCR rounds produced fragments A (789 bp), B (114 bp) and C (198 bp) that served as a template in the second PCR stage, where only primers to amplify G*13*P_WT, identical in the G*13*P-AT constructs were included. This PCR product, containing copies of full length G*13*P-V10-I35AT construct, was digested with NdeI and XhoI and then ligated into pET28 vector for subsequent sequencing and expression.

(a) Anti-TRAP amino acid sequence

```
        10         20         30         40         50
MVIATDDLEV ACPKCERAGE IEGTPCPACS GKGVILTAQG YTLLDFIQKH LNK
```

(b) Anti-TRAP DNA sequence

```
        10         20         30         40         50         60
ATGGTTATCG CTACCGACGA CCTGGAAGTT GCTTGCCCGA AATGCGAACG TGCTGGTGAA
        70         80         90        100        110        120
ATCGAAGGTA CCCCGTGCCC GGCTTGCTCC GGTAAAGGTG TTATCCTGAC CGCTCAGGGT
       130        140        150        160
TACACCCTGC TGGACTTCAT CCAGAAACAC CTGTGGAAAT AA
```

(c) Partial HK97 gp5 amino acid sequence

```
       310        320        330        340        350        360
QAFTSNIMWG LPVVPTKAQA AGTFTVGGFD MASQVWDRMD ATVEVSREDR DNFVKNMLTI
       370        380
LCEERLALAH YRPTAIIKGT FSSGS
```

(d) Partial SPP1 G*13*P amino acid sequence

```
       190        200        210        220        230        240
GMHSATMASA VKQDLIEFVK DSQSGIRFPT YMNKRVIVDD SMPVETLEDG TKVFTSYLFG
       250        260        270        280        290        300
AGALGYAEGQ PEVPTETARN ALGSQDILIN RKHFVLHPRG VKFTENAMAG TTPTDEELAN
       310        320
GANWQRVYDP KKIRIVQFKH RLQA
```

(e)

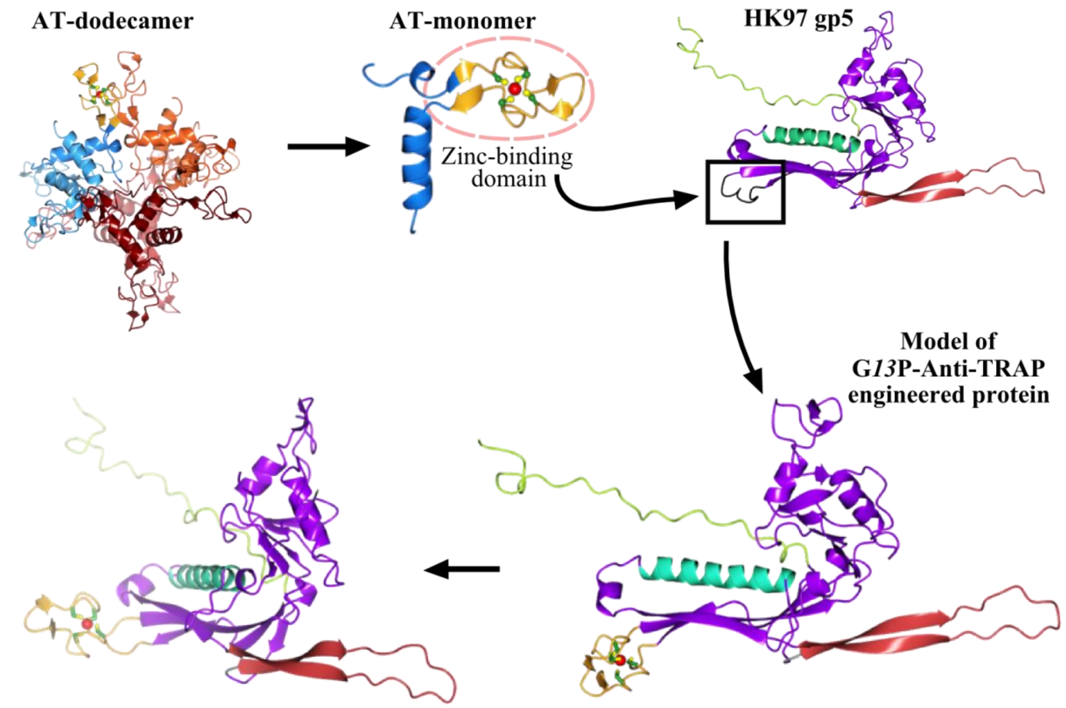


Figure 32. Engineered G*13*P protein containing the Zn-binding domain from Anti-TRAP

(a), (b) Anti-TRAP amino acid and DNA sequences. Sequences highlighted in yellow and brown correspond to the portions of AT that constructs G*13*P-V10-I35AT and G*13*P-A11-V34AT incorporate, respectively. (c), (d) Partial amino acid sequence of gp5 and G13P, respectively. Substituted sequences are highlighted in black/grey background. (e) Model for G*13*P-V10-I35AT. The Zn-binding domain is coloured in yellow and gp5 is coloured following the scheme used in Figure 6. PDB entries: AT: 2BX9, gp5: 1OHG.

(a)

```
   1  ATGGCATACA CAAAAATTTC AGATGTTATC GTACCGGAGT TATTTAACCC GTACGTCATT
  61  AACACAACAA CACAACTTTC TGCCTTCTTC CAGTCAGGAA TTGCGGCAAC AGATGACGAA
 121  TTGAATGCAC TTGCAAAAAA AGCGGGCGGC GGTAGCACTT TAAACATGCC GTACTGGAAT
 181  GACCTAGACG GAGATTCCCA AGTGTTGAAC GACACTGACG ACCTTGTACC GCAAAAAATC
 241  AACGCTGGAC AAGATAAAGC TGTCCTTATC CTTCGCGGTA ACGCTTGGAG TTCTCACGAT
 301  TTAGCGGCAA CACTTTCCGG TTCTGACCCA ATGCAGGCTA TCGGCTCCCG TGTAGCGGCA
 361  TACTGGGCGC GCGAAATGCA AAAGATTGTT TTCGCTGAAC TTGCAGGTGT GTTCTCTAAC
 421  GATGATATGA AAGACAACAA ACTCGATATC TCTGGAACGG CTGACGGTAT TTATTCAGCG
 481  GAAACTTTCG TTGATGCATC TTACAAGCTT GGAGATCATG AAAGCTTACT TACAGCTATC
 541  GGTATGCATT CTGCTACGAT GGCAAGCGCA GTTAAACAAG ACTTGATTGA GTTTGTCAAA
 601  GATTCCCAAA GTGGTATCCG  TTTCCCGACA TACATGAATA AGCGTGTAAT CGTAGATGAT
 661  TCTATGCCAG TAGAAACGCT TGAAGATGGA ACTAAGGTAT TCACATCTTA CTTGTTCGGA
 721  GCTGGTGCTC TAGGATACGC AGAAGGACAA CCGGAAGTAC CAACAGAAAC AGTTGCTTGC
 781  CCGAAATGCG AACGTGCTGG  TGAAATCGAA GGTACCCCGT GCCCGGCTTG CTCCGGTAAA
 841  GGTGTTATCG ATATTCTTAT CAACCGTAAA CACTTTGTTT TACACCCGCG CGGCGTAAAA
 901  TTCACAGAAA ACGCTATGGC GGGAACAACG CCTACGGACG AAGAACTTGC TAACGGTGCG
 961  AACTGGCAAC GCGTGTACGA CCCTAAGAAA ATCCGTATCG TTCAATTCAA ACACAGACTA
1021  CAAGCATAA
```

(b)

```
         10         20         30         40         50         60
MAYTKISDVI VPELFNPYVI NTTTQLSAFF QSGIAATDDE LNALAKKAGG GSTLNMPYWN
         70         80         90        100        110        120
DLDGDSQVLN DTDDLVPQKI NAGQDKAVLI LRGNAWSSHD LAATLSGSDP MQAIGSRVAA
        130        140        150        160        170        180
YWAREMQKIV FAELAGVFSN DDMKDNKLDI SGTADGIYSA ETFVDASYKL GDHESLLTAI
        190        200        210        220        230        240
GMHSATMASA VKQDLIEFVK DSQSGIRFPT YMNKRVIVDD SMPVETLEDG TKVFTSYLFG
        250        260        270        280        290        300
AGALGYAEGQ PEVPTETVAC PKCERAGEIE GTPCPACSGK GVIDILINRK HFVLHPRGVK
        310        320        330        340
FTENAMAGTT PTDEELANGA NWQRVYDPKK IRIVQFKHRL QA
```

Figure 33. Engineered protein G*13*P-V10-I35AT

(a), (b) Nucleotide and amino acid sequence of G*13*P-V10-I35AT. The region corresponding to AT is highlighted in yellow. Areas in red correspond to the annealing sites for the primers.

Table 7. Primers for G*13*P-V10-I35AT

Template	Fragment Amplified	Designed Primers
G*13*P_FL	A: 1-789 (G*13*P)	1_F 5’ GG AAT TCT CAT ATG GCA TAC ACA AAA ATT TCA G 3’ II-R 5` GCA TTT CGG GCA AGC AAC TGT TTC TGT TGG TAC TTC 3`
Anti-TRAP	B: 754-867 (G*13*P -AT- G*13*P)	III-F 5` GAA GTA CCA ACA GAA ACA GTT GCT TGC CCG AAA TGC 3` IV-R 5` ACG GTT GAT AAG AAT ATC GAT AAC ACC TTT ACC GGA 3`
G*13*P_FL	C: 832-1029 (G*13*P)	V-F 5` TCC GGT AAA GGT GTT ATC GAT ATT CTT ATC AAC CGT 3` 1_R 5’ CCA GAA CTC GAG TTA TGC TTG TAG TCT GTG TT 3’

Areas highlighted in yellow correspond to the AT sequence. NdeI and XhoI restriction sites are highlighted in cyan and salmon, respectively.

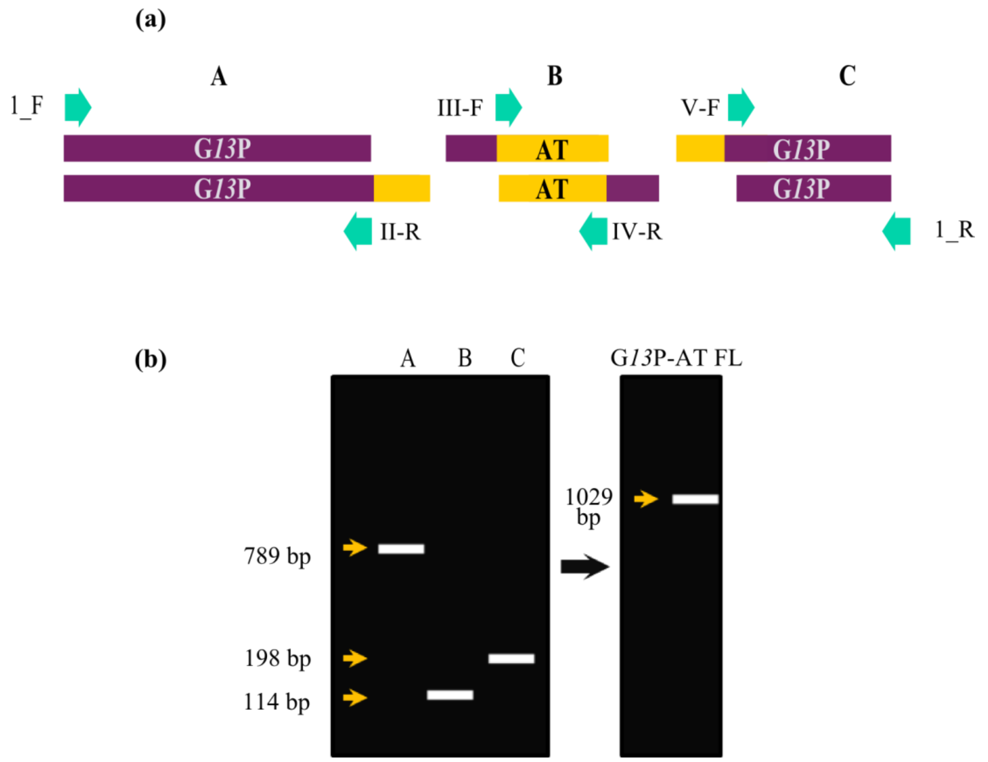


Figure 34. Generation of G*13*P-V10-I35AT by PCR

(a) Fragments A-C. Scheme of possible fragments generated on the first PCR when using the designed primers (not to scale). It is important to note that after several cycles all fragments would have blunt ends and both sequences would be incorporated on them. We are interested in any fragment that could have complementarity with other fragments. The arrows indicate the primers that generated that chain, not the one to which they anneal to. (b) PCR scheme of the predicted DNA fragments for the first (left) and second PCR (right). Length is indicated at the right of each gel.

3.4.2 Solubility test

Both engineered proteins were successfully cloned as described above (Figure 35a). Sequencing results showed that the AT fragment was properly incorporated into G*13*P and so were the resulting full length G*13*P-AT constructs into pET28 (appendix **8** and **9**). Sequencing was performed by synthesis, with T7 promoter and terminator primers by GATC-BIOTECH (Germany)

BL21 pLysS and B834 cells were transformed with the recombinant DNA to be further used to induce protein at 16 and 37°C. Proteins were expressed in high amounts only at 16°C in B834 cells (Figure 35b). In spite of being soluble under most of the conditions used in the experiment, two bands on SDS-PAGE of slightly different size were present where the FL constructs were expected (G*13*P-V10-I35AT: 39.19 KDa, G*13*P-A11-V34AT: 38.98 kDa). At this early stage, large scale production of protein was started.

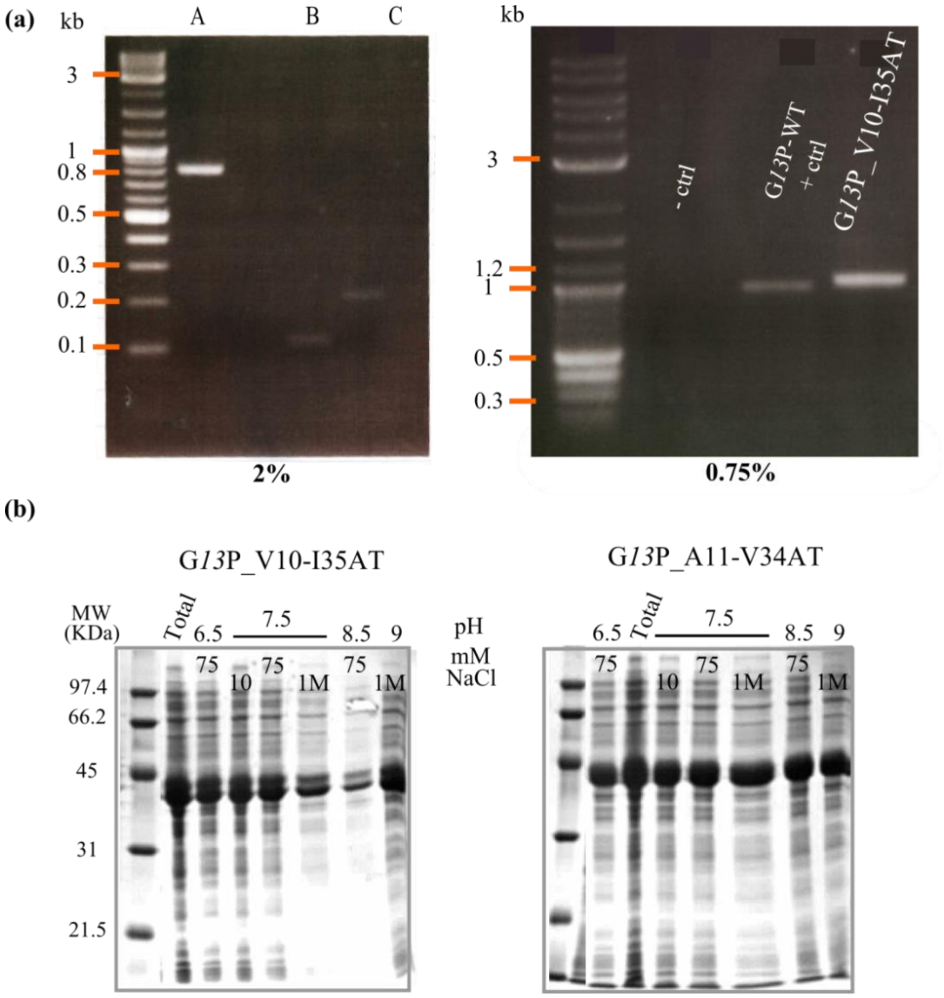


Figure 35. Production of G*13*P-V10-I35AT and G*13*P-A11-V34AT

(a) PCR reactions. The left gel (2% agarose) shows fragments A-C from the first PCR. The right gel presents the 2nd PCR, in order, - control (no template), + ctrl (G*13*P_WT), G*13*P-V10-I35AT. The 2-log ladder is at the first lane in every gel. (b) Solubility screen in B834 cells.

3.4.3 Purification of G13P-AT constructs

NAC and SEC followed large scale production in B834 cells at 16°C. Taking the aforementioned double mutants of G*13*P as a reference point it is possible to say that both G*13*P-AT constructs exist predominantly as monomers or at least low-molecular weight species in solution since they elute at ~13.5 mL without the formation of significantly bigger species (Figure 36). Even though only one peak is observed for the monomer in each run, two protein bands are still visible in the SDS-PAGE in fractions taken from the SEC eluent.

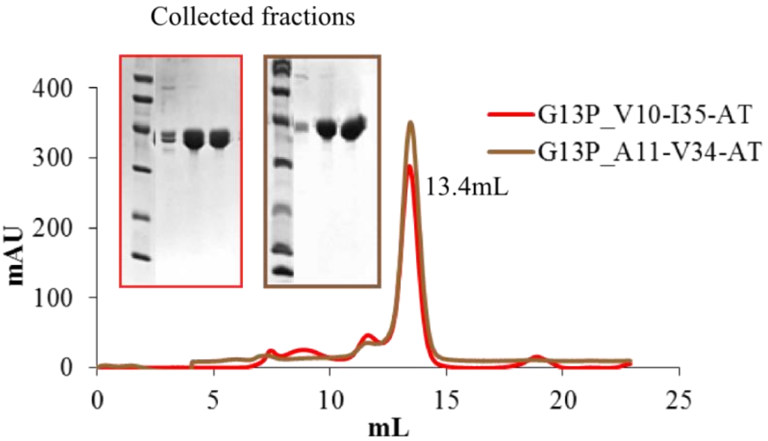


Figure 36. SEC of G*13*P-V10-I35AT and G*13*P-A11-V34AT

To determine where the possible degradation giving rise to these two distinct bands was taking place, several approaches including Ni-affinity pull-down assays, thrombin-digestion, and His-tag stain were carried out (Figure 37). The two constructs were cloned into pET28a that include the His-tag at the N-terminus. If degradation was taking place at the C-terminus, all of the sample would still bind to the Ni-beads. On the contrary, with the absence of His-tag at the N-terminus, the sample would not bind to the Ni-beads. It was confirmed that the N-terminus is intact as both constructs were shown to be present in the samples containing washed beads when run on SDS-PAGE, and in the eluted sample (as it was initially observed in NAC) (Figure 37a).

When digested with thrombin, G*13*P-V10-I35AT showed a discreet shift in size, the same shift observed when the His-tag is removed from the N terminus in other samples (Figure 37b). The His-tag stain that selectively dyes His-tagged protein also confirmed that the His-tag is present in the two constructs. This suggested that degradation was occurring at the C-terminus, as evident by the presence of the two species both possessing His-tags (Figure 37c).

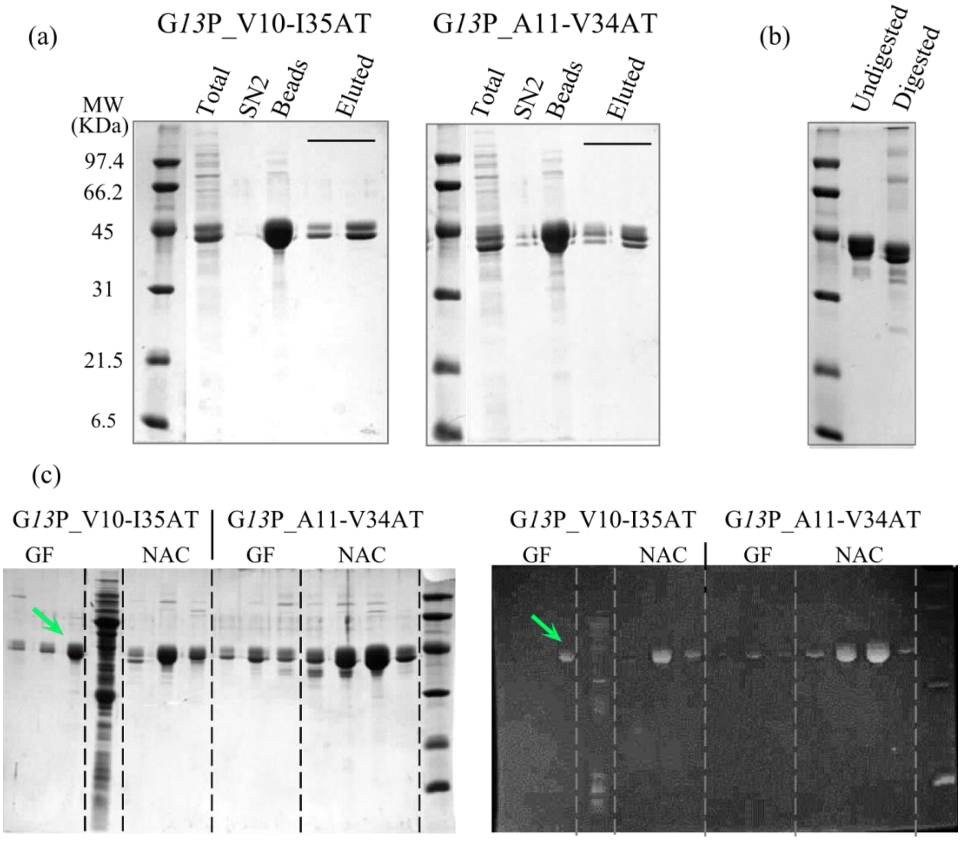


Figure 37. Detecting the His-tag in G*13*P-AT constructs.

(a) Pull-down assays. Total fraction, supernatants from the second wash (SN2), concentrated beads and two dilutions of the eluted sample are shown for every construct. Samples expressed in BL21 Star^{TM} cells. (b) G*13*P-V10-I35AT before and after thrombin-digestion. (c) His-Tag stain of samples from NAC and SEC from both constructs. The green arrow points towards the POI.

Attempts to stop degradation consisted of expressing protein in BL21 Star^{TM} pLysS cells. BL21 Star^{TM} pLysS cells carry a mutated *rne* gene that results in a phenotype which does not degrade mRNA and hence is expected to continue translating mRNAs without interruptions until the product is finished. This would rule out premature termination as the reason for two species with two different sizes appearing in samples. The protein was expressed as routine at 16°C to be used in solubility screens. No change from the behaviour in BL21 pLysS or B834 cells was observed. The constructs were expressed in high amount and were soluble but two bands were still visible on SDS-PAGE. It is clear from the SDS-PAGE that the truncation happens early on in the purification, possibly in the cell or during cell lysis.

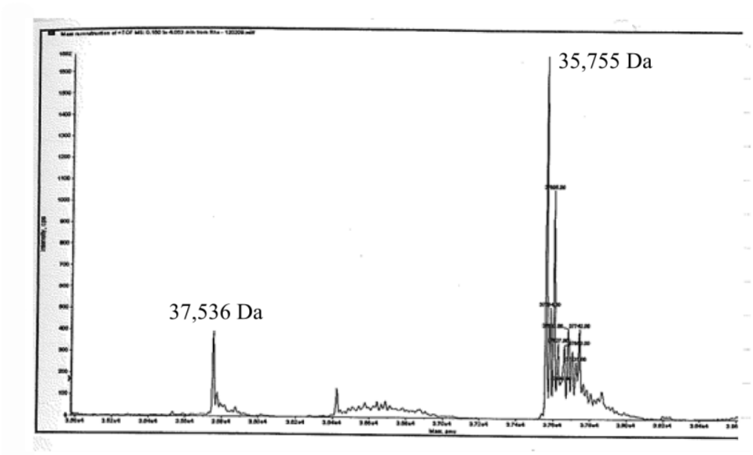
Versions of both constructs with the His-tag at the C-terminus were subsequently cloned into the pET22b vector. Their analysis demonstrated that degradation still occurs (Not shown). Concentrated samples of C-tagged G*13*P-V10-I35AT and G*13*P-A11-V34AT were used to set up INDEX and PACT crystal screens.

SEC samples from previous runs to purify G*13*P-V10-I35AT and G*13*P-A11-V34AT were concentrated to be used in crystallographic studies as the spontaneous degradation might signify the removal of naturally unstable portions (See section 3.4.4).

The determined molecular weight of G*13*P-V10-I35AT and G*13*P-A11-V34AT determined by SEC-MALLS were 30.8 and 29.78 kDa respectively (Not shown). Both are lower than the expected theoretical molecular weight presumably due to degradation. Electrospray ionisation (EI) was performed as a more accurate technique to measure molecular weight.

Because G*13*P-A11-V34AT is essentially identical to G*13*P-V10-I35AT, only the second one was used for EI. Two species of 35,755 and 37,563 Da were detected, representing the two species observed by SDS-PAGE. These molecular weights can be explained if 29 or 14 residues were removed from the C-terminus, respectively (Figure 38), indicating that both species observed by SDS-PAGE were potentially missing C-terminal segments.

(a)



(b)

```
        190        200        210        220        230        240
GMHSATMASA VKQDLIEFVK DSQSGIRFPT YMNKRVIVDD SMPVETLEDG TKVFTSYLFG

        250        260        270        280        290        300
AGALGYAEGQ PEVPTETVAC PKCERAGEIE GTPCPACSGK GVIDILINRK HFVLHPRGVK

        310         320         330        340
FTENAMAGTT PTD EELANGA NWQRVYDPK K IRIVQFKHRL QA

               35,755 Da         37,536 Da
```

Figure 38. Truncations at G*13*P-V10-I35AT

(a) Electrospray ionisation for G*13*P-V10-I35AT. The molecular weight for the two main species is indicated (b) partial G*13*P-V10-I35AT amino acid sequence. The missing segments and resulting constructs are indicated by purple arrows. Sequence highlighted in yellow corresponds to AT. Sequences in red correspond to the annealing points for the primers to produces fragments ABC.

The residues E314 and K329 were identified as the point of the truncations by relating the missing mass with the amount of residues that would account for that loss. In an attempt to produce more stable constructs, truncated versions were designed based on this information. Two reverse primers were designed to produce the truncations in G*13*P-V10-I35AT and G*13*P-A11-V34AT that signifies four new constructs: G*13*P_V10-I35AT_ΔC330, G*13*P_V10-I35AT_ΔC314, G*13*P_A11-V34AT_ΔC328 and G*13*P_A11-V34AT_ΔC312.

3.4.4 Production of truncated G13P-AT constructs and on-column refolding

The four new truncations were successfully cloned into a modified version of vector pET28a using the full length G*13*P-AT DNAs as templates. B834 and BL21 PLysS cells were transformed to express the protein at 16 and 37 °C. Even though constructs were produced in high amount, none of them were soluble under the tested conditions (See Figure 39a for the solubility test of G*13*P_V10-I35AT_ΔC330). Nevertheless, constructs with these truncations produced only one visible band in SDS-PAGE, demonstrating that no degradation exists when expressed in the cell and thus validating this approach. The previous experience with G*13*P_ΔC, for which a thorough solubility screen was carried out, suggested that the C-terminus of G*13*P is essential for solubility and therefore it was likely that further efforts with the truncated G13P-AT truncated constructs would make no improvements to solubility despite difference in amino acid sequence.

Often recombinant proteins that are overproduced in *E. coli* cluster in insoluble aggregates known as inclusion bodies (93). It was decided to implement an alternative, on-column refolding, to take advantage of the availability of the four cloned constructs for G*13*P-AT which exhibited insolubility, and to continue the extensive work made on the SPP1 capsid protein.

The on-column protein refolding protocol consisted of the denaturation of the inclusion bodies and then refolding the protein whilst still bound to the Ni-beads in the column (93).

G*13*P_V10-I35AT_ΔC330 was used as a pilot test to perform this technique. Briefly, the bacterial pellet of B834 cells induced at 16°C was resuspended in cold denaturing buffer (6M guanidine hydrochloride) to be manually loaded to the His-trap column. The column was washed with denaturing buffer plus 20 mM imidazole to remove all non-specifically bound proteins. Next, 0.1% Triton X-100 detergent buffer was flowed through the column to prevent aggregation. The remaining detergent was removed with 5mM β-cyclodextrin and the attached protein was eluted with an imidazole gradient (Figure 39b).

G*13*P_V10-I35AT_ΔC330 eluted in two peaks that were kept separately to perform SEC with each one. A small volume from peak 2 was centrifuged and then loaded on SDS-PAGE. It was observed that most of the protein remains in solution after centrifugation in comparison with a non-centrifuged sample, indicating that the on-column refolding was successful in producing soluble protein (not shown).

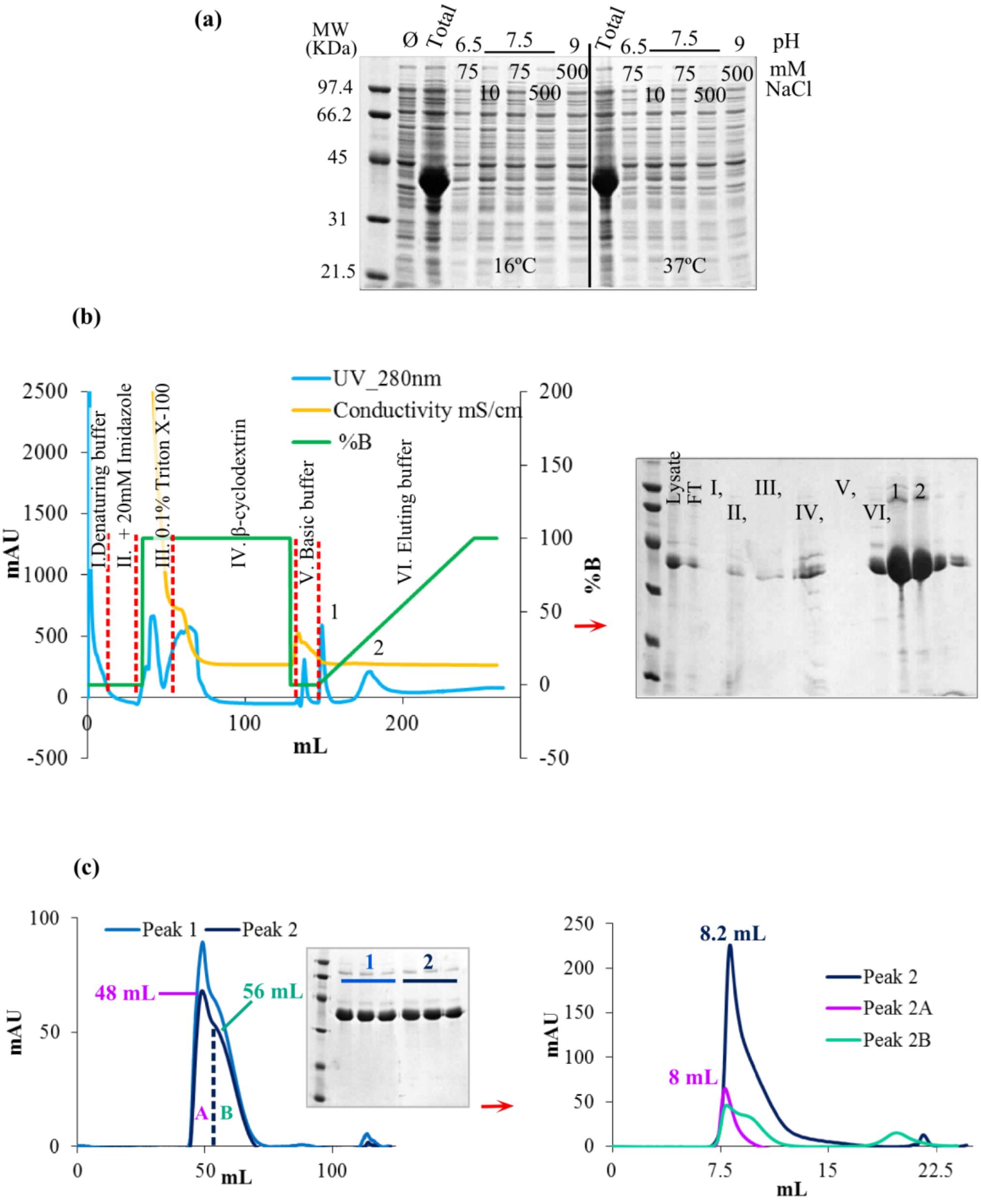


Figure 39. Production of G*13*P_V10-I35AT_ΔC330

(a) Solubility screen in B834 cells at 16 and 37°C. (b) On-column refolding. The solutions used at each stage are indicated. Samples in the gel are numbered according to the solutions flowed over. (c) SEC profile of samples from refolding. Elution volumes of P2A and P2B are shown. The chromatograms at the left and right correspond to runs in S200 10 60 and S200 10 30 columns, respectively.

The SEC profile for peaks 1 and 2 showed that the protein present in both of them eluted close to the void volume, suggesting that G*13*P_V10-I35AT_ΔC330 forms two types of aggregates. Both chromatograms consisted of one main peak and a second one, visible as a shoulder on the major peak. Therefore, these two different species might have very similar molecular weights or hydrodynamic radii. Peaks observed in the SEC for peak 2 from NAC are classified as P2A (elution volume 49 mL) and P2B (elution volume 56 mL) from here on. Both of them have elution volumes considerably higher than that observed for the full length G*13*P_V10-I35AT in the S200 16 60 column (~87 mL), confirming that they are not monomers.

Samples from P2A and P2B were loaded onto a SEC S200 10 30 column to investigate their individual behaviour (Figure 39c). One sample of peak 2 from NAC was included as control. The three samples eluted close to the void volume, at ~8.1mL, in comparison with the full length G*13*P_V10-I35AT that in S200 10 30 columns elutes at 13.4 mL, as monomer. Sample 2A exists as single species while sample 2B is present as two different species.

3.5 Crystallisation of G*13*P

3.5.1 Crystallisation of native G13P

Protein constructs that behaved like stable homogeneous monomers and those that could be concentrated to >10 mg mL^{-1} were used for crystallisation. Crystallisation experiments were performed by sitting drop vapour diffusion using commercial screens with 96 conditions in each. The protein sample was kept in the original SEC buffer or buffer-exchanged to new conditions. When a condition was observed to yield crystals, hanging drop vapour diffusion optimisation experiments were set up manually in 24-well plates, by varying the concentration of the main precipitant in small steps by probing the influence of various additives.

Table 8 lists the crystallisation experiments set up for eight of the constructs produced during this project. Single mutant G*13*P_FL_G64W, G13P-AT engineered proteins or G*13*P_FL_G64W; T104Y did not yield any crystals, even though the latter two were monomers.

G*13*P_FL_T104Y; A261W was the only double mutant that initially formed crystalline material (Figure 40). Clusters of small needles were observed in condition C11 from PACT screen. Condition C11 contained 0.1 M Hepes pH 7, 0.2 M $CaCl_2$, 20 % (w/v) PEG 6K. Manual optimisation experiments did not yield larger crystals (Figure 40a).

Low-quality crystals were also grown in the INDEX screen, condition H5: 0.1 M succinic acid pH 7, 20 % (w/v) PEG 3.35 K (Figure 40b). These crystals represented triangular clusters of smaller single crystals, however, their borders were not well defined. Subsequent optimisation produced similar crystals, which quickly dissolved when fished by a nylon loop for subsequent freezing and X-ray diffraction testing.

One example of a crystal of G*13*P_FL_T104Y; A261W is shown in Figure 40c. Small triangular and rhomboidal crystals were grown in condition D10 of the Morpheus screen: 0.1 M MES pH 6.5, 12% (w/v) PEG 20K. Although the crystal edges were more well-defined than those described previously, they were too small to test and instead a seed stock was made out of them for optimisation experiments by crystal seeding.

After working with the G*13*P-AT constructs, untagged versions of the monomeric double mutants were cloned, purified and crystallised. Approximately three weeks later several screens were set up for u-G*13*P_FL_T104Y; A261W no crystals were observed and the trays were transferred to 10°C. Three days later octahedral crystals were grown in a drop where only microcrystalline precipitate was observed before the transfer to the lower temperature. The microcrystalline precipitate formed at higher temperature, probably served as the nucleation points for growing larger crystals (Figure 40d,e). These crystals grew from G12 PACT condition: 0.1 M bis-tris propane pH 7.5, 0.2 M Sodium malonate, 20% (w/v) PEG 3350, with 10 mg mL^{-1} protein concentration. Three of these crystals were fished at 4°C, cryo-cooled and tested using the in-house diffractometer. No cryo-protectant was added as original conditions served well to prevent ice formation. Crystals exhibiting diffraction using the in-house MSC Micromax 007HF X-ray generator (Rigaku) with rotating anode and MAR345 detector (Mar Research) were sent to the I04 beamline at the Diamond Light Source, UK, for collection of a high resolution data set for structure determination.

Further work consisted of optimising condition PACT G12, but this did not produce any crystals. However, long rectangular crystals grew in 0.1 M PCB (propionic acid, cacodylate, bis-tris propane), 20 % (w/v) PEG 1.5 K at room temperature with 10 mg mL^{-1} protein concentration (Figure 40f). This condition was derived from C5 PACT screen: 0.1 M PCB, 20 % (w/v) PEG 1.5 K. When fishing was attempted, crystals dissolved upon contact with the loop and exposure to air.

A Se-Met version of u-G*13*P_FL_T104Y; A261W was purified to set up crystal trays in parallel with the native protein. Additives like Zn^{+2}, crystals seeds and TCEP were included.

3.5.2 Crystallisation of Se-Methionine labelled G13P

Se-Met labelled u-G*13*P_FL_T104Y; A261W was buffer-exchanged to 10 mM Hepes pH 7.5, 100 mM $(NH_4)_2SO_4$ and concentrated to 10 mg mL^{-1} to setup sitting drop vapour diffusion experiments at RT using the crystal screen PACT. After seven weeks of incubation, small crystals appeared in conditions D12 (0.1 M Tris pH 8, 10 mM $ZnCl_2$, 20 % PEG 6K) and E9 (0.2 M Na/K Tartrate, 20 % PEG 3.35 K) (Figure 40g,h). Since crystals from both conditions were too small to be tested, they were mixed together in 100 µL of solution containing 0.1 M Tris pH 8, 10 mM $ZnCl_2$, 20 % PEG 6K, to create a seed stock. This seed stock was used to set up the INDEX and PACT screens for Se-Met u-G*13*P_FL_T104Y; A261W diluted to 11 mg mL^{-1} protein in 10 mM Tris pH 7.5, 200 mM NaCl at RT. Crystals grew after seven weeks of incubation (Figure 40i,j) in the INDEX screen in conditions A3 (0.1 M Bis-Tris pH 5.5, 2.0 M $(NH_4)_2SO_4$) and A5 (0.1 M Hepes pH 7.5, 2.0 M $(NH_4)_2SO_4$). The cryo-cooled crystals on Figure 40i,j, diffracted to ~ 5.0 and 8.2 Å, respectively, when tested using in-house equipment.

Both crystals were sent to the I04 beamline at the Diamond Light Source, UK, for collection of a high resolution data set for structure solution.

Table 8. Cristallisation of G*13*P

Protein	Concentration mg mL^{-1}	Buffer	Crystal screen	Observations
G*13*P_FL_G64W	3 / 5	25mM Tris pH 7.5 10mM NaCl	INDEX, PACT	No crystals
G*13*P_FL_T104Y; A261W	5.9	25mM Tris pH 7.5 10mM NaCl	INDEX, PACT	Small needles in PACT
	7.7		INDEX, PACT	
	5.8	25mM Tris pH 8, 100mM NaCl	INDEX, PACT	
	6.4	25mM Tris pH 7.5 10mM NaCl	INDEX, PACT	
	10.7		CSS 1&2, MPD, Hampton	
	11.7		INDEX, PACT	Small needles in PACT
	10 / 7		$(NH_4)_2SO_4$	
	7.5		24-Well tray	Micro and macro seeding
	9.7	10mM Tris pH 7.5 50mM NaCl	CSS 1&2*	*Bis-Tris pH 5.5 and 6.5 used
	9.7		INDEX	Seeding in 96-well plate
	7	25mM Tris pH 7.5 500mM NaCl	INDEX	His-Tag removed
	10.9		INDEX, PACT	
	15	10mM Tris pH 7.5, 150mM NaCl,	MPD, CSS 1&2*, Hampton	
	15	150mM NaCl, 10mM Tris pH 7.5	CSS 1&2*	*Tris pH 6.5 and 8.5 used.
G*13*P_FL_G64W; T104Y	Similar experimental setting as for G*13*P_FL_T104Y; A261W, no crystals were grown.			
G*13*P_V10-I35AT	8.95	10mM NaCl, 25mM Tris pH 7.5	INDEX, PACT	
	30	50mM NaCl, 10mM Tris pH 7.5	INDEX	
	13.5	150mM NaCl, 10mM Tris pH 7.5	MPD, CSS 1&2*, Hampton	Tris pH 6.5 and 8.5 used
A) G*13*P_V10-I35AT-C'_tag B) G*13*P_A11-V34AT-C'_tag	25 12.5	25mM Tris pH 7.5 500mM NaCl	INDEX, PACT	No crystals grown
u-G*13*P_FL_T104Y; A261W	10 15	10 mM Tris pH 7.5 200 mM NaCl	INDEX, PACT	Optimisation of PACT A4, A5, A6. C4, C5, C6, F8, G12 at RT and 10°C
	15		MPD, Hampton, CSS 1&2, Morpheus	
	10		PACT	Set up in parallel at RT and 10°C
	11		INDEX, PACT	Plus: Zn^{+2}, seeds, NA
u-G*13*P_FL_G64W; T104Y	15	10 mM Tris pH 7.5 200 mM NaCl	INDEX, PACT, MPD, Hampton, CSS ½, Morpheus	Optimisation of PACT A4, A5, A6. C4, C5, C6, F8, G12 at RT and 10°C
Se-Met u-G*13*P_FL_T104Y; A261W	10	10 mM Tris pH 7.5 200 mM NaCl	$(NH_4)_2SO_4$	Optimisation at Optimisation PACT G12, F8 at RT and 10°C
	10	10 mM Tris pH 7.5 100 mM NaCl	PACT	
	11	10 mM Tris pH 7.5 200 mM NaCl	INDEX, PACT	Plus: Zn^{+2}, seeds, TCEP, NA

RT- room temperature
NA- no additive

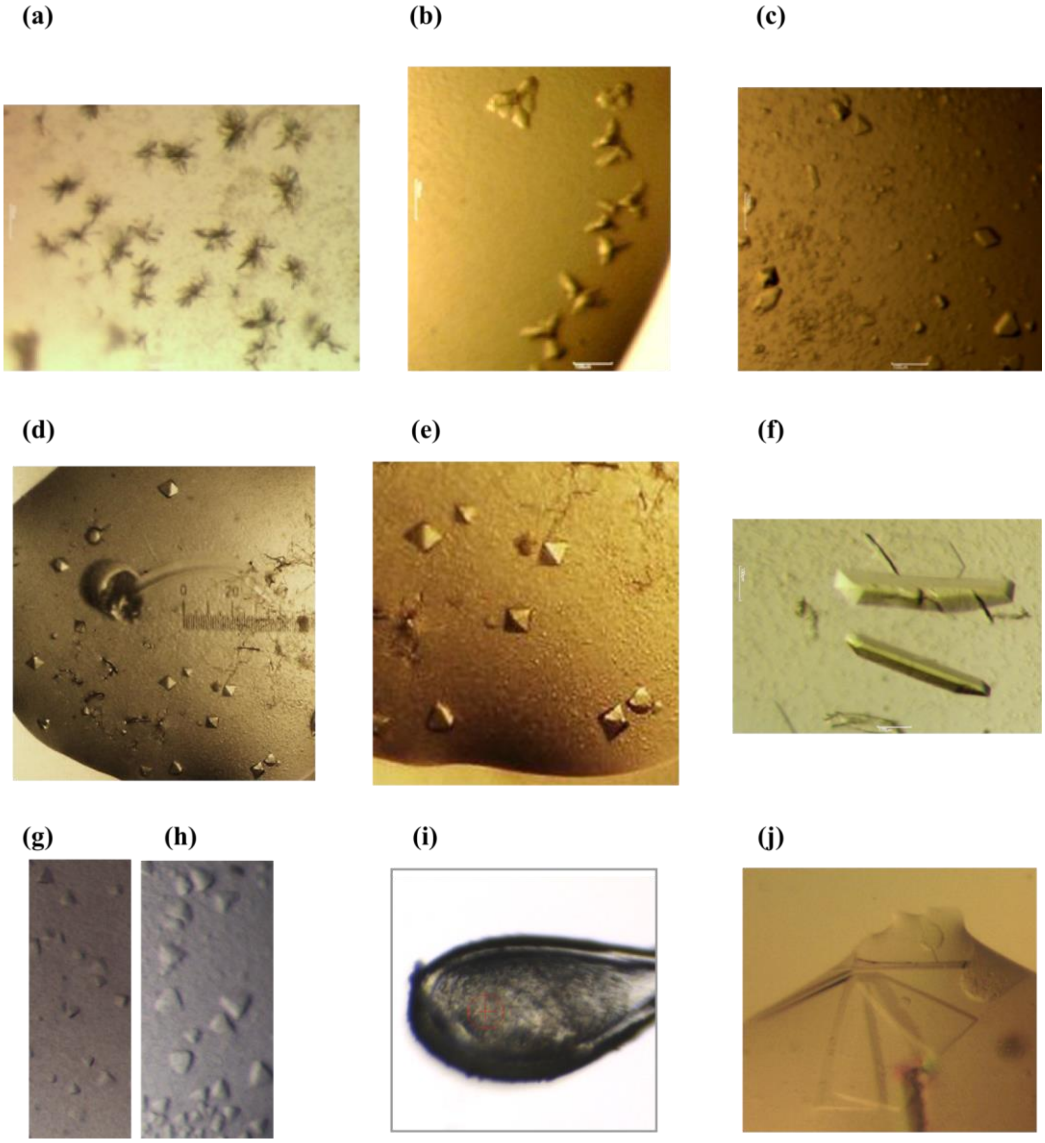


Figure 40. Crystallisation of G*13*P_FL_T104Y; A261W

(a-c) Low quality crystals of tagged G*13*P_FL_T104Y;A261W. (d-f) Crystals of native u-G*13*P_FL_T104Y;A261W. (g-h) Crystals of Se-Met u-G*13*P_FL_T104Y;A261W used as seeds. (i-j) Crystals of Se-Met u-G*13*P_FL_T104Y;A261W used for structure solution. Pictures not to scale. See text for screen conditions.

3.6 Conclusions

In total, efforts to crystallise the major capsid protein G*13*P, comprised the cloning and expression of thirty-two different constructs that include truncations at the N- or C-terminus, single or double mutants and engineered fusion proteins. Out of these, only six constructs were monomers. These six constructs and the G*13*P_G64W mutant protein were subjected to crystallisation experiments. The rest of the produced constructs were insoluble from the cell pellet, formed insoluble aggregates during purification or produced soluble protein in very low yield unfit for the purpose of crystallography.

All six monomeric constructs, including double mutants (G*13*P_FL_G64W;T104Y, G*13*P_FL_T104Y;A261W) and G*13*P-AT chimeras (full length N- or C-terminally tagged), have in common the alteration of the P-loop (Figure 27 and Figure 32e). Alteration of the P-loop consisted in the introduction of two bulky amino acids or the AT Zn-binding domain, suggesting that this area of the capsid protein features residues essential in establishing contacts with other subunits.

To date, from the references reviewed, SPP1's G*13*P from this study is the second protein, besides T4's gp24 in being crystallised from the overexpression of the encoding gene alone (37). The rest of capsid proteins are available from the crystallisation or cryo-EM reconstruction of assembled purified capsids (section1.5.1). SPP1 G*13*P is the first major capsid protein characterised by X-ray crystallography in the monomeric state since T4 gp24 self-assembles into pentamers. No information about gp24's oligomeric state is provided in the original reference, but analysis of the corresponding PDB entry shows that it does not appear to form oligomers.

The identification of the double mutant of G*13*P for which diffraction quality crystals could be obtained and X-ray data could be collected (see Chapter 4) signified an important step forward, enabling determination of the X-ray structure of this protein in a stand-alone monomeric form.

Although modification of two residues abolished G*13*P's propensity to aggregate, we do not expect this to affect protein's conformation as the side chains of these two residues are exposed and not involved in intra-subunit interactions stabilising capsid assemblies of other viruses. Thus the crystallised species is likely to be truly representative of the WT protein, although influence of G*6*P, G*7*P, G*11*P or G*12*P if any, might not be reflected in the recombinantly produced protein.

Chapter 4

4. Elucidation of structure of SPP1 capsid protein

4.1 Native protein crystals

4.1.1 Data collection

A complete data set was collected from a cryo-cooled crystal of uG*13*P_FL_T104Y;A261W at the I04 beamline at the Diamond light Source UK at a wavelength of 0.97950 Å. The distance crystal-to-detector was 533 mm. 1,100 images were collected, each one at a 0.2° rotation of the crystal, to cover a 220° overall rotation range. The highest diffraction where weak reflections were observed was ~ 3.8 Å (Figure 41).

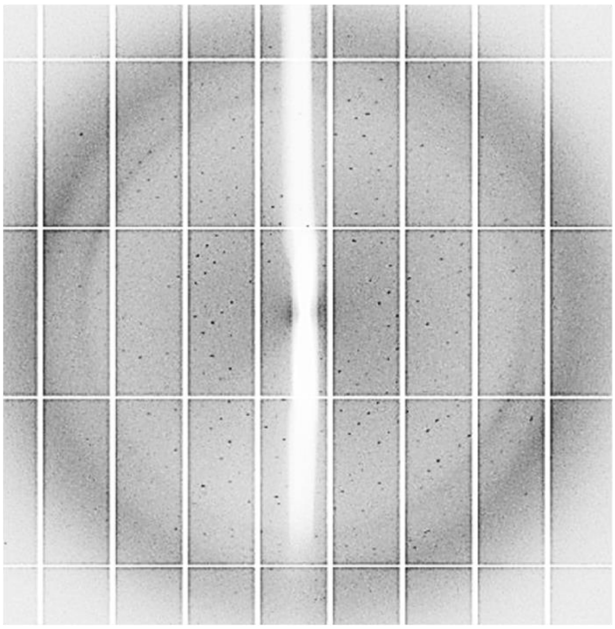


Figure 41.- Data collection from SPP1 uG*13*P_FL_T104Y;A261W native crystals

4.1.2 Structure determination

Images were processed with XDS (107) software using the resolution range 50-3.6 Å. Images were indexed in the space group *P* 4_12_12 with the cell dimensions of *a*=*b*= 79.8 Å and *c*= 134.8 Å, α=β=γ= 90° (See Table 9). The relevant parameters of the collected data set can be found in Table 9, in summary, the R-factors R_{merge} and R_{meas} (all I^+ & I^-) had values of 17.8 % (118 %) and 18.4 % (122.1 %), respectively. The completeness and multiplicity values of the data set were 99.6 % (96.1 %) and 12.3 (12.0), respectively (108,109).

The calculated solvent content was 59.6 % (Matthew's coefficient 3.0 $Å^3$/Da) (119), corresponding to one subunit per asymmetric part.

Molecular replacement was performed to determine phases using both the cryo-EM model of SPP1 G*13*P (PDB 4AN5), and a poly-Ala version of HK97 gp5 (PDB 1OHG) as search models. None of the models could successfully be used as source of phases. The solutions obtained showed that with the SPP1 EM model as search model, the long helix at the P-domain matches well the electron density; however, for the rest of the structure, there is no agreement between the provided model and the electron density. A similar case resulted when HK97 gp5 poly-Ala model was fed onto Phaser (112). Some features (e.g. a few β-strands at the A-domain and helix α3') were found in SPP1 G*13*P, but the rest of the structure could not be traced.

Indexing the reflections in the enantiomorph space group P 4_32_12 did not yield reliable results when the SPP1 EM model was used to recover the phases. Subsequent efforts focused on the production of Se-Met labelled uG*13*P_FL_T104Y;A261W to solve the phase problem experimentally (Section 4.2.3).

4.2 Se-Met protein crystals

4.2.1 Data collection

Se-met crystals of uG*13*P_FL_T104Y;A261W were sent to the I04 beamline at the Diamond Light Source, UK, to collect a data set at the wavelength of 0.97934 Å. This wavelength was established from the X-ray fluorescence scan, which found a peak at ~ 12,660 eV (Figure 42a). The crystal's total rotation range was 180°, with data collected at a 542 mm crystal-to-detector distance and 0.2° crystal rotation per image. Diffraction spots were observed up to ~3.2 Å (Figure 42b).

4.2.1 Structure determination

Using XDS for data reduction led to the determination of the space group and cell dimensions (Table 9). The used reflections were in the resolution range 49.28-3.00 Å. The protein crystallised in the space group *R* 32 with cell dimensions *a*, *b* = 176.2 Å; *c* = 356.7 Å and angles α, β = 90° , γ = 120° (hexagonal setting). Data processing statistics are in Table 9.

For the heavy atom detection and the subsequent phasing of the structure, the SHELXCDE suite was used.

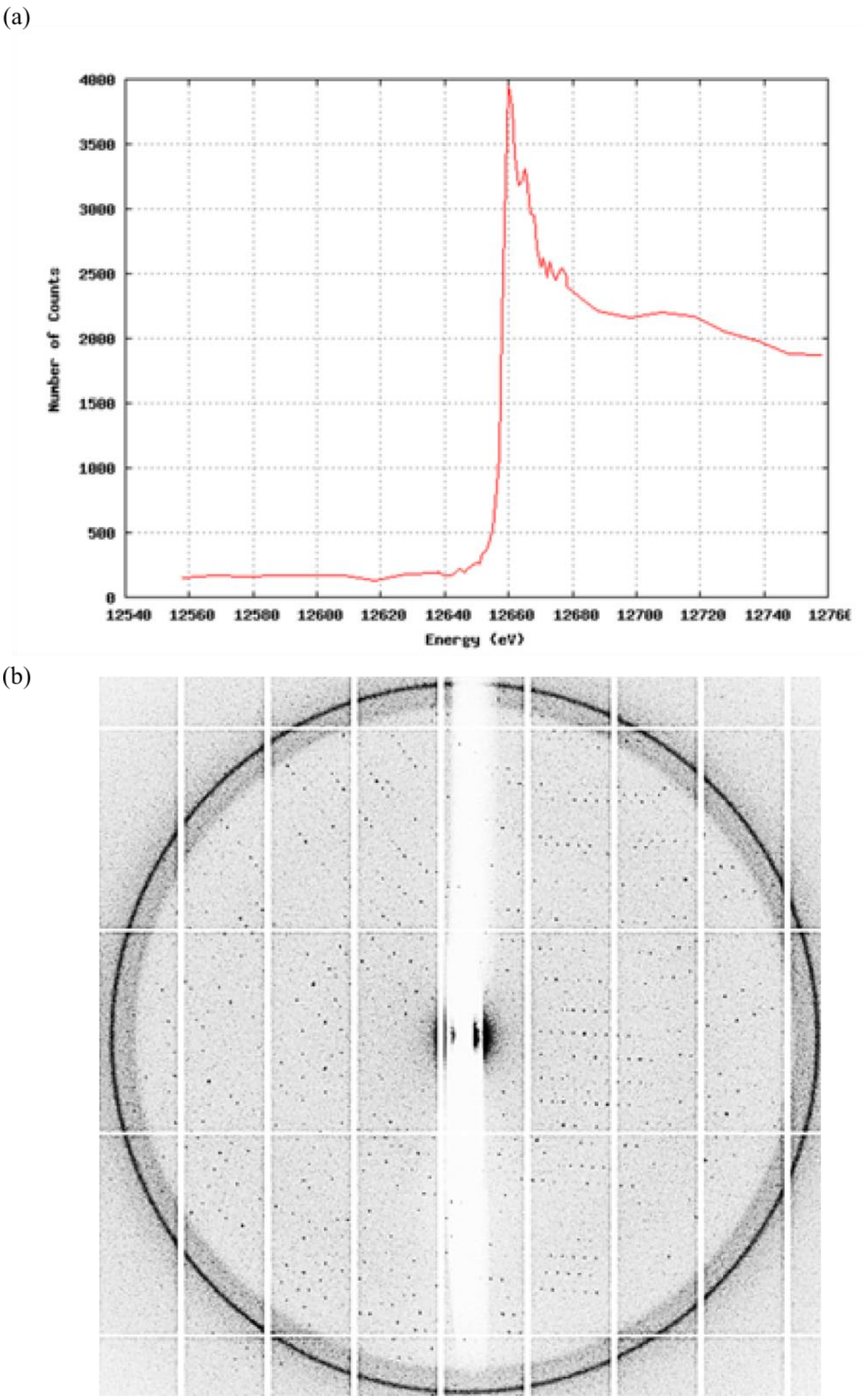


Figure 42.- Data collection from SPP1 uG*13*P_FL_T104Y;A261W Se-Met crystals

(a) X-ray fluorescence scan showing the characteristic peak of absorption for Selenium atoms. (b) Diffraction image. The resolution at the edge of the plate is ~ 3.5 Å.

Table 9. X-ray data statistics for SPP1 G*13*P

	Native crystal	**Se-Met crystal**
X-ray source	Beamline I04, Diamond LS, UK	Beamline I04, Diamond LS, UK
Detector	Pilatus 6M-F	Pilatus 6M-F
Wavelength (Å)	0.97950	0.97936
Space group	$P\ 4_1\ 2_1\ 2$	*R* 32 (hexagonal setting)
Cell dimensions (Å)	*a*,*b* = 79.8; *c* = 134.8 α, β, γ = 90°	*a*, *b* = 176.2; *c* = 356.7 α, β = 90, γ = 120°
Resolution range (Å)	45.01-3.60 (3.94-3.60)	49.28-3.00 (3.11-3.00)
No. of unique reflections	5477 (1254)	42732 (4265)
R_{merge}† (%)	17.8 (118.0)	18.1 (207.8)
R_{meas} (%)	18.4 (122.1)	18.9 (217)
Average *I*/σ(*I*)	14.6 (2.9)	12.4 (1.3)
Completeness (%)	99.7 (99.1)	99.6 (96.1)
Multiplicity	14.9 (15.4)	12.3 (12.0)
CC (1/2) *	0.998 (0.898)	0.990 (0.503)
CC (1/2) DelAnom	NA	0.620 (0.027)
Wilson B-factor (Truncate) (Å^2)	84.9	70.2
Number of atoms	2319 (1 chain x 308 residues)	12097 (4 chains x 323 residues, 1 x 308 residues)
Solvent content (%)	59.6	59.2
Number of reflections used in refinement		40571
R-factor (%)		24.4 (37.1)
Number of reflections used for R_{free}		2097
R_{free} (%)		26.5 (37.6)
Average atomic B-factor (Å^2)		56.1
Rms BondLength (Å)		0.007
Rms BondAngle (°)		1.27

Values in parentheses are for the highest resolution shell.

† $R_{merge} = \Sigma_{hkl} \Sigma_i |I_i(h) - \langle I(h) \rangle| / \Sigma_{hkl} \Sigma_i I_i(h)$, where $I(h)$ is the intensity of reflection h, $\langle I(h) \rangle$ is the average value of the intensity, the sum Σ_{hkl} is over all measured reflections and the sum Σ_i is over i measurements of a reflection.

* CC1/2 - Correlation coefficient between the average intensities of random unmerged half-data sets. The value of CC1/2 is closer to +1 when there is a good correlation between the half-data sets and 0 when there is no correlation. At low resolution CC 1/2 is generally close to 1 at low resolution and tends to get closer to zero at higher resolution since the intensities are weaker (109).

The integrated intensities were merged in a .HKL file that was used as input file by SHELX*C* *(110)*. SHELX*C* created the run files for SHELX*D*, which uses the merged intensities to identify and locate the heavy atoms (Figure 43). SHELX*D* identified 47 heavy atoms, indicating that the asymmetric unit would contain a pentamer or hexamer of subunits (there are ten Se-methionine residues per monomer). Figure 43b shows the solutions found for the substructure of uG*13*P_FL_G64W;T104Y. The heavy atom positions were then used to determine the phases of the protein using Phaser option SAD (111,112). Inspection of the heavy atoms coordinates showed a symmetrical arrangement, where a possible pentamer was observed (Figure 43c,d).

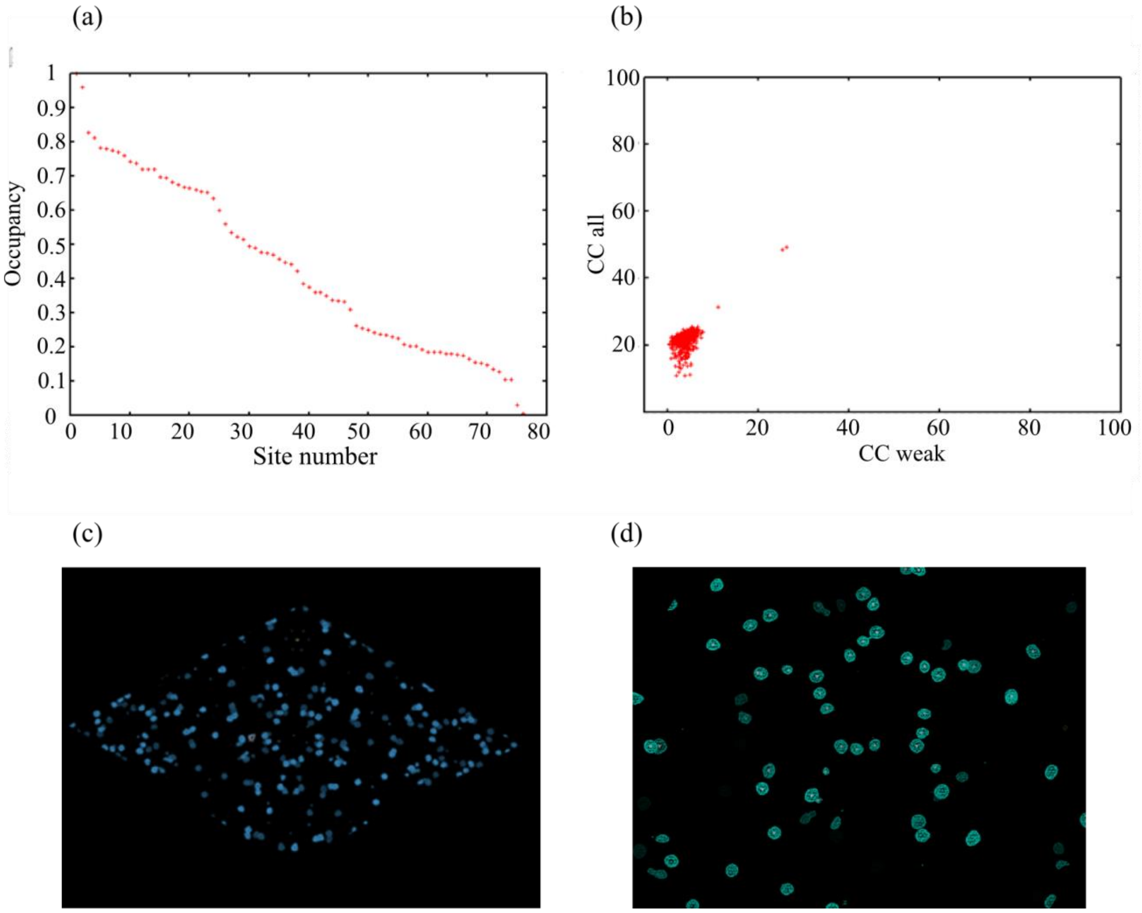


Figure 43.- Location of heavy atoms by SHELX*C/D*

(a) Occupancy graph for the heavy atoms. (b) Substructures of heavy atoms located by SHELXD. (c), (d) visualization of the heavy atoms in Coot. Figures in this chapter featuring initial versions of G*13*P's structure were prepared by Coot (116).

Only the inverted hand output map from Phaser had discernible features (Figure 44a, left) and it was therefore the choice to continue with the structure solution process. After density modification was performed, the electron density map was interpretable to an extent that α-helices and loops were in continuous arrangement (Figure 44b, right).

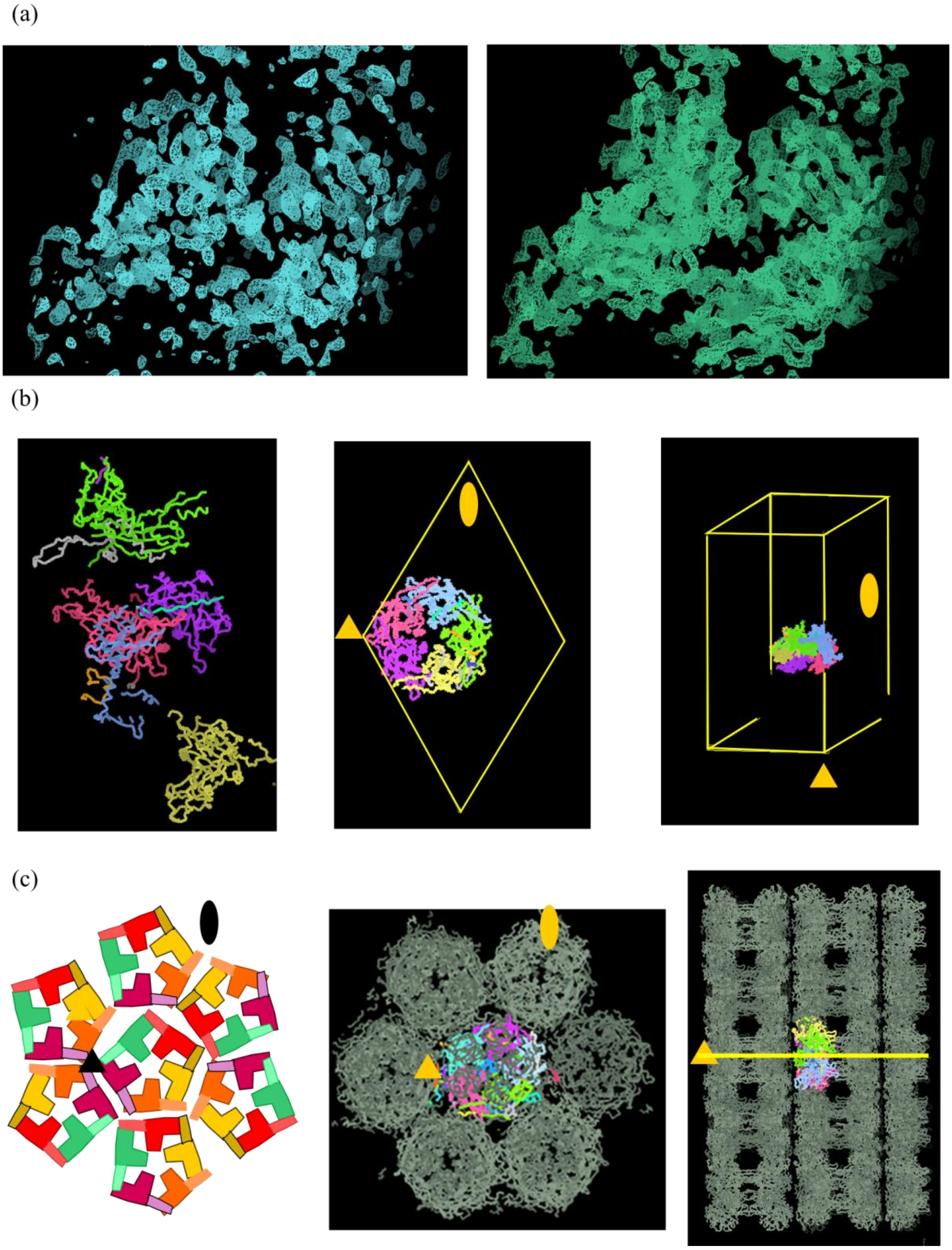


Figure 44. Solvent flattening, model-building and crystal arrangement

(a) Same view of the electron density maps before (left) and after (right) density modification. Contoured at 2 σ. (b) Automatic model-building. Initial (left) and final (centre and right) models built by Buccaneer. The unit cell is drawn with yellow lines and the 3- and 2-fold crystallographic axes are indicated with a triangle and an oval. (c) Views along the Z and Y axis of the crystal.

Automatic model building was performed with Buccaneer (114). The first generated .pdb file in the first cycle of building, displayed several subunits with no apparent symmetrical arrangement, although some of them resembled the fold of HK97 gp5 (Figure 44b, left). By the end of the final cycle, Buccaneer built 1,562 residues (out of 1,620 for a pentamer) in 39 fragments arranged in 5 chains in the asymmetric unit. The completeness of every chain was 86.4 % (Figure 44b, centre and right). The final model involved five copies of G*13*P arranged in a circular structure.

Interestingly, every pentamer is tightly surrounded by six pentamers, as it was a hexamer, producing a hexagonal lattice when viewed along the Z axis (Figure 44c and Figure 46c). This observation points out that the subunits from the pentamer establish different contacts with the neighbouring pentamers in order to be packaged as a hexamer.

Pentamers exhibited a head-to-head and tail-to-tail packaging in the unit cell since the top of each pentamer (tower loops from the A-domain) face the top region of another pentamer. Likewise, the tail (bottom of the P-domains) faces the tail of another pentamer (Figure 44c).

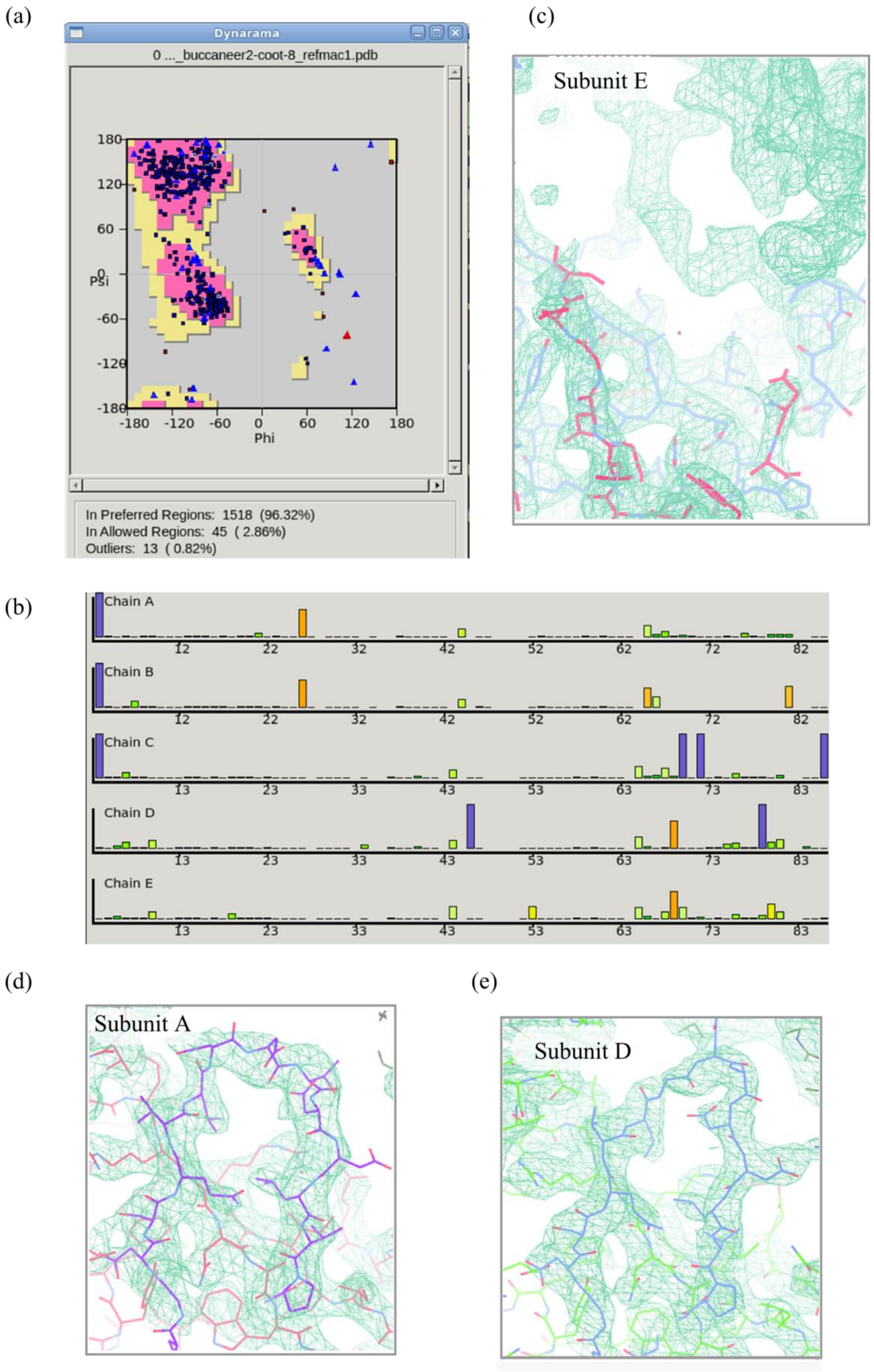
4.2.2 Refinement and validation

Several modifications of the model built by Buccaneer involving the removal of inserted residues or random fragments, merging fragments from the main chain and sequence renumbering were carried out in Coot (Figure 45a,b). The side chain rotamers and Ramachandran outliers were adjusted using the validation tools from Coot followed by refinement in reciprocal space with Refmac5 (115). The final value of the R was 24.4 (37.1) % (R_{free} = 26.5 (37.6) %). The complete refinement statistics can be found in Table 9. At the current resolution of 3.0 Å, some side chains atoms, especially the loops residues, were not clearly defined in the electron density maps (Figure 45c-e).

The extended loop (E-loop) represented the most difficult part to conform with the Ramachandran plot (Figure 45 c-e). Electron density was observed for most of the E-loop Cα-backbone but not all the side chains were defined. This portion was removed from the initial model and manually rebuilt using omit maps to improve the fit. After a few cycles of refinement, some positive density buds indicated where some side chains could be positioned.

Using NCS during model building, the sequence of the loop could be assigned and most of the loops were completed. Electron density for a large fraction of the E-loop from subunit E was not observed (details on Chapter 5). Subunit E differs from the rest of the subunits only at this region. There was only electron density for the start and end of the E-loop.

The final model showed good stereochemistry with 96.32 % of the residues located in the preferred regions of the Ramachandran plot, 2.86 % in the additionally allowed regions and 0.82 % were outliers (Figure 45a).



```
In Preferred Regions:  1518  (96.32%)
In Allowed Regions:  45  ( 2.86%)
Outliers:  13  ( 0.82%)
```

Figure 45. Model validation

(a) Ramachandran plot. The blue triangles correspond to glycine residues. (b) Rotamer analysis. The purple bars indicate residues with truncated side chains. (c-e) E-loops from the pentamer assembly. The E-loop in (c) belongs to subunit E. The electron density maps from the 3.0 Å resolution G*13*P Se-Met structure are contoured at 1 σ.

4.2.3 Structure determination of the native protein by Molecular replacement

The refined structure of Se-Met G*13*P was used to determine the structure of the “native” protein by MR (Figure 46). Electron density was visible for the A- and P-domains but not for the E-loop (Figure 46a).

A remarkable difference between native and Se-Met crystals is the arrangement of molecules (Figure 46). While in the Se-Met crystals, subunits are organised in symmetric pentamers, in the native crystals no hexameric or pentameric arrangement is detected. Instead, every subunit uses different surfaces to interact with several neighbours. Part of one P-domain rests in the cavity formed by the A- and P-domains from another subunit, whilst other area at the bottom of the same P-domain faces another P-domain.

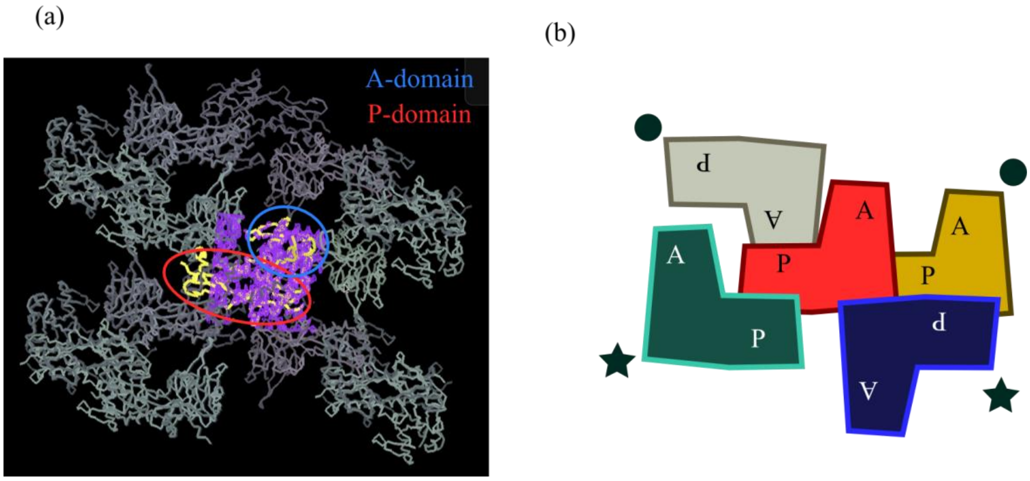


Figure 46. Molecular replacement on native crystals

(a) MR performed with subunit E as search model. (b) Representations of the crystal contacts for the subunit in red. Circles and stars indicate subunits at the back and front of the subunit in red, respectively.

Refinement of the native structure was in progress at the time when this chapter was written, a finished version of it will be available in a corresponding publication.

4.3 Conclusions

The structure of SPP1 u-G*13*P_FL_T104Y;A261W was solved at 3.0 Å resolution by SAD. Crystals belonged to the space group R 32.

The Se-Met model enabled structure determination of the lower-resolution native crystals using MR. The structure of the native protein was solved at 3.6 Å resolution. The proteins differ in the conditions and space group they crystallised in. There appear to be differences in the structures, with the most significant being the absence of interpretable electron density corresponding to the E-loop in the native crystals.

The conditions where both crystal forms grew suggested that high salt could shift the oligomerisation state of u-G*13*P_FL_T104Y;A261W from monomer (crystals grown in 0.2 M Sodium malonate) to pentamer (crystal grown in 2.0 M $(NH_4)_2SO_4$). No additional X-ray data sets for Se-Met or native crystals were collected to establish an oligomerisation pattern, although several crystals of Se-Met protein grew in very similar conditions. Based on visual inspection, it is expected that these crystals are ordered in a similar way to those from which the X-ray structure was determined.

At the time this thesis was submitted, X-ray structures of the capsid proteins were available only for bacteriophages HK97 (39) and T4 (37), while cryo-EM studies identified the HK97-fold in a number of viral capsids. Until now, no X-ray structure of a stand-alone native capsid protein was reported. The limited X-ray structural information on capsid proteins is most probably due to difficulties associated with the crystallisation. The resolution of the SPP1 G*13*P X-ray structure (this study) is comparable to the resolution achieved during structural studies of HK97 and T4 capsid proteins (Table 10) and other capsid-like structures.

Table 10. Comparison of X-ray statistics of capsid/capsid-like proteins structures

	SPP1¥	HK97*	T4	Bb_PhRP	PfV	Encapsulin *T. maritima**
Resolution (Å)	3.0	3.45	2.90	2.05	3.6	3.1
Mean ((I)/sd(I))	2.9	1.9	8.42	1.91	1.85	1.32
R-factor	0.247	0.373	0.27	0.186	0.268	0.22
R_{free}	0.265	0.374	0.3	0.227	0.267	0.239
Ramachandran outliers (%)	0.82	2.8	3.6	0.1	2.2	2.2
Rms BondLength (Å)	0.007	0.010	NA	0.013	0.010	0.009
Rms BondAngle (°)	1.27	1.40	NA	1.40	1.37	1.33
PDB	Pending	1OHG	1YUE	3BJQ	2E0Z	3DKT

¥ Latest model version. * Values for the complete capsid or capsid-like structure.
The mean ((I)/sd(I)) corresponds to the outershell values.

Chapter 5

5. Structure of SPP1 capsid protein

Chapter 5 describes in detail the 3.0 Å resolution X-ray structure of the SPP1 capsid protein uG*13*P_FL_T104Y;A261W construct (referred to as G*13*P henceforth). G*13*P was compared with (i) capsid proteins from phages HK97 and T4 and (ii) capsid-related proteins, e.g., PfV and encapsulins.

5.1 SPP1 G*13*P monomer and crystallographic oligomer

5.1.1 Monomer

The G*13*P monomer has a mixed α/β structure that comprises nine α-helices, one 3_{10} helix, eighteen β-strands and several loops of varying length (Figure 47 and Figure 48a). G*13*P folds into three spatially distinct regions: the axial (A-) and peripheral (P-) domains and the extended loop (E-loop), all named following the same convention as for the HK97 phage.

The main core of the globular A-domain comprises seven clustered antiparallel β-strands with helices α7 and α8 at each side of the cluster. The principal features of the elongated P-domain are the N-terminus which makes up helix α1, the long helix α5, and the antiparallel strands β5 and β15, which span half of the P-domain length. The E-loop is a 28 residues-long curved polypeptide chain that accounts for half of the P-domain length. The E-loop can be considered as extension of the P-domain that connects strands β2 and β4. Both domains are connected by three short linkers (Figure 47 and Figure 48).

The introduced mutations T104Y and A261W were located exactly where predicted by the PSI-BLAST alignment, validating the use of the HK97 structure for initial analysis of possible mutations. The T104Y mutation is located on helix α4 in P-loop 1 (PL-1), while A261W is located at the PL-2, which connects strands β14 and β15.

5.1.2 Oligomer and inter-subunit contacts

Although the G*13*P double mutant exists as monomer in solution (Figure 31b), in the crystal, five copies of G*13*P form pentameric assemblies (Figure 48b,d). The overall atom-to-atom diameter of the pentamer is 110 Å. The diameter of the internal tunnel is ~ 8 Å.

Within the pentamer, five A-domains are in close proximity, positioned at the centre of the pentamer. Helices α7 and α9, enclose the antiparallel β-strands forming a ring of intercalating helices. The P-domains are situated at the periphery of the structure, defining the sides of the pentamer. The E-loop of one monomer sits over the P-domain of the right-hand neighbouring subunit. Specifically, every E-loop covers the PL-1 and part of helix α5 and strands β5 and β15. Other characteristics of the assembly are the tower of loops on top of the pentamer formed by the loops connecting β6- α7 and β11- β12 (Figure 48b,d).

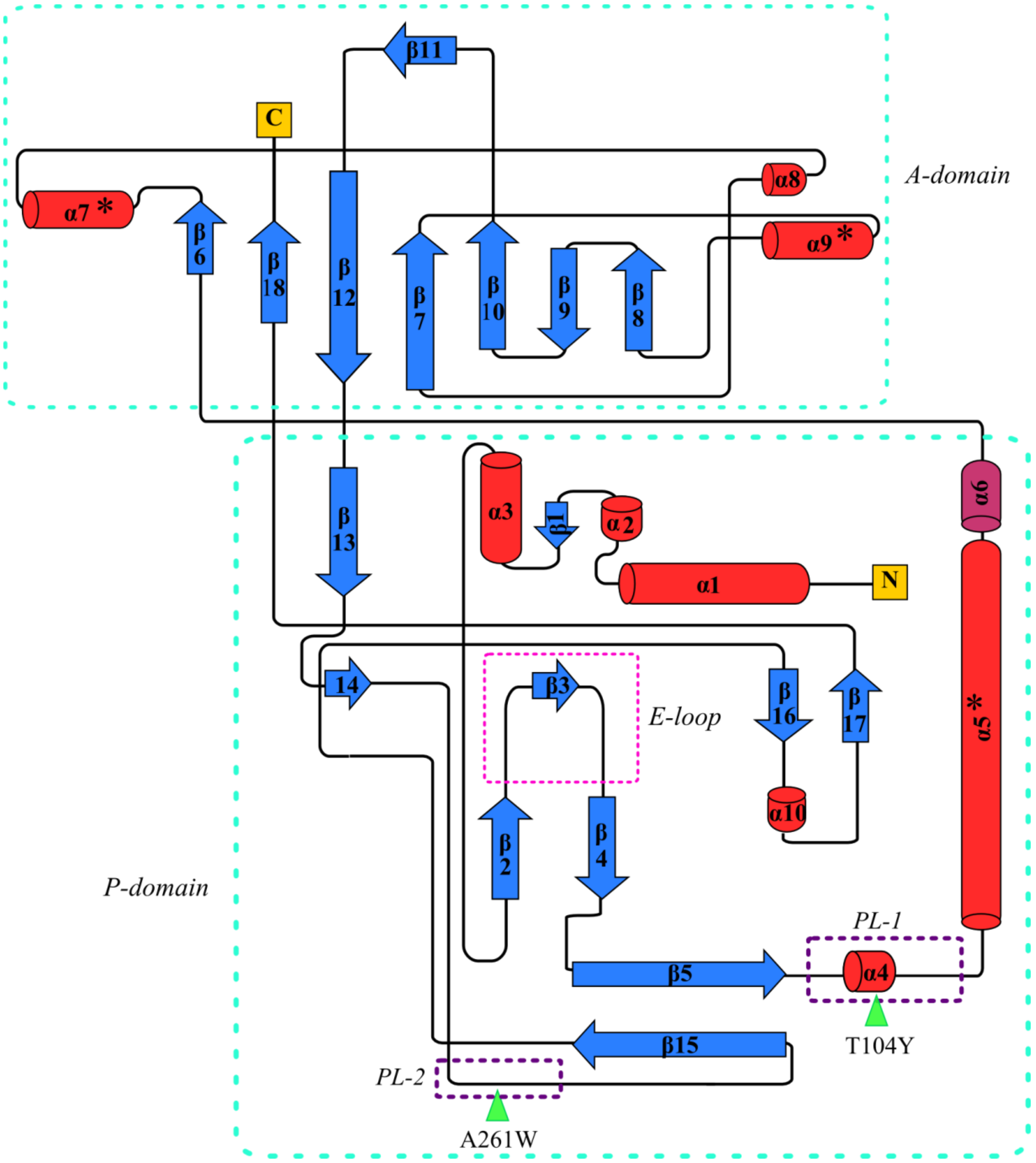


Figure 47. G*13*P topology

Structure cartoon of G*13*P generated with the Pro-origami server (104). The α-helices are coloured in red, 3_{10} helices in magenta, β-strands in blue and N- and C-termini in yellow. The A- and P-domains are contoured with cyan dashed lines. The E-loop and PL-1&PL-2 are squared using pink and purple lines, respectively. The introduced mutations are indicated with green triangles. Helices α5, α7 and α9 are labelled with a star (*).The length of α-helices and β-strands drawn approximately to scale.

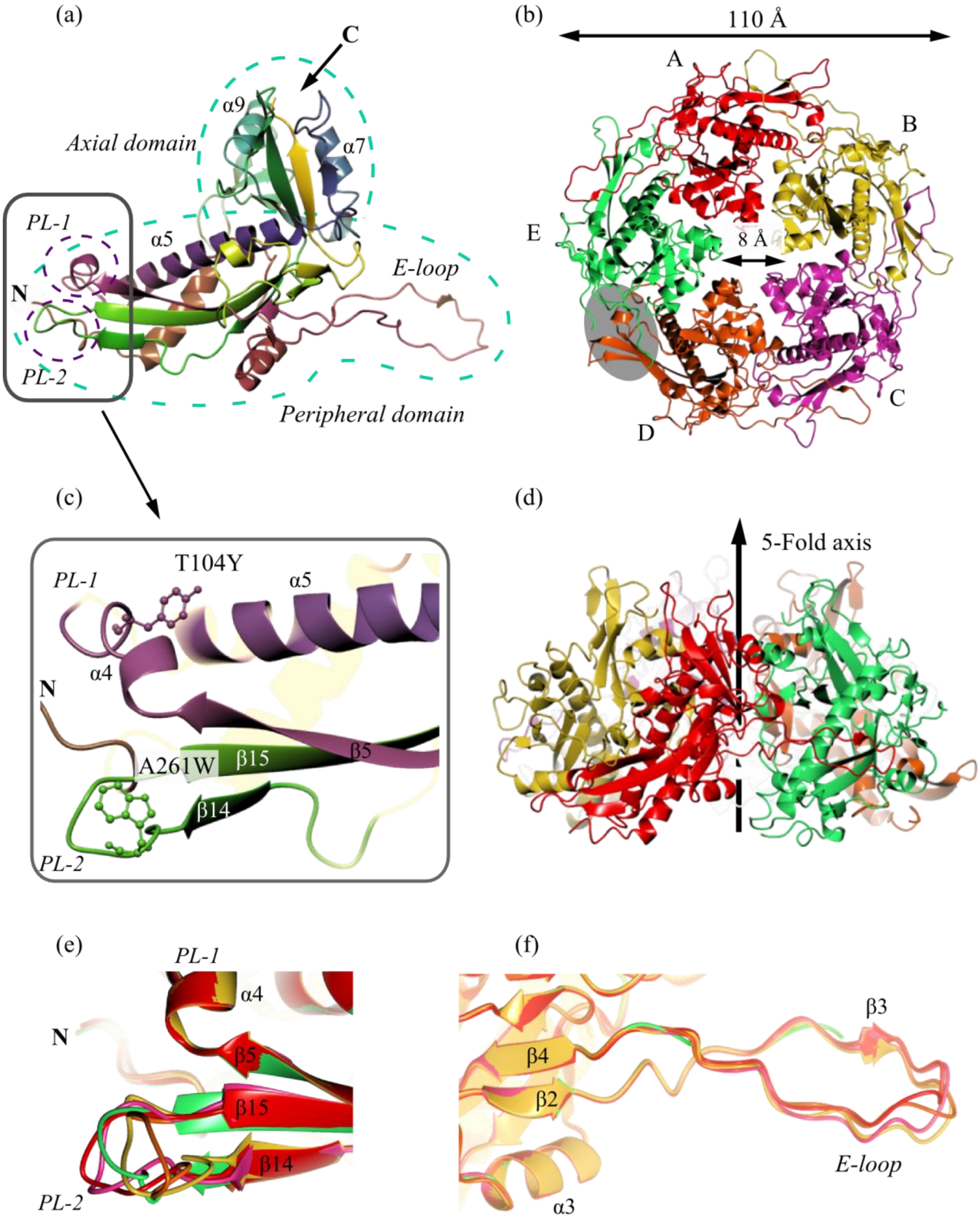


Figure 48. SPP1 G*13*P X-ray crystal structure

(a) Ribbon diagram of G*13*P. Domains and main secondary structure elements are labelled. The area where mutations were introduced is enlarged in (c). Mutations are shown in ball and stick representation. Top (b) and side (d) views of the crystallographic pentamer with every subunit in a different colour. Diameters of the pentamer and tunnel are indicated (atom-to-atom distance). The location of the missing E-loop in subunit E is shaded in grey. (e-f) Flexible loops in G*13*P.

The position and length of the secondary structure elements is the same in the five subunits of the pentamer. Conversely to the folded regions, loops PL-2, E-loop, loops β11-β12, β8-β9 and β6-α7 are highly flexible (Figure 48e,f).

The extent of flexibility of PL-2 confirmed that the adopted approach to substitute it with a folded Zn-binding domain from AT was in theory feasible, although it proved to be unsuccessful, presumably due to affecting the fold of the protein.

One unexpected difference in fold between subunits, was observed in subunit E (between A and D), which lacks most of the E-loop. As aforementioned in section 4.2.2, electron density for residues 60-80 was missing. The portion left of the E-loop, superimposes perfectly with its counterparts in other subunits. Since only one subunit lacks the E-loop, it is proposed that fitting a fifth subunit in the pentamer results in the increased flexibility of the E-loop of this subunit. Furthermore, the loop's flexibility suggests that during capsid assembly, in each copy of G*13*P the loop adjusts its conformation to fulfil the required roles of both pentamers and hexamers. The intrinsic flexibility is reflected in the missing density in subunit E in the Se-Met structure and in the G*13*P monomer present in crystals of the native construct. Since in the Se-Met structure, loop's flexibility (lack of electron density) is observed only for one subunit, proteolytic cleavage or degradation are unlikely.

The pentamer is held together by combination of H-bonds and salt bridges (SB) that repeat in every inter-subunit interface (Table 11,Table 12, Figure 49). On average, every subunit-subunit interface buries 1,379.4 Å^2 of surface and contains nineteen H-bonds and seven salt bridges (SB). G*13*P contains not Cys residues hence there are no disulphide bridges to stabilise the pentamer. It is deduced that every subunit establishes 38 H-bonds with its two neighbours. In comparison, subunits from the hexamer in the HK97 asymmetric unit, establish 64 H-bonds.

Table 11. Pentameric G*13*P interfaces*

Structure 1				**Structure 2**				**Interface area Å^2**	**Δ^{i}G**	**N$_{HB}$**	**N$_{SB}$**	**N$_{DS}$**
Subunit	^{i}N$_{at}$	^{i}N$_{res}$	Surface Å^2	Subunit	^{i}N$_{at}$	^{i}N$_{res}$	Surface Å^2		Kcal/mol			
A	174	52	16,808	B	153	44	16,701	1,469.3	-7.2	20	7	0
B	169	48	16,701	C	150	43	16,610	1,499.3	-10.1	20	7	0
C	171	50	16,610	D	142	39	16,560	1,476.8	-8.6	15	5	0
D	127	36	16,560	E	100	28	15,253	1,027.0	-2.1	19	10	0
E	172	53	15,253	A	136	35	16,808	1,424.7	-12.0	21	7	0
							Average	1,379.4	-8.0	19	7	0

*. Table calculated with the PDBePISA server (105). All definitions were extracted from the output file.
^{i}N$_{at}$ - Number of interfacing atoms in the corresponding structure.
^{i}N$_{res}$ - Number of interfacing residues in the corresponding structure.
Surface Å^2 - Total solvent accessible surface area in square Angstroms.
Interface area Å^2 - Difference in total accessible surface areas of isolated and interfacing structures divided by two.
Δ^{i}G - Solvation free energy gain upon formation of the interface, in Kcal/M. Negative Δ^{i}G corresponds to hydrophobic interfaces, or positive protein affinity.
N$_{HB}$ - Number of potential H-bonds across the interface.
N$_{SB}$ - Number of potential salt bridges across the interface.
N$_{DS}$ - Number of potential disulphide bonds across the interface.

Table 12 lists the H-bonds established by chain B at the AB and BC interfaces (Figure 49). Twenty H-bonds are formed in both interfaces mainly where the A-domains converge and where the E-loop sits over the adjacent P-domain.

At the A-domain interface, helices α7-α9 and strands β8-β9 from different subunits are responsible for ~35 % of the H-bonds, however the area of extensive interaction occurs at the E-loop-P-domain region. In both areas, several charged and polar residues, predominantly Arg, Asp and Ser, create a network of H-bonds that stabilises the pentamer and is reproduced in every interface. On Table 12 it is shown that the residues forming H-bonds from chain A at the interface AB are mostly the same that chains B uses to form H-bonds with chain C at the interface BC, with slight variation.

In the HK97 asymmetric unit (head II), the same areas of contact are observed though the number of H-bonds is larger due to the presence of the N-arm, which slides below the start of the E-loop of the left-hand adjacent subunit (39). In G*13*P no N-arm is present (See section 5.2).

Table 12. H-bonds established by subunit B*

	Interface AB				Interface BC		
##	Subunit A	Dist. [Å]	Subunit B	##	Subunit B	Dist. [Å]	Subunit C
1	ARG 117[NH1]	2.4	ASN 60[O]	1	ASP 100[O]	3.5	ASN 60[ND2]
2	ARG 117[NH2]	2.9	LEU 62[O]	2	GLU 125[OE2]	2.7	SER 66[N]
3	TRP 96[NE1]	2.7	ASP 63[OD1]	3	GLY 93[O]	3.3	GLN 67[N]
4	ARG 117[NE]	3.3	ASP 63[OD1]	4	LEU 91[O]	2.9	LEU 69[N]
5	ARG 117[NH2]	3.0	ASP 63[OD1]	5	THR 291[OG1]	2.2	ASN 70[ND2]
6	TYR 121[OH]	2.6	GLY 64[O]	6	ARG 207[O]	3.7	ASN 213[ND2]
7	ARG 92[NH1]	3.5	SER 66[OG]	7	SER 106[OG]	3.5	ASP 39[OD2]
8	E:ARG 92[NH1]	3.4	SER 66[O]	8	ARG 117[NH1]	2.8	ASN 60[O]
9	GLY 93[N]	3.4	GLN 67[O]	9	ARG 117[NH2]	3.7	LEU 62[O]
10	LYS 192[NZ]	2.4	GLU 161[OE2]	10	TRP 96[NE1]	3.1	ASP 63[OD1]
11	ARG 207[NH1]	3.7	LYS 200[O]	11	ARG 117[NH2]	3.1	ASP 63[OD1]
12	LYS 200[NZ]	2.8	SER 202[OG]	12	TYR 121[OH]	2.9	GLY 64[O]
13	GLU 125[OE2]	2.4	SER 66[N]	13	ARG 92[NH1]	3.6	SER 66[OG]
14	E:GLU 125[OE2]	2.6	SER 66[OG]	14	GLY 93[N]	3.2	GLN 67[O]
15	E:GLY 93[O]	3.2	GLN 67[N]	15	THR 291[N]	3.8	ASN 70[OD1]
16	E:ASN 94[OD1]	3.3	GLN 67[NE2]	16	LYS 192[NZ]	3.7	GLU 161[OE1]
17	E:LEU 91[O]	3.3	LEU 69[N]	17	ARG 124[NH2]	3.0	ASP 172[OD2]
18	E:ASP 220[OD2]	2.5	HIS 173[NE2]	18	ARG 124[NH2]	3.1	SER 175[OG]
19	E:SER 184[OG]	3.9	HIS 173[NE2]	19	LYS 200[NZ]	3.9	SER 202[OG]
20	E:ARG 207[O]	3.6	ASN 213[ND2]	20	ARG 207[NE]	3.8	ASN 213[OD1]

*. Table calculated with the PDBePISA server (105).

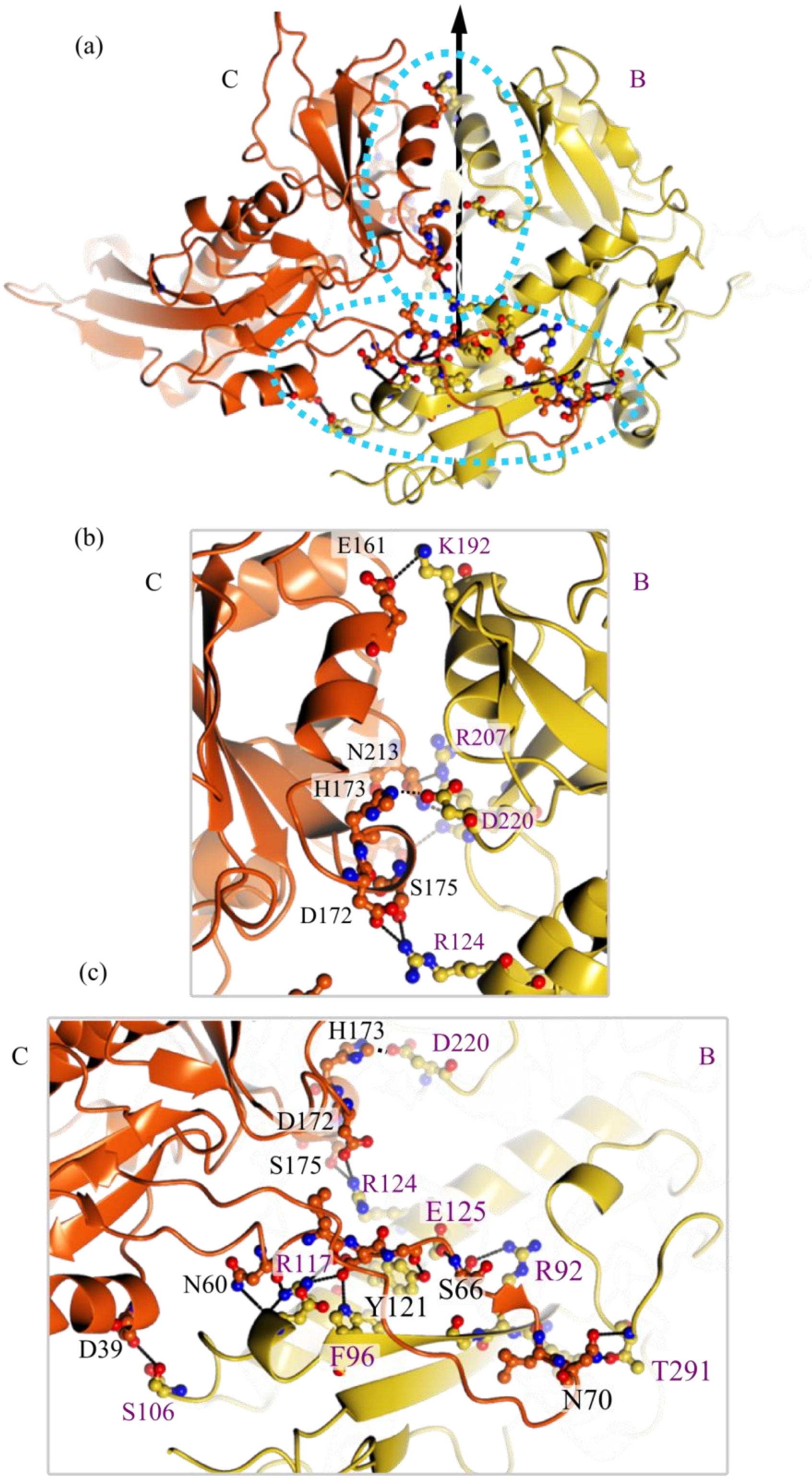


Figure 49. Intersubunit H-bonds in pentameric G*13*P

(a) Side view of the interface BC. (b) A-domain interface. Residues forming H-bonds are shown in ball and stick representation. Residues from subunit B are labelled in purple. (c) E-loop-P-domain interface.

The top view of the electrostatic potential representation of the pentamer, assigned as the outer capsid surface, exposes several positively charged residues that are arranged with 5-fold symmetry (Figure 50). It may be expected that DNA containers such as capsids would display many positively charged side-chains at their inner surface to interact with DNA; nevertheless, the inner surface of G*13*P pentamer has few positive residues at its centre and periphery. Residues at the periphery of the pentamer are located at the bottom part of the P-domain, therefore they are likely to directly face other P-domains within a complete capsid particle and they might not be exposed to the inner surface. The overall negative charge inside the capsid must create an energy barrier during translocation of the negatively charged DNA into the procapsid. This repellent force offered by the inner surface could potentially contribute to drive the DNA out of the capsid during cell infection.

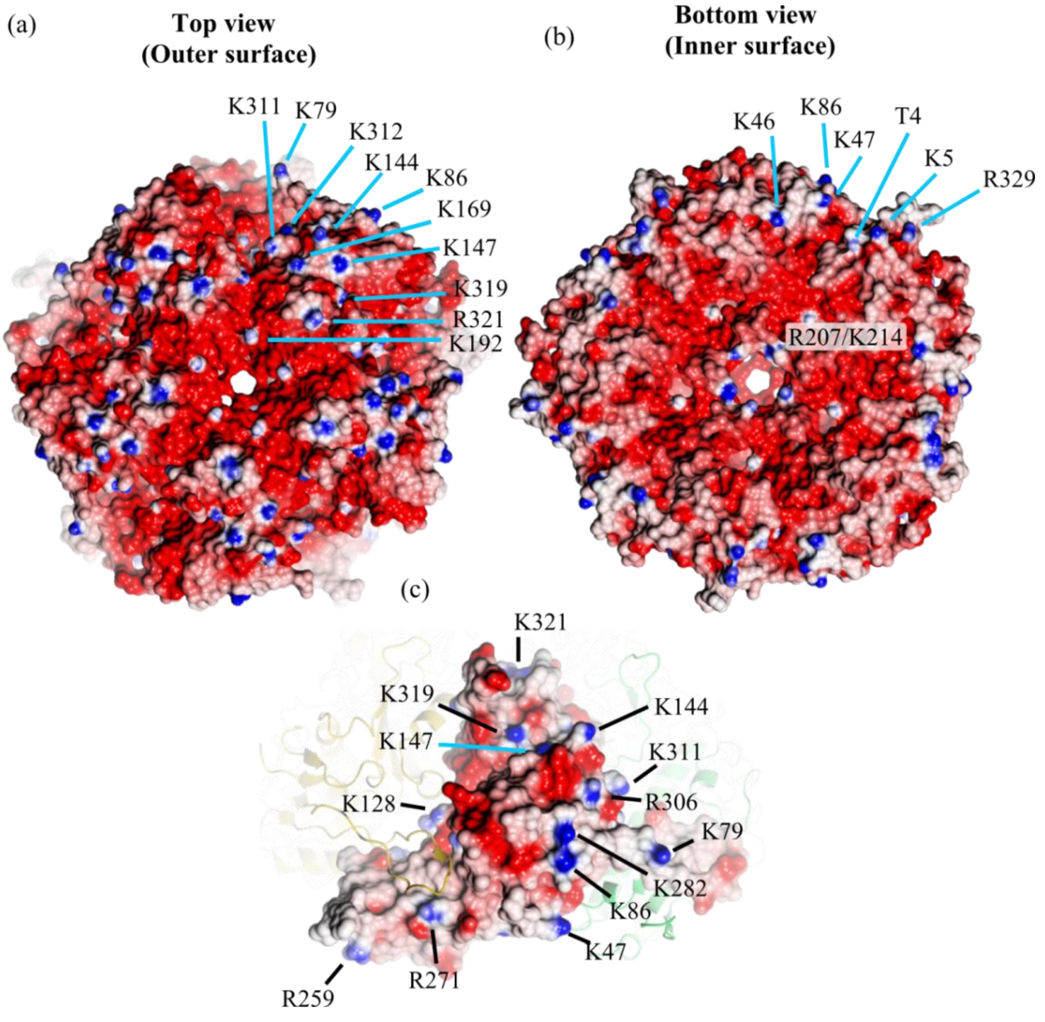


Figure 50. Electrostatic surface representation of G*13*P pentamer

Top (a), bottom (b) and side (c) views of the G*13*P pentamer. Exposed residues are indicated. The five-fold symmetry breaks where the E-loop is absent on chain E

Electrostatic surface representations of the G*13*P and Bb_PhRP pentamers and of the icosahedral asymmetric units of HK97 and *T. maritima* encapsulin showed that their outer and inner surfaces do not follow a pattern of charge (not shown).

All these oligomers expose a large number of positively charged residues at the outer surface (top view). In contrast, no pattern of charge is observed at the inner surface: HK97 and *T. maritima* encapsulin have more positively charged residues in comparison with the G*13*P pentamer and Bb_PhRP, that have a reduced number of positively charged residues. The biological significance for the presence of positive charge covering the virus it is not known.

5.2 Structural comparison with HK97 and T4 capsid proteins and with other capsid-related proteins

Despite the low sequence homology between capsid proteins or capsid-related proteins that exhibit the HK97-fold the adopted tertiary structure is remarkably similar (Table 13, Figure 51 and Figure 53). Monomers from the hexameric and pentameric HK97 gp5 assembly superpose onto SPP1 G*13*P monomer, with a Cα rmsd of 3.03 Å (173 residues) and 2.99 Å (174 residues), respectively.

Whilst SPP1 G*13*P has seven antiparallel β-strands at the core of the A-domain, HK97 gp5 has six. One additional common feature in this domain is the presence of the two α-helices (α7 & α9 in SPP1, α5 & α6 in HK97 (Figure 6)) from the A-domain, which together with helix α5 (α3' in HK97) and the E-loop represent the most emblematic features of the HK97-fold. These three α-helices are rapidly identified from cryo-EM maps of complete capsids to produce cryo-EM protein models. The organisation of the P-domains is also comparable between G*13*P and gp5. In SPP1 G*13*P, helix α5 finds its counterpart in helix α3' in HK97. Most of the long β-strands that define the bottom of the P-domain are found in G*13*P, although β2 & β4, equivalent to βB & βC in HK97, are shorter.

Table 13. Structural superposition with G*13*P

Moving model	rmsd (Å)	Residues aligned	Total residues of moving model
HK97 gp5-Hexamer-capsid	3.03	173	281
HK97 gp5-Pentamer-capsid	2.99	174	281
HK97 gp5-Hexamer-procapsid I	3.10	186	265
HK97 gp5-Hexamer- procapsid II	3.29	187	265
T4 gp24	2.98	201	409
Bb_PhRP	3.23	189	300
PfV	3.06	173	236
T. maritima encapsulin	3.95	177	264
HK97 pentamer capsid*	2.89	194	1405
HK97 pentamer procapsid I*	1.269	34	1325
Bb_PhRP *	5.73	770	1502
T. maritima encapsulin*	5.33	770	1320
PfV*	3.67	821	1180

G*13P*'s structure length = 323 residues, Crystallised pentamer= 1600 residues.

Star (*) indicates crystallised or extracted from capsid pentamer assembly.

The principal differences with HK97 gp5 are found in the N-terminus and the E-loop. The N-terminus in G*13*P folds into two α-helices behind helix α5. The same region in HK97 exists as an extended polypeptide chain located above α3'. When viewed from top, the E-loop remains lined up with the P-domain in HK97 gp5 (hexamer). On the contrary, in G*13*P the E-loop creates an angle with the P-domain bending towards the five-fold axis. G*13*P bends more than the E-loop from pentameric gp5, suggesting that the conformation of the crystallised G*13*P might slightly vary at this region from the G*13*P present in the capsid (considering that the folds are rather alike and the *T* numbers the same) (Figure 51).

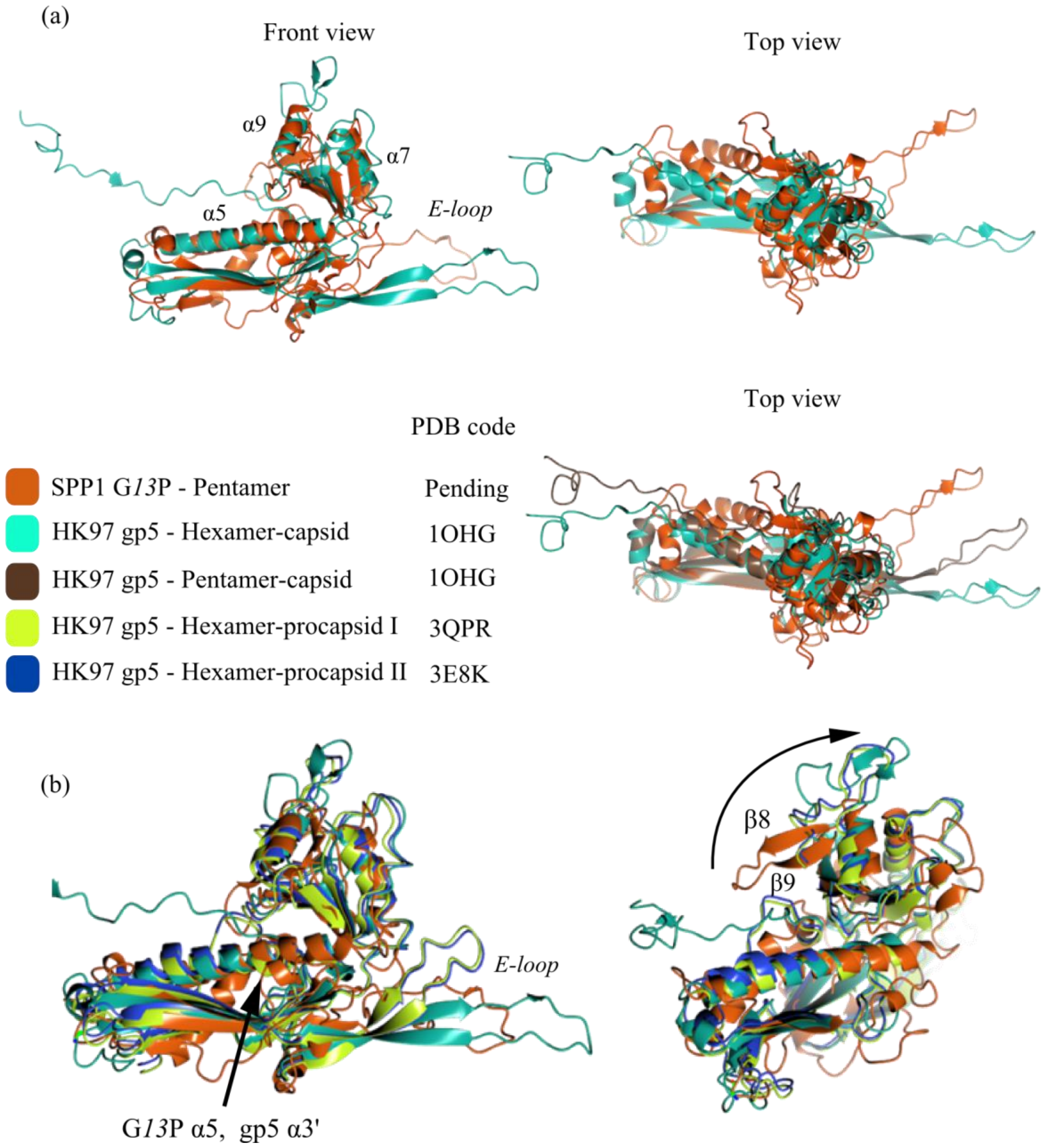

(a)
Front view
Top view
α9
α7
α5
E-loop
Top view
PDB code
🟧 SPP1 G13P - Pentamer | Pending
🟦 HK97 gp5 - Hexamer-capsid | 1OHG
🟫 HK97 gp5 - Pentamer-capsid | 1OHG
🟩 HK97 gp5 - Hexamer-procapsid I | 3QPR
🟦 HK97 gp5 - Hexamer-procapsid II | 3E8K
(b)
E-loop
β8
β9
G13P α5, gp5 α3'

Figure 51. Structural comparison of SPP1 and HK97 capsid proteins.

It is known that subunits of HK97 in the pentamer and hexamer structures only vary in conformation of the E-loop. It is expected that the SPP1 G*13*P follows the same behaviour, e.g. E-loop is lined up with the P-domain in the hexameric arrangement and acquires a more bent conformation in pentamers. The possibility of other differences between the crystallised G*13*P and that in the capsid cannot be ignored since the fold of the N-terminus and loops at the top of the A-domain is distinct from their counterpart in HK97. The N-terminus of G*13*P could be involved in interactions with the scaffolding protein, G*11*P, like the N-terminus from HK97 gp5 is involved in directing the assembly of the procapsid (40). Additionally, the β8-β9 loop could adopt a position close to the top of the A-domain near helices α7 & α9 and thus close to the centre of the hexamer (or pentamer) as in HK97 (Figure 51b right). A spike of electron density of ~ 15 Å in diameter, attributed to G*12*P, was observed at the centre of the hexamers in decorated SPP1 capsids (41). The potential binding sites for G*12*P on G*13*P are the β8-β9 loop and helices α7 and α9 since they would be the elements closer to the 6-fold axis in the hexamers.

Information concerning conformational changes in G*13*P during viral maturation is not available, although it is known that at the late stages of the assembly, the capsid does not undergo major changes in size (41). Based on the structural rearrangement that HK97 gp5 undergoes during the transition from procapsid to mature capsid it is speculated that the crystallised G*13*P is likely to represent the state found in the mature capsid for two reasons. First, in the HK97 procapsid, the E-loops are closely packed against the A-domains, while in the mature capsid (like in G*13*P) they extend away (Figure 51). The second reason is the conformation of the long helix spanning half of the P-domain. In HK97 procapsids I and II, helix α3' acquires a bent conformation while in the mature capsid it is not bent. In G*13*P, helix α5 is also not bent.

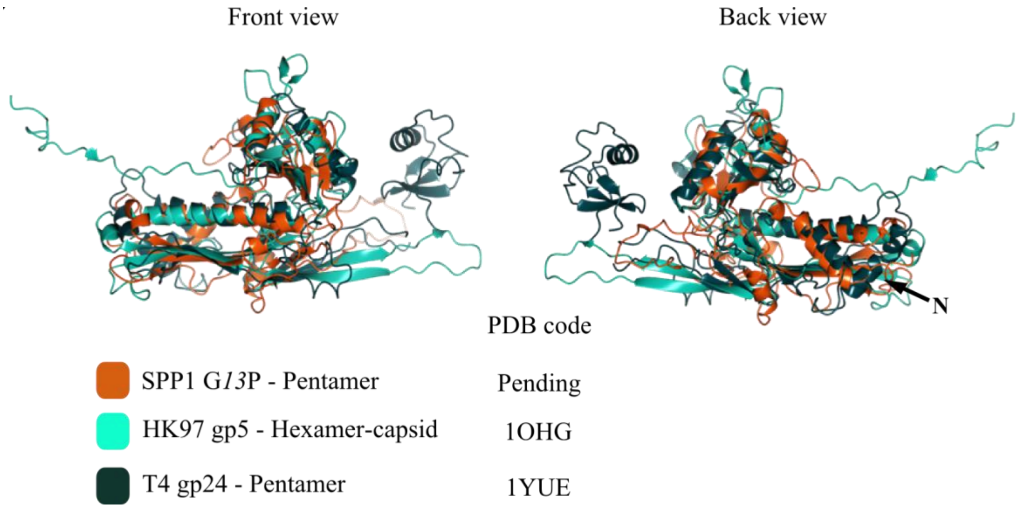


Figure 52. Structural comparison of SPP1 and T4 capsid proteins.

T4 gp24 superposes onto G*13*P with a Cα rmsd of 2.98 Å for 201 residues. The same characteristics at the A- and P- domains that G*13*P shares with HK97 gp5 are also found in T4 gp24. gp24 includes an insertion domain (chitin binding-like domain) absent in G*13*P.

In T4 the insertion domain is suggested to play an analogous role in virus particle stabilisation to the covalent bond between residues K169 and N356 in HK97 since they occupy the same position that the E-loop in HK97, where K169 is located. Moreover, the interface created between the insertion domain and P-domain of adjacent subunits was found to have complementary hydrophobic and polar patches (37). In SPP1 there is no insertion domain analogous to those present in T4 (37), P22 (43) or φ29 (38) capsid proteins, or cementing protein like gpD in λ phage (53) or Soc in RB69 (T4-like) phage (120). Covalent bonding between G*13*P subunits has not been identified. From the PSI-BLAST used to design mutants in this study, it is clear that the HK97 residues responsible for the covalent bond, are absent in SPP1 (Figure 26). In G*13*P's structure the sequence of the region where K169 in gp5 is located, is 72-TDDLV-76 (E-loop), and the sequence of the PL-1, where N356 is located, is 263-GSQDI-267. The mechanism SPP1 uses to keep subunits together to bear the high pressure inside the capsid remains to be investigated. Further information will be available once G*13*P's structure is fitted onto the EM-map of the complete capsid.

A set of capsid-related proteins exhibiting the HK97-fold was superposed onto G*13*P. The Cα rmsd values for the superposition of A-subunits are as follows: Bb_PhRP (3.2 Å for 189 residues), PfV (3.1 Å/173), *T. maritima* encapsulin (4.0 Å/177), *M. xanthus* encapsulin (3.4 Å/180) and Ec_Pcp (3.6 Å/204) (Figure 53). (Not all the structures were included to keep the illustration clear). All structures display an α-helix at the N-terminus and the same overall organisation in A- and P-domains. The region where most fold discrepancies are detected is the E-loop. The E-loop is absent in Ec_Pcp and it is shorter in PfV. Although such loop in Bb_PhRP and *T. maritima* encapsulin is of comparable size to that in SPP1 and HK97, the fold exhibits some differences. In Bb_PhRP there is an extra α-helix which disrupts continuity of the loop and in *T. maritima* encapsulin it protrudes away from the P-domain, sticking out of the pentamer.

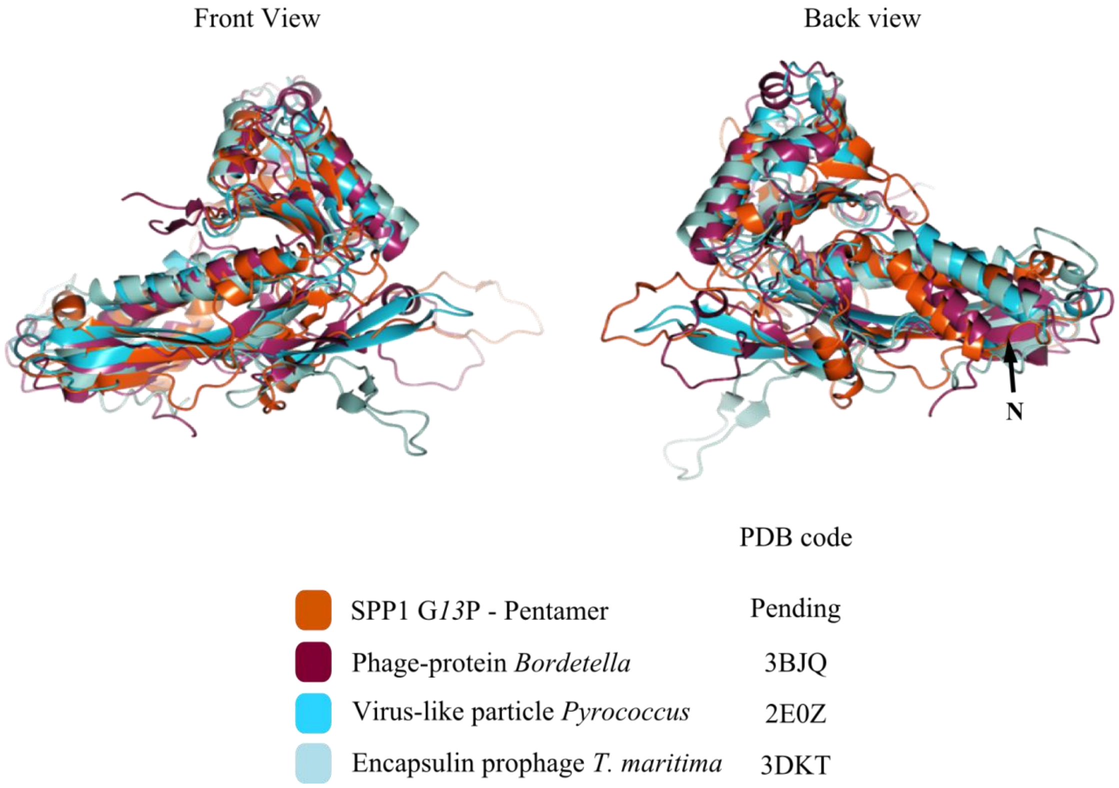

Front View
Back view
N
PDB code
SPP1 G*13*P - Pentamer
Pending
Phage-protein *Bordetella*
3BJQ
Virus-like particle *Pyrococcus*
2E0Z
Encapsulin prophage *T. maritima*
3DKT

Figure 53. Structural comparison of proteins with the HK97-fold

5.3 Structure-based analysis of modifications made to G*13*P

In section 3.3 the logic to design single and double mutants was explained. This approach validated by the G*13*P's X-ray structure since all mutated residues are located where initially predicted. Furthermore, in section 3.3.5 a potential mechanism by which mutations T104Y and A261W prevented aggregation was proposed.

Truncated G*13*P versions illustrate the importance of secondary structure elements in protein solubility and inter subunit interactions. G*13*P_ΔC, which lacked strands β17&β18 and helix α10, was among the completely insoluble constructs. Its reduced solubility suggests that this portion promotes the correct folding of the adjacent elements, e.g. without strand β18, sandwiched between β6 and β12, the A-domain would collapse.

Another subset of truncations at the N-terminus comprising G*13*P_ΔN15, ΔN21 and ΔN35 was insoluble. These constructs lacked a portion or all of the two helices, α1 and α2 (like ΔN35). The N-terminal segment could stabilise helix α5 and the β-strands at the bottom of the P-domain. Constructs G*13*P_ΔN2, ΔN6 and ΔN10, lacking short segments of the polypeptide chain before helix α1 where soluble but aggregated (G*13*P_ΔN2 yield was extremely low after NAC and no SEC was performed), which suggests that the N-terminus is vital for the protein fold, but does not participate in inter-subunit contacts as only double mutants, with no truncations, did not aggregate. The X-ray structure of G*13*P has shown that the truncation of complete secondary structure elements both at the N- or C-termini produced insoluble protein probably because these elements are essential for protein fold.

5.4 The capsid protein as a marker of evolution

dsDNA bacteriophages are the most abundant microorganisms on earth (1). Several studies point out that the HK97-fold is the archetype in the *Caudovirales* order (Figure 7), hence it could be the most widespread fold in viruses (121). This fold has not only been found in viruses infecting members of the *Bacteria* domain but also members of the Archaea (44) and Eukarya domains (122).

The presence of the HK97-fold in Archaea was demonstrated by the finding of virus-like particles in *Pyrococcus furiosus* (44). The HK97-fold has also been identified in the eukaryotic viruses, specifically in the *Herpesvirales* order. From the cryo-EM structure of HSV-1 determined at 8.5 Å resolution (122), it is thought that the floor domain of the capsid protein VP5, has a fold similar to HK97 gp5. This prediction is based on comparable dimensions and capsomers spacing and similar pattern of sequence-predicted secondary structure elements (123). However, since an X-ray structure is available only for the upper domain of VP5 (124), confirmation of the presence of the HK97-fold should await the X-ray structure determination of a full-length VP5.

The HK97-fold is not, however, the only fold found in dsDNA viruses adopted to construct their capsids (Figure 54) (reviewed in references (42,125)). Other common folds include the “viral jelly roll” or “viral β-barrel”, with the PRD1 virus as prototype (126). The BTV-fold (127) is another widespread fold, found in dsRNA viruses. Additionally, ssRNA Sindbis (128) and MS2 (129,130) viruses use capsid proteins with alternative folds to build their capsids.

The fact that capsid proteins of dsDNA phages have a common fold has led to a hypothesis that these viruses have evolved from a common ancestor, although deviations from the fold, like in the case of insertion domains, e.g. T4 (37), P22 (43) and φ29 (38) capsid proteins, or folded N-termini, e.g. SPP1 (this study) and T4 capsid proteins, can occur (42,121,125,131-133).

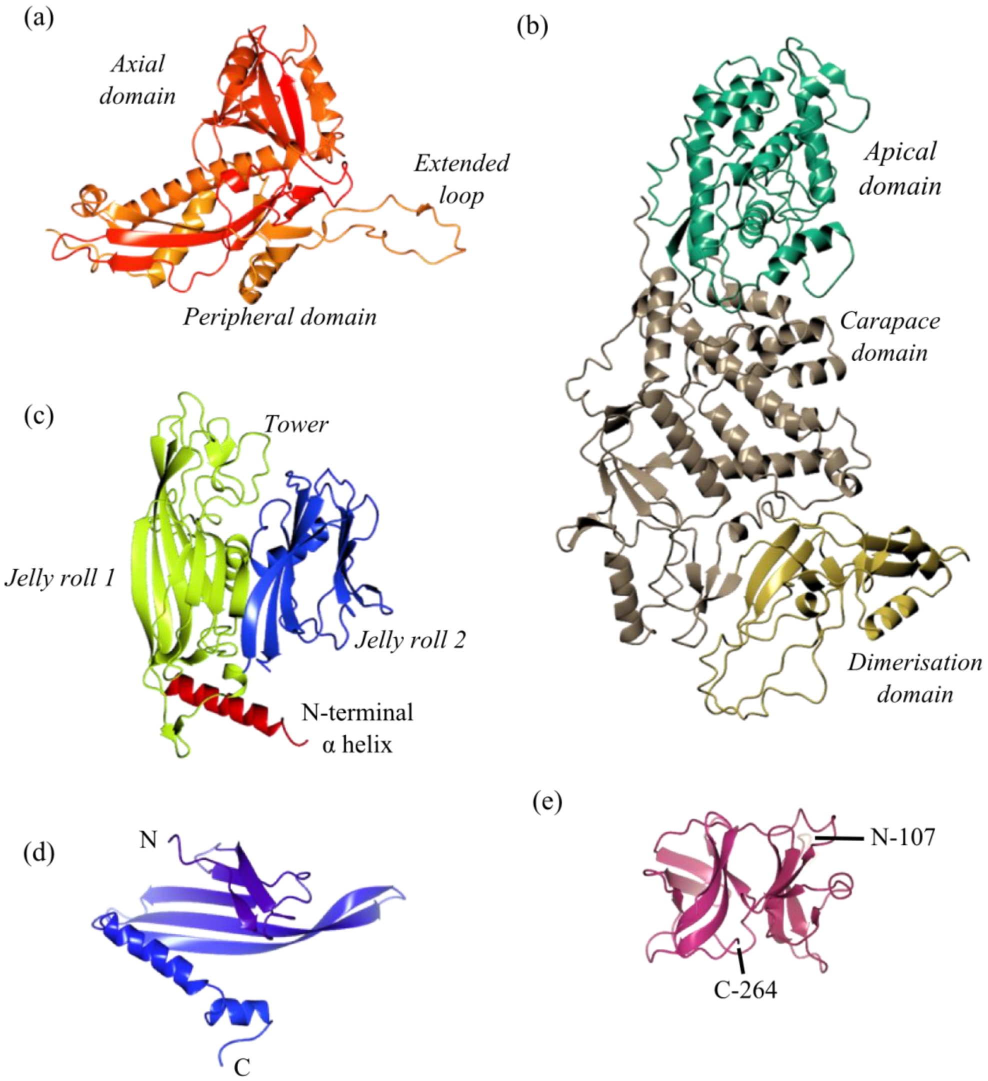


Figure 54. Folds observed in viral capsid proteins

(a) SPP1 G*13*P representing the HK97-lineage. (b) VP3 BTV representing the BTV-lineage, PDB 2BTV. (c) PRD1 P3 representing the PRD1-like lineage, PDB 1HX6. (d) MS2 capsid protein, PDB 2MS2. (e) Sindbis virus capsid protein, PDB 2SNW.

5.5 Conclusions

The X-ray structure of the SPP1 capsid protein, G*13*P revealed that it has the HK97-fold, comprising an axial- (A) and peripheral (P) domains and extended (E-) loop. G*13*P crystallised in pentameric assemblies with the A-domains at the centre of the pentamer and the P-domains at its periphery. The E-loop rests on top of the neighbouring subunit. Both the fold and arrangement of subunits in the crystallised pentamer resemble those adopted by the HK97 capsid protein gp5. The same fold has been observed in T4 capsid protein and capsid-related proteins of encapsulins and virus-like particles found in *Pyrococcus furiosus*.

Structural analysis revealed that the pentamer is stabilised by several H-bonds and salt bridges and that the inner and outer surfaces have a partial negative and positive electrostatic charges, respectively. The produced structure explained why truncated versions of G*13*P were either insoluble or aggregated. The mutagenesis approach carried out to prevent aggregation into aberrant capsids was validated by this structure since both targeted residues where located where predicted on the basis of secondary structure prediction and multiple iteration PSI-BLAST alignment with HK97.

To date, G*13*P is the third capsid protein from a dsDNA bacteriophage for which the X-ray structure has been elucidated. As expected, it exhibits the HK97-fold further confirming the evolutionary relationship between dsDNA bacteriophages and evolution of *Caudovirales* from a common ancestor.

Chapter 6

6. Production of full length and truncated G20C small terminase protein

In order to grow crystals and assess the DNA-binding activity of the WT and truncated ST protein the WT protein and its Se-Met derivative were purified by chromatography. As both WT and Se-Met proteins were purified using identical protocols, only purification of the Se-Met protein (used for structure determination) is described in this chapter. The WT protein was used to perform the DNA-binding experiments. The three truncated versions of the WT protein were designed once the ST's X-ray structure became available, to assess which structural domains are responsible for DNA-binding. Purification and crystallisation of these truncated versions is described in this chapter.

6.1 Production of the G20C WT small terminase

The following paragraphs describe how the purification of the Se-Met ST_WT, used for crystallographic studies, was carried out. The native protein was purified in the same way with the exception that AEC was omitted. The protocol by which the native G20C ST_WT was purified and crystallised has been published (134).

It was possible to locate the ST gene based on sequence homology with the hypothetical STs from other thermophilic phages. The ST is usually encoded by a gene immediately preceding the LT and portal protein genes (motor operon). Such a gene has been identified in the G20C genome, with a size typical for a ST protein. This gene corresponds to the ORF P23p84 (UniProtKB/TrEMBL A7XXB6) and to ORF P74p83 (UniProtKB/TrEMBL A7XXR0) in the closely related phages P23-45 and P74-26, with which G20C ST shares 99 and 98 % sequence identity, respectively. (Appendix **10**) (7).

The full-length gene encoding the hypothetical G20C ST was cloned into the expression vector pET28a between NdeI and HindIII restriction sites. B834 cells were transformed with the recombinant DNA and grown in Se-Met medium at 16°C to produce the POI at large scale. The induction was made with 1 mM IPTG.

To proceed with NAC, the pelleted bacteria were resuspended in 25 mM Tris pH 7.5, 500mM NaCl, 10 mM imidazole and sonicated according to the described protocol. The NAC profile consisted of two peaks (Figure 55). Peak 1 contains numerous proteins that bound to the nickel beads with low affinity and a small amount of the ST_WT protein. Peak 2 contains less contaminants and the ST_WT is present in considerably higher amount than in peak 1. Fractions from peak 2 were pooled together to remove the N-terminal His-tag by thrombin protease-cleavage; at the same time the sample was buffer exchanged to remove imidazole. A second NAC was performed to separate the digested and non-digested samples. The digested sample, which included no tag, eluted in the flow-through of the purification. This flow-through was buffer exchanged to low salt-buffer in order to perform AEC (Figure 55b). The adsorbed sample was eluted with 25 mM Tris pH 7.5, 1 M NaCl. The AEC profile contains a single peak. SDS-PAGE analysis of this major peak revealed that the sample still contains some contaminants. A final separation by size was made to remove other contaminants. The SEC profile shows one main specie eluting at 62 mL (Table 14) (Figure 55c). Fractions from SEC were pooled together and concentrated to setup crystallisation experiments (section 6.3.1).

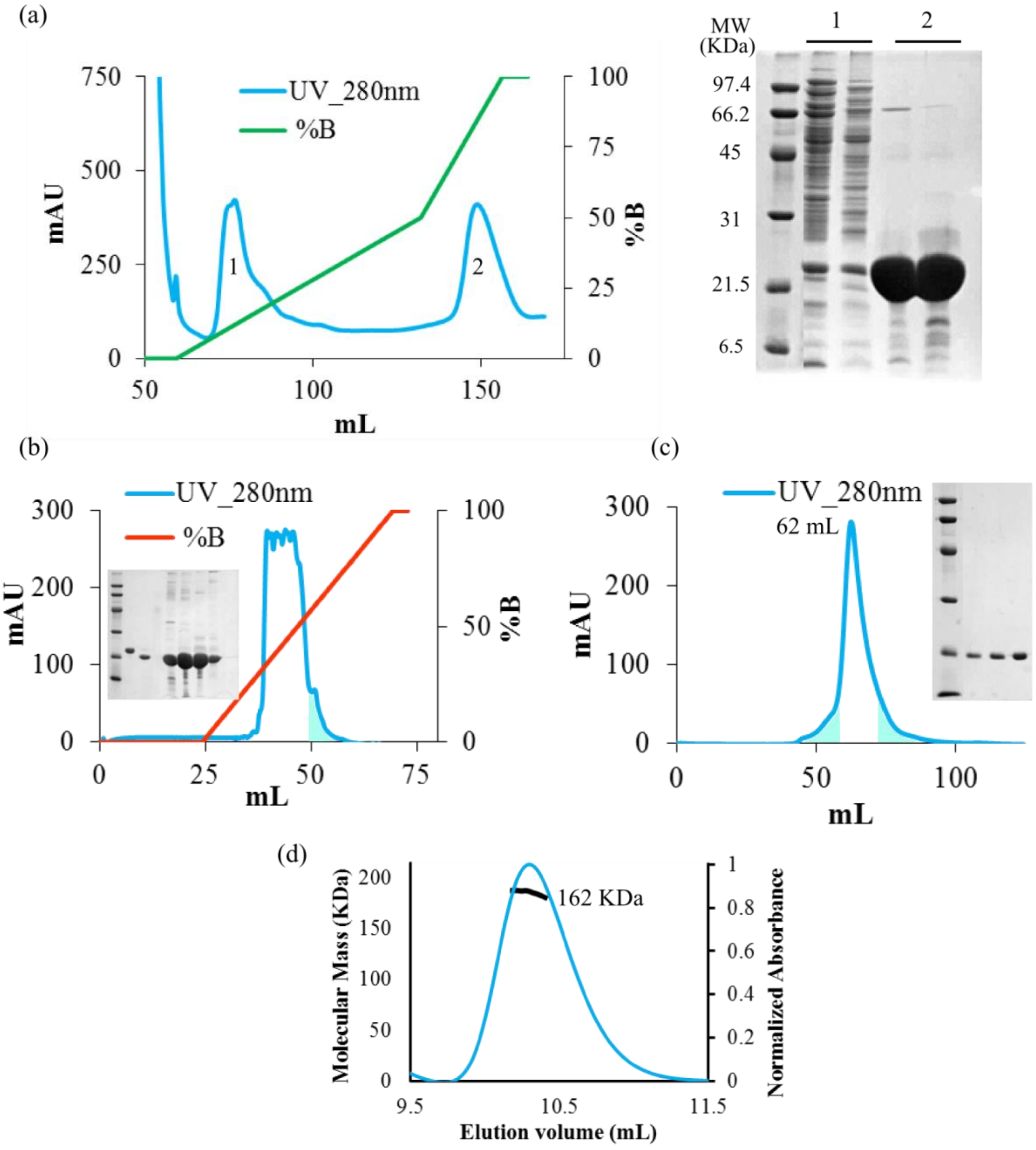


Figure 55. Purification of ST_WT

(a) NAC profile and its corresponding SDS-PAGE. Only few fractions from peaks 1 and 2 are shown. (b) AEC for thrombin digested ST_WT sample. (c) SEC using a S200 16/60 column. Shaded areas were not concentrated. For (b) and (c) gels illustrate how the purification process evolved at every stage. (d) SEC-MALLS for ST_WT.

ST_WT was loaded onto a SEC-MALLS column (Figure 55d) to determine MW. The calculated MW (167 kDa) is in good agreement with the expected MW of a 9-subunit oligomer (171.7 kDa). A comparison of oligomerisation states between ST will be made in Chapter 8.

Table 14. SEC and SEC MALLS of ST_WT and constructs

Construct	Elution volume 16/60 columns (mL)	Elution volume 10/30 columns (mL)	Expected MW (kDa)	Determined MW SEC-MALLS (kDa)
ST_WT	62	10.4	19.1	162.7/9 = 18.1
ST_ΔC	74	12	15.8	133.1/9= 14.8 133.1/15.8= 8.4
ST_ΔN	67	10.6	13.1	120.4/9= 13.4
ST_NTD	82.3	17.4	6.2	7

6.2 Design and purification of ΔC, ΔN and NTD constructs

Several truncations based on the G20C ST_WT X-ray crystal structure were introduced to investigate the effect of removing individual domains or segments on DNA-binding compared to the ST_WT (Figure 56). Construct ST_ΔC (residues 1-138) included the NTD and OD, but it lacked the sequence-predicted α-helix that was disordered in the structure. Construct ST_ΔN (residues 54-171) included the predicted C terminal α-helix and OD. The third construct consisted of the NTD only (residues 1-52). The properties of each construct can be found in the appendix **3**.

ST_ΔC and ST_NTD constructs were produced by introducing a stop codon in primers so that both constructs remained in the same pET28a vector, containing the same restriction sites as the WT construct with an N-terminal His-tag (Appendix **1**). Construct ST_ΔC was amplified by PCR and inserted into the LIC-3C vector that also includes an N-terminal His-tag cleavable by the 3C protease.

All constructs were expressed in large scale using B834 cells at 16°C, with induction by 1 mM IPTG. The first NAC was performed in 25 mM Tris pH 7.5, 250mM NaCl. The eluted samples were digested with either thrombin or 3C proteases to carry out a second NAC followed by SEC (Figure 56b). The constructs ST_ΔC, ΔN and NTD eluted as one main species at ~ 74, 7 and 82.3 mL, respectively (Table 14).

Intriguingly, ST_ΔC and ST_ΔN did not migrate in the SEC column as expected. Assuming that both constructs keep the same oligomerisation state than the WT, nine copies of ST_ΔC (9 x

15.8 kDa = 142.1 kDa) are expected to elute faster than nine copies of ST_ΔC (9 x 13.1 Da= 118.2 kDa) as the latter would be delayed for longer because of its smaller size.

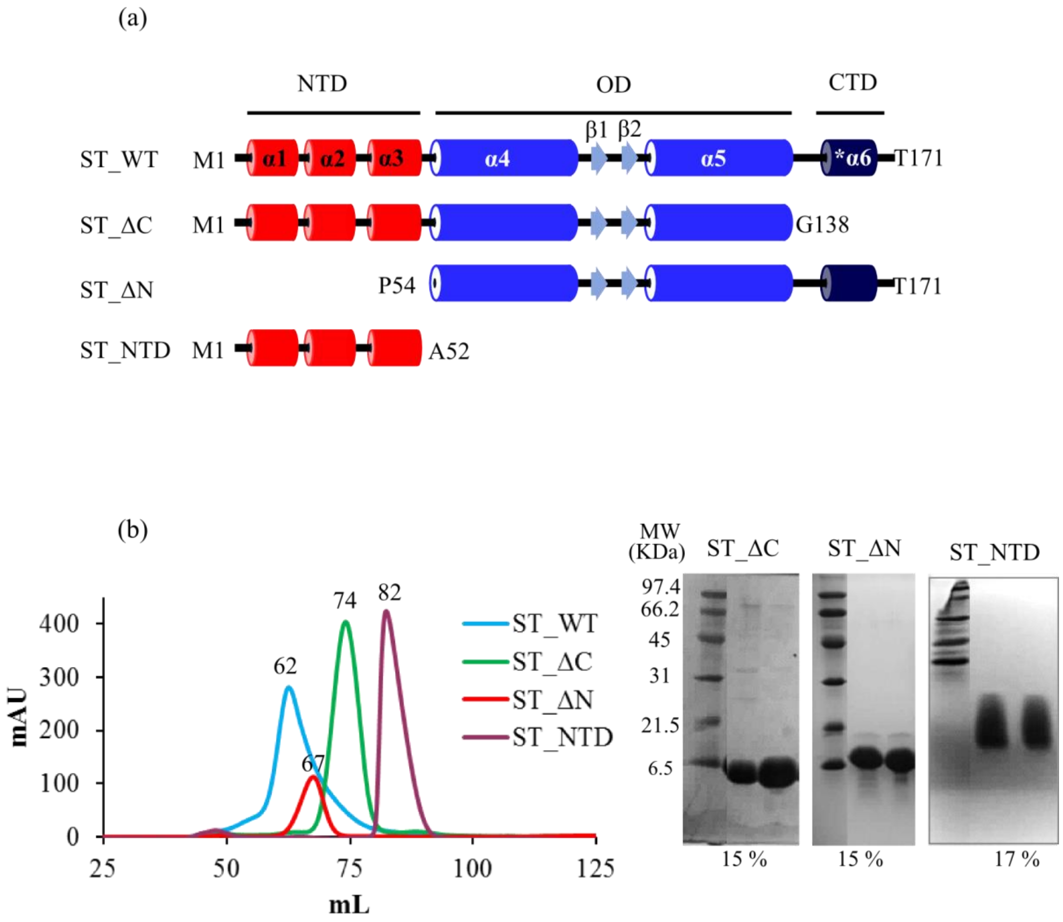


Figure 56. Constructs of G20C ST that have been purified

(a) Secondary structure of ST constructs. The first and last residues of every construct are indicated. Star (*) corresponds to the JPRED prediction of α6-helix. (b) SEC performed in a S200 16/60 column for ST_WT, ST_ΔC, ST_ΔN and a S75 16/60 column. Fractions from every run are displayed at the right side of the chromatogram. The percent (w/v) of polyacrylamide is indicated below the gels.

SEC was repeated for all the constructs in a S200 10/30 column equilibrated with 20 mM Tris pH 7.5, 1 M NaCl to further assess their behaviour (Figure 57a). The NTD eluted at 17.4 mL. ST_ΔN (12 mL) eluted later than ST_ΔC (10.6 mL) suggesting a change in the conformation or oligomerisation state of ST_ΔC.

Fractions from the run in the S200 10/30 SEC column were loaded onto a 15 % (w/v) polyacrylamide gel. While the WT protein migrated similarly to the 21.5 kDa standard, ST_ ΔC and NTD migrated faster than the 6.5 kDa standard, as expected for smaller proteins. More intriguing is the finding that even in polyacrylamide gels ST_ ΔC behaved as a protein smaller than ST_ ΔN.

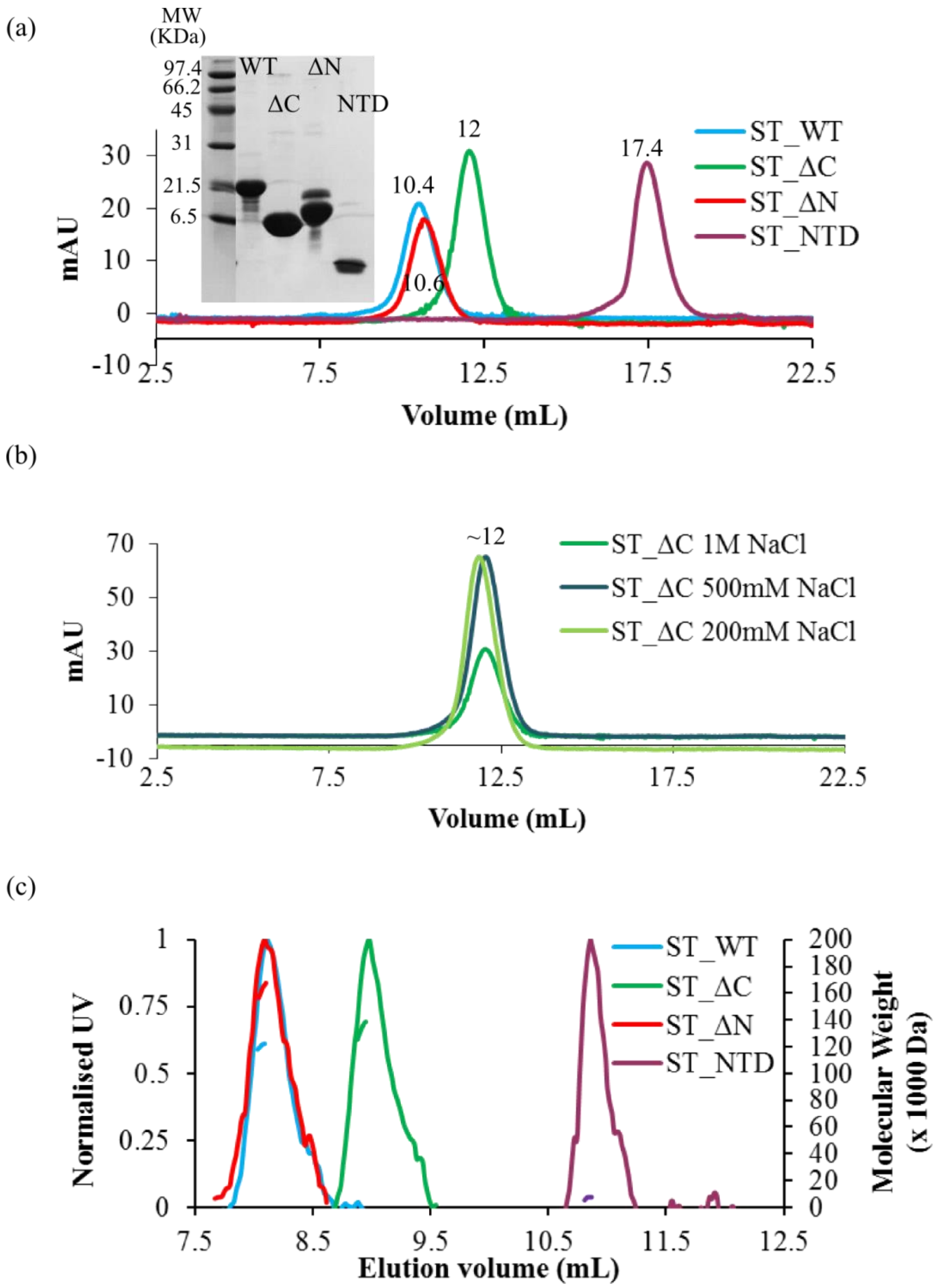


Figure 57. Characterisation of G20C ST constructs

(a) Analytical SEC in a S200 10/30 column. Fractions from individual runs were analysed in SDS-PAGE (left side). (b) Analytical SEC in varying NaCl concentration for ST_ΔC. (c) SEC-MALLS for ST_WT and constructs.

Another round of analytical SEC was carried out to investigate whether the oligomerisation state of ST_ΔC changed in response to the salt concentration (Figure 57b). The reasoning to test this approach is the observation that the removal of 33 residues (ST_ΔC) had a significant impact in the elution profile in comparison with ST_WT, while the removal of 53 residues (ST_ΔN) did not.

Figure 57b shows that the oligomerisation state of ST_ΔC did not change when eluted in 250 mM, 500 mM or 1 M NaCl since the elution volume of the construct is ~12 mL in all tested conditions. Considering that the oligomerisation state of ST_ΔC did not change with increasing salt concentrations, it is assumed that ST_ΔN would not undergo any changes under the same conditions. Therefore the described SEC experiment was carried out only with ST_ΔC.

Other approaches were used to study constructs ST_ΔC and ST_ΔN in depth. The first experiment performed consisted of the analysis of the recombinant DNAs to rule out the possibility of sample-swapping (Figure 58). ST_ΔC was cloned by introducing a stop codon after residue 138, therefore the 5' and 3' ends remained identical to the ST_WT. Therefore, using the primers used to amplify the WT protein it would be possible to amplify a fragment that differs in size to the WT gene by only one codon. ST_ΔN lacks the first 53 codons at the 5' end, hence using the primers for the ST_WT would not lead to the amplification of any product (Figure 58a). It was demonstrated that the DNA samples were not swapped during cloning since using the mentioned primers, it was possible to amplify a ST_ΔN fragment of the same length as the ST_WT (Figure 58b). Conversely, the same primers failed to amplify any product using the ST_ ΔN DNA as template.

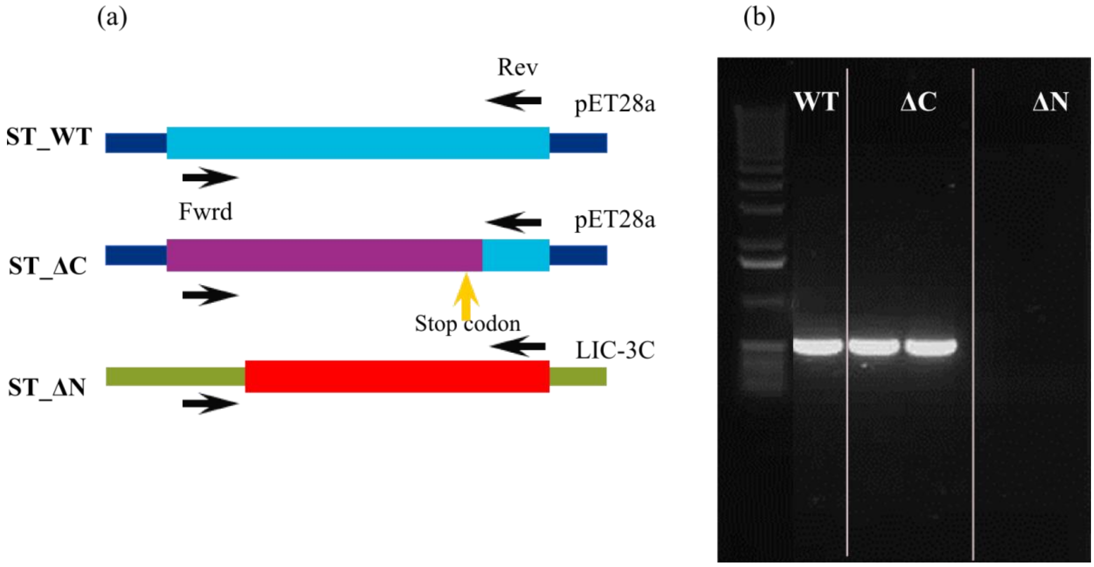


Figure 58. Investigation of ST_ ΔC and ST_ ΔN recombinant DNAs

(a) General approach to amplify a WT-size gene. (b) PCR of ST_ΔC and ST_ΔN using full length primers. The ST_WT was included as positive control.

The ST_ΔC and ST_ΔN constructs were analysed by electrospray ionisation (EI) mass spectrometry to calculate their MW (per subunit) and estimate the agreement with the predicted one. Calculation of the MW allowed determining whether the constructs were correctly expressed and purified or degrading. The calculated MW for ST_ΔC and ST_ΔN were 15 782

Da and 13 136 Da, respectively, both in good agreement with the predicted 15 785 Da for ST_ΔC and 13 138 for ST_ΔN.

SEC-MALLS was carried out to determine the MW of all ST constructs in solution and deduce the number of subunits (Figure 5.3c). For ST_WT and ST_ΔN, the calculated MW agreed well with that expected for a nine-subunit assembly (Table 14). The calculated MW of the NTD corresponded to that expected for a monomer. The estimated MW of ST_ ΔC was 133.1 kDa, which divided by nine potential subunits, yields 14.8 kDa per subunit. The calculated MW of the ST_ ΔC oligomer accounts for 8.4 copies, when divided by 15.8 kDa (one subunit of ST_ΔC). This technique suggests that ST_ΔC exists in solution as either eight- or nine-subunit assembly.

6.3 Crystallisation of ST constructs

6.3.1 Crystallisation of Se-Met ST_WT

Se-Met crystals of ST_WT were grown using microseeding. Initially, native protein crystals grew in condition G8 from the INDEX screen (0.1 M Hepes pH 7.5, 0.2 M Ammonium acetate, 25 % (w/v) PEG 3.35 K) after two weeks of setup. The protein concentration was adjusted to 15 mg mL^{-1} with 20 mM Tris pH 7.5, 250 mM NaCl. Further optimisation experiments produced a variety of crystals in four similar conditions. The crystals from different drops were mixed together in 0.1 M Hepes pH 7.5, 0.4 M ammonium acetate, 30 % PEG to produce a seed stock.

The conditions from which the crystals were fished were:

a) 0.1 M Hepes pH 7.5, 0.4 M ammonium acetate, 30 % (w/v) PEG 3.35 K.
b) 0.1 M Hepes pH 7.5, 0.3 M ammonium acetate, 25 % (w/v) PEG 3.35 K.
c) 0.1 M Hepes pH 7.5, 0.2 M ammonium acetate, 24 % (w/v) PEG 3.35 K.
d) 0.1 M Hepes pH 7.5, 0.4 M ammonium acetate, 27 % (w/v) PEG 3.35 K.

The seed stock was used to setup hanging drop vapour diffusion experiments in 24-well trays. Se-Met protein purified by SEC was diluted to 21 mg mL^{-1}. Hexagonal crystals of ~ 100 µm wide x 200 µm high grew in 0.1 M Hepes pH 7.5, 0.4 M ammonium acetate, 18 % (w/v) PEG3.35 K after two weeks of incubation at room temperature (Figure 59). The crystals were clustered in pairs, joined by their ends with an angle of ~140 °. The crystals were separated by breaking the contact surface and then cryo-cooled in liquid nitrogen using no cryo-protectant solution. The crystals diffracted to 3.28 Å resolution when tested with in-house equipment. The data collection was carried out at the I02 beamline at the Diamond Light Source, UK.

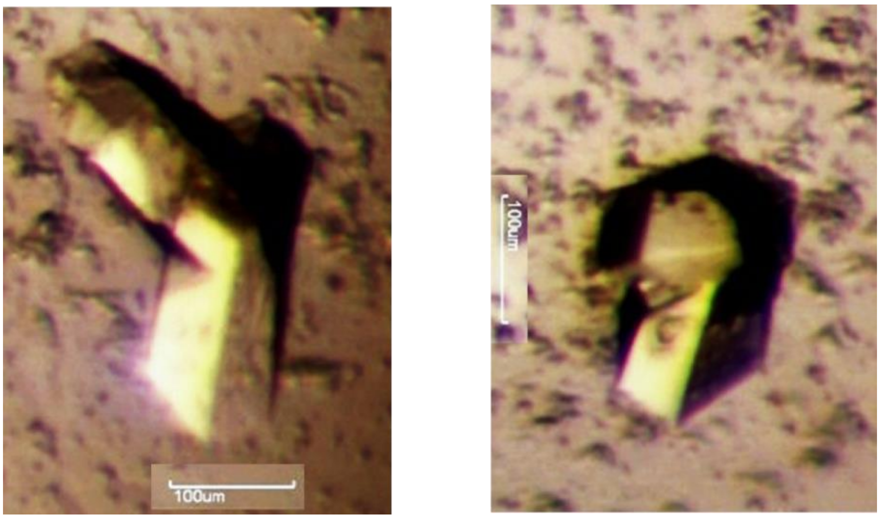


Figure 59 G20C ST_WT Se-Met crystals

6.3.2 Crystallisation of ST_ΔC and ST_ΔN

Constructs ST_ΔC and ST_ΔN were diluted to 15 mg mL^{-1} to setup sitting drop vapour diffusion experiments using the commercial screens PACT and INDEX. Well shaped crystals of ST_ ΔN were observed after 3 days of incubation in conditions numbered 25 and 37 from the INDEX and PACT screens, respectively (Figure 60). The shape of the crystals belonged predominantly to the cubic (Figure 60a) and triclinic systems (Figure 60c). Clusters of crystals, long needles and thin plates were also present in other conditions.

After the crystals were observed and fishing was attempted, it was noticed that some crystals disappeared, suggesting that changes in temperature (from 18°C to room temperature) contributed to dissolve them.

Crystals in Figure 60a grew in condition G10 from the INDEX screen: 0.2 M Na citrate, 20 % (w/v) PEG 3.35 K. These crystals dissolved after being observed under the microscope. Crystals from Figure 60b grew on condition H7, INDEX screen: 0.15 M DL-Malic acid pH 7, 20 % (w/v) PEG 3.35 K. crystals grown on this condition diffracted up to 10 Å resolution when tested in-house. No cryo-protectant was utilised. The crystals from Figure 60c were grown in 0.2 M $NaNO_3$, 20 % (w/v) PEG 3.35 K, conditions from the PACT screen, condition E5. Cryo-protectant solution containing 20mM Tris pH 7.5, 100 mM NaCl, 0.2 M $NaNO_3$, 32 % (w/v) PEG 3.35 K was used to fish these crystals. Diffraction to 8Å was observed. The crystals in fFigure 60d grew in condition F9, PACT screen: 0.1 M Bis-Tris propane pH 6.5, 0.2 M Na/K tartrate, 20 % (w/v) PEG 3.35 K. Crystals fished from this condition diffracted to ~ 4.1 Å resolution. Cryo-protectant solution containing 0.1 M Bis-tris propane pH 6.5, 20 mM Tris pH 7.5, 0.2M Na/K tartrate, 0.1 M NaCl, 20 % (w/v) PEG, 20% glycerol was used.

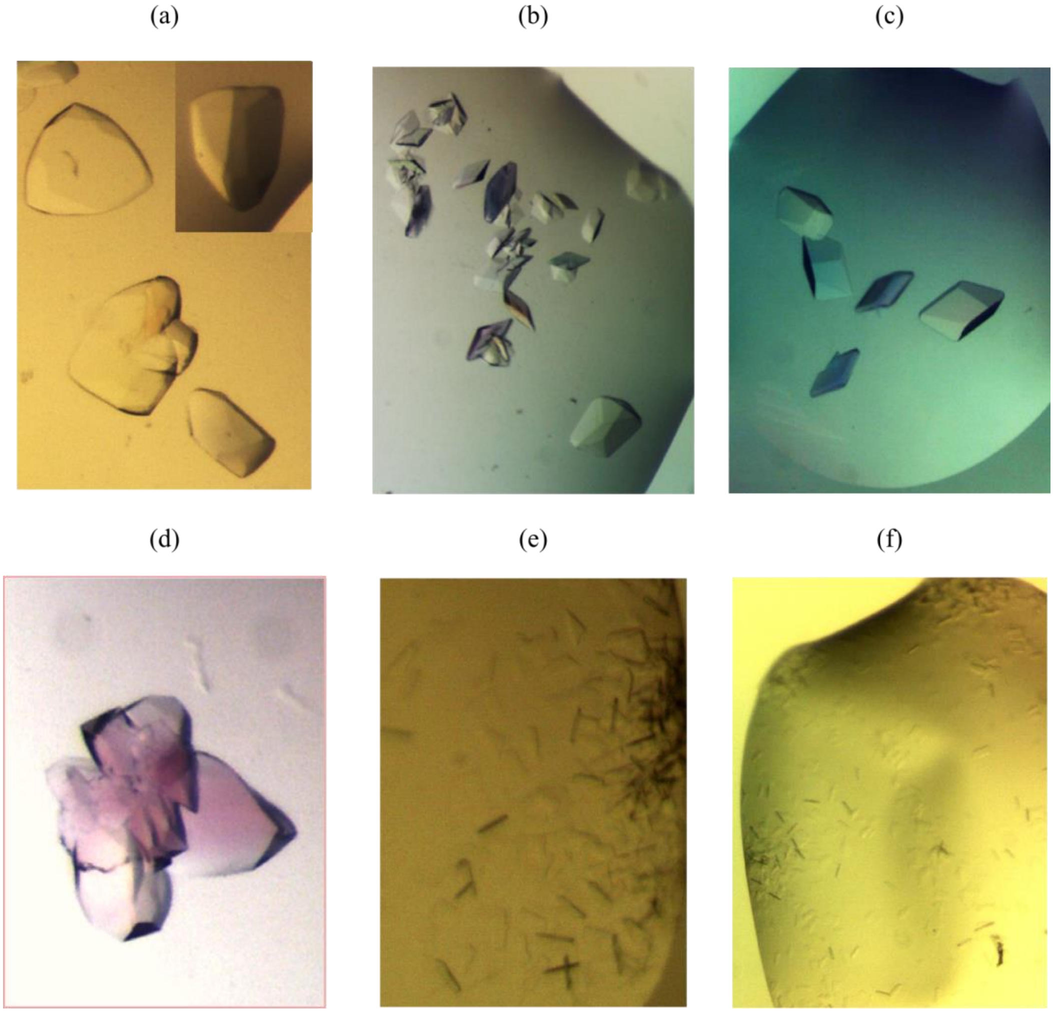


Figure 60. Crystals of G20C ST_ΔC and ST_ΔN

Crystals of ST_ ΔC in (a) INDEX G10, (b) INDEX H7, (c) ST_ ΔC , PACT E5 and (d) PACT F9. (e),(f) ST_ ΔN in INDEX C9. Crystals were colourless; the colour observed is due to the light polariser used to take the pictures.

Small needles of ST_ΔN were observed after 8 weeks of setup in condition C9 from the INDEX screen: 0.1 M Hepes pH 7.0, 1.1 M Na Malonate pH 7, 0.5 % (v/v) Jeffamine ED-2001® pH 7. The needles were too small to attempt fishing to further expose them to X-rays, therefore they were used as seeds for subsequent crystallisation experiments.

6.4 Conclusions

Se-Met ST_WT was purified by a variety of chromatographic techniques to carry out crystallisation experiments. Crystals that diffracted to ~ 3.3 Å resolution were grown optimising those conditions that were shown to produce crystals of native protein. It was determined that G20C ST exists as nine-mer in solution by SEC-MALLS.

Three constructs of G20C ST were purified to perform DNA-binding activity assays: ST_ΔC, ST_ΔN and NTD. Even though it was observed that ST_ΔC migrates as if it was smaller than ST_ΔN in SEC and SEC-MALLS columns and in SDS-PAGE, it was confirmed by SEC-MALLS and EI-MS that the MW of both constructs corresponds to the predicted one. The reason of this behaviour is likely to lie in the presence or absence of the C-terminal domain, as the NTD is not expected to participate in inter-subunit contacts that determine the oligomerisation state. Removal of the C-terminus, which is highly negatively charged from the WT protein, may induce conformational changes on ST_ΔC that lead to retarded elution in comparison with ST_ΔN.

ST_ΔC was the subject of extensive experimental work because of the difference in MW observed in comparison with the WT protein (~ 6 kDa). The consideration that nine subunits are contributing to the difference in MW does not explain such difference in the SEC profiles. The crystallisation of ST_ ΔC and ST_ ΔN was carried out in order to gain insight into the organisation of such constructs. Crystallisation of these constructs is in progress.

Chapter 7

7. Elucidation of structure of G20C small terminase

7.1 Native protein crystals

7.1.1 Data collection

X-ray data were collected from a single cryo-cooled crystal at the I04-1 beamline at the Diamond Light Source, UK. Data were collected at a wavelength of 0.9200 Å with a crystal-to-detector distance of 325.16 mm, 0.2° crystal rotation per image and 180° total crystal rotation (Figure 61). Crystal data are published in reference (134).

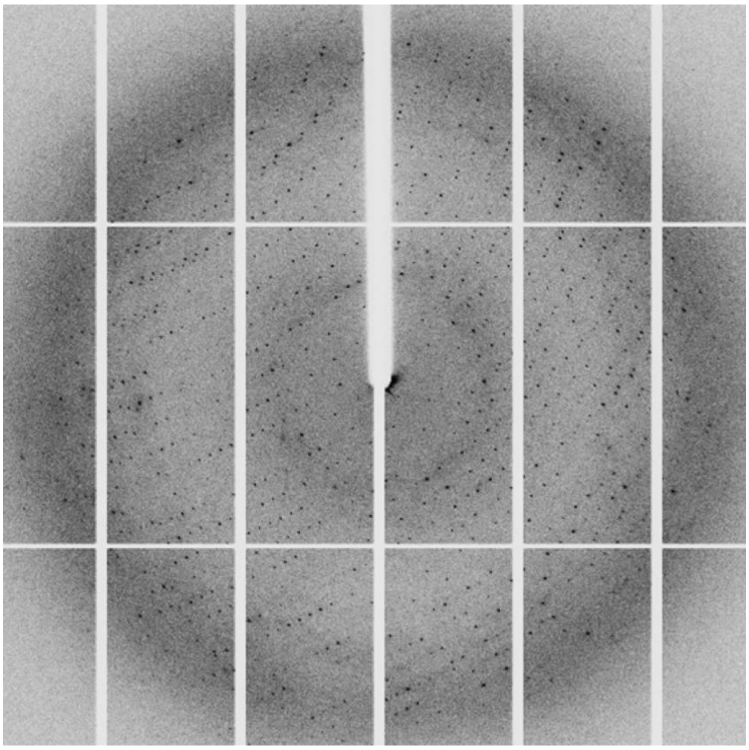


Figure 61. Diffraction pattern of a native crystal of ST_WT

7.1.2 Structure determination

The 900 diffraction images were indexed using reflections in the resolution range 49.8-2.80 Å using the XDS (107) software. The crystal belonged to the *P* 2_1 2_1 2 space group with unit cell dimensions *a*= 94.3 Å, *b*= 125.6 Å and *c*=162.8 Å, and angles α, β, γ = 90°.

Table 15 summarises the statistics of the X-ray data set.

The calculated solvent content (119) for nine subunits in the asymmetric unit was 56.1 %.

At the time the diffraction images of the native crystal were collected, X-ray crystal structures for SF6, P22 and 44RR ST were available. However, as it will be explained in Chapter 8, even though they share similar domains, the spacing between domains varies. MR did not work in this case and hence the Se-Met version of G20C ST_WT was produced.

7.2 Se-Met protein crystals

7.2.1 Data collection

900 images were collected at the I02 beamline at the Diamond Light Source, UK from a flash-cooled crystal at a wavelength of 0.97950 Å (Figure 62, Table 15). The crystal was rotated by 180° with 0.2° per image. The chosen wavelength corresponds to the energy absorption peak of Selenium atoms: 12,657.5 eV, as derived from a fluorescence scan performed before data collection (Figure 62a). Faint ice rings were observed at around 3.2 Å while strong diffraction spots corresponding to the protein crystal, were visible up to ~ 2.8 Å resolution.

7.2.2 Structure determination

Processing of diffraction images was carried out with XDS (107). XDS indexed reflections in the 45.39-2.51 Å resolution range in the space group H 3 (Table 15). The unit cell dimensions are *a*, *b* = 152.6 Å; c = 108.8 Å and angles α, β = 90, γ = 120° (hexagonal setting).

The values for Rmerge and Rmeas values of the 32,286 unique reflections calculated by the Aimless software were 8.5 (138.2) % and 9.7 % (156.4) %, respectively (Table 15).

SHELX*C/D* were used to identify and locate the selenium atoms (Figure 63). SHELX*D* produced a list of coordinates featuring 28 heavy atoms in the asymmetric unit. The coordinates of the 21 atoms with the highest occupancies (above 0.55) were separated for subsequent steps.

Table 15. X-ray data statistics for G20C ST

	Native crystal	**Se-Met crystal**
X-ray source	Beamline I04-1, Diamond LS, UK	Beamline I02, Diamond LS, UK
Detector	Pilatus 2M	Pilatus 6M
Wavelength (Å)	0.92000	0.97950
Space group	$P\ 2_1\ 2_1\ 2$	*H* 3
Cell dimensions (Å)	*a = 94.3, b* = 125.6 *c* = 162.8 α, β, γ = 90°	*a*, *b* = 152.6; *c* = 108.8 α, β = 90, γ = 120°
Resolution range (Å)	49.8-2.8 (2.89-2.80)	45.39-2.51 (2.61-2.51)
No. of unique reflections	48371 (4371)	32286 (3586)
$R_{merge}{}^{\dagger}$ (%)	9.6(153.2)	8.5 (138.2)
R_{meas} (%)	10.4 (167.7)	9.7 (156.4)
Average *I*/σ(*I*)	16.3 (1.3)	12.6 (1.2)
Completeness (%)	99.9 (100.0)	99.7 (98.1)
Multiplicity	6.8 (6.1)	4.6 (4.6)

Table continues on the next page.

Table 14. X-ray data statistics of G20C ST. Continuation

	Native crystal	**Se-Met crystal**
CC (1/2)	0.999 (0.510)	0.996 (0.368)
CC (1/2) DelAnom	NA	0.671 (0.014)
Wilson B-factor (Truncate) (Å ^{2})	62.3	63.9
Number of atoms	9718 (9 chain x 137 residues)	6420 (3 chains x 137 residues, 3 chains x 143 residues)
Solvent content (%)	56.1	43.1
Number of reflections used in refinement		31224
R-factor (%)		21.9 (36.5)
Number of reflections used for R_{free}		1056
R_{free} (%)		28.5 (37.4)
Average atomic B-factor (Å ^{2})		81.0
Rms BondLength (Å)		0.010
Rms BondAngle (°)		1.4

Values in parentheses are for the highest resolution shell.

† $R_{merge} = \Sigma_{hkl} \Sigma_i |I_i(h) - \langle I(h) \rangle| / \Sigma_{hkl} \Sigma_i I_i(h)$, where $I(h)$ is the intensity of reflection h, $\langle I(h) \rangle$ is the average value of the intensity, the sum Σ_{hkl} is over all measured reflections and the sum Σ_i is over i measurements of a reflection.

Subsequent data-processing involved density modification using the Parrot software (113). The output .mtz file and the amino acid sequence were used in the Buccaneer pipeline to perform model building after experimental phasing (114). Buccaneer built 726 residues, allocating 662 of them in 6 nearly identical chains. The completeness of the model in terms of built residues was 84.4 % and the completeness of the chains 63.4 %.

The chains were nearly identical, varying in length due to the absence of electron density at the C-terminus (see next chapter).

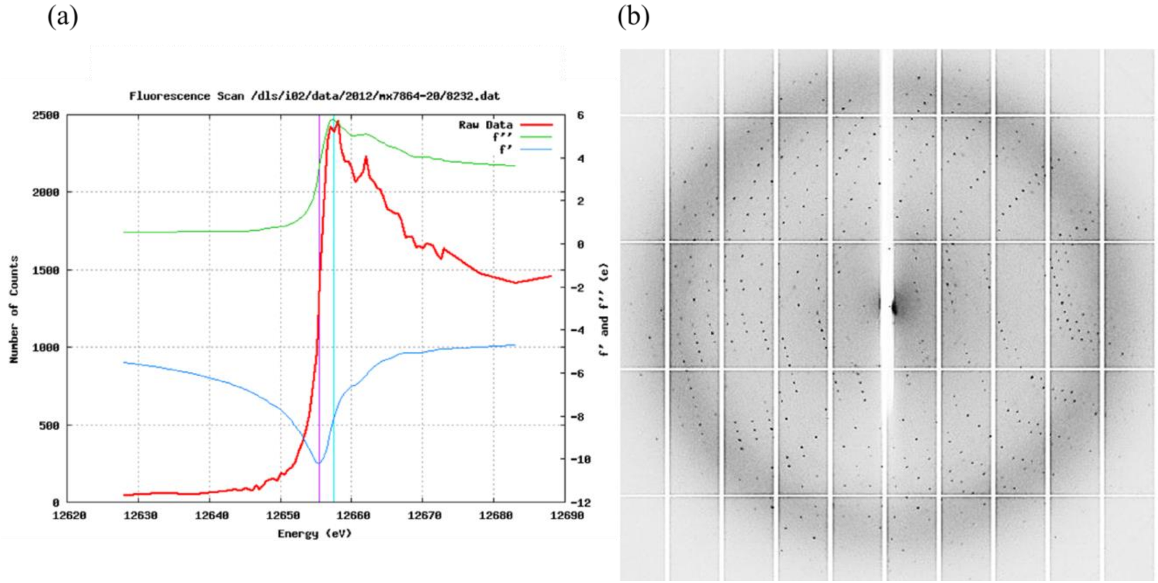
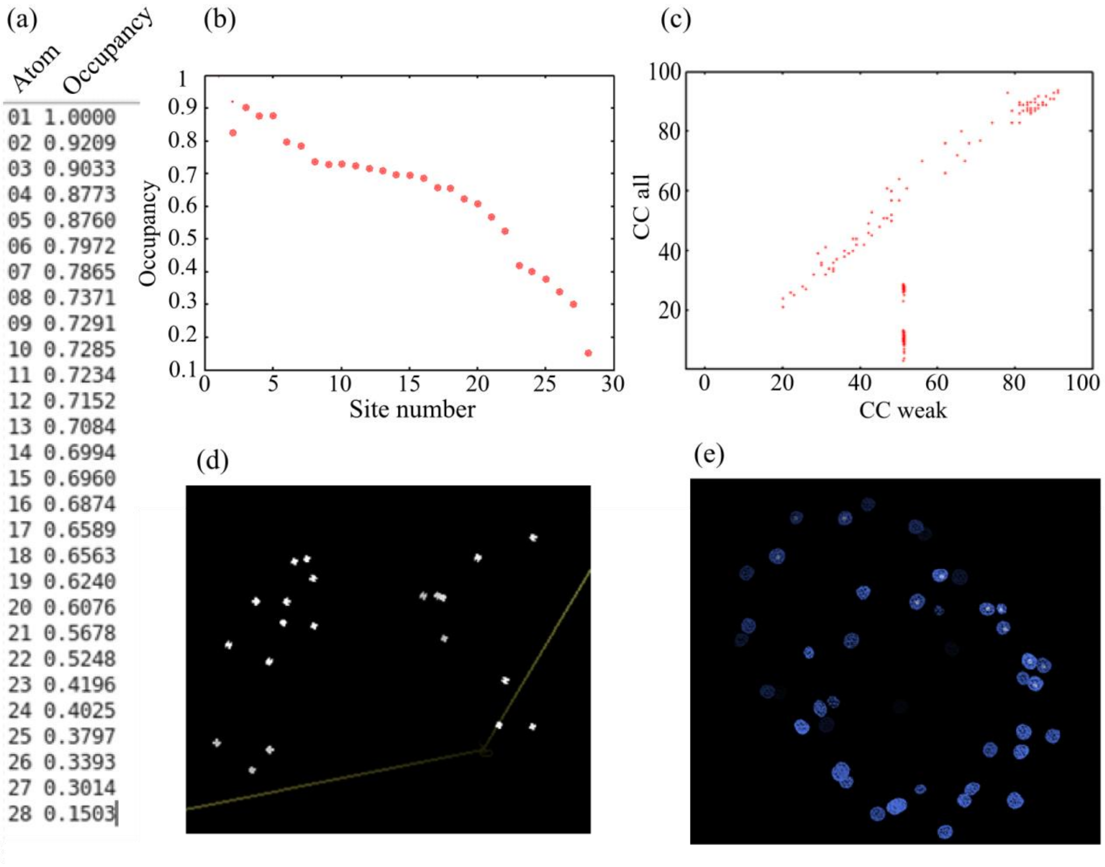


Figure 62. Data collection from G20C ST_WT Se-Met crystals

(a) X-ray fluorescence scan. The peak (cyan line) and inflection (pink line) wavelengths are indicated. (b) Diffraction image. The resolution at the edge of the image is ~ 2.5 Å.



Atom	Occupancy
01	1.0000
02	0.9209
03	0.9033
04	0.8773
05	0.8760
06	0.7972
07	0.7865
08	0.7371
09	0.7291
10	0.7285
11	0.7234
12	0.7152
13	0.7084
14	0.6994
15	0.6960
16	0.6874
17	0.6589
18	0.6563
19	0.6240
20	0.6076
21	0.5678
22	0.5248
23	0.4196
24	0.4025
25	0.3797
26	0.3393
27	0.3014
28	0.1503

Figure 63. Location of heavy atoms by SHELX*C*/*D*

List (a) and graph (b) of occupancies of the heavy atoms identified by SHELX*D*. In (b) the horizontal and vertical axis plot the site number and occupancy, respectively. (c) Substructure solutions by SHELX*D.* (d) and (e) heavy atoms coordinates.

The initially built chains belonged to different oligomers. Individual chains were re-grouped into two groups of three chains each. Complete oligomers, containing 9 subunits are generated by the crystallographic 3-fold axis

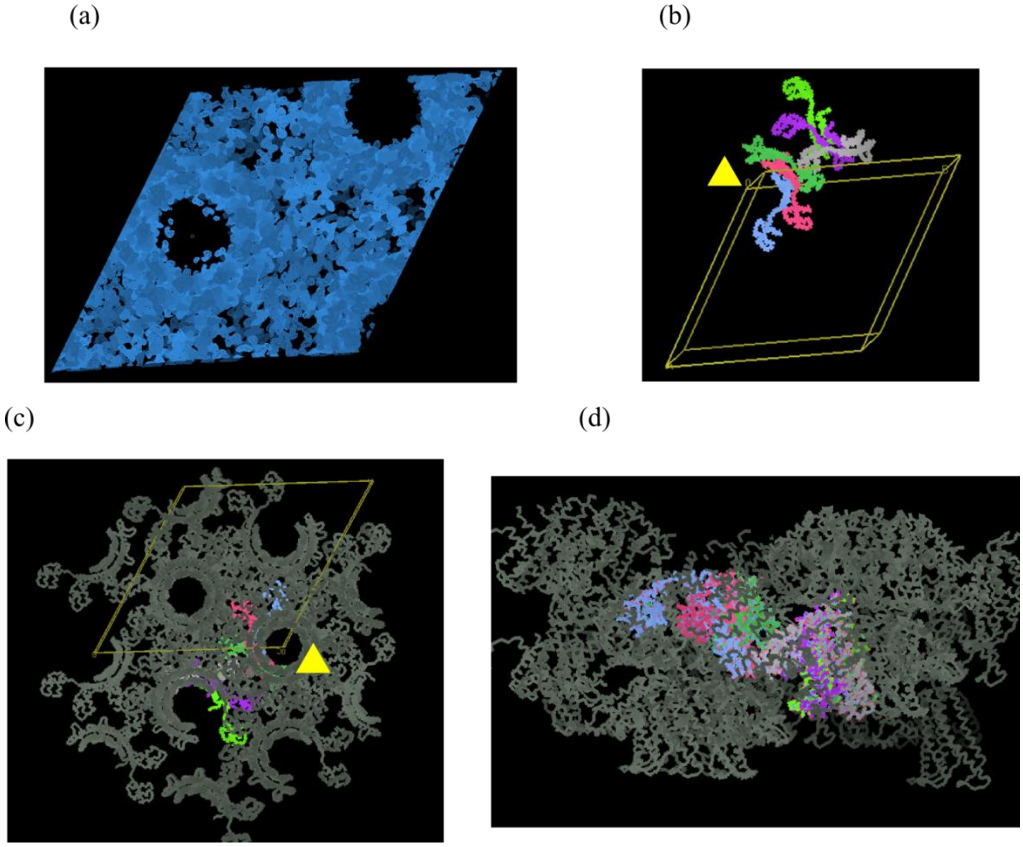
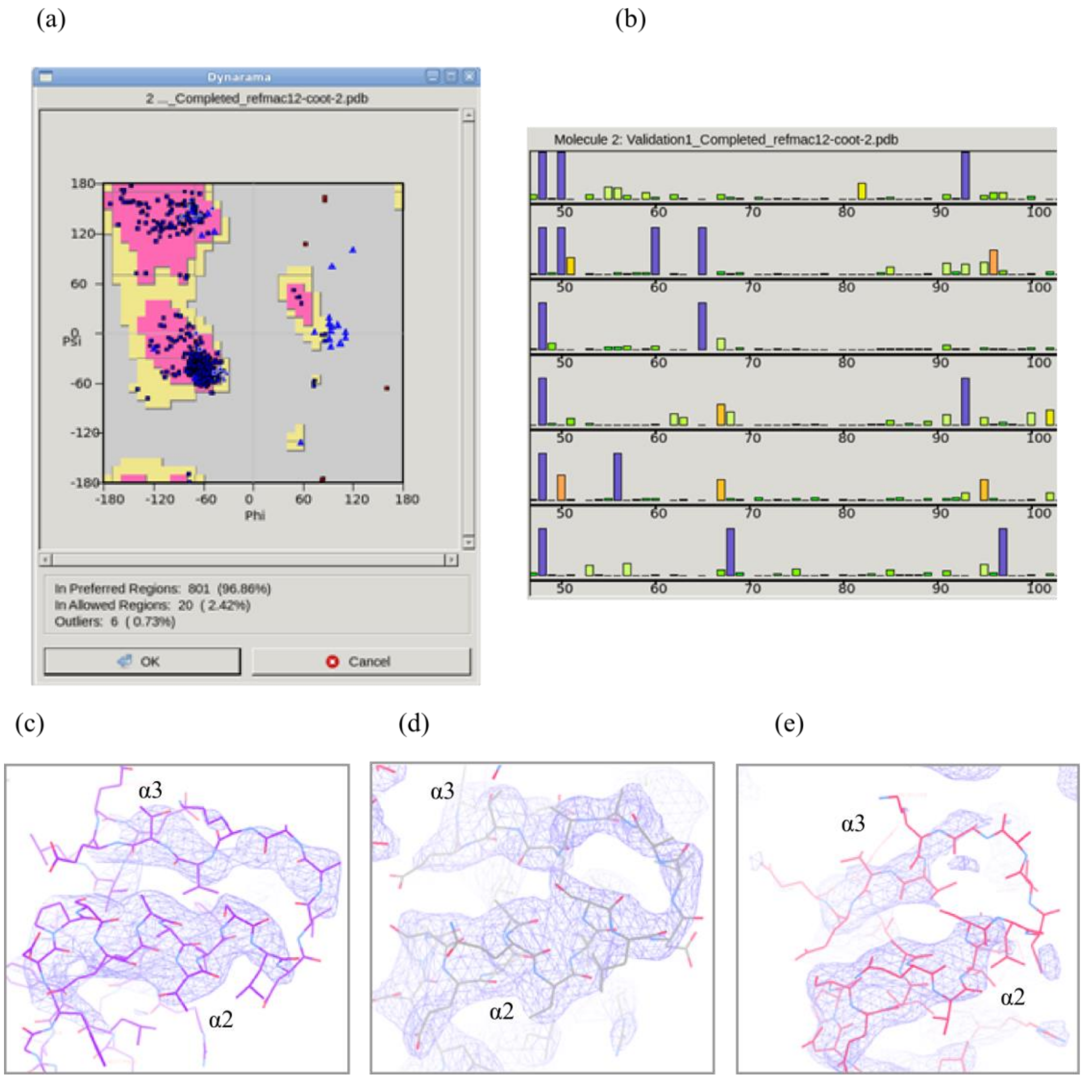


Figure 64.- Crystal arrangement in crystals of Se-Met labelled G20C ST

(a) Electron density map contoured at 1σ after solvent flattening by Parrot. (b) Asymmetric unit containing two sets of trimers, each representing one third of two separate 9-mers. View along the Z (c) and Y (d) axes.



In Preferred Regions: 801 (96.86%)
In Allowed Regions: 20 (2.42%)
Outliers: 6 (0.73%)

Figure 65. Validation and final structure limitations

(a) Ramachandran plot. (b) Rotamer analysis. Purple bars represent stubbed side chains. (c-e) α2- α3 connecting loop from chains B, C and E, respectively. Electron density maps correspond to the 2.51 Å resolution Se-Met G20C ST's structure and are contoured at 1σ.

7.2.3 *Refinement and validation*

The model underwent several cycles of refinement and rebuilding which included rotamer analysis (Figure 65). The final value of R was 21.7 (36.5) % and that of $R_{free,}$ 28.5 (37.4) % (calculated from a test data set of 1056 reflections) (Table 15).

In the current model, 96.86 % of the residues are in the preferred regions of the Ramachandran plot, 2.42 % in the allowed regions and 0.73 % are outliers. The outliers correspond to residues located at the loop connecting helices α2 and α3 where the electron density is somewhat unclear or available only for protein's backbone (Figure 65c-e).

7.2.4 Structure determination of the native protein by molecular replacement

One complete 9-subunit oligomer was generated using the crystallographic 3-fold symmetry and used for structure solution of native crystals by MR (Figure 66). No major difference in fold between native and Se-Met proteins was observed in spite of crystals belonging to different space groups. Oligomers do not pile on top of each other, instead they adopt slightly inclined mutual orientations (Figure 66).

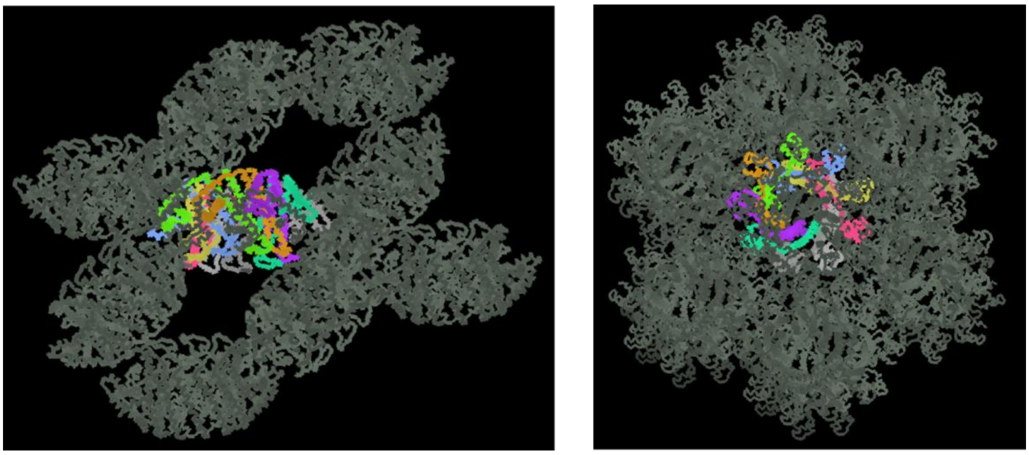


Figure 66. Crystal packing of native protein

7.2.5 Conclusions

Crystal structures of the G20C ST_WT native and Se-Met derivative were determined to 2.50 and 2.80 Å by SAD and MR, accordingly. First, anomalous difference data were used to solve the Se-Met structure. The Se-Met structure was then used for generating initial phases for the native data set by molecular replacement. Both structures have been refined.

In contrast to the capsid protein of dsDNA bacteriophages, for the ST there is a bigger collection of X-ray structures available, implying higher tendency to be crystallised. The propensity to grow crystals could be in part due to the formation of circular oligomers ranging from 8-12 subunits. Oligomerisation is facilitated by tight interaction of subunits via their ODs. In general, full-length STs or truncated versions tend to crystallise in a wide variety of conditions, producing crystals that often diffract to high resolution. Table 16 offers a comparison of structure-quality indicators for available X-ray structural information on STs from several viruses (compiled using data available in September 2014).

Table 16. X-ray statistics of ST structures

	Construct / PDB	**Resolution (Å)**	**Space group**	**I/σ(I)**	**R-factor (%)**	**R_{free} (%)**	Ramachandran outliers (%)	Rms BondLength (Å)	Rms BondAngle (°)
G20C[¥]	WT/ Pending	2.5	*H* 3	1.2	21.7	28.5	0.73	0.0100	1.4
SF6 (SPP1-like)	G*1*P_WT(1-164) / 3ZQQ	4.0	*R* 32:H	4.1	20.0	28.2	3.3	0.002	0.66
	53-120 10-mer / 3ZQM	1.9	$P\ 2_1$	4.4	17.2	21.4	0	0.017	1.58
	53-120 10-mer / 3ZQN	2.2	$P\ 4_12_12$	4.0	19.7	22.2	0	0.018	1.09
	53-120 9-mer / 3ZQO	1.7	$P\ 2_1$	3.1	18.3	23.4	0	0.015	1.42
	65-141 9-mer / 3ZQP	3.0	$P\ 3_2$	2.7	18.7	22.6	0	0.017	1.24
Sf6 (P22-like)	gp1_WT(1-139) /3HEF	1.7	$P\ 4\ 2_1\ 2$	3.4	22.4	25.1	0	NA	NA
P22	gp3 1-140 / 3P9A	1.8	$P\ 2_1$	1.8	17.7	21.6	0	0.006	0.92
	gp3_WT- 1-162	3.4	$P\ 2_12_12$	3.3	23.2	26.5	NA	0.009	1.21
44RR (T4-like)	gp16 26-112[¶], 11-mer / 3TXQ	2.8	$P2_12_12_1$	4.8	31.0	34.0	0.3	0.005	0.81
	gp1625-114, 12-mer / 3TXS	1.8	*R*3	5.0	23.3	27.7	0.3	0.008	1.02
Prophage phBC6A51	2AO9	1.9	$P\ 2_12_12$	2.75	20.5	23.6	0.1	0.01	1.07

NA Not available

¥ Latest model version.

¶ gp16 FL is 154 residues long

The mean I/σ(I) corresponds to the outer shell values.

See

Table 2 for references.

Chapter 8

8. Structure of G20C ST

The current chapter contains a description of the Se-Met G20C ST's X-ray structure and a detailed comparison with structures of other STs (section 1.8). The structural comparison attempts to identify fold similarities and potential role in DNA recognition.

8.1 G20C ST monomer and oligomer

8.1.1 Monomer

G20C ST has a ring-like structure made up of nine subunits that surround a central channel (Figure 67a,b). Each individual subunit has an NTD and a central OD. Electron density was observed for residues 1-134, although for some subunits the density is interpretable for longer segments, up to residue 141 or 145, depending on the subunit. The last 25-37 C-terminal residues are disordered as no clear electron density was observed either in the Se-Met or native electron density map.

The NTD, situated at the periphery of the assembly, comprises residues 1-52, which are organised into three short α helices: α1 (residues 3-15), α2 (19-30) and α3 (36-51) (Figure 67c). The NTD is connected to the OD by a 4-residue Ala-Leu-Pro-Ser linker. Two long antiparallel α-helices, α4 (residues 56-92) and α5 (107-137), connected by a short β-hairpin, comprising β1 and β2 strands, form the OD.

8.1.2 Oligomer and inter-subunits interactions

The OD helices pack closely to form two concentric rings, the outer formed by helices α4 and the inner by helices α5. The central channel inner diameter at its most constricted part is ~ 25 Å, measured as the atom-to-atom distance between the Nη atoms of opposing R108 side-chains. This corresponds to a vdW radius of ~ 22 A. The maximum overall diameter and height of the assembly are 112 Å and 62 Å, respectively; measured for atoms protruding the most from the assembly. The chirality of the subunit packing within the oligomer is similar to that found in other ST oligomers, showing a left-handed twist (when the oligomer is viewed from the outside), so that every subunit is tilted by about 60° with respect to the central channel axis (Figure 67). The NTDs establish two H-bonds with the OD of the adjacent subunit but not with its own OD. Superposition of the six subunits in the asymmetric unit using the OD as reference shows that each NTD occupies a slightly different position in the oligomer (Figure 67d). While the tertiary structure of the DBD (HTH-motif) remains the same, movement of the DBD is observed in all the directions, with a maximum displacement of 6.7 Å in helix α1 and of 6.2 Å in helix α3.

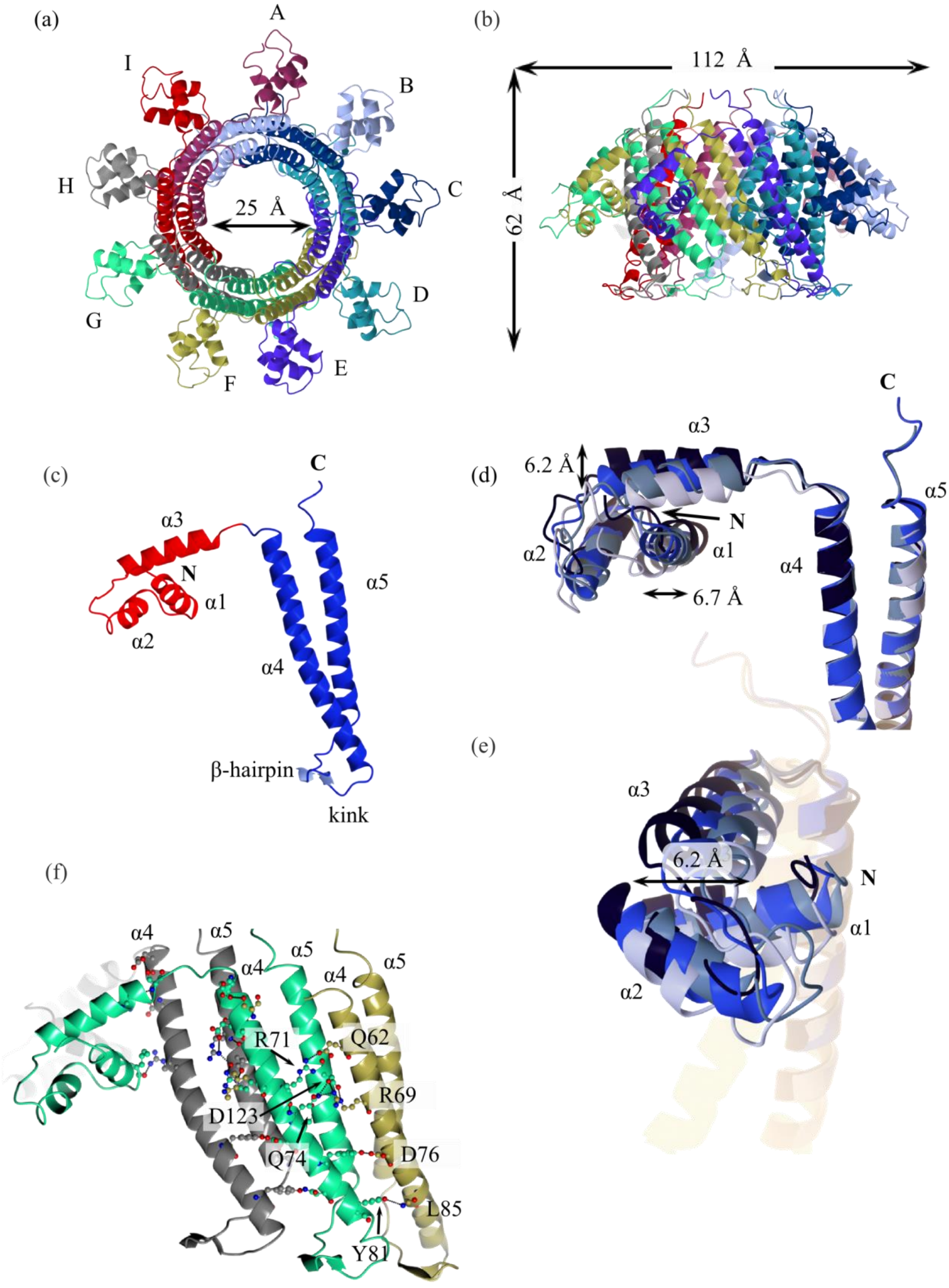


Figure 67. G20C ST structure

(a) Top and (b) side view of the oligomer. Subunits are displayed in different colours to illustrate their organisation (c) Monomer. The NTD and OD, shown in red and blue, respectively, with their secondary structure elements numbered. (d) and (e) Displacement of the NTD with respect the OD. For clarity only four NTDs out of the six that form the asymmetric unit are displayed, the omitted two are located in intermediate positions. The total displacement of helices α1 and α4 are indicated. (f). Inter-subunit interactions. The residues participating in H-bonds are shown in ball and stick representation, with the carbon atoms coloured as the main chain.

The 9-mers are tightly packed against each other in the crystal, keeping the DBD in a fixed place, however it is unlikely that this rigidity is observed in solution, where the flexible linkers bridging the DBDs to the core would permit some adjustment in their positioning. This is consistent with observations made on the crystal structure obtained for SF6 ST (16).

The ODs create a network of H-bonds and electrostatic interactions (Figure 67e,f) with adjacent ODs. In average, every subunit establishes 16 direct H-bonds with adjacent subunits (Table 17 and Table 18). All chains use the same residues to establish the H-bonds with their neighbours, creating a circular network of H-bonds throughout the oligomer that contributes to the stability of the ST. Besides H-bonds, the complementarity of electrostatic charge between the inter-subunit surfaces also plays an important role in the formation of the oligomer.

Table 17. Interfaces at the G20C ST 9-mer*

Structure 1				**Structure 2**				**Interface area Å^2**	**Δ^iG**	**N_{HB}**	**N_{SB}**	**N_{DS}**
Subunit	$^{i}N_{at}$	$^{i}N_{res}$	Surface Å^2	Subunit	$^{i}N_{at}$	$^{i}N_{res}$	Surface Å^2		Kcal/mol			
A	150	44	10662	B	162	44	10318	819.3	-12.1	8	8	0
B	141	44	10318	C	161	43	10548	820.6	-11.7	10	5	0
C	129	37	10548	D	154	43	10669	785.8	-11.7	7	6	0
D	150	44	10669	E	164	44	10313	818.5	-12.1	8	8	0
E	142	44	10313	F	156	43	10534	817.6	-11.6	10	5	0
F	130	37	10534	G	155	43	10668	784.9	-11.7	7	6	0
G	149	44	10668	H	162	44	10314	818.0	-12.2	8	8	0
H	143	44	10314	I	160	43	10539	818.7	-11.7	10	5	0
I	128	37	10539	A	153	43	10662	784.6	-11.7	7	6	0
							Average	807.6	-11.8	8	6	0

*. Parameters calculated with the PDBePISA server (105).

$^{i}N_{at}$ - Number of interfacing atoms in the corresponding structure.

$^{i}N_{res}$ - Number of interfacing residues in the corresponding structure.

Surface Å^2 - Total solvent accessible surface area in square Angstroms.

Interface area Å^2 - Difference in total accessible surface areas of isolated and interfacing structures divided by two.

Δ^iG - Solvation free energy gain upon formation of the interface, in Kcal/M. Negative Δ^iG corresponds to hydrophobic interfaces, or positive protein affinity.

N_{HB} - Number of potential H-bonds across the interface.

N_{SB} - Number of potential salt bridges across the interface.

$_{DS}$ - Number of potential disulphide bonds across the interface.

Table 18. H bonds formed by subunit G*

	Interface FG				**Interface GH**		
##	**Subunit F**	**Dist. [Å]**	**Subunit G**	##	**Subunit G**	**Dist. [Å]**	**Subunit H**
1	LEU 14[O]	2.90	ARG 69[NH1]	1	LEU 14[O]	3.00	ARG 69[NH1]
2	GLU 62[OE2]	2.67	ARG 71[NH2]	2	GLU 62[OE2]	2.36	ARG 71[NH2]
3	ASP 76[OD2]	2.89	TYR 81[OH]	3	ASP 76[OD2]	2.88	TYR 81[OH]
4	SER 51[OG]	3.58	GLU 57[OE2]	4	ARG 47[NH2]	3.71	GLU 62[OE2]
5	ARG 69[NH1]	2.86	ASP 123[OD2]	5	ARG 69[NE]	2.93	ASP 123[OD2]
6	ARG 69[NH1]	2.67	ASP 123[OD1]	6	ARG 69[NH2]	3.30	ASP 123[OD1]
7	F:ASN 83[ND2]	3.17	TYR 88[OH]	7	ASN 83[ND2]	3.33	TYR 88[OH]
				8	ARG 118[NH1]	3.36	ALA 110[O]

*. Table calculated with the PDBePISA server (105).

8.1.3 The disordered C-terminus

The predicted secondary structure for the last 34 residues at the C-terminus features an α-helix (*α6, with LVSVEELVAEVVD sequence) from residues 154-166, connected to α5 by a 16-residue linker (GHVGSTTAGALPSATE) which has no predicted secondary structure. The reason of this disorder is likely to be the flexibility afforded by the long connecting region which is rich in simple amino acids such as Gly, Ala, or Thr. This helix is similar to that observed at the C-terminus of the P22 ST structure, but differs markedly from the β-barrel found in SF6 ST. More information about the C-terminus is provided later in this chapter and the following one, including a discussion of its potential functional role.

8.2 Electrostatic surface of G20C ST

The electrostatic potential surface of the complete oligomer was calculated in CCP4mg, showing an area of positive charge around the outside of the oligomer. A close analysis revealed that this is created by the side chains of residues K11 and K31, R32, K33 (Figure 68a,c), all of them located at the loop that connects α2 with α3 in the NTD. These patches create a ring of positive charge around the periphery of the molecule, offering a potential binding site for DNA through formation of complementary electrostatics interactions.

The cross-section of the oligomer exposing the central channel shows that side chains of residues R108, R118, K121 and K135 which are exposed into the tunnel, form four rings of positive charge spanning the height of the assembly (Figure 68b). In contrast to the observations supporting a peripheral binding mode of DNA, this observation makes it difficult to completely discard the channel as potential binding site for DNA, given that the diameter of the channel is consistent with the diameter of dsDNA (~ 23 Å vdW diameter).

Figure 68d shows a side view of the electrostatic surface of the oligomer calculated with NTDs omitted from the PDB file, to illustrate patches of positive charge at the outer surface of the OD. A protein construct missing the NTD (putative DBD) was designed and purified to test its capability to bind DNA (See Chapter 10).

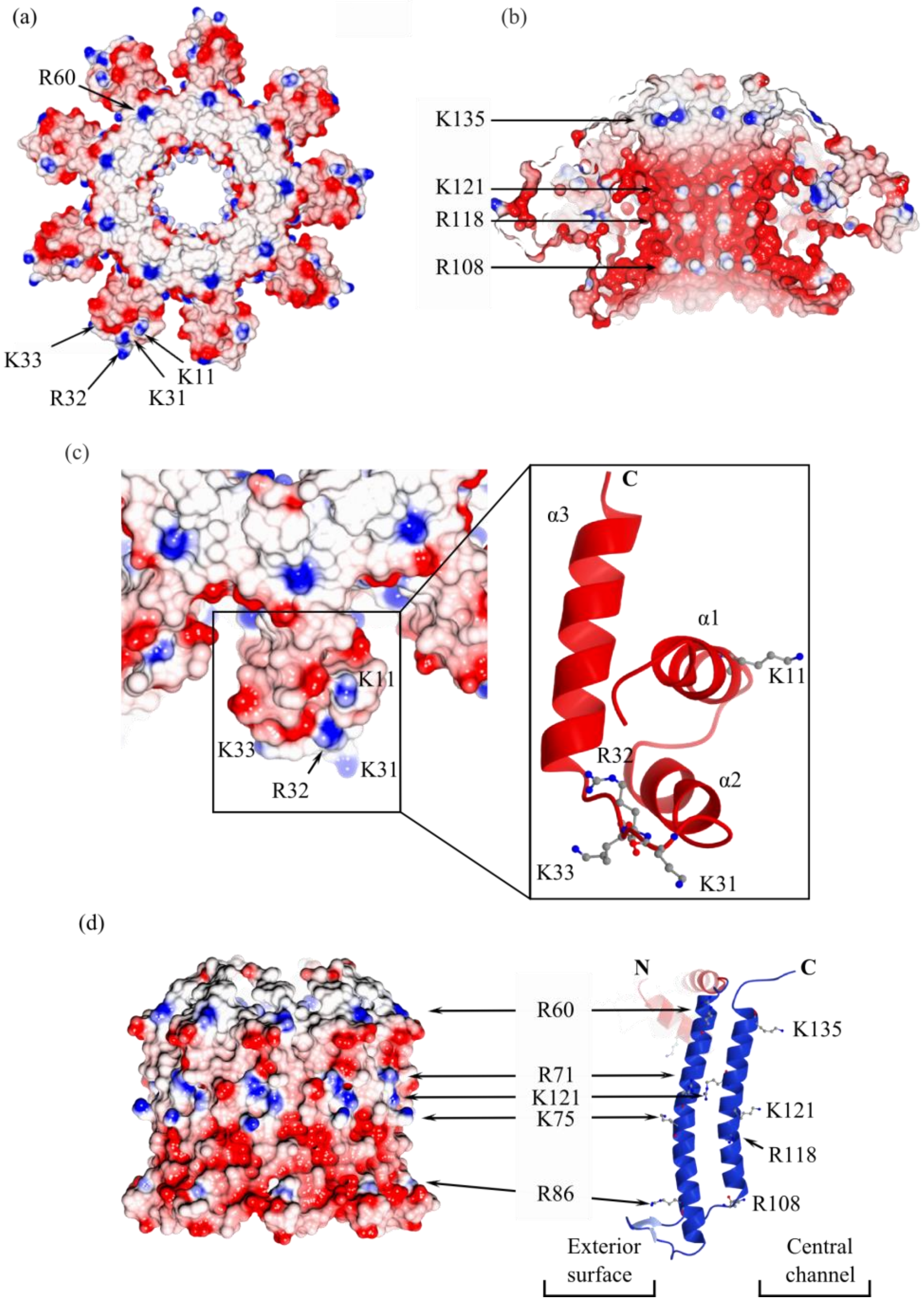


Figure 68. Electrostatic surface of G20C ST

(a) Top and (b) cross section views of the oligomer. (c) Close up of the NTD. (d) Side view of the barrel formed by ODs. Residues responsible for the positive charge patches are shown in ball and stick. Figures not to scale.

8.3 Comparison of G20C fold with other ST

From the available ST crystal structures it is evident that despite low sequence homology some architectural features are shared between them even when they belong to viruses from different families (Figure 12 and Figure 70). The number of subunits observed in crystal structures is 9 in the case of G20C, SF6 (which also forms 10-mers) (16), P22 (14) and PhBC6A51 (70), 8 subunits in Sf6 (68) and 11 or 12 subunits in 44RR (69). At the monomer level, all ST share features, namely a HTH-motif (NTD), a helical core (OD) and a variable CTD, which can be flexible or adopt a folded conformation.

G20C ST exhibits an NTD almost identical to that of SF6 (16,135) and Sf6 (68) (Figure 69a), where a helix-turn-helix (HTH) motif has been identified in the segment comprising helices α2 and α3. In G20C and Sf6, the NTD consists of three helices, while in SF6 a fourth helix is observed. In the case of phBC6A51 (70), the first four N terminal α helices superpose well with those of SF6, suggesting that a HTH motif is also likely to exist to carry out a similar function as in SF6 and Sf6 (16,136). Previous biochemical data on SF6 and Sf6 ST demonstrated that this domain is the DNA-binding domain (DBD).

P22 does not exhibit a clear boundary between the NTD and OD. Instead of an NTD characterised by a HTH motif, it has a purely helical core assembly that resembles the OD of other STs (10). When structural superposition was performed, it was clear that helices α1 & α2 of G20C, SF6 and Sf6 superpose well with helices α2 & α3 from P22 ST, respectively (Figure 69a). P22 α4 does not superpose with helix α3 from other STs but it follows a similar direction and is possible that a motif able to serve the same function as the NTD also exists in P22 between α3 and α4 and their connecting loop.

G20C OD superposes well onto that from other STs (Figure 69b left). G20C, Sf6 and 44RR present two long antiparallel helices separated by a turn at the bottom of the assembly, as the core of the OD. The length and orientations of the helices are comparable, however, α4 in Sf6 is considerably bent (Figure 69b centre). One difference to the described OD from G20C, Sf6 and 44RR, is observed in SF6 and P22, which present a bent β-hairpin composed of β1 and β2 strands (Figure 69b right). PhBC6A51 only exhibits two short helices exactly in the same orientation as those in SF6 and P22. The structural feature common to ODs of all STs is the presence of two helices, suggesting that it is essential for oligomer assembly. The length of the helix or the presence of β-hairpins appear to be non-essential features that may influence the oligomer stability.

Except for SF6 and 44RR, the overall outer diameter is similar for all the assemblies (Figure 69c). Nonetheless, because of the flexibility of the linker connecting the DBDs with the ODs, it is possible for SF6 ST to adopt a more compact configuration. In the case of 44RR, the OD is the only feature ordered in the structure since a large portion of the NTD appears as a chain with no obvious secondary structure and the CTD is not present. Secondary structure predictions for those segments missing in the crystal structures of the STs are provided in the paragraphs below.

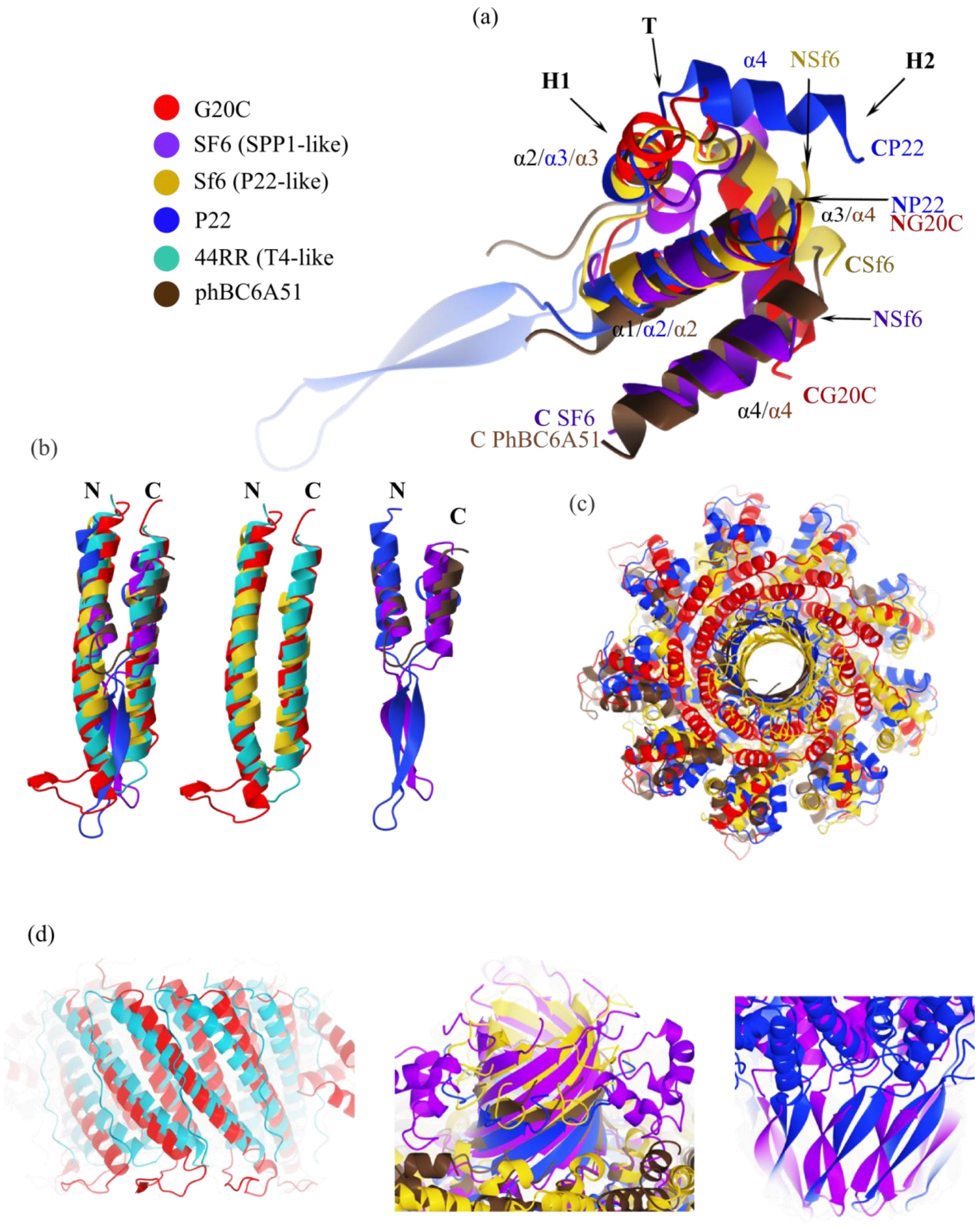


Figure 69. Structural comparison of ST

(a) Superposition of NTDs. Each ST is shown in different colour. The numbering of secondary structure elements is the same as in figure 1.12. (b) OD superposed by secondary structure. Left –all ST and right two different subgroups of ST. (c) Superposed G20C, Sf6, P22 and PhBC6A51 oligomers. (d) Conserved structural features between the diverse oligomers, from left to right: OD, β-barrel and β-hairpin.

The distribution of the NTDs in G20C, SF6 (not included in the superposition of the complete oligomers), Sf6, PhBC6A51 ST and the helical cores in P22 ST is similar (Figure 69c). The average distance between NTDs calculated for SF6 is ~34 Å (Table 19). In G20C the average spacing between NTDs is ~36.5 Å, in good agreement with the spacing observed in SF6.

Likewise, ST from Sf6 and PhBC6A51 exhibit a similar separation between the NTDs. The measured distances correspond to the distance between the Cα of one residue at the loop of the HTH motif and its counterpart in the adjacent NTD.

Table 19. Distance between NTDs

	G20C	SF6	Sf6	P22	44RR	PhBC6A51
Distance (Å)	36.5	34	33.3	28.7	---	34.4
Residue	R32	Published	G39	L88	---	G59

The residue of which the Cα was used as reference in every ST is indicated.

--- For 44RR the NTD is disordered.

Besides the circular organisation and spacing between NTDs, other characteristics are common between the STs (Figure 70). Of particular note is the arrangement of ODs between G20C and 44RR, where the spacing is very alike in spite of the larger diameter of the channel on 44RR ST (Figure 69d left). SF6, Sf6, P22 and PhBC6A51 ST present a β-stranded barrel on top of the main core (Figure 69d centre). Such β-barrel was shown to be important in defining the oligomerisation state of SF6 ST (16).

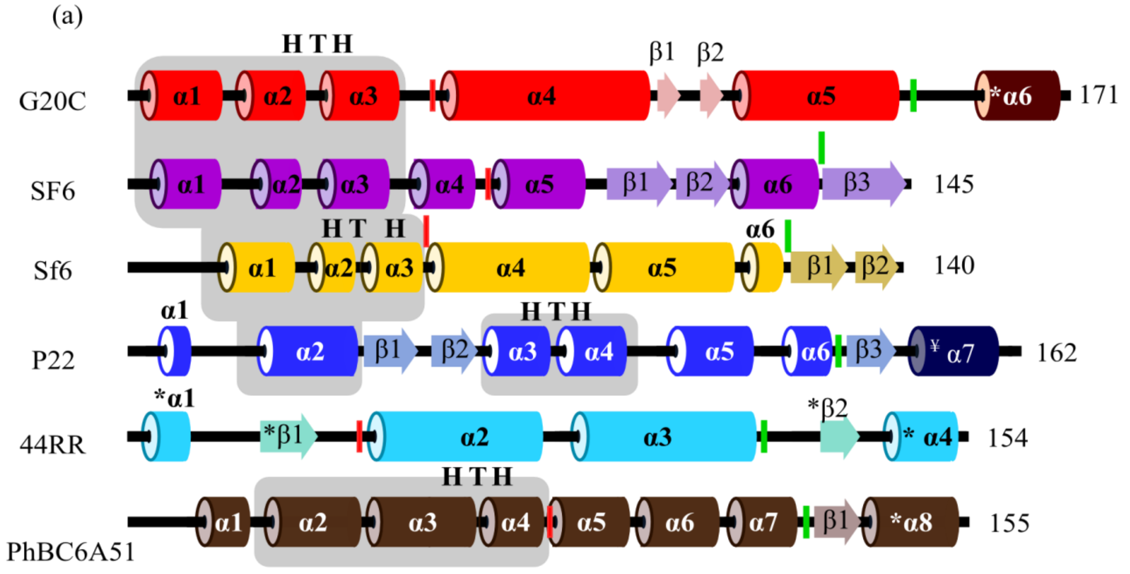
8.3.1 Secondary structure and amino acid sequence analysis: Domain prediction and domain mosaic array.

The alignment of the secondary structure elements and amino acid sequences of all the full length ST proteins shows the generally good conservation of domains (refer to Figure 70 throughout this section).

The sequence length of the STs is analogous, ranging from 140 (Sf6) to 171 (G20C) residues. With exception of P22, the domain organisation of all the STs is similar, comprising the NTD, OD and CTD (See the domain boundaries on Figure 70). Three to four α-helices form the NTD, two α-helices, sometimes accompanied by a β-hairpin, form the OD, and the C-terminal segment contains a β-strand, accompanied in some cases by an α-helix. Several STs have deviations from the common fold. The first example is offered by the Sf6 OD, where the OD comprises the long helices α4, α5 and the short α6. Helices α5 and α6 are separated by only a 1-residue kink (Figure 12 and Figure 70).

Helices α5 and α6 in PhBC6A51 represent the second example of structural diversity among STs. Two α-helices are separated from each other by a distorted helical turn that prevents the formation of a single longer helix.

(a)



(b)

```
G20C        MSVSFRDRVLKLYLLGFDPS-EIAQTLSLDVKRKVTEEEVLHVLAEARELLSALPSLEDI 59
P22         ----MAAPKGNRFWEARSSHGRNPKFESPEALWAACCEYF---EWVEANPLWEMKAFSYQ 53
PhBC6A51    MPFSISGRKGSEMMAKLDELKQKLTAKQIQAAYLLVENEL---MESNNEEKRTQDEMANE 57
SF6         MKEPKLSPKQERFIEEYFINDMNATKAAIAAGYSKNSASA---IGAEN---LQKPAIRAR 54
Sf6         MATEPKAGRPSDYMPEVADDICSLLSSG-ESLLKVCKRPG---MPDKS---TVFRWLAKH 53
44RR        MNDVLDFTQLKDLNGIEGIHGEDVQVYAPLVLRDPVSNPNNRKIDQDDDYELVRRNMHYQ 60
                      .                                            :

G20C        RAEVGQALERARIFQKD----LLAIYQNMLRNYNAMMEGLTEHPDGTPVIGVRPADIAAM 115
P22         GEVIQEPIAKMRAMTITGLTLFIDVTLETWRTYRLREDLSEVVTRAEQVIYDQKFSGAAA 113
PhBC6A51    LGINRTTLWEWRTKNQD----FIAFKSEVADSF-LAEKREQVYSKLMQLILGPQPSVKAM 112
SF6         IDARLKEINEKKILQAN-------EVLEHLTRIALGQEKEQVLMGIGKGAETKTHVEVSA 107
Sf6         EDFRDKYAKATEARADS-------IFEEIFEIADNAIPDAAEVAKARLRVDTRKWALARM 106
44RR        SQMLLDMAKIALENAKN------ADSPRHVEVFAQLMGQMTTTNKEMLKMHKEMKDLAGA 114
                                       .

G20C        ADRIMKIDQERITALLNSLKVLGHVGSTTAGALPSATELVSVEELVAEVVDEAPKT 171
P22         DLLNANIIARDLGLKEQSQVEDVTPDKGDRDKRRSRIKELFNRGTGRDS------- 162
PhBC6A51    QLY-----MQRFGLLTDKKVIEG--DLGNATRTNAEIEEQLQKLKKLTGE------ 155
SF6         KDR-----IKALELLGKAHAVFT-------DKQKVETNQVIIVDDSGDAE------ 145
Sf6         NPR-------KYGDKVTNELVGK-------DGGAIQIETSPMSTLFGK-------- 140
44RR        ATVAIDG---QVQKDADGEFIEFE------GSPDELLDLELADEDIGDD------- 154
                                                 .
```

Figure 70. Comparison of the STs from different phages

(a) Secondary structure elements extracted from the X-ray structures. Star (*) indicates features predicted by JPred. $^¥$ indicates modelled features based on secondary structure prediction. The areas shaded in grey represent the part of the NTD that superpose onto one another. The vertical red and green lines delimit the NTD and CTD, respectively. The length of each element is proportional to the number of residues (b) Alignment of amino acid sequences for every ST included in the discussion. The areas shaded in yellow represent the α-helices of the HTH motifs at the NTD and the predicted α-helices at the CTD, the blue areas, either the observed or predicted secondary structure elements at the CTD.

In contrast to the alignment of the secondary structure elements, aligning the amino acid sequences is almost impossible due to very low sequence identity (<20 %), exemplifying how the sequence changes faster during evolution than the three dimensional structure (Figure 70b) (137).

The available X-ray structures of STs are the most detailed source of structural information regarding the oligomerisation state, domain organisation and identification of potential sites for DNA-binding. In G20C, P22, 44RR and PhBC6A51 ST the CTD is incomplete or not present in the crystal structure. Prediction of the secondary structure for these disordered regions revealed that the CTD segment is likely to comprise either a β-strand (which is present in the structures of SF6, Sf6 and P22 and PhBC6A51 STs) followed by one α-helix or only an α-helix as predicted for G20C and 44RR STs.

For P22 ST it was previously proposed that α7 in solution exists in the form of either an unstructured polypeptide or folded α-helix, both transient conformations being present in an equilibrium state (14). This observation was reinforced by the finding of longer assemblies on negatively stained samples that included a feature consistent with an α-helix (found above the β-barrel).

The absence of helix *α6 in the G20C ST X-ray crystal structure could be explained by two different hypotheses. It is possible that helix *α6 is structured in solution but the flexibility of the long linker directs it away from other *α6 helices. In the second scenario the putative helix *α6 could fold upon binding to DNA with other parts of the molecular motor.

Both hypotheses are supported by the absence of electron density for the CTD in native and Se-Met crystals, and by crystallisation experiments performed with ST_ΔC and ST_ΔN. ST_ΔC crystallised in multiple conditions (around 45 % of the tried conditions) while ST_ΔN only produced crystals in one condition. Although helix *α6 could be ordered, the amino acid sequence reinforces the existence of an unfolded helix *α6 in two ways:

a) As in P22, the G20C CTD is predicted to contain one α-helix as seen in 44RR and PhBC6A51 STs. The similar architecture of the NTD and OD between the cited STs, points out that the fold of the C-terminus could also be alike. There is room for the potential formation of an α-helical barrel on top of the assemblies; nonetheless, important differences arise when looking closely at the sequence of the predicted helix: in P22, helix α7 is rich in positively charged residues, while in G20C and 44RR ST the last helix is rich in negatively charged residues. In PhBC6A51 the number of positive residues equals the number of negative residues.

   Moreover, G20C ST *α6 presents the signature characteristics of intrinsic disorder (138), which are (*i*) low sequence complexity (in *α6 50% of the residues correspond to Val/Leu), with low content of bulky hydrophobic residues and (*ii*) high content of polar or charged residues (50% are polar or charged residues: His/Glu/Lys). Additionally, intrinsic disordered regions suffer from the lack of hydrophobic amino acids that help to bury a hydrophobic core to acquire tertiary structure. The sequences of the C terminal α helices from other STs follow a similar pattern.

b) Although the G20C α5-*α6 linker is quite long, in phages P22 and PhBC6A51 the linkers β3-α7 and β1-α8, are not long enough to account for the disorder of the complete missing fragment. The sequence of G20C α5-*α6 connection confirms its nature of flexible linker as it contains the traditional features of linkers (138): high content in polar uncharged (37.5%) or small amino acids (31.3%) and Pro (6.3%).

Disordered or unstructured regions are not uncommon features in proteins. Dyson and Write reviewed several proteins that present coupled folding and binding, which is the property of intrinsically disordered regions or complete proteins to fold upon binding to their target(138). A classic example is the kinase-inducible domain (pKID) of the transcription factor cyclic-AMP-response-element-binding protein (CREB), which contains the KID-binding (KIX) domain of the CREB-binding protein (CBP) (Figure 71a). To function as a regulator of the expression target genes, KID must be phosphorylated at Ser133 by the protein kinase A (PKA) to then form complexes with KIX. KID is unstructured in isolation, but the solution structure of the phosphorylated KID complexed with the hydrophobic core region of KIX, showed that KID undergoes a conformational change to fold into two perpendicular α-helices. Further experiments demonstrated the pSer133 is essential to stabilise the pKID-KIX complex (139). Phosphorylated CREB regulates a variety of genes in response to cAMP by interacting with components of the basal transcription machinery. CREB is involved in diverse physiological processes, for instance, in the central nervous system it modulates development, plasticity, circadian entrainment, neuroprotection and resiliency (140).

Crane-Robinson *et al* revised the importance of N- and C-terminal portions present in several DNA binding motifs and showed examples of how some extended segments contribute to interaction with either the minor or major groove (141). The protein Antennapedia (Antp) offers one example of terminal extensions interacting with DNA (Figure 71b). The N terminus of Antp contains four positively charged residues, including the Arg at position five, highly conserved in other arms or tails from DNA binding proteins. The X-ray structure determined to 2.4 Å resolution, and NMR structure of the Antp homeodomain (~60 residues organised into three α-helices with the second and third helices form a HTH motif that interacts with DNA) showed that the recognition helix (α3) of the HTH motif is inserted into the major groove of the DNA, while part of the N-terminus extends along the minor groove (142).

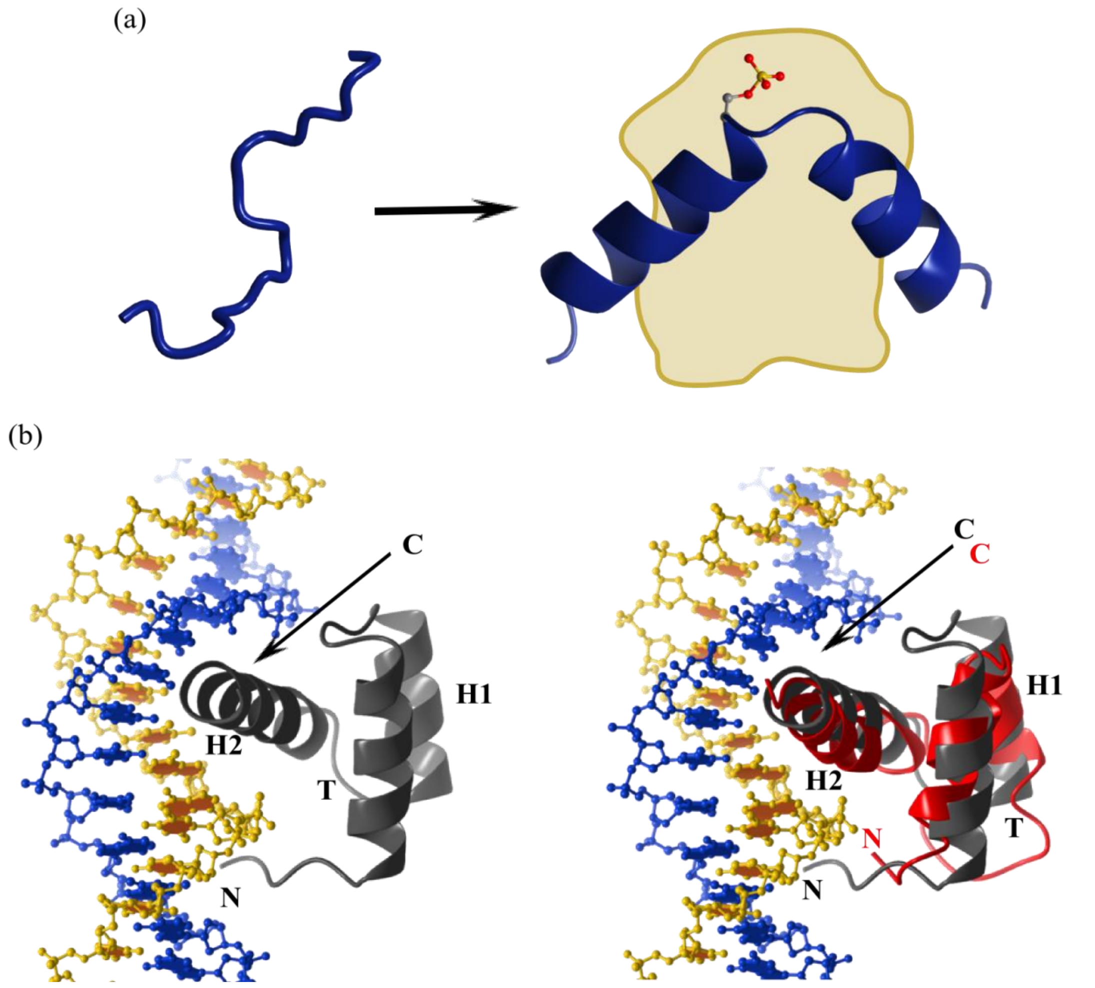


Figure 71. Examples of unstructured protein regions

(a) Coupled binding-folding of KID (blue ribbon) upon binding to KIX. Although the hydrophobic core of KIX consists of three α-helices, for simplicity reasons, the complete KIX is shown as a protein contour. The phosphorylated Ser133 is shown in ball and stick representation. (b) Complex of Antp-DNA (left). Right, Superposition of G20C_NTD (red) with Antp-DNA complex. PDB code: KID: 1KDX, Antp: 9ANT.

Antp is one of the homeotic genes clustered in the Antennapedia complex (ANT-C) that control the identity of body segments in *Drosophila melanogaster*. Some dominant mutations in the *Antp* gene produce abnormalities at the thorax or head of the flies which include the transformation of antennae into legs or vice versa (143).

The identified HTH motif of G20C ST superposes onto Antp with an rmsd of 2.4 Å for 44 residues of G20C NTD. More details about the superposition with the HTH motif on Antp homeodomain and other proteins are provided in section 8.4.

8.3.2 *Mosaic boundary in small terminases*

It is recognised that many species of virus possess genomes which are composed of mosaics or modules. These modules, which may be single or collections of genes, are defined as regions with a degree of similarity found when genomes from two or more related phages are compared. These regions are considered to be genome sections that were horizontally exchanged among phages and potentially other species during their evolution (144). Analysis of 152 genome sequences of P22-like phages demonstrated that mosaic boundaries correlate with protein domain boundaries within the ST gene, specifically between the N-terminal region (NTD + OD) and CTD in the ST (24). Moreover, the studied sequences in these genomes were classified according to eight mosaic boundaries that are the result of the combination of four different N-terminal regions with four different C-terminal regions. Similarly, the C-terminal region of the ST relate to a number of NTDs of the LT (24,144).

Figure 72 shows the manual alignment of the C-terminus of the STs considered in this project based on the X-ray structure and secondary structure prediction, without insertions. It is important to note that alignment by commonly employed software does not reveal conserved patterns. It is evident that the CTD of the six ST can be classified into a three different subgroups according to the secondary structure elements displayed. The first group (SF6, Sf6) features a β-strand seen in two ST crystal structures. The second group (PhBC6A51, P22, 44RR), features one β-strand followed by one α-helix. The last group, which includes only G20C, exhibits one α-helix.

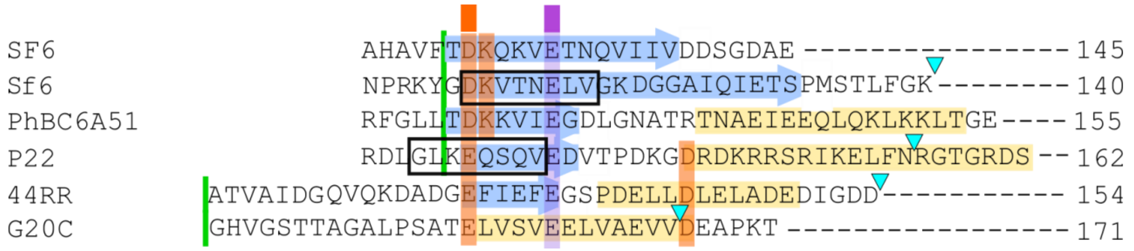

SF6 AHAVFTDKQKVETNQVIIVDDSGDAE ---------------- 145
Sf6 NPRKYGDKVTNELVGKDGGAIQIETSPMSTLFGK --------- 140
PhBC6A51 RFGLLTDKKVIEGDLGNATRTNAEIEEQLQKLKKLTGE ---- 155
P22 RDLGLKEQSQVEDVTPDKGDRDKRRSRIKELFNRGTGRDS -- 162
44RR ATVAIDGQVQKDADGEFIEFEGSPDELLDLELADEDIGDD ----------- 154
G20C GHVGSTTAGALPSATELVSVEELVAEVVDEAPKT ------------------ 171

Figure 72. Alignment of CTDs

Manual alignment of CTDs. The vertical green line delimits the CTD. The areas shaded in yellow represent the predicted α-helices, while the blue areas, either the observed or predicted β-strands at the CTD. The domain boundaries of P22 and Sf6 are squared in black-lined boxed. Areas shaded in purple and orange indicate residues shared by all the STs or a subgroup of them, respectively. The cyan triangles point the start of the LT protein.

Concerning the interaction of the ST CTD with the LT NTD, it has been reported for the P22 ST that removal of the CTD impairs the formation of ST-LT complexes, pointing out that its role in the DNA translocation is to generate communication between the ST and LT. The close relationship between P22 and Sf6 ST CTD with their LT NTD is evident from the sequence analysis performed by Leavitt *et al* (24), where the ST CTD is part of the mosaic that contains the LT gene, suggesting that both regions are exchanged together to ensure interaction. Interestingly, the mosaic boundaries in P22 and Sf6, which lie within regions G126-V133 and D113-V120 align with the initial portion of SF6 and PhBC6A51 CTD.

The relevance of the overlap between the ST and LT genes is potentially another way to ensure that both proteins remain together, however, in Sf6 and 44RR no overlap occurs since the LT gene continues immediately after the ST stop codon.

8.3.3 The central channel

The role of the central channel in DNA binding has been the subject of much debate. Earlier in section 1.8 the diameter at the narrowest part of the channel was introduced to assess whether it is sufficiently wide to accommodate DNA. These channels, however, are not regular cylinders with the same diameter throughout (Figure 73). For instance in G20C, the central part of helices α4 and α5 creates an area of ~25 Å diameter, while their top part and connecting small β-hairpin define areas of 27.9 and 54.1 Å diameter, respectively. These are interatomic distances measured between opposing atoms protruding the most towards the tunnel axis and not van der Waals (vdW) distances. The vdW diameter would be ~ 3 Å narrower. For comparison, the vdW diameter of DNA is ~23 Å .

Similar analysis of other STs demonstrates that the β-strand at the C-terminus of SF6, Sf6, P22 and PhBC6A51 STs, defines the most constricted part of the channel, and that there is no strict correlation between the number of subunits and the diameter of the channel. As example, G20C, SF6 and P22 assemble into nonamers, however their narrowest diameter is 25, 14.3 and 19.4 Å respectively. 44RR is one exception to the previous assumption as it is formed by the largest number of subunits and also exhibits the widest channel, 27.8 Å for the 12-mer.

From the displayed distances, it is fair to hypothesise that only 44RR ST could accommodate B-form DNA inside the central channel and potentially G20C. Although for phages P22 and Sf6 it is proposed that the central channel is the binding site for DNA, several aspects refute this hypothesis. On one hand, in P22 there is only an apparent remnant of a HTH motif to bind DNA, while on the other hand, the channel hast two ~ 19.6 Å constrictions. No experimental evidence is available to establish whether the channel is involved in DNA-binding. Likewise, for Sf6 it has been proposed that the body region of the assembly nicely accommodates DNA, however, analysis of the published structure showed that no part on this assembly would have a vdW diameter big enough to host DNA. Moreover, at the intersection of the neck and body regions in Sf6 ST, the atom-to-atom distance of the most protruding residues is 15.5 Å. This part would be wider if the distance between Cα was considered, which raises questions about where these long side chains would reside if DNA is present in the tunnel. It has been speculated that the ST could disassemble or undergo conformational changes to open and incorporate DNA or to expand; nonetheless, experimental evidence is yet to be presented to support such theory.

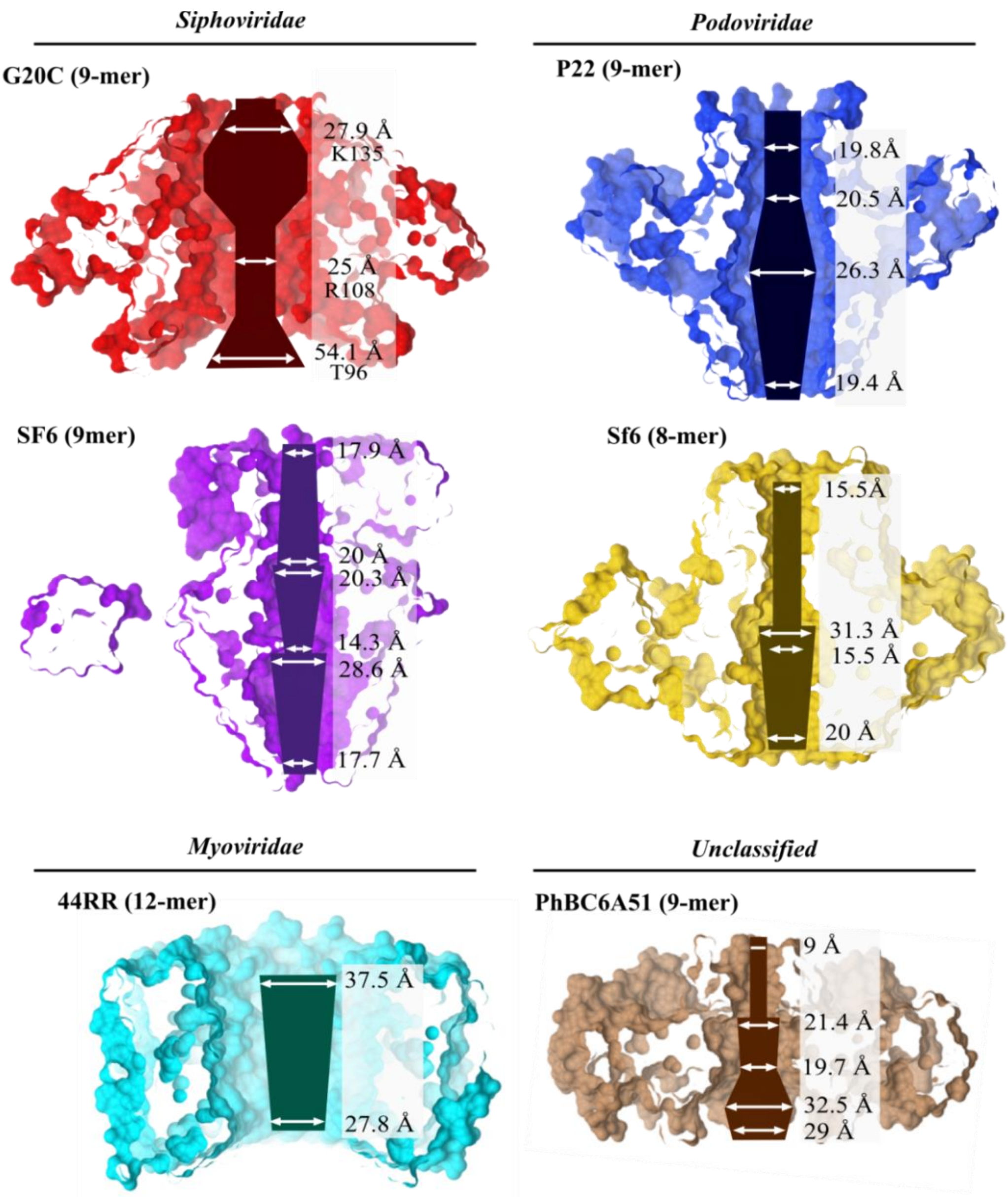


Figure 73. The central channel in ST from dsDNA phages

Representation of the central channel for every structurally characterised ST. The drawn architecture of the channel is approximately to scale. The displayed distances correspond to atom-to-atom distances between the side chains protruding towards the inner channel. Where two distances are provided for the same area of the channel, the largest one corresponds to the distance between atoms of the Cα backbone and not to side chains.

8.4 Model of interaction of a single NTD with dsDNA

In section 8.3.1, it was discussed how G20C ST_NTD superposed onto Antp-DNA complex, a protein that helped to exemplify that unstructured termini can interact with DNA. Even though the rmsd is 2.4 Å (calculated using CCP4mg), the coordinates of G20C ST_NTD were submitted to the PDBeFold server to search for structures in complex with DNA that could align to ST_NTD with higher quality. The PDBeFold server searches through the RCSB Protein Data Bank for structures with similar secondary structure elements. The target structures are organised according to the Quality (Q)-score that is an estimate of the quality of the Cα-alignment, maximised by the secondary structure matching (SSM) alignment algorithm. Identical structures have a Q-score value of 1, the higher the score, the better the match.

The first match, the *Mycobacterium tuberculosis* hypoxic response regulator DosR C-terminal domain (DosR CTD) is shown just as the reference as it was not crystallised in complex with DNA. The second best match for the purpose of this project is the structure of the human telomeric repeat-binding factor 2 (HTRF2) in complex with its dsDNA telomeric site (145). The Q-score is 0.49, compared to 0.52 of DosR, and the rmsd of the superposition is 1.8 Å (Figure 74, Table 20). Other parameters provided by PDBeFold are the P score that represents the minus logarithm of the P value, which is probability to score a certain value of S, by chance, when matching two structures *A* and *B*. The S score is defined as a sum of quality scores Q. S varies from 0 to *k* (ideal alignment). k is the number of secondary structure elements present in a structure. The Z score measures statistical significance of the match, the higher the score, the higher the statistical significance (106).

Table 20. Protein domains with similar fold to G20C ST_NTD identified by PDBeFold

##	Q	P	Z	rmsd	N_{algn}	N_{gaps}	Q-$_{\%sse}$	Match	T-%sse	Target
1	0.5178	3.502	5.345	1.80	43	2	100	3C57:B	100	*M. tuberculosis* hypoxic response regulator DosR CTD
7	0.4874	2.915	4.872	1.80	44	3	100	1W0U:A	100	HTRF-DBD-DNA complex
9	0.4624	3.163	5.077	1.95	43	2	100	3N97:A	100	RNA_Pol_σ70-D4 –DNA complex
31	0.4212	1.628	3.653	2.36	45	3	100	9ANT:A	100	Antennapedia Homeodomain-DNA complex

- Root mean square deviation, **rmsd**, calculated distance between Cα atoms of matched residues.
- Length of alignment, $\mathbf{N_{algn}}$, length of alignment. Number of matched residues.
- Number of gaps, $\mathbf{N_{gaps}}$.
- Percent of matched secondary structure elements, **Q- %**$_{\mathbf{sse}}$, percent of the secondary structure of query identified on the match. T-**%**$_{\mathbf{sse}}$, percent of the secondary structure of the target identified on the query.

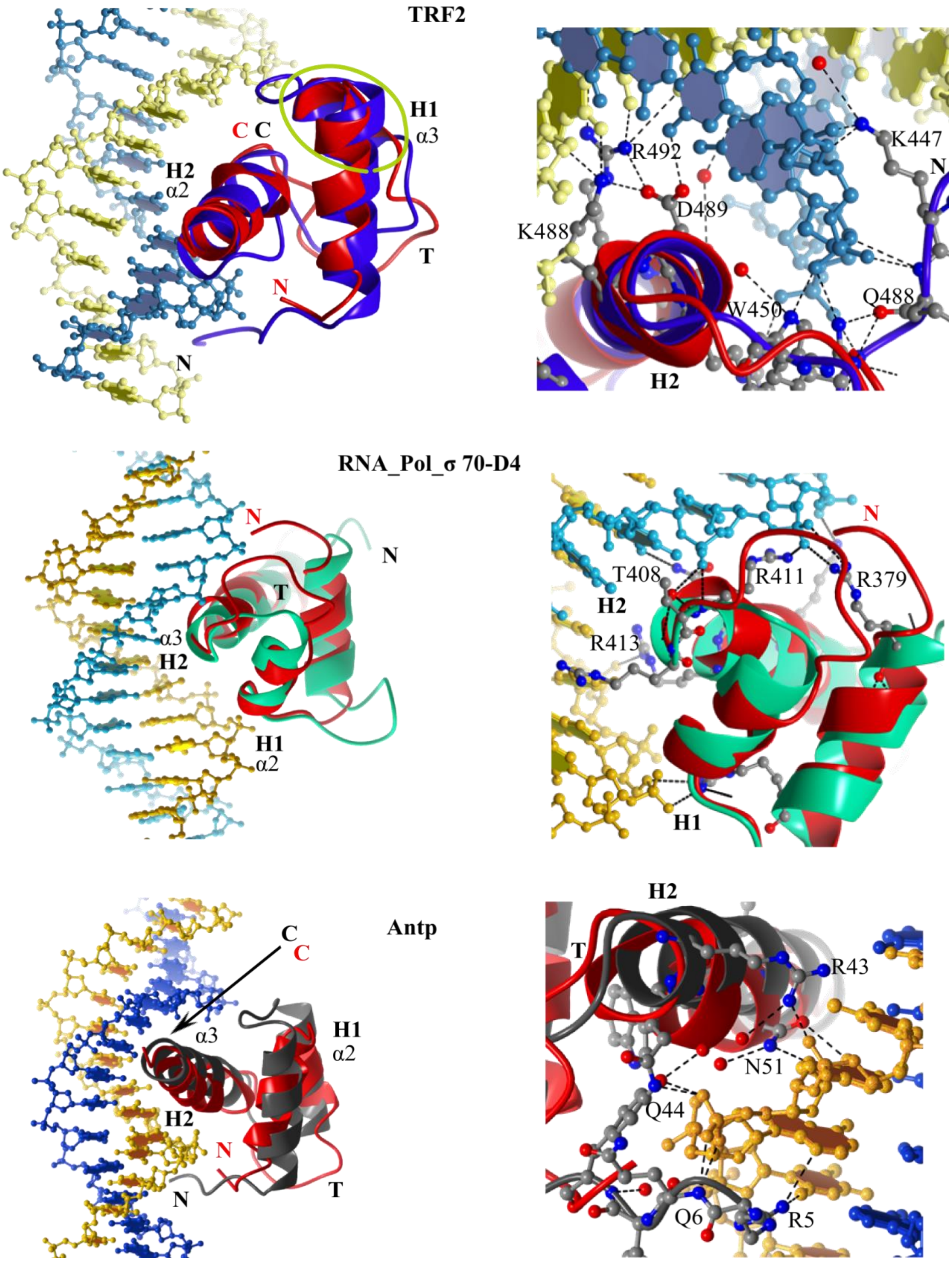


Figure 74. Superposition of G20C ST_NTD onto HTH-DNA complexes

Matching protein-DNA complexes found by the PDBeFold search. G20C NTD (red) superposed onto the HTRF2, RNA polymerase sigma 70-factor domain 4 and Antp. DNA is in different tones of blue/yellow as the DNA sequences are different. Elements of the HTH motif are indicated as H1-T-H2, being H2 the recognition helix. Helices α2 and α3 are shown only for the G20C NTD. H1= α2 and H2=α3. The G20C ST_NTD N- and C-termini are labelled in red. Close up view at protein-DNA interactions is on the right. In each case the residues in ball and stick representation correspond to residues from the protein crystallised with DNA and not to the G20C NTD.

Besides the HTRF CTD and Antp homeodomain, G20C ST_NTD also superposed onto the σ70 factor domain 4 of the *Thermus aquaticus* RNA polymerase (146) (Figure 74). The three mentioned matches present the HTH motif, with the recognition helix, H2, fitted into the major groove of the DNA. Charged and polar side chains projecting out of H2 establish H bonds with the phosphate backbone or bases of DNA. G20C ST_NTD superposes well onto the complete HTH motif present in the HTRF, Antp and RNA_Pol σ70-D4, with helices α2 and α3 representing H1 and H2 from the HTH motif, respectively.

Differences in the fold are observed at the N- or C-termini, and at the H1-H2 connecting loop. In the HTRF2 and Antp the N-terminus creates several H-bonds with DNA, in contrast, in the G20C ST, the initial N-terminal portion does not form an extended polypeptide chain since it is part of helix α1. Moreover, according to the superposition, residues 1-5 are situated far away from DNA, discarding the possibility of the N-terminal segment playing a role in interacting with DNA. Concerning the HTRF2, the crystallised species corresponds to residues 446-500. While the N-terminus of the construct may have been free in solution, the H-bonds established with DNA are likely to be a consequence of the truncation since the conformation adopted by the full length protein is likely to be different (145).

In all identified homologues the H1-H2 connecting loop is less extended than in G20C. Although H2 is the main element that interacts with DNA, in G20C the loop is rich in positively charged residues, suggesting that it may contribute to interaction with DNA (Figure 68).

8.5 Conclusion

Structure and sequence-based comparison of G20C ST with other STs showed that in spite of low sequence homology, several shared structural features are conserved, both at monomer and oligomer levels. For example, the presence of the HTH motif at the NTD, the OD consisting of two α-helices and the arrangement of several subunits into a circular assembly with comparable spacing between NTDs. The flexibility of the CTD can be interpreted as another characteristic in common between some STs since this region is disordered in G20C, P22, 44RR and PhBC6A51 phages. Structural differences arising between STs can be used to classify the structures according to their individual domains: subsets of STs contain an HTH motif in the NTD, others incorporate β-hairpins in their ODs and other STs are expected to have one β-strand followed by one α-helix at the CTD.

Furthermore, structural analysis revealed that the NTDs are the potential binding sites for DNA. The degree of variation in the channel diameter and number of subunits in the assembly while keeping a conserved spacing between individual NTDs, suggests that the channel may play a non-essential role in DNA recognition. Close structural homology with HTH motifs of several characterised DNA-binding proteins further reinforces the view that the HTH motif of G20C is directly involved in DNA interaction. A model for this protein-nucleic acid interaction will be proposed in chapter 10.

Chapter 9

9. Production of DNA for DNA-binding activity experiments

The current experimental evidence describing direct *in vitro* binding of the ST to DNA has utilised several sources of DNA. As described in the introduction, for some dsDNA phages the recognition site for the ST is known and well characterised. Such information allowed the design of appropriate DNA fragments to include in DNA-binding experiments, e.g. in the case of SF6, three variants of the SPP1 *pac* site were generated to test binding by surface plasmon resonance (SPR) (16). For Sf6, a systematic search of the recognition site was carried out. Initially EMSA including a fragment of 1,836 bp, selected because it features the ST and LT genes, were performed (68). Further experiments included random DNA fragments of varied lengths, or supercoiled DNA (136). The latest research on this phage has delimited the region with *pac* site activity to a 30 bp sequence by determining its ability to support the initiation of packaging by a P22 prophage that carried a hybrid ST. This hybrid featured the NTD and OD from Sf6, but the C-terminus from P22 ST (24). The DNA sequence used on PaP3 ST was a 239 bp fragment that included the *cos* site (15). The *cos* site was first identified in λ phage and used for DNA-binding experiments (147). For P22 ST, the gene 3, which codes for the ST but also includes the *pac* site, was used for biochemical characterisation (14).

The approach followed to investigate the G20C ST interaction with DNA is described in the present and following chapters.

9.1 Design of viral and non-viral 1kbp DNA fragments: G20C_F1 and pUC18_F1

To evaluate the capability of G20C's ST to bind DNA, 1 kbp DNA fragments of viral and non-viral origin were amplified. At the beginning of this project, Severinov's lab in Moscow, kindly provided us with G20C's genomic DNA. At that time they had just sequenced the genome and partially annotated it, however, no further information was available about whether a recognition site for the ST existed to start the translocation process, its sequence or location.

Originally, the G20C's ST WT was identified by its sequence homology with the putative small subunit from thermophilic dsDNA bacteriophages P23-45 and P74-26 (7) and by its genomic context. In many dsDNA phages the organisation of the molecular motor is a recurrent arrangement that comprises the gene encoding the ST overlapping downstream with a short portion of the LT gene. The LT gene is separated from the portal protein gene by a sequence of varying length.

As it was discussed in section 1.10.2, for several phages, the ST recognition site lies within the gene encoding the ST itself or a region close to this gene. The region comprising (Figure 75a,c) positions 42,098-43,097 in G20C's genome was amplified using the pair of primers: forward 5' CGC CCT CCT CTT TTC CAC CGC GGA TGA GCT 3' and reverse: 5' GGG AGC AAT AAT CCA GCC CTG GCT TCC AGG 3'. This region includes 516 bp that correspond to the ST gene, at the centre of the 1 kbp fragment (only G20C_F1 henceforth) plus 300 bp upstream and 184 bp downstream. The downstream region corresponds to a fraction of the LT gene (Figure 75a,c and Figure 76a). The overlapping region between the ST and LT genes is 20 bp long.

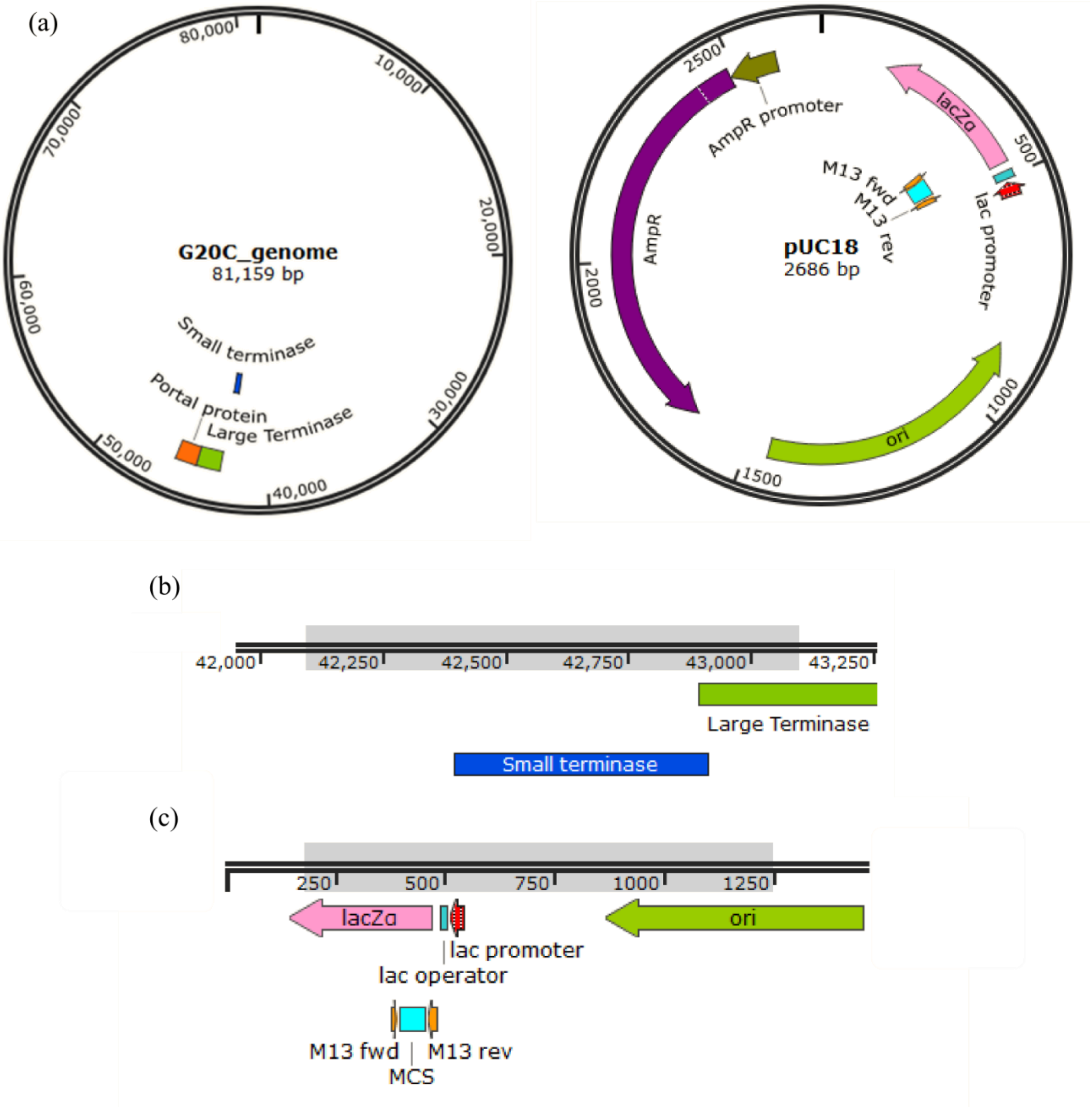


Figure 75. 1 kbp DNA fragments: G20C_F1 and pUC18_F1

(a) G20C genome (left) and pUC18 plasmid (right). Structural and functional elements of the G20C genome and pUC18 are indicated. (b) Scheme of G20C_F1 comprising the ST and portion of the LT genes. (c) Scheme of pUC18_F1 comprising the illustrated functional elements.

The fragment pUC18_F1 amplified covered the region 185-1,185 on the original plasmid (Figure 75a, c and Figure 76a). The designed primers were: Forward: 5’ ATG CGG TGT GAA ATA CCG CAC 3’ and Reverse: 5’ GTT GGA CTC AAG ACG ATA CTT 3’. Some functional elements of the plasmid, like the N-terminal part of the β-galactosidase (*lac*Za), the *lac* operon and promoter, the multiple cloning site (MCS) region and a portion of the replication origin are included in pUC18_F1.

(a)

```
ACATGCCGAC GTACAAGGAG CGGCAAAACC ACGTCATCGC CCTCCTCTTT TCCACCGCGG     42120
ATGAGCTTAA AGAGTTTGAC GACCTGGTTC AGAGCCTGGC GACTCCAGAA GAGCGGAAGC     42180
TCCCTAGGGG TCAACGTAGG CTTAGCGTTC TTAAAAGAGC TTTGCAAGCG CTCAAAACCT     42240
AGGAGACGTC TTATGGAGTA CGAGACCCTT TGCGACGTGT GCATTCGTAA ATACGAGGGA     42300
TGCTACCTTT ACGAGTTCTT CATCGCTCCC CTATCCGCCG ATGGTGTGCG CCTTGAGGTC     42360
AAAGACTGCG TGCAGTACCT GTCTGAGGTA GAAAGCGATG AGCGTGAGTT TTAGGGACAG     42420
GGTGCTCAAG CTTTACCTGC TTGGCTTTGA CCCTAGCGAG ATTGCGCAGA CCCTTAGCCT     42480
GGACGTCAAG CGCAAGGTTA CAGAGGAGGA AGTTCTACAC GTCCTAGCCG AGGCTAGAGA     42540
GCTTCTTTCC GCCTTGCCTT CCCTTGAAGA CATCCGGGCC GAGGTGGGTC AGGCTCTAGA     42600
GCGCGCCCGG ATTTTCCAGA AAGACCTGCT AGCGATTTAC CAGAACATGC TCCGCAACTA     42660
CAACGCCATG ATGGAAGGCT TGACCGAGCA TCCAGACGGC ACCCCGGTGA TTGGCGTAAG     42720
ACCGGCGGAT ATAGCCGCTA TGGCCGACCG GATTATGAAG ATTGACCAGG AGCGCATCAC     42780
CGCTCTGCTC AATAGCCTCA AGGTACTAGG CCATGTCGGG TCCACAACCG CCGGAGCTCT     42840
CCCCTCCGCT ACAGAGCTAG TGAGCGTGGA GGAGCTGGTG GCGGAGGTGG TGGATGAAGC     42900
GCCTAAGACC TAGCGACAAG TTCTTTGAGC TTCTAGG CTA CAAGCCGCAT CACGTACAGC     42960
TAGCCATCCA CCGGAGCACG GCCAAGCGCC GTGTGGCTTG CTTGGGGCGC CAGTCGGGCA     43020
AGTCCGAGGC GGCTAGCGTA GAGGCCGT GT TTGAGCTCTT CGCCCGG CCT GGAAGCCAGG     43080
GCTGGATTAT TGCTCCCACG TACGACCAGG CGGAGATTAT CTTTGGTCGC GTGGTGGAGA     43140
AGGTGGAGCG TCTATCTGAG GTCTTCCCCA CCACCGAGGT TCAGCTTCAA CGTAGACGCT     43200
```

(b)

```
ACCATATGCG GTGTGAAATA CCGCACAGAT GCGTAAGGAG AAAATACCGC ATCAGGCGCC       240
ATTCGCCATT CAGGCTGCGC AACTGTTGGG AAGGGCGATC GGTGCGGGCC TCTTCGCTAT       300
TACGCCAGCT GGCGAAAGGG GGATGTGCTG CAAGGCGATT AAGTTGGGTA ACGCCAGGGT       360
TTTCCCAGTC ACGACGTTGT AAAACGACGG CCAGTGCCAA GCTTGCATGC CTGCAGGTCG       420
ACTCTAGAGG ATCCCCGGGT ACCGAGCTCG AATTCGTAAT CATGGTCATA GCTGTTTCCT       480
GTGTGAAATT GTTATCCGCT CACAATTCCA CACAACATAC GAGCCGGAAG CATAAAGTGT       540
AAAGCCTGGG GTGCCTAATG AGTGAGCTAA CTCACATTAA TTGCGTTGCG CTCACTGCCC       600
GCTTTCCAGT CGGGAAACCT GTCGTGCCAG CTGCATTAAT GAATCGGCCA ACGCGCGGGG       660
AGAGGCGGTT TGCGTATTGG GCGCTCTTCC GCTTCCTCGC TCACTGACTC GCTGCGCTCG       720
GTCGTTCGGC TGCGGCGAGC GGTATCAGCT CACTCAAAGG CGGTAATACG GTTATCCACA       780
GAATCAGGGG ATAACGCAGG AAAGAACATG TGAGCAAAAG GCCAGCAAAA GGCCAGGAAC       840
CGTAAAAAGG CCGCGTTGCT GGCGTTTTTC CATAGGCTCC GCCCCCCTGA CGAGCATCAC       900
AAAAATCGAC GCTCAAGTCA GAGGTGGCGA AACCCGACAG GACTATAAAG ATACCAGGCG       960
TTTCCCCCTG GAAGCTCCCT CGTGCGCTCT CCTGTTCCGA CCCTGCCGCT TACCGGATAC      1020
CTGTCCGCCT TTCTCCCTTC GGGAAGCGTG GCGCTTTCTC ATAGCTCACG CTGTAGGTAT      1080
CTCAGTTCGG TGTAGGTCGT TCGCTCCAAG CTGGGCTGTG TGCACGAACC CCCCGTTCAG      1140
CCCGACCGCT GCGCCTTATC CGGTAACTAT CGTCTTGAGT CCAACCCGGT AAGACACGAC      1200
```

Figure 76. Nucleotide sequence of 1kbp fragments

(a) G20C_F1 and (b) pUC18_F1. The first and last residues of each fragment are shaded in purple. Every structural or functional element is shaded following the same colouring scheme from the previous figure. A-tracts are shown in bold letters.

Several A-tracts are present in both DNA fragments (Figure 76). Although A-tracts contribute to the bending of DNA, making it more prone to interact with DNA-binding proteins, the purpose of the experiments carried out on this project addressed sequence specificity rather than the influence of the DNA’s conformation. The degree of curvature of the designed fragments was not determined. Both sequences are expected to have a similar degree of curvature in spite of pUC18_F1 having more A-tracts than G20C_F1. The irregular spacing between the A- tracts might attenuate the effect of the A-tracts since they are not continuous, and therefore do not reinforce each other to produce bending.

9.2 Amplification of G20C_F1 and pUC18_F1

The experimental stage of the production of G20C_F1 and pUC18_F1 consisted of using recombinant DNA from the motor operon (cloned from the genomic DNA) as the template to amplify G20C_F1. The full-length plasmid pUC18 was the template to produce pUC18_F1. The manufacturer’s instructions for the DreamTaq DNA polymerase (Thermo SCIENTIFIC) were followed to set up the PCR (Figure 76), using the program described below.

G20C_F1 / pUC18_F1	
94°C	2 min
94°C	30 s
60°C	30 s
72°C	45 s
72°C	5 min

36 cycles

100 µl PCR reactions were set up and run in a 0.75 % agarose gel stained with SYBR Safe (Life technologies). The main band approximately 1 Kbp in size was excised from the gel and purified with the GeneJET gel extraction kit (Fermentas, Life Sciences). DNA concentration was measured in ng·µL^{-1} and converted to µM to follow the procedure described in section 2.16.

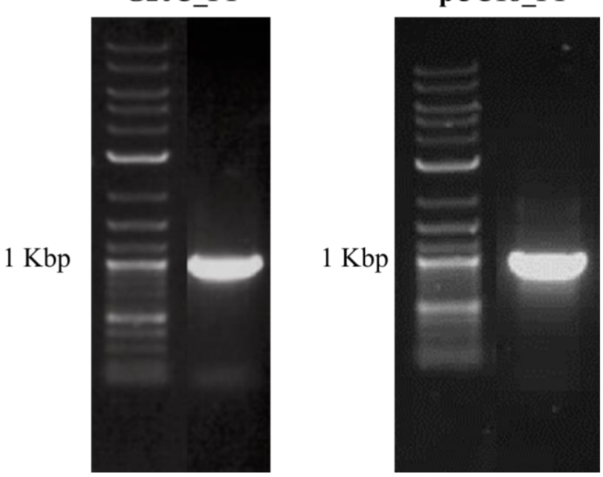


Figure 77. PCR for G20C_F1 and pUC18_F1

9.3 Amplification of 150 bp fragments

In addition to the two 1 kbp DNA fragments described above, eight 150 bp fragments were amplified from the G20C genome, named G20C_F2 – F9 (Figure 78). Some of these segments lay either within the G20C_F1 sequence or in regions adjacent to its 3’ end. The purpose of producing shorter DNA fragments was to track the region recognised by the ST to start with DNA translocation. The regions comprised by each fragment and corresponding primers are listed in Table 21.

The proposed strategy consisted of detecting binding to a long DNA and then to screen the binding to shorter fragments within or adjacent to G20C_F1. The region covered by the 150 bp fragments on G20C genome was 42,098 – 43,197. The fragments overlapped each other by 20 bp at 5’ and 3’ ends.

All forward and reverse primers included 21 bp linker regions which were complementary to another subset of primers labelled with different fluorophores. After nested PCRs, every fragment would be labelled with a pair of fluorophores suitable for fluorescence resonance energy transfer (FRET). This technique uses a biomolecule, DNA in our case, covalently linked to fluorescent probes, the energy donor (D) and acceptor (A). Exciting probe D by a light source leads to the release of energy, which is transferred to probe A. For this energy transfer to happen the probes need to be close together.

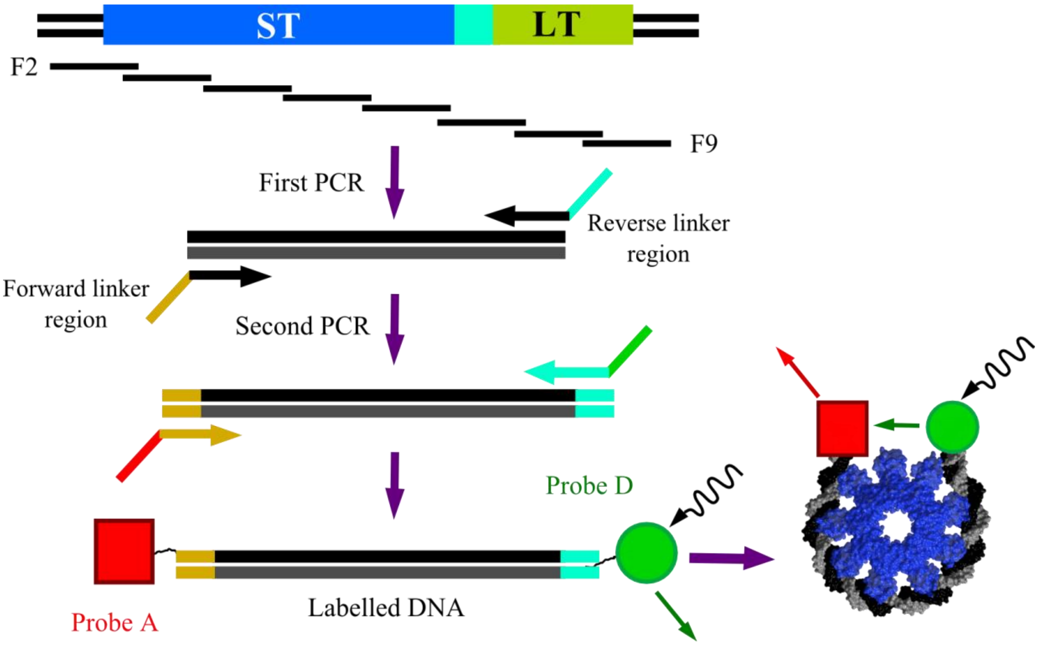


Figure 78. Strategy to locate the packaging initiation site in the G20C genome

In the hypothetical case that DNA uses the NTDs to wrap around the ST, the ST would be able to discriminate the fragment containing the recognition region from the rest. Then the DNA would acquire a bent conformation due to the presence of several NTDs, bringing both labelled ends to proximity, which would allow measuring the amount of energy transferred to probe A from probe D. Short distances, high FRET states, would only be achieved by fragments that bend around the ST. Fragments that did not interact with the ST would not bend and the probes would remain away from each other in a low FRET state.

Fragments were amplified by PCR using DreamTaq^{TM} DNA polymerase following the manufacturer’s instructions and the program below:

94°C	2 min
94°C	30 s
60°C	30 s
72°C	10 s
72°C	5 min

During this thesis project only the fragment G20C_F3 was partially characterised in the presence of the construct ST_NTD due to time limitations.

Table 21. Primers to amplify 150 bp fragments from G20C’s genome

Fragment	Region amplified	Primers
G20C_F1	42,098 – 43,097	Forward: 5’ CGC CCT CCT CTT TTC CAC CGC GGA TGA GCT 3’ Reverse: 5’GGG AGC AAT AAT CCA GCC CTG GCT TCC AGG 3’
G20C_F2	42,138 – 42,287	Forward: 5’ CCA GGA GGA CGC AGG GCA GGT GAC GAC CTG GTT CAG AGC CTG GCG ACT CC 3’ Reverse: 5' TAA GGA ACC CCC TAT CCG GGG CGA ATG CAC ACG TCG CAA AGG GTC TCG TA 3'
G20C_F3	42,268 – 42,417	Forward: 5’ CCA GGA GGA CGC AGG GCA GGT CTT TGC GAC GTG TGC ATT CGT AAA TAC GA 3’ Reverse: 5' TAA GGA ACC CCC TAT CCG GGG TCC CTA AAA CTC ACG CTC ATC GCT TTC TA 3'
G20C_F4	42,398 – 42,547	Forward: 5’ CCA GGA GGA CGC AGG GCA GGT ATG AGC GTG AGT TTT AGG GAC AGG GTG CT 3’ Reverse: 5' TAA GGA ACC CCC TAT CCG GGG AAG AAG CTC TCT AGC CTC GGC TAG GAC GT 3'
G20C_F5	42,528 – 42,677	Forward: 5’ CCA GGA GGA CGC AGG GCA GGT CCG AGG CTA GAG AGC TTC TTT CCG CCT TG 3’ Reverse: 5' TAA GGA ACC CCC TAT CCG GGG CTT CCA TCA TGG CGT TGT AGT TGC GGA GC 3'
G20C_F6	42,658 – 42,807	Forward: 5’ CCA GGA GGA CGC AGG GCA GGT CTA CAA CGC CAT GAT GGA AGG CTT GAC CG 3’ Reverse: 5' TAA GGA ACC CCC TAT CCG GGG AGT ACC TTG AGG CTA TTG AGC AGA GCG GT 3'
G20C_F7	42,788 - 42,937	Forward: 5’ CCA GGA GGA CGC AGG GCA GGT CTC AAT AGC CTC AAG GTA CTA GGC CAT GT 3’ Reverse: 5' TAA GGA ACC CCC TAT CCG GGG CCT AGA AGC TCA AAG AAC TTG TCG CTA GG 3'

Table continues on the next page

Table 20. Primers to amplify 150 bp fragments from G20C’s genome. Continuation

Fragment	Region amplified	Primers
G20C_F8	42,918 – 43,067	Forward: 5’ CCA GGA GGA CGC AGG GCA GGT AAG TTC TTT GAG CTT CTA GGC TAC AAG CC 3’ Reverse: 5' TAA GGA ACC CCC TAT CCG GGG CCG GGC GAA GAG CTC AAA CAC GGC CTC TA 3'
G20C_F9	43,048 – 43,197	Forward: 5’ CCA GGA GGA CGC AGG GCA GGT TGT TTG AGC TCT TCG CCC GGC CTG GAA GC 3’ Reverse: 5' TAA GGA ACC CCC TAT CCG GGG GTC TAC GTT GAA GCT GAA CCT CGG TGG TG 3'
pUC18_F1*	185 - 1,185	Forward: 5’ ATG CGG TGT GAA ATA CCG CAC 3’ Reverse: 5’ GTT GGA CTC AAG ACG ATA CTT 3’

* Region corresponding to pUC18 vector

🟨 Forward linker, 🟦 Reverse linker

Chapter 10

10.DNA-binding activity of the G20C small terminase

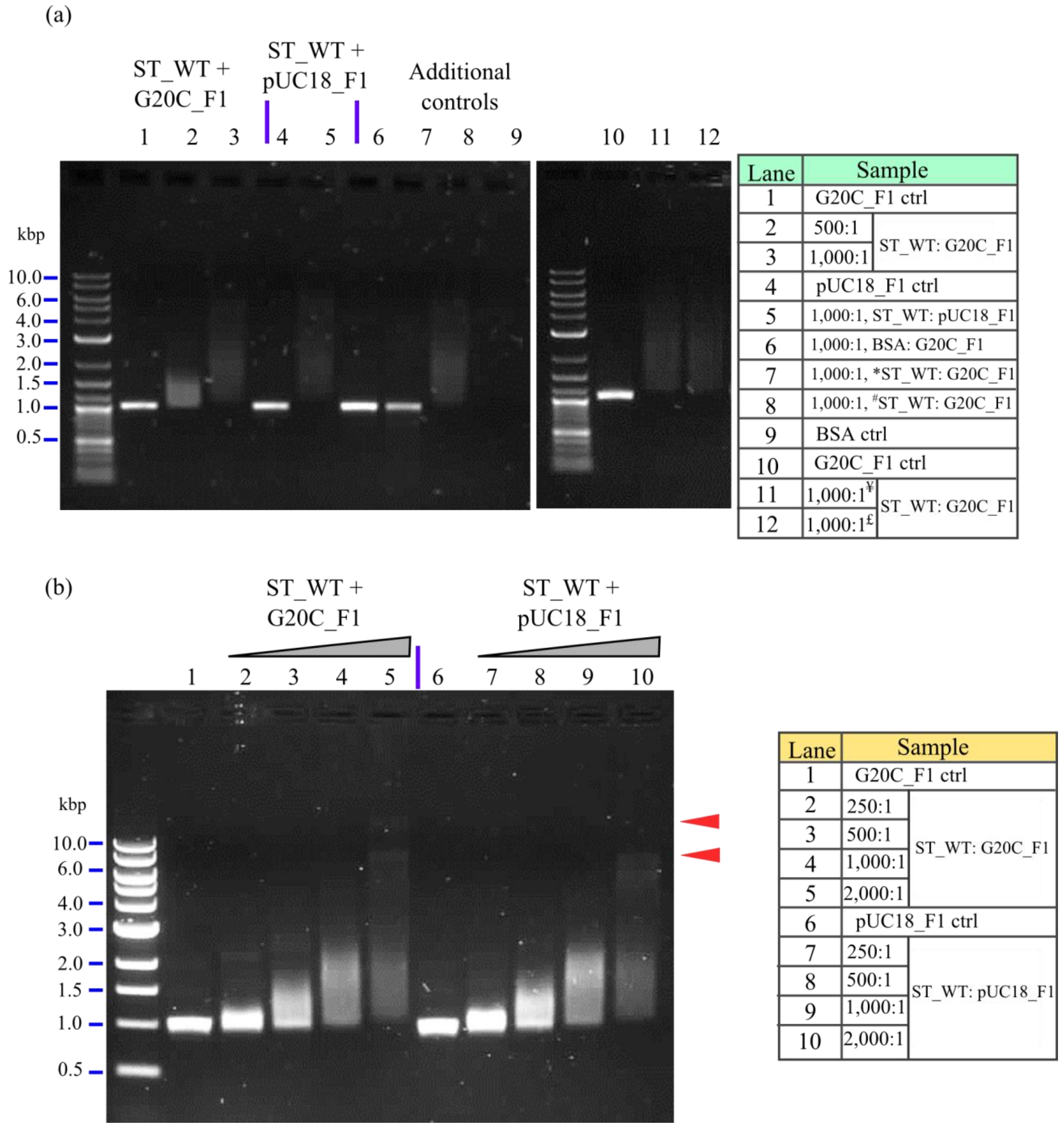
This chapter summarises the experiments carried out to test whether G20C ST_WT exhibited DNA-binding activity and the conditions under which activity is evident. Likewise, the constructs ST_ΔC, ST_ΔN and ST_NTD, missing individual domains, were tested to assign a role in DNA binding to every domain. The available experimental information about the function of the ST from dsDNA phages, namely recognition of DNA was used to compare with that of G20C ST. As a literature review, section 10.4 condenses the data available that describe the regulation of the LT ATPase and nuclease activities by the ST. Finally, the structure and functional data produced on this project for G20C ST were combined to produce a model of the potential G20C ST-DNA interaction.

10.1 DNA-binding by the G20C ST_WT

The initial screening to detect any DNA-binding activity in G20C ST_WT consisted of investigating its capability to bind G20C_F1 and pUC18, which correspond to DNA of viral and non-viral origin (Figure 79a).

Bovine serum albumin (BSA) on its own, and in the presence of G20C_F1 was included in the initial screen as the negative protein control to demonstrate how the binding exhibited by the ST_WT is a property of DNA-interacting proteins that is not present in random proteins such as BSA. Temperature was another variable tested. The optimal growth temperature of *Thermus thermophilus*, host of G20C phage, is 65 ºC. Since the extent of binding observed at 37 ºC appears to be the same at 60 ºC, it was decided to characterise G20C ST at 37 ºC. At 37 ºC proteins from thermophilic microorganisms are expected to be highly stable. Further biochemical characterisation could involve incubation at the G20C's physiological temperature. Samples of Se-Met labelled protein and a 2nd batch of DNA were included to screen the properties of the reagents available for the experimental work. The Se-Met ST_WT did not exhibit binding to DNA. The reason for such behaviour could either be due to the presence of Selenium atoms, or the time elapsed between the protein purification and the setup of the DNA-binding experiments. All other experiments were carried out with native protein freshly purified. G20C ST_WT exhibited binding to the second batch of DNA. Batches of DNA were produced following the same protocol and tested before being included in the shown experiments.

Once binding to DNA was detected, a more detailed experiment was set up to test sequence specificity using several protein:DNA molar ratios (Figure 79b). Binding to viral and non-viral DNA was investigated in order to identify whether G20C ST exhibits any preference towards its own DNA since this is the first structural and biochemical characterisation performed on G20C ST and no information about the specific recognition site is available.



Lane	Sample	
1	G20C_F1 ctrl	
2	500:1	ST_WT: G20C_F1
3	1,000:1	
4	pUC18_F1 ctrl	
5	1,000:1, ST_WT: pUC18_F1	
6	1,000:1, BSA: G20C_F1	
7	1,000:1, *ST_WT: G20C_F1	
8	1,000:1, #ST_WT: G20C_F1	
9	BSA ctrl	
10	G20C_F1 ctrl	
11	1,000:1¥	ST_WT: G20C_F1
12	1,000:1£	

Lane	Sample	
1	G20C_F1 ctrl	
2	250:1	ST_WT: G20C_F1
3	500:1	
4	1,000:1	
5	2,000:1	
6	pUC18_F1 ctrl	
7	250:1	ST_WT: pUC18_F1
8	500:1	
9	1,000:1	
10	2,000:1	

Figure 79. EMSA to investigate the DNA-binding activity of G20C ST_WT

(a) Initial screen and additional controls. * G20C Se-Met Se-Met labelled, # second batch of G20C_F1, ¥ Incubation at 37 ºC, £ Incubation at 60 ºC. Final reaction conditions: 25 mM Tris pH 7.5, 113 mm NaCl, 20 mM Mg^{+2}, 10 mM ATP (b) Binding of G20C ST_WT to G20C_F1 (viral) and pUC_F1 DNA (non-viral) fragments. Final reaction conditions: 25 mM Tris pH 7.5, 100 mm NaCl, 20 mM Mg^{+2}, 10 mM ATP. The red arrowheads indicate the positions of high MW species.

G20C ST_WT bound in a concentration dependent manner to both G20C_F1 and pUC18_F1 (Figure 79b). DNA molecules 1 kbp long, mixed with G20C ST_WT in the ratios 250:1 to 2,000:1 (protein to DNA), exhibited retarded migration by EMSA as compared to DNA alone, indicating that complexes with DNA were formed. The extent of retardation increased when the molar ratio of ST_WT was higher, suggesting that multiple ST_WT molecules could bind to a single DNA molecule in a potentially cooperative manner.

G20C ST_WT did not form tightly bound complexes of specific stoichiometry with DNA since defined bands, that would form in the case of strong protein:DNA interactions, were replaced by diffuse DNA bands that represent transient complexes that re-equilibrate during electrophoresis. In other words, several 9-mers could possibly interact with a single DNA molecule. Having more 9-mers in the reaction means that more complexes can form on each DNA molecule therefore increasing the retardation of migration. Since the interaction with DNA seems not to form only one type of complex, and the interaction is relatively weak, it is likely that these complexes dissociate into their protein or nucleic acid components, leaving every component free to bind to other molecules or complexes.

According to the experimental data, G20C ST_WT does not have preference for a specific DNA sequence. The binding to either viral or non-viral DNA fragments shows no difference despite having different origins and a sequence homology of 21.2 %. The same behaviour is observed for both fragments at the same molar ratios. The only difference observed is the formation of two faint, but discrete, bands when G20C ST_WT is incubated with viral DNA at ratio 2000:1. The two bands could correspond to two stable high MW species, and migrate similar to the 10 kbp standard. Such bands are not observed when non-viral DNA is mixed with G20C ST_WT.

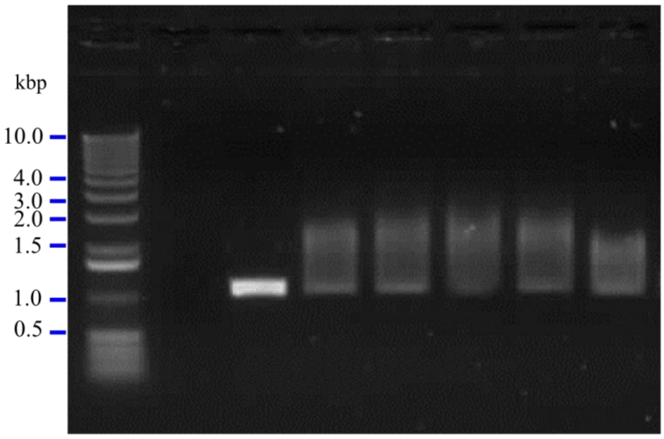
10.1.1 Adjusting the buffer reaction conditions

At the same time that the preliminary experiments (Figure 79a) to investigate binding to DNA, optimal temperature and random proteins binding to DNA as control were run, the buffer conditions were varied in order to determine what reagents were necessary for complex formation (Figure 80a). Our experiments follow a similar protocol used for observing the binding of Sf6 ST to DNA (68). In these assays DTT as well as ATP and Mg^{+2} was included in the reaction buffer. For G20C ST, it was observed that DTT does not affect the formation of ST:DNA complexes. However, the presence of either ATP or Mg^{+2} resulted in DNA shifts that were similar to the shifts observed in the presence of all components.

Several Mg^{+2} concentrations were tested to identify whether variations in Mg^{+2} or its replacement by Mn^{+2}, Ca^{+2} or Zn^{+2} cations produced differences in the DNA-binding. Reactions including 5-30 mM Mg^{+2} and 10 mM Mn^{+2}, Ca^{+2} or Zn^{+2} were setup using molar ratios 2,000:1 and 1,000:1. To discount any effect of the increasing ionic strength, the total ionic strength of the final reaction volume was adjusted to be the same in all reactions by varying the NaCl concentration. It was observed that the binding to DNA follows the same behaviour when the concentrations of Mg^{+2} changed. No significant differences were observed when 10 mM of another divalent cation was included in the reaction. For all subsequent experiments and for gel b in Figure 79a 20 mM Mg^{+2} and 1 mM ATP were used.

(a)

	1	2	3	4	5	6	7
ATP	✓	✓	✓	✓	✓	x	x
Mg^{+2}	✓	✓	✓	✓	x	✓	x
DTT	✓	✓	✓	x	x	x	x
DNA	x	✓	✓	✓	✓	✓	✓
Protein	✓	x	✓	✓	✓	✓	✓



(b)

Lane	1	2	3	4	5	6	7	8	9
Mg^{+2}		1 mM		5 mM		10 mM		20 mM	
ST_WT: G20C_F1		2000:1	1000:1	2000:1	1000:1	2000:1	1000:1	2000:1	1000:1
NaCl		176 mM		160 mM		140 mM		100 mM	

10	11	12	13	14	15	16	17	18
30 mM			10mM Mn^{+2}		10mM Ca^{+2}		10mM Zn^{+2}	
2000:1	1000:1		2000:1	1000:1	2000:1	1000:1	2000:1	1000:1
60 mM			140 mM		140 mM		140 mM	

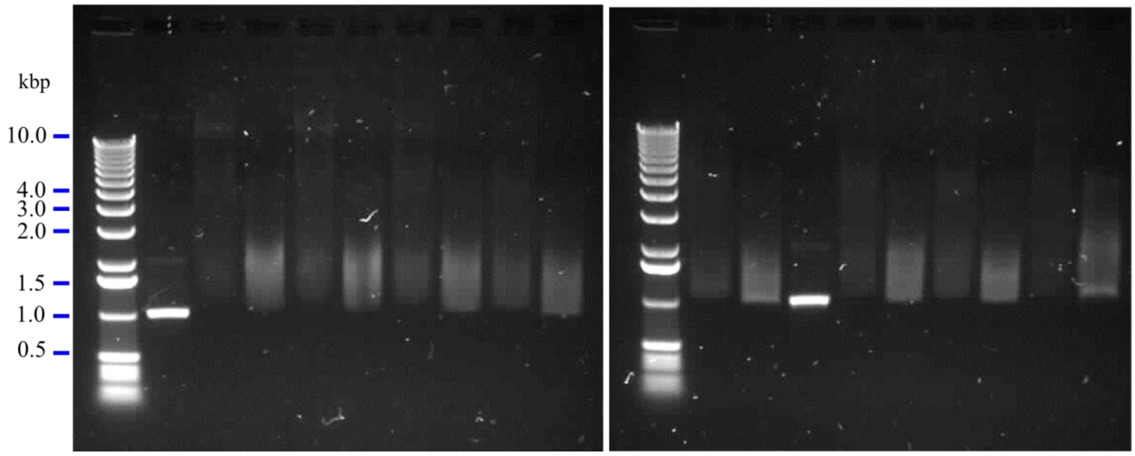


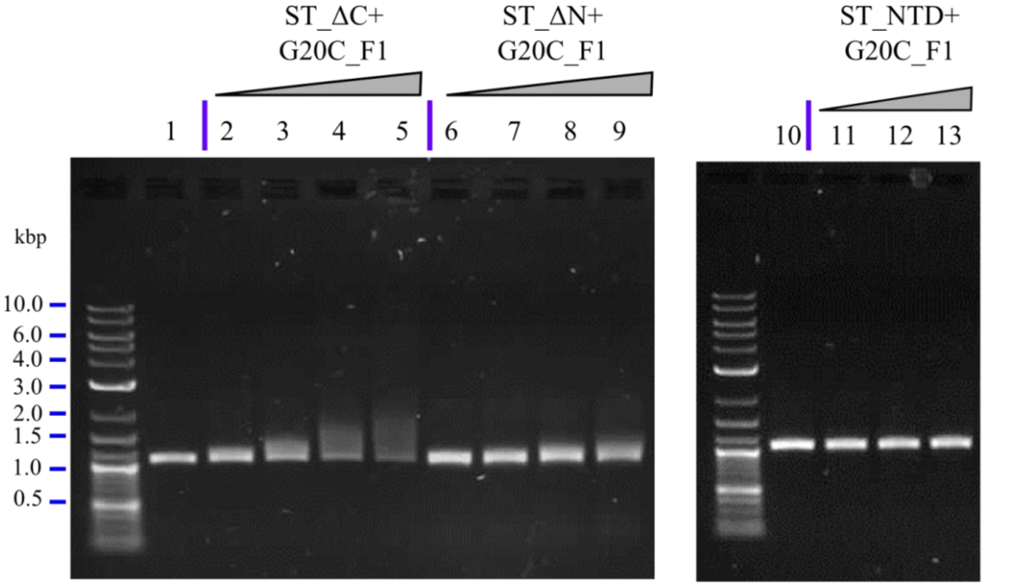
Figure 80. Effect of ATP, DTT and metal in binding to DNA

a) EMSA to test the effect of 1 mM ATP, 20 mM Mg^{+2} and 1mM DTT in DNA-binding. Lanes 1 and 2 contain the protein- and DNA-only controls. Lanes 3-7 include both components in 500:1 molar ratio. (b) Effect of Mg^{+2} concentration and other metals in DNA-binding. Lanes 1 and 12 contain DNA controls in 20 mM Mg^{+2}. Reactions from both lanes also contain 25 mM Tris pH 7.5 and 1 mM ATP.

10.2 DNA-binding of the G20C ST_ΔC, ΔN and NTD constructs

Interaction with DNA was clearly evident for the ST_WT. The purpose of designing and producing truncated constructs was to assess the role of the terminal and core domains in binding to DNA. Figure 81 shows the screen carried out for each construct mixed with G20C_F1 in ratios ranging from 500:1 to 3,000 for ST_ΔC and ΔN constructs, and 4,500–13,500 for ST_NTD. ST_ΔC and ΔN were demonstrated to be 9-mers in solution, hence the ratios are calculated for 9-mers of truncated protein to 1 kbp DNA molecules. Since ST_NTD exists as a monomer in solution, the molar ratios used were chosen considering this in determining the ratio of ST_WT: 1 kbp molecule, where every molecule of 9-mer contributes 9 NTDs, for instance, the molar ratio 4,500:1 (NTD:DNA) includes the same number of NTD molecules as does the ratio 500:1 for 9-mer:1kbp DNA. The same logic was followed to include ratios 9,000:1 and 13,500:1, ST_NTD: 1 kbp.

ST_ΔC showed increased binding to G20C_F1 in molar ratios 1,000:1 - 3,000:1, while ST_ΔN only showed residual binding. The ST_NTD construct did not show any binding (Figure 81). All experiments were carried out in 25 mM Tris pH 7.5, 100 mM NaCl, 1 mM ATP, 20 mM Mg^{+2}.

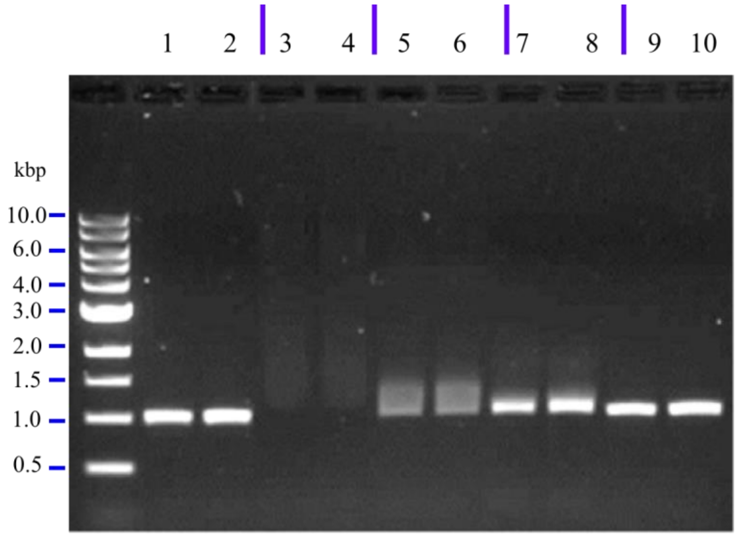


Lane	Sample	
1	G20C_F1 ctrl	
2	500:1	ST_ΔC: G20C_F1
3	1,000:1	
4	2,000:1	
5	3,000:1	
6	500:1	ST_ΔN: G20C_F1
7	1,000:1	
8	2,000:1	
9	3,000:1	

Lane	Sample	
10	G20C_F1 ctrl	
11	4,500:1	ST_NTD: G20C_F1
12	9,000:1	
13	13,500:1	

Figure 81. EMSA of DNA binding activity of the G20C ST_ΔC, ΔN and NTD constructs

The gel on Figure 82 compares the DNA-binding activity of all G20C ST constructs produced on this study when incubated with G20C_F1 and pUC18_F1. With the exception of ST_NTD, all the reaction show ratios of 2,000:1 as it is the only ratio where ST_ ΔN exhibits residual binding. The equivalent ratio, 18,000:1, for ST_NTD was set up. In contrast to ST_WT, which clearly bound to both DNA fragments, constructs ST_ΔC and ST_ΔN exhibited dramatic loss of binding at the same ratios. Similarly to ST_WT, no difference is observed for the constructs when incubated with DNA of viral and non-viral origin, suggesting that the recognition of DNA, under the tested experimental conditions, is non-sequence specific.

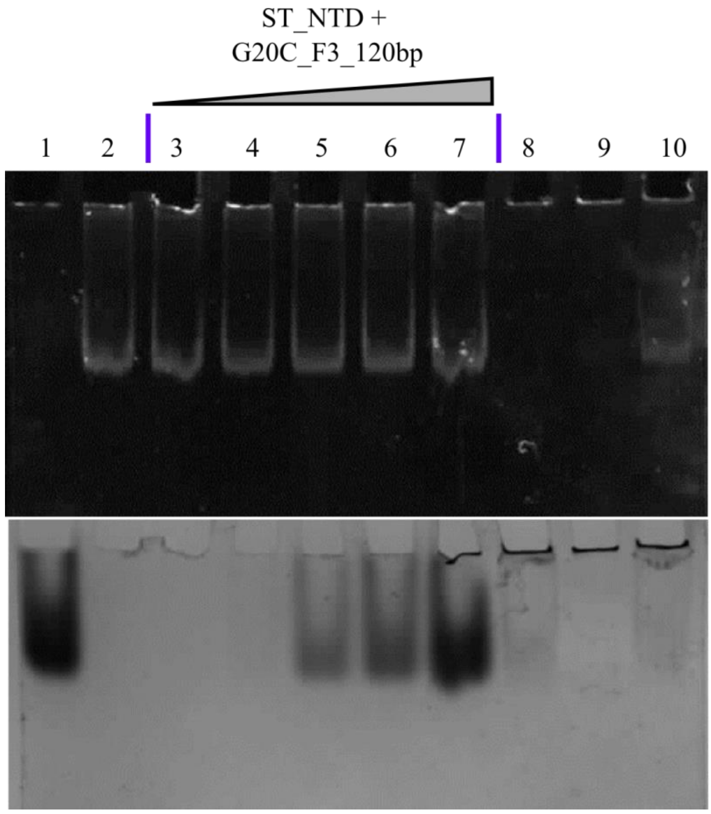


Lane	Sample
1	G20C_F1 ctrl
2	pUC18_F1 ctrl
3	2,000:1, ST_WT: G20C_F1
4	2,000:1, ST_WT: pUC18_F1
5	2,000:1, ST_ΔC: G20C_F1
6	2,000:1, ST_ΔC: pUC18_F1
7	2,000:1, ST_ΔN: G20C_F1
8	2,000:1, ST_ΔN: pUC18_F1
9	18,000:1, ST_NTD: G20C_F1
10	18,000:1, ST_NTD: pUC18_F1

Figure 82. EMSA of the binding of G20C_WT and constructs to viral and non-viral DNA

To further investigate the DNA-binding activity of ST_NTD in presence of G20C_F3 native EMSA experiments were carried out (Figure 83). 7.5 % polyacrylamide gels were prepared and run in [0.5X tris borate EDTA (TBE) [5X TBE: 13.5 g Tris, 6.88 boric acid, 1.163 g EDTA]]. Reactions including ST_NTD (monomer in solution) to 150 bp DNA molar ratio of 50:1–2000:1 were prepared in 25 mM Tris pH 7.5, 50 mM NaCl and 10 mM Mg^{+2}. An EMSA was carried out in native polyacrylamide gels, as opposed to agarose, as this made it feasible to detect differences in protein migration with Coomassie blue stain as well as DNA with SYBR® Safe.

No binding of G20C ST_NTD to G20C_F1 was observed at any molar ratio of the experimental set up.



Lane	Sample	
1	ST_NTD ctrl	
2	G20C_F3_120bp ctrl	
3	50:1	ST_NTD : F3
4	100:1	
5	500:1	
6	1,000:1	
7	2,000:1	
8	BSA ctrl	
9	Empty	
10	1,000:1 BSA : F3	

Figure 83. EMSA testing the binding of G20C ST_NTD to 150 bp fragments

The top picture corresponds to the SYBRsafe® stain. At the bottom, the same gel stained with Coomassie blue.

10.3 Functional comparison of ST:DNA binding activity from different viruses

Binding to DNA has been proved for G20C (this study), SF6 (16), Sf6 (68,136), P22 (14,73), PaP3 (15), and Lambda (147) STs, however the nature of the recognised sequence and how the structural domains achieve binding to DNA are still important questions in order to understand the DNA translocation process.

G20C ST is not the only ST that interacts with both viral and non-viral DNA. Sf6 ST besides interacting with ~1,800 bp linear DNA fragments that comprise Sf6 ST-LT region (68), also showed binding to random linear DNA fragments from a DNA ladder, with apparent increased affinity towards long DNA. Moreover, Sf6 formed more stable complexes with commercial supercoiled DNA than with long linear DNA, suggesting that the information that this ST recognises lies in the supercoiling of the DNA (conformation) rather than in the sequence itself (136). Nevertheless, Sf6 possesses a unique *pac* site, close to the ST gene, that is able to start DNA packaging *in vitro* (24).

P22 ST also demonstrated binding to viral and random DNA (73). The DNA binding activity of SF6 was only assessed with viral DNA that included the region corresponding to the ST and LT genes. For SF6, the SPP1 recognition site was used as it has been shown that both SPP1 and SF6 ST proteins, and probably the *pac* sites are homologous (148). No experiments including non-viral DNA were published to demonstrate whether SF6 ST discriminates the *pac* site from another source of DNA. However, the NTD of SF6 does not show any preference for sequences of viral origin (Maria Chechik, unpublished data). PaP3 ST only binds to *cos*-containing PaP3 genomic DNA but fails to bind to fragments of the same length taken downstream of the *cos* site.

The set of experiments carried out for G20C ST allowed the identification of DNA-binding activity towards both viral and non-viral DNA. Nevertheless, these experiments were not designed to identify or allocate any specific ST recognition site. In the structurally similar Sf6 ST, from *in vivo* experiments, the *pac* site has been assigned to a short sequence located at the centre of the region coding for the DBD and OD, between bp 154-183. When this sequence was compared with P22 *pac* site, it was found that Sf6 has seven of the thirteen bp that are essential for recognition in P22 (Figure 84). Moreover, the encoded amino acids of both *pac* sites reside at the top of the external α-helix, immediately after the DBD, in case of Sf6. It is of note that both P22 and Sf6 *pac* sites are in the same position in the oligomer, pointing towards the fact that the recognition sequence remains within the region coding for the protein sequence responsible for DNA binding, possibly to avoid being separated during horizontal gene exchange (24). The alignment of Sf6 or P22 *pac* sites with the region around G20C ST did not show significant sequence homology. Nonetheless, when G20C OD was superposed onto that of Sf6, it was found that some of residues encoded by the *pac* region on Sf6 and P22, which are the same in these two ST, are also present in G20C (Figure 84). Additionally, some similarities (charged or nonpolar residues) between G20C OD and that of either Sf6 or P22 are observed.

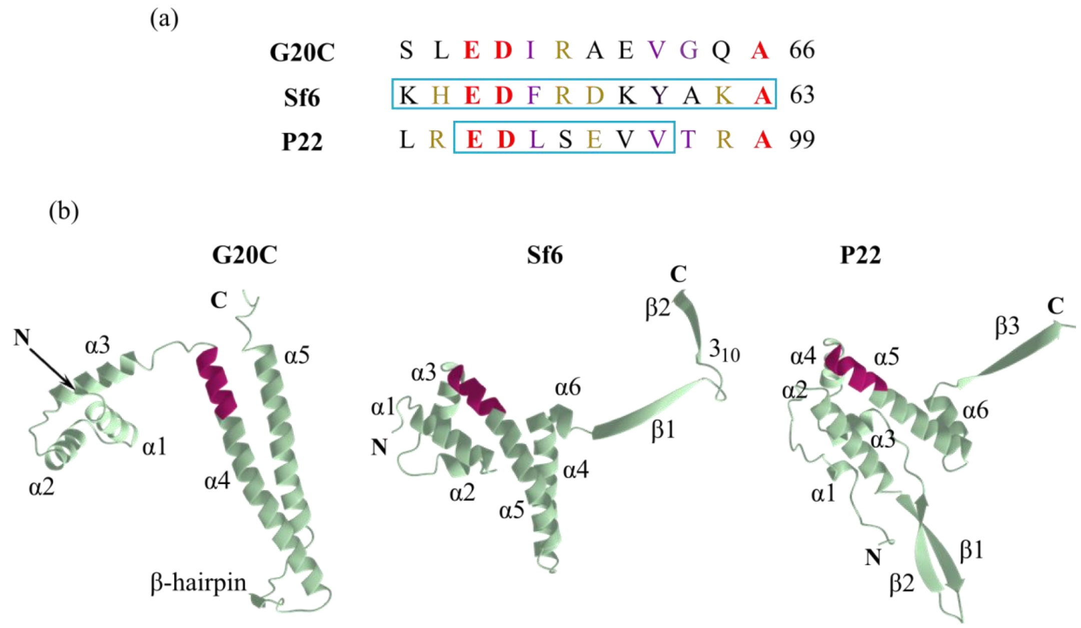

(a)
G20C S L E D I R A E V G Q A 66
Sf6 K H E D F R D K Y A K A 63
P22 L R E D L S E V V T R A 99
(b)
G20C
N
C
α3
α5
α1
α2
α4
β-hairpin
Sf6
C
β2
3₁₀
α3
α6
α1
N
β1
α4
α2
α5
P22
C
β3
α4
α5
α2
α6
α3
α1
N
β1
β2

Figure 84. Location of Sf6 and P22 *pac* sites within G20C ST sequence

(a) Comparison of the amino acids encoded by Sf6 and P22 *pac* sites with those of G20C. The colouring of amino acids is: red, for identical, purple for hydrophobic residues, and gold for residues with the same charge. Squared areas correspond to the residues encoded by the minimum DNA sequence with *pac* activity. (b) Location of residues from (a) on the X-ray structure. Figure format adapted from (24).

Although G20C belongs to the *Siphoviridae* family and Sf6 and P22 to *Podoviridae*, this manual alignment based on secondary structure raises the question about how related are the DNA-recognition mechanisms used by phages from different families. More experiments should be performed to confirm the presence and location of a *pac* site in phage G20C, which could be “hidden” somewhere in the genome and due to the lack of homology with other *pac* sites could not be detected.

The experiments performed in this study establish the G20C ST NTD as the DNA-binding domain (G20C ST NTD will be referred as G20C DBD henceforth) since its removal leads to only residual binding activity in comparison with ST_WT. It is important to notice that the removal of the CTD dramatically reduces the binding activity too, though not to the same extent that removal of the DBD. The residual binding of ST_ΔN is attributed to positively charged residues (Figure 68d) at the outer surface of the OD that become exposed when the DBD is removed. The same residual binding activity was observed for SF6 ST constructs missing the DBD. SF6 ST_ΔN constructs, like G20C ST_ΔN expose positive residues at the outer surface as well, in fact, mutation of one of these residues further diminished the DNA-binding activity (16).

In SF6 and Sf6 STs, the NTD has been established as the DBD because its removal in SF6 (16), or mutations of positively charged residues, in Sf6, greatly decreases or abolishes the DNA binding activity (136). Functional studies on Sf6 revealed that mutation of positively charged residues at the start of the NTD (unstructured in the crystal structure) and at helices α3, α4 and α6 to negatively charged residues, significantly reduced the DNA binding activity. Only the first of the above described mutations was introduced at the NTD, while the rest, in helices α3 and α4 were at immediately adjacent areas (136). For SPP1 ST, constructs missing the CTD but keeping the NTD or featuring the CTD of SF6 ST bound to DNA; conversely the construct missing the first 62 residues did not (148).

The behaviour of the CTD in DNA-binding appears to be more versatile. On SF6 ST, removal of the CTD in constructs missing the DBD did not produce decreased DNA-binding activity but produced 10-subunit assemblies, in contrast to 9-subunit assemblies that formed when the CTD was present (16). For Sf6 there is no published information about the role of the CTD in binding to DNA. The DNA-binding activity of P22 ST was ablated in comparison to the WT when the C-terminal 22 residues, where the unstructured helix α7 is predicted to be, were removed. This portion of P22 ST is speculated to be the DBD since the HTH motif, present in G20C, SF6 and Sf6, is absent. Experimental evidence that discards the participation of other portions of P22 ST, for instance some α helices from the helical core, in binding to DNA are still necessary to assign such function to the P22 ST CTD (14,73).

G20C ST presents unique characteristics in the sense that both DBD and CTD contribute to different extents in DNA-binding activity. The DBD accounts for most of the binding, but the presence of the CTD greatly increases the interaction with DNA. G20C ST shares similar architecture for the DBD with SF6 and Sf6 ST, and for the three of them the DBD seems to be the most essential domain, however, G20C ST differs in that the CTD is also required for efficient DNA binding, while in SF6 and Sf6, no role in DNA-binding has been given to the CTD. G20C and P22 ST share the involvement of the CTD in recognition of DNA although in G20C ST, its deletion does not lead to the complete loss of activity. For 44RR or PhBC6A51 ST no functional studies on individual domains were published.

10.4 Revising the ST-LT interaction: Regulation of the LT ATPase and nuclease activities and complex formation

10.4.1 Regulation of the LT ATPase and nuclease activities by the ST

In addition to the DNA-binding activity, the ST also modulates the LT enzymatic activities.

SPP1, T4, P22 and PaP3 ST increase the ATPase activity of their respective LT though some peculiarities apply to each system. In SPP1 it was observed that the presence of the ST (G*1*P) increases the amount of hydrolysed ATP (13) and further studies demonstrated that the SPP1 portal (G*6*P) also stimulates this activity, although to a lesser extent than the ST. ATP hydrolysis was increased even higher when the LT (G*2*P), ST and portal were combined (149). What is interesting is the fact that keeping concentrations of G*2*P and G*6*P constant while increasing the amount of G*1*P leads to the increase in the ATPase activity, however it decreases when the G*1*P reaches a certain ratio with respect to G*2*P. The same behaviour is observed when both G*1*P and G*2*P concentrations are constant but that of G*6*P increases (149). T4 gp17-ATPase (LT) is only stimulated by the ST_WT (gp16) and by a gp16 NTD-CTD construct but not by the individual central or terminal domains (17,69,150). For P22 ATPase (gp2) it was demonstrated that the ST (gp3) and *pac*-containing DNA are required to produce a significant rise in ATP hydrolysis in comparison to gp2 with DNA alone. Moreover, removal of the last 22 residues in gp3, abolished ATP hydrolysis (14). PaP3 ST (p01) exists in solution as either monomer or dimer, but it is the dimer species that increases the ATPase (p03) activity the most in a ratio p03:p01 of 1:1 or 1:2 (15).

SPP1 (13) and T4 (69) STs bind to DNA but do not hydrolyse ATP, while P22 (14) ST displays residual ATPase activity on its own, which is absent when DNA is supplied. PaP3 p01 does not exhibit any ATPase activity on its own (15).

For the DNA-translocation process to be successful it is indispensable that the LT nonspecific endonuclease activity is reduced after cutting at the packaging initiation site, such inhibition in some phages is provided by the ST.

The endonuclease activity of SPP1 G*2*P is not inhibited when G*2*P and G*6*P concentrations are constant but G*1*P's is low. As the amount of G*1*P rises, the function of G*2*P endonuclease is permanently inhibited. G*6*P in the other hand is not able to prevent such inhibition when G*2*P and G*1*P concentrations are constant, but G*1*P's is high, even with increasing amount of G*6*P (149). In T4 phage, WT gp16 and a NTD-CTD construct inhibit gp17 endonuclease whereas very high concentrations of the central OD are required to achieve the same degree of inhibition. Individual NTDs or CTDs alone are not enough to stop digestion of DNA when incubated with gp17 (69). P22 gp2 endonuclease activity is reinforced by the presence of ATP, indicating how both ATPase and endonuclease cross-talk to regulate the latter activity. Addition of gp3 inhibits DNA digestion but such inhibition is not overcome even if ATP is present (151). In a similar example of domain cross-talk, the Sf6 LT endonuclease activity is also increased by a higher concentration of ATP. This study, involving the Sf6 ST and LT, did not address the effect of DNA in the experiment (152). Studies on PaP3 p03 showed increased endonucleolytic activity towards commercial DNA when p01 concentration was raised to a p01:p03 ratio of 4:1 (74).

The ST has been shown to be involved in *in vitro* DNA packaging. T4 gp16 stimulates DNA packaging in *in vitro* experiments that mimic *in vivo* packaging due to the inclusion of crude lysate of *E.coli* cells infected with phage terminase-deficient phage (*16am17am23amrII),* T4 *r*II^{+} phage DNA, and phage completion gene products (neck-tail and head completion proteins required for phage assembly) in addition to the recombinant gp16 and gp17. *am* stands for amber mutations, which generate the stop codon UAG. Phages carrying this mutation are generally able to infect only one type of bacterial strain, known as the amber suppressor since in these strains phages recover the altered function. *r* stands for rapid, which corresponds to a mutation that causes faster lysis of bacteria than the WT. *r*II loci refers to the position in the T4 chromosome (153). In contrast, gp16 inhibits the DNA packaging produced by functional systems that only feature expanded procapsids, gp17_ΔC (lacking the last 33 residues) and linear plasmid DNA, pointing out that at least *in vitro* gp16 is dispensable for the translocation process (154). It was shown that inhibition of DNA packaging only occurs when the central OD is present as gp16 WT, ΔN, ΔC and OD were able to inhibit the packaging while individual domains and the ΔN-ΔC construct stimulated the DNA packaging (69).

10.4.2 ST-LT Complex formation

The mechanism the two terminase proteins use to interact with each other before or while attached to the procapsid is less well characterised than their individual functions or combined behaviour. The presented functional data on ST and LT on phages SPP1, T4 and P22 indicate that both components of the molecular motor communicate to increase the rate of ATP hydrolysis, in the presence of DNA for T4 and P22, and to diminish the endonuclease activity of LT. The formation of ST-LT complexes has been partially characterised for phages SPP1, T4, P22 and λ.

Investigation of the SPP1 molecular motor components suggests that both terminases and portal proteins form functional complexes *in vitro*. Glycerol gradients performed on SPP1 G*1*P and G*6*P showed that both proteins physically interact to form a complex of estimated mass of ~ 1,000 kDa. Moreover, the same study, based on the ATPase experiments carried out described above, indicates that the adduct with the stoichiometry $G\mathit{2}P_1{\cdot}G\mathit{1}P_2{\cdot}G\mathit{6}P_1$ presents the highest rate of ATP hydrolysis (149). Experiments with truncated SPP1 G*1*P and truncated versions immobilised by polyclonal antibodies on protein A-sepharose columns showed that the DBD accounts for most of the interaction with G*2*P since the WT and one construct missing segment III (CTD) retained G*2*P in the column after washes with increasing NaCl concentrations, while the ΔN construct did not retain G*2*P even at low NaCl concentration. These experiments were performed in the absence of DNA. The domain in G*1*P recognised by the polyclonal antibodies is not mentioned to discard steric hindrance that could have occurred if G*1*P uses the same domain that the antibody recognises to interact with G*2*P (148).

Early research on P22 showed that gp3 copurified with gp2 in an apparent ~ 200 KDa complex, when it was intended to purify gp2 from cells containing the corresponding gene for both proteins (155). Later it was proven that gp3 CTD is essential to produce species of higher MW than the separate gp3 oligomer and gp2 monomer. Whereas FL gp3 forms complexes with gp2, one construct missing the last C-terminal 22 residues that also prevented binding to viral or random DNA, failed to produce complexes with gp2 (14,73).

The ST (gpNu1) and LT (gpA) in λ phage associate into two different active species. The first one consists of a homogeneous 114.2 kDa heterotrimer composed of one copy of gpA and two of gpNu1. The second assembly is a 528 (± 34) kDa, referred as 13.3S species, for which stoichiometry is not clear, though a pentamer of heterotrimers ($[\text{gpA-gpNu1}_2]_5$) would have a similar molecular weight (570.9 kDa) (156).

10.5 Proposed model of interaction for G20C ST-DNA complex

The combination of structural and biochemical data points towards a wraparound model where DNA surrounds the G20C ST (Figure 85). The model is proposed in view of the following evidence:

- The exposure of several positively charged residues at the DBD creates a positive electrostatic potential around the ST that could attract the negatively charged DNA phosphate backbone and induce bending in DNA.
- The spacing between DBDs in G20C, SF6, Sf6 ST is ~ 34 Å irrespective of differences in the oligomerisation state, suggesting that the number of subunits is less relevant to produce a functional assembly than the spacing between DBDs.
- The identification on G20C ST of the HTH motif that is part of the DBD in SF6 and Sf6 ST, and which was crystallised in complex with DNA in HTRF2, RNA_Pol_σ_70_D4 and Antp protein also indicates that the HTH motif on G20C ST could be responsible for the interaction with DNA.
- G20C ST DBD is able to adjust its position, as evidenced by up to 6.7 Å positional differences in individual subunits of the crystal structure. Hypothetically, in solution the DBDs will have more freedom to move and adjust their position. In the absence of precipitating agents or high salt concentrations, G20C ST is not forced to precipitate in an ordered way, hence the space restraints in the crystal are unlikely to occur *in vivo*.
- Even though the central channel in G20C is almost wide enough to accommodate DNA, we speculate that this channel is a consequence of the need to array the DBDs at regular distances in circular arrangement. In agreement with this notion is the observation that SF6

ST did not show binding to DNA when the DBDs were removed and that both SF6 and Sf6 ST bind to DNA even though the channel is too narrow to host DNA.

In the proposed complex the DNA wraps around the G20C ST with the recognition helix (α2) establishing contacts with the major groove. The H-bonds between the DBD and DNA could be mediated by polar and charged residues coordinated with the phosphate-sugar backbone and exposed edges of nucleotide bases, there could be also water-mediated H-bonding interactions.

One difference with the wraparound model proposed for SF6 ST is the position of DNA. While on SF6, the recognition α helices face the outwards of the assembly and DNA is exactly at the periphery, in G20C ST the recognition α helices are located pointing towards the top of the assembly. DNA would be therefore seated on top of the assembly. Although on Sf6 ST, the HTH motif clearly exists, the recognition helix is less exposed suggesting that DNA binding, may be dependent on more prominent angular and positional adjustments of individual DBDs.

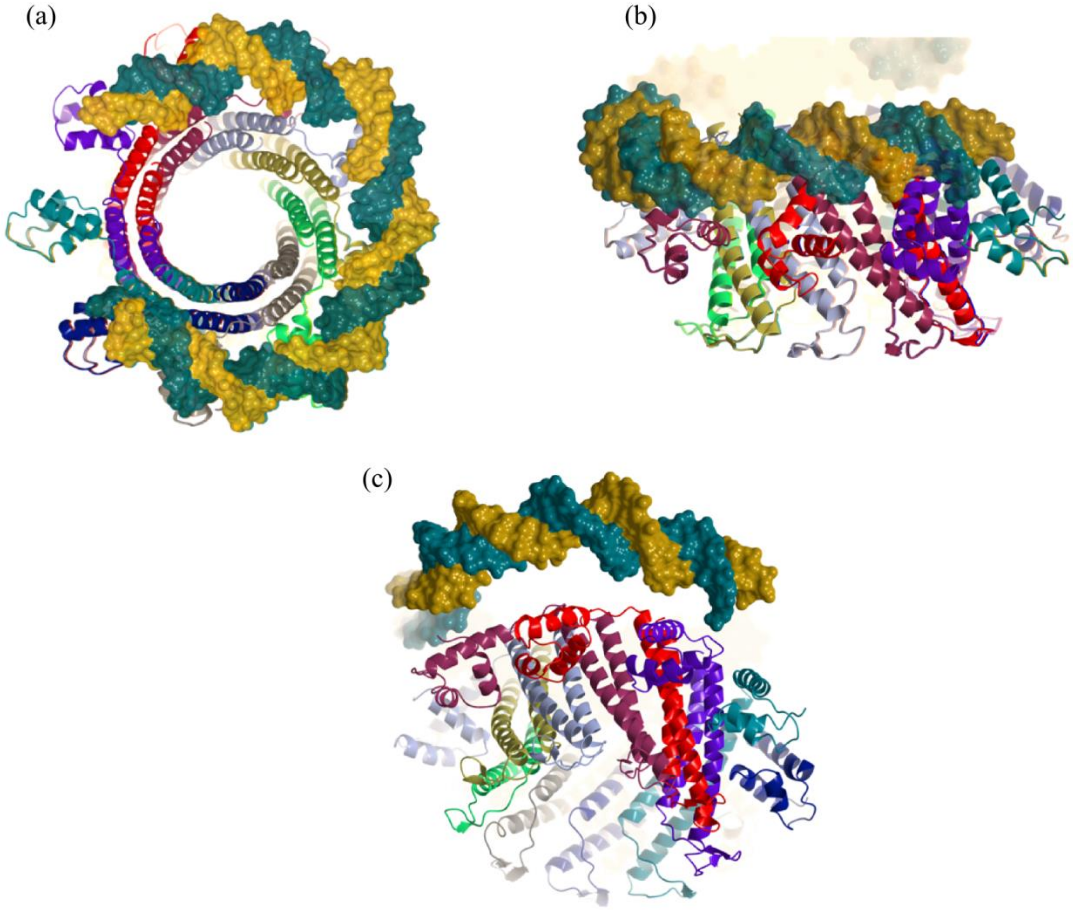


Figure 85. Model of G20C ST-dsDNA interaction

Top (a) and side (b) views of the wraparound G20CS ST-dsDNA model. (c) Complementarity of the G20C DBD recognition helix with the major groove. The DNA was extracted from the X-ray structure of the *Drosophila* nucleosome, PDB 2NQB (157) .

The assumption that the DNA is seated on top of the G20C assembly fits well with the characteristics of the DBD-OD linker, which is not as long as in SF6 ST, but still it can provide the required up-and-down movement of the DBD to fit with DNA. From the structure it is known that the DBD can also move from side to side, which may be required to fit the recognition helices into the major grooves. Restriction in the movement of the DBD is also demonstrated by the high number of conditions from the commercial screens that produced crystals of G20C ST_ΔC, suggesting that removing the CTD reduces disorder in the molecule to some extent and that the DBD's location is not subject to a high degree of variation.

A second difference with the SF6 ST-DNA model is the role assigned to the CTD. In SF6 it was demonstrated that the CTD is not required for DNA recognition but it is involved in defining the oligomeric state (16). In G20C ST the removal of the CTD significantly decreased binding to DNA. It is proposed that on the basis of the EMSA data presented and the high proportion of negatively charged residues, the CTD's role in binding is to repel the DNA and direct it to the surface provided by the DBD, or alternatively facilitate positional rearrangement of individual DBDs during binding to DNA.

Although the model presented is based on the X-ray structure and data from experiments described above, it should be looked at with some caution as there are some aspects that limit its reliability. One of the principal limitations of this model is the width (Atom-to-atom: 25 Å, vdW: ~ 22 Å) of the central channel and the presence of four rings of positive charge along its height. B-form DNA has an atom-to-atom distance and vdW diameters that are ~ 20 and ~ 23 Å, respectively, implicating that if any conformation rearrangement to widen the channel occurred, DNA would fit inside. Another limitation of our model is that none of the experiments performed in this project were designed to assess DNA binding within the central channel and therefore discarding this option is based on the observations that binding to DNA is achieved principally by the DBD.

In the past, models of protein structures involved in binding to DNA or models for ST-DNA interactions have been proposed, however further experimental work has made evident a misinterpretation. Both examples presented below are a reminder that the elucidation of the mechanism for recognition of DNA by the ST requires the X-ray crystal structure of the ST in complex with DNA.

In 2008 the 3D reconstruction of gp3 FL (1-162) at 18 Å resolution was used to map the location of the CTD by comparing it with a construct (1-127) which lacked the CTD. The difference between two maps, the narrower part of the assembly, was assumed to be the location of the CTD, the domain being responsible to bind to DNA. Later, in 2012, the X-ray structure of construct gp13 Δ1-139 solved at 1.75 Å and reconstruction of FL gp3, combined with Cryo-EM data revealed the location of the C-terminus. The C-terminus was shown to be situated above the widest part of the assembly, in the opposite direction of where it was first proposed, with the β1-β2 β-hairpin occupying the position where the CTD was speculated to be. Moreover, in the latter study, a population with an extended barrel on top of the widest part, formed by the α-helical cores, was identified by cryo-EM. This extended barrel was demonstrated to account for the disordered helix α7.

According to the structural superposition performed in this study, using CCP4mg and G20C DBD (residues 1-52), Sf6 DBD (residues 10-53) and SF6 DBD (residues 5-60), the HTH motif on Sf6 that aligns with that from G20C and SF6 involves helices α2 and α3, the latter being the recognition helix. The proposed model to describe the Sf6 NTD-dsDNA interaction in 2010 presented helix α1 inserted into the major groove and based on this study a new publication included a wraparound model with DNA. In view of our observations, a wraparound model directly from the X-ray structure, using the identified recognition helix on Sf6 DBD, helix α3, would not be accurate as it would imply significant changes in the orientation of this helix to adjust to the major grooves on the DNA.

10.6 Conclusions

The set of experiments carried out identified that G20C ST is a functional protein which binds to 1 kbp DNA fragments of viral and non-viral origin. The viral DNA comprised a region of the G20C's genome that includes the ST gene, a region upstream of this gene and another short portion of the LT gene. Interestingly, although the NTD was demonstrated to be the DBD, the CTD was shown to also contribute to the observed binding of the WT protein to DNA. Combined with the X-ray crystal structure, the experimental data described in this chapter are consistent with a wraparound model that describes the way G20C ST interacts with DNA. This model and those available in the literature are subject to change as only a crystal structure in complex with DNA will provide the degree of accuracy necessary to understand the DNA-recognition process during early stages of bacteriophage assembly.

Chapter 11

11.Conclusions and suggested future approaches

Molecular motors in dsDNA bacteriophages translocate viral genomes into preformed capsids. The molecular motor is composed of the portal protein, embedded in a single vertex of the procapsid, and the small and large terminase proteins. The small terminase recognises DNA and presents it to the large terminase, which generates cuts to produce genome-sized DNA fragments. The large terminase in addition can act as an ATPase to provide the driving-force of the molecular motor.

The main outcomes of this project are the X-ray structures of the SPP1 capsid protein and G20C's ST, determined by SAD at 3.0 and 2.5 Å resolution, respectively (Table 9 and Table 15).

G*13*P, the SPP1 capsid protein, is organised into A- and P-domains and E-loop (Figure 48). G*13*P crystallised as 5-subunit assemblies that provided information about the potential arrangement of subunits in the capsid as they resemble the organisation of the HK97 capsid morphological units. G*13*P adopts the HK97-fold (39), similar to the capsid protein of T4 (37) and several other capsid-related proteins, e.g. encapsulins (Figure 51 and Figure 53).

The G20C ST monomer is comprised of two domains: the NTD and OD. G20C ST oligomerises into 9-subunit circular structures (Figure 67). Structural comparison with SF6 (16), P22 (14), Sf6 (68) and other STs revealed that the G20C NTD has a HTH motif and that the OD shares some architectural characteristics with that from other STs. Shared characteristics between oligomers include the circular array, formation of a central channel, overall diameter and spacing between individual NTDs. DNA-binding experiments with G20C ST_WT and truncated versions demonstrated binding to both viral and non-viral DNA. Moreover, the NTD was established as the most essential portion for DNA binding, although the CTD was also demonstrated to be important. The involvement of both the NTD and CTD in DNA-binding had not been previously observed in STs. A wrap-around model to describe the interaction of G20C ST with DNA is proposed on the basis of the performed biochemical experiments and structural analysis (section 8.4).

Both proteins contribute to the structural knowledge of viral capsid construction and the viral assembly process. As aforementioned, as of September 2014, G*13*P is the third crystal structure for a dsDNA bacteriophage capsid protein and the second for a Siphovirus, in spite of the fact that *Caudovirales* outnumber all other organisms in our planet by a factor of ~ 10. Moreover the elucidation of the SPP1 capsid structure positions this bacteriophage as the most extensively characterised by structural analysis, since there are X-ray structures available for all the components of the molecular motor (11,16,22), and there are also NMR structures of the connector proteins (61) and cryo-EM structures of the tail (60) and complete capsids at late stages of morphogenesis (41).

G20C ST is the first ST characterised for a thermophilic phage and the first protein from G20C phage for which the X-ray structure was elucidated and biochemical characterisation was carried out. It was not until 2012 when the X-ray structures of SF6, P22 and 44RR ST were published that a significant amount of structural information became available. G20C ST posits a dilemma when it comes to address the potential mechanism of interaction with DNA since the atom-to-atom diameter of the central channel is almost wide enough to accommodate B-form DNA and this would suggest that the DNA could be threaded through it. The only phage for which it is reasonable to hypothesise that the DNA could be threaded through the channel is 44RR, where the channel diameter is wide enough to accommodate DNA. However, the variability in subunit number of ST oligomers refutes channel involvement in DNA-binding. Intriguingly, G20C ST also presents an NTD that was identified as DBD. Moreover the CTD also plays an important role in binding to DNA.

The structural conservation of most of the SPP1 capsid protein with other capsid proteins and some shared characteristics of G20C ST with the available X-ray structures of ST, provide further evidence toward the hypothesis that all dsDNA tailed phages have evolved from a common ancestor.

Further work will focus on the fitting of the SPP1 capsid protein’s X-ray structure into the available cryo-EM maps of the complete capsid to investigate how G*13*P interacts with the G*6*P portal protein. Mutations will be introduced, at the identified surfaces of P- and A-domains, to promote formation of complexes involving the portal protein surrounded by a single layer of capsid proteins. The longer term goal may be to attach the capsid-portal complexes onto surfaces, where after assembly of the molecular motor, it would be possible to deliver DNA through these barriers. Combining X-ray and EM data to produce a pseudo-atomic model of a complete capsid would be invaluable for understanding inter- and intra-capsomer interactions during different stages of capsid assembly.

Future work with the G20C ST involves the production of higher-quality crystals. Further approaches to characterise the DNA-binding activity may include mutagenesis in the putative DBD to investigate residues relevant for binding. Additionally, shorter DNA fragments could be used to identify a packaging-initiation site with a minimal size that could be used for crystallisation of the ST-DNA complex.

The available constructs can be used, together with G20C LT, to determine whether G20C ST CTD is the LT-binding domain. Other experiments including the ST and LT could be performed in the presence of both proteins to investigate their properties including DNA-binding, ATPase and nuclease activities. Some biochemical characterisation may have to be performed at higher temperatures to account for the high optimal growth temperature of *Thermus thermophilus*.

Appendices

Appendix 1. Primers for SPP1 G*13*P and G20C ST constructs

Clone	N-Residue	C-Residue	Tag- Vector	Forward primer 5' → 3'	Reverse 5' → 3'
G*13*P_WT	M1-A2	A324	N' His -pET28a	GG AAT TCT **CAT ATG** GCA TAC ACA AAA ATT TCA G	CCA GAA **CTC GAG** TTA TGC TTG TAG TCT GTG TT
G*13*P_WT-GST	M1-A2	Ala 324	N' GST -pGEX	CC CTG **GGA TCC** ATG GCA TAC ACA AAA ATT TCA	CCA GAA **CTC GAG** TTA TGC TTG TAG TCT GTG TT
G*13*P_ΔC	M1-A2	N286	N' His -pET28a	GG AAT TCT **CAT ATG** GCA TAC ACA AAA ATT TCA G	CCA GAA **CTC GAG** TTA GTT TTC TGT GAA TTT TAC GCC
G*13*P_ΔN2	M1-T4	A324	N' His -pET28a	ATTCT **CAT ATG** ACA AAA ATT TCA GAT GTT	CCA GAA **CTC GAG** TTA TGC TTG TAG TCT GTG TT
G*13*P_ΔN6	M1-D8	A324	N' His -pET28a	ATTCT **CAT ATG** GAT GTT ATC GTA CCG GAG	CCA GAA **CTC GAG** TTA TGC TTG TAG TCT GTG TT
G*13*P_ΔN10	M1-P12	A324	N' His -pET28a	GGAATTCT **CAT ATG** CCG GAG TTA TTT AAC CCG TAC G	CCA GAA **CTC GAG** TTA TGC TTG TAG TCT GTG TT
G*13*P_ΔN15	M1-P17	A324	N' His -pET28a	AT TCT **CAT ATG** CCG TAC GTC ATT AAC ACA ACA	CCA GAA **CTC GAG** TTA TGC TTG TAG TCT GTG TT
G*13*P_ΔN15	M1-P17	A324	N' GST -pGEX	CCCTG **GGA TCC** CCG TAC GTC ATT AAC ACA ACA	CCA GAA **CTC GAG** TTA TGC TTG TAG TCT GTG TT
G*13*P_ΔN21	M1-T23	A324	N' His -pET28a	AT TCT **CAT ATG** ACA ACA CAA CTT TCT GCC TTC	CCA GAA **CTC GAG** TTA TGC TTG TAG TCT GTG TT
G*13*P_ΔN21	M1-T23	A324	N' GST -pGEX	CC CTG **GGA TCC** ACA ACA CAA CTT TCT GCC	CCA GAA **CTC GAG** TTA TGC TTG TAG TCT GTG TT
G*13*P_ΔN35	M1-T37	A324	N' His -pET28a	GGA ATTCT **CAT ATG** ACA GAT GAC GAA TTG AAT GCA C	CCA GAA **CTC GAG** TTA TGC TTG TAG TCT GTG TT
G*13*P_FL_G64W	M1-A2	A324	N' His -pET28a	G TAC TGG AAT GAC CTA GAC TGG GAT TCC CAA GTG TTG AAC	GTT CAA CAC TTG GGA ATC CCA GTC TAG GTC ATT CCA GTA C
G*13*P_FL_D100R	M1-A2	A324	N' His -pET28a	AGT TCT CAC CGT TTA GCG	CGC TAA ACG GTG AGA ACT
G*13*P_FL_T104Y	M1-A2	A324	N' His -pET28a	TTA GCG GCA TAC CTT TCC GGT	ACC GGA AAG GTA TGC CGC TAA
G*13*P_FL_D194W	M1-A2	A324	N' His -pET28a	CGC AGT TAA ACA ATG GTT GAT TGA	TCA ATC AAC CAT TGT TTA ACT GCG
G*13*P_FL_A261W	M1-A2	A324	N' His -pET28a	CA ACA GAA ACA GCC CGT AAC GCT TGG GG TTC CCA AGA TAT TC	G AAT ATC TTG GGA ACC CCA AGC GTT ACG GGC TGT TTC TGT TG
G*13*P_FL_L262W	M1-A2	A324	N' His -pET28a	CGT AAC GCT TGG GGT TCC CAA	TTG GGA ACC CCA AGC GTT ACG

Appendix 1. Primers for SPP1 G*13*P and G20C ST constructs. Continuation

Clone	N-Residue	C-Residue	Tag- Vector	Forward primer 5' → 3'	Reverse 5' → 3'
G*13*P_FL_T104Y; A261W	M1-A2	A324	N' His -pET28a	Primer containing 2nd mutation	Primer containing 2nd mutation
G*13*P_FL_G64W;T104Y	M1-A2	A324	N' His -pET28a	Primer containing 2nd mutation	Primer containing 2nd mutation
G*13*P_FL_G64W;A261W	M1-A2	A324	N' His -pET28a	Primer containing 2nd mutation	Primer containing 2nd mutation
* G*13*P_V10-I35AT	M1-A2	A342	N' His -pET28a	GG AAT TCT CAT ATG GCA TAC ACA AAA ATT TCA G	CCAGAA CTC GAG TTA TGC TTG TAG TCT GTG TT
* G*13*P_A11-V34AT	M1-A2	A340	N' His -pET28a	GG AAT TCT CAT ATG GCA TAC ACA AAA ATT TCA G	CCAGAA CTC GAG TTA TGC TTG TAG TCT GTG TT
G*13*P_V10-I35AT_ΔC330	M1-A2	K329	N' His -LIC 3C	GG AAT TCT CAT ATG GCA TAC ACA AAA ATT TCA G	CCA GAA CTC GAG TTA CTT AGG GTC GTA CAC
G*13*P_V10-I35AT_ΔC314	M1-A2	D313	N' His -LIC 3C	GG AAT TCT CAT ATG GCA TAC ACA AAA ATT TCA G	CCA GAA CTC GAG TTA GTC CGT AGG CGT TCT
G*13*P_A11-V34AT_ΔC328	M1-A2	K327	N' His -LIC 3C	GG AAT TCT CAT ATG GCA TAC ACA AAA ATT TCA G	CCA GAA CTC GAG TTA CTT AGG GTC GTA CAC
G*13*P_A11-V34AT_ΔC312	M1-A2	D311	N' His -LIC 3C	GG AAT TCT CAT ATG GCA TAC ACA AAA ATT TCA G	CCA GAA CTC GAG TTA GTC CGT AGG CGT TCT
G*13*P_V10-I35AT-C'_tag	M1-A2	A342	C' His -pET22b (Non cleavable)	GGA GAT ATA CAT ATG GCA TAC ACA AAA ATT TCA	GTG GTG GTG CTC GAG TGC TTG TAG TCT GTG TTT
G*13*P_A11-V34AT-C'_tag	M1-A2	A340	C' His -pET22b (Non cleavable)	GGA GAT ATA CAT ATG GCA TAC ACA AAA ATT TCA	GTG GTG GTG CTC GAG TGC TTG TAG TCT GTG TTT
G*13*P_ ΔN6_T104Y; A261W	M1-D8	A324	N His -LIC- (Non cleavable)	ATC ACC ACC ACC AC ATG GAT GTT ATC GTA CCG GAG	TGA GGA GAA GGC GCG TTA TGC TTG TAG TCT GTG
G*13*P_FL_T104Y; A261W	M1-A2	A324	pET22b (Untagged)	GGA GAT ATA CAT ATG GCA TAC ACA AAA ATT TCA	GTG GTG GTG CTC GAG TTA TGC TTG TAG TCT
G*13*P_FL_G64W;T104Y	M1-A2	A324	pET22b (Untagged)	GGA GAT ATA CAT ATG GCA TAC ACA AAA ATT TCA	GTG GTG GTG CTC GAG TTA TGC TTG TAG TCT
G*13*P_V10-I35ATΔ330-HMBP	M1-A2	K329	N' His -HMBP	CAG GGG CCC GGA TCC ATG GCA TAC ACA AAA	TGC CTG GAC GTC GAC TTA CTT AGG GTC GTA
Clone	**N-Residue**	**C-Residue**	**Tag**	**Forward primer 5' → 3'**	Reverse 5' → 3'
G20C_ST_WT	M1-S2	T171	pET28a	GGA CAA CAT ATG AGC GTG AGT TTT AGG GAC	GGC AAG CTT CTA GGT CTT AG GCG CTT CAT C
G20C_ST_WT-C'_tag	M1-S2	T171	C' His- pET22b	GAA GGA GAT ATA CAT ATG AGC GTG AGT TTT AGG G	GTG GTG GTG CTC GAG GGT CTT AGG CGC TTC

Appendix 1. Primers for SPP1 G*13*P and G20C ST constructs. Continuation

Clone	N-Residue	C-Residue	Tag	Forward primer 5' → 3'	**Reverse 5' → 3'**
G20C_ST_NTD	M1-S2	A52	N' His- pET28a	G CTT CTT TCC GCC TAG TTG CCT TCC CTT GAA G	TTC AAG GGA AGG CAA CTA GGC GGA AAG AAG

G20C_ST_ΔC	M1-S2	G138	N' His-pET28a	C AAG GTA CTA GGC TAG CAT GTC GGG TCC AC	GT GGA CCC GAC ATG CTA GCC TAG TAC CTT
G20C_ST_ΔN	M1-P54	T171	N' His-LIC 3C	TTC CAG GGA CCA GCA ATG CCT TCC CTT GAA GAC	TGA GGA GAA GGC GCG CTA GGT CTT AGG CGC TTC

Unless otherwise stated, all the tags were cleavable.
Codons in Cyan correspond to introduced stop codons, codons in yellow highlight the introduced mutation.
Restriction sites: NdeI in blue, XhoI in green, BamHI in purple, SalI in orange, HindIII in pink
* Primers that amplify the FL protein are mentioned, for details see section 3.4.1

Appendix 2. List of buffers to purify SPP1 G*13*P and G20C ST constructs

Clone	Affinity chromatography buffer	SEC buffer	Final result
G*13*P_WT	25 mM Tris pH 7.5, 75mM NaCl + 10/500 mM Imidazole/ HT	25mM Tris pH 7.5, 75mM NaCl §	Aggregation into capsid-like particles
G*13*P_WT-GST	25 mM Tris pH 7.5, 75mM NaCl + 50mM glutathione (to elute) / glutathione resin	25mM Tris pH 7.5, 75mM NaCl	Soluble Too low yield in NAC, SEC
G*13*P_ΔC	X	X	Soluble when induced in Artic Express cells
G*13*P_ΔN2	25 mM Tris pH 7.5, 150mM NaCl + 10/500 mM Imidazole / FF	X	Too low yield Aggregation
G*13*P_ΔN6	25 mM Tris pH 7.5, 150mM NaCl + 10/500 mM Imidazole / HT	25 mM Tris pH 7.5, 150mM NaCl	Soluble Aggregation
G*13*P_ΔN10	25 mM Tris pH 7.5, 150mM NaCl + 4/500 mM Imidazole / FF	25 mM Tris pH 7.5, 150mM NaCl	Aggregation

Appendix 2. List of buffers to purify SPP1 G*13*P and G20C ST constructs. Continuation

Clone	Affinity chromatography buffer	SEC buffer	Final result
G*13*P_ΔN15	X	X	Insoluble
G*13*P_ΔN15-GST	X	X	Insoluble
G*13*P_ΔN21	X	X	Insoluble

G*13*P_ΔN21-GST	X	X	Insoluble
G13P_ΔN35	X	X	Insoluble
G*13*P_FL_G64W	25 mM Tris pH 7.5, 10mM NaCl + 10/500 mM Imidazole/ HT	25 mM Tris pH 7.5, 10mM NaCl	Soluble Formation of 1x10^6 Da aggregates No crystals grown
G*13*P_FL_D100R	25 mM Tris pH 7.5, 10mM NaCl + 10/500 mM Imidazole/ HT	X	Soluble Too low yield
G*13*P_FL_T104Y	25 mM Tris pH 7.5, 10mM NaCl + 10/500 mM Imidazole/ HT	25 mM Tris pH 7.5, 10mM NaCl	Soluble - Too low yield Monomers and aggregates in SEC
G*13*P_FL_D194W	25 mM Tris pH 7.5, 10mM NaCl + 10/500 mM Imidazole/ HT	X	Soluble Too low yield
G*13*P_FL_A261W	25 mM Tris pH 7.5, 10mM NaCl + 10/500 mM Imidazole/ HT	25 mM Tris pH 7.5, 10mM NaCl	Soluble - Too low yield Monomers and aggregates in SEC
G*13*P_FL_L262W			Soluble - Too low yield Apparently monomer
G*13*P_FL_T104Y; A261W			Soluble – Monomer - Needles
G*13*P_FL_G64W;T104Y			Soluble – Monomer - No crystals
G*13*P_FL_G64W;A261W			Aggregation
G*13*P_V10-I35AT G*13*P_A11-V34AT	25 mM Tris pH 7.5, 10 or 500mM NaCl + 10/500 mM Imidazole/ HT	25 mM Tris pH 7.5, 10mM NaCl 25 mM Tris pH 7.5, 500mM NaCl	Soluble – Monomer - Degradation - No crystals
G*13*P_V10-I35AT_ΔC330 G*13*P_V10-I35AT_ΔC314 G*13*P_A11-V34AT_ΔC328 G*13*P_A11-V34AT_ΔC312	X	X	Insoluble
G*13*P_V10-I35AT-C'_tag G*13*P_A11-V34AT-C'_tag	25mM pH 7.5, 500mM NaCl + 10/500 mM Imidazole/ HT	25mM pH 7.5, 500mM NaCl	Soluble – Monomer - Degradation - No crystals

Appendix 2. List of buffers to purify SPP1 G*13*P and G20C ST constructs. Continuation

Clone	Affinity chromatography buffer	SEC buffer	Final result
G*13*P_ ΔN6_T104Y; A261W	50 mM Tris pH 7.5, 500mM NaCl + 10/500 mM Imidazole/ HT	25 mM Tris pH 7.5, 500mM NaCl	Soluble – Mix of aggregate and monomer
u-G*13*P_FL_T104Y; A261W	A: 25 mM Tris pH 7.5, 10 mM NaCl B: 25 mM Tris pH 7.5, 500mM NaCl /FFQ	10 mM Tris pH 7.5, 200mM NaCl	Soluble – Monomer – Crystals Se-Met purified with same conditions
u-G*13*P_FL_G64W;T104Y			Soluble – Monomer – No crystals
G13P_V10-I35ATΔ330-HMBP	25 mM Tris pH 7.5, 200mM NaCl + 10/500 mM Imidazole/ HT	X	Too low yield
Clone	**Affinity chromatography buffer**	**SEC buffer**	**Final result**

G20C_ST_WT	25 mM Tris pH 7.5, 500mM NaCl + 10/500 mM Imidazole/ HT	25 mM Tris pH 7.5, 250mM NaCl	Soluble, 9-mer, crystallised
G20C_ST_WT-C'_tag G20C_ST_NTD G20C_ST_ΔC G20C_ST_ΔN	25 mM Tris pH 7.5, 200mM NaCl + 10/500 mM Imidazole/ HT	25 mM Tris pH 7.5, 250mM NaCl	Soluble, used for activity assays.

X indicates that no chromatography was carried out

HT- His Trap column, FF- FF crude column, FFQ- Fast flow Q column

§ More details in the text

Note: For most of the constructs several purifications with varying NaCl concentrations were made, however no changes were observed in the chromatography profiles.

Appendix 3. Properties of SPP1 G*13*P and G20C ST constructs

Clone	**Residues No tag**	**MW * kDa**	**pI (- tag)**	**Abs 0.1 % (= 1 g/L) ***	**Residues With tag**	**MW (+ tag) kDa**	**pI (+ tag)**	**Abs 0.1 % ¥ (= 1 g/L)**
G*13*P_WT	327	35.636	4.95	1.035	344	37.518	5.49	0.984
G*13*P_WT-GST	329	35.766	4.89	1.032	555	62.178	5.21	1.283
G*13*P_ΔC	289	31.299	4.70	0.956	306	33.181	5.19	0.901
G*13*P_ΔN2	325	35.401	4.95	1.000	342	37.283	5.49	0.950
G*13*P_ΔN6	321	34.972	4.88	1.013	338	36.854	5.40	0.961
G*13*P_ΔN10	317	34.545	4.95	1.025	334	36.427	5.49	0.972
G*13*P_ΔN15	312	33.945	5.01	1.043	329	35.827	5.58	0.988
G*13*P_ΔN15-GST	313	33.944	4.93	1.043	539	60.356	5.25	1.301

G*13*P_ΔN21	306	33.257	5.01	1.020	323	35.139	5.58	0.965
G*13*P_ΔN21-GST	307	33.256	4.93	1.020	533	59.668	5.25	1.287
G13P_ΔN35	292	31.833	5.01	1.066	309	33.715	5.58	1.006
G*13*P_FL_G64W	327	35.765	4.95	1.189	344	37.647	5.49	1.126
G*13*P_FL_D100R	327	35.677	5.11	1.034	344	37.559	5.68	0.982
G*13*P_FL_T104Y	327	35.698	4.95	1.075	344	37.580	5.49	1.022
G*13*P_FL_D194W	327	35.707	5.02	1.187	344	37.589	5.59	1.128
G*13*P_FL_A261W	327	35.751	4.95	1.186	344	37.633	5.49	1.127
G*13*P_FL_L262W	327	35.709	4.95	1.187	344	37.591	5.49	1.128
G*13*P_FL_T104Y; A261W	327	35.813	4.95	1.226	344	37.695	5.49	1.164
G*13*P_FL_G64W;T104Y	327	35.827	4.95	1.225	344	37.709	5.49	1.164
G*13*P_FL_G64W;A261W	327	35.880	4.95	1.335	344	37.762	5.49	1.268
G*13*P_V10-I35AT	345	37.396	4.94	0.993	362	39.278	5.42	0.946
G*13*P_A11-V34AT	343	37.183	4.94	0.999	360	39.065	5.42	0.951
G*13*P_V10-I35AT_ΔC330	332	35.721	4.61	1.040	351	37.867	4.87	0.981
G*13*P_V10-I35AT_ΔC314	316	33.849	4.61	0.891	335	35.995	4.89	0.838
G*13*P_A11-V34AT_ΔC328	330	35.508	4.61	1.046	349	37.655	4.87	0.987
G*13*P_A11-V34AT_ΔC312	314	33.636	4.61	0.897	333	35.783	4.89	0.843
G*13*P_V10-I35AT-C'_tag	NC	NC	NC	NC	350	38.179	5.20	0.963

Appendix 3. Properties of SPP1 G*13*P and G20C ST constructs. Continuation

Clone	**Residues** **No tag**	**MW *** **kDa**	**pI (- tag)**	**Abs 0.1 %** **(= 1 g/L) ***	**Residues** **With tag**	**MW (+ tag)** **kDa**	**pI (+ tag)**	**Abs 0.1 %** **(= 1 g/L) ¥**
G*13*P_A11-V34AT-C'_tag	NC	NC	NC	NC	348	37.967	5.20	0.978
G*13*P_ ΔN6_T104Y; A261W	---	---	---	---	328	36.053	5.23	1.176
u-G*13*P_FL_T104Y; A261W	324	35.531	4.89	1.235	---	---	---	---
u-G*13*P_FL_G64W;T104Y	324	35.545	4.89	1.235	---	---	---	---
G*13*P_V10-I35ATΔ330-HMBP	334	35.922	4.61	1.027	724	78.846	5.03	1.310
Clone	**Residues** **No tag**	**MW *** **kDa**	**pI (- tag)**	**Abs 0.1 %** **(= 1 g/L) ***	**Residues** **With tag**	**MW (+ tag)** **kDa**	**pI (+ tag)**	**Abs 0.1 %** **(= 1 g/L) ¥**
G20C_ST_WT	174	19.074	4.94	0.234	190	20.825	5.66	0.215
G20C_ST_WT-C'_tag	NC	NC	NC	NC	179	19.858	5.32	0.225
G20C_ST_NTD	55	6.198	5.55	0.240	72	8.080	6.70	0.184
G20C_ST_ΔC	141	15.785	5.39	0.283	158	17.667	6.29	0.253
G20C_ST_ΔN	122	13.138	4.72	0.227	141	15.284	5.37	0.195

--- Not applicable, for example no tag was included
*   Includes the additional residues left after tag removal
¥ Assuming all pairs of Cys residues form cysteines
NC- Non-cleavable tag
Note: only G*13*P-AT constructs included Cys residues (four in the Zn-binding domain)

Appendix 4. H-bonds between subunits F-A from HK97 gp5

##	Structure 1	Dist. [Å]	Structure 2
1	F:MET 119[N]	3.20	A:ALA 147[O]
2	F:ILE 121[N]	2.87	A:GLU 149[O]
3	F:ILE 125[N]	2.84	A:VAL 151[O]
4	F:ALA 187[N]	2.93	A:ASP 161[O]
5	F:HIS 188[ND1]	3.30	A:SER 172[OG]
6	F:TRP 189[N]	2.99	A:PRO 170[O]
7	F:TRP 189[NE1]	3.19	A:LYS 169[O]
8	F:TYR 206[OH]	2.54	A:GLU 153[OE1]
9	F:ARG 210[NE]	2.90	A:GLU 153[OE1]
10	F:ARG 210[NH2]	3.05	A:GLU 153[OE2]
11	F:ARG 280[NH2]	3.04	A:SER 246[OG]
12	F:ARG 294[NE]	3.44	A:ASP 290[OD2]
13	F:ARG 294[NH2]	3.29	A:ASP 290[OD2]
14	F:TYR 295[OH]	2.54	A:ASP 256[OD2]
15	F:GLN 301[NE2]	3.90	A:THR 304[OG1]
16	F:GLN 301[NE2]	3.75	A:PHE 297[O]
17	F:LYS 317[NZ]	3.07	A:GLU 267[O]
18	F:GLN 117[O]	3.47	A:SER 145[OG]
19	F:MET 119[O]	3.04	A:GLU 149[N]
20	F:MET 119[SD]	3.32	A:THR 143[OG1]
21	F:ILE 125[O]	3.53	A:GLU 153[N]
22	F:MET 126[O]	3.06	A:ARG 372[NH2]
23	F:THR 185[O]	2.65	A:VAL 163[N]
24	F:ALA 187[O]	3.26	A:ASP 161[N]
25	F:TRP 189[O]	2.86	A:SER 172[N]
26	F:GLY 214[O]	3.45	A:ASN 158[ND2]
27	F:ASN 291[O]	3.15	A:ASN 291[ND2]
28	F:GLU 292[O]	3.08	A:ASN 291[N]
29	F:PRO 300[O]	2.70	A:TRP 309[N]
30	F:GLN 301[O]	2.73	A:GLY 310[N]

Appendix 5: H-bonds between subunits G-F from HK97 gp5

##	Structure 1	Dist. [Å]	Structure 2
1	G:PHE 176[N]	3.83	F:ASP 110[O]
2	G:THR 185[OG1]	3.29	F:GLN 195[OE1]
3	G:ARG 338[NH2]	3.00	F:ASP 198[O]
4	G:ARG 338[NH2]	3.07	F:ASP 199[OD1]
5	G:ARG 365[NH1]	2.96	F:ASP 199[OD1]
6	G:ARG 365[NH2]	3.17	F:ASP 199[OD2]
7	G:ARG 350[N]	3.80	F:GLU 348[O]
8	G:SER 346[OG]	2.49	F:GLU 348[OE1]
9	G:ARG 347[N]	3.49	F:GLU 348[OE1]
10	G:GLU 348[N]	3.31	F:GLU 348[OE1]
11	G:ARG 350[N]	3.22	F:ARG 350[O]
12	G:ASP 173[OD2]	3.09	F:SER 104[N]
13	G:GLU 171[OE2]	3.50	F:SER 104[N]
14	G:GLU 344[OE2]	2.58	F:ARG 194[NH2]
15	G:GLU 344[OE1]	3.18	F:ARG 347[NH2]

Appendix 6. Sequence and primers of engineered protein G*13*P-A11-V34AT

(a)

```
   1 ATGGCATACA CAAAAATTTC AGATGTTATC GTACCGGAGT TATTTAACCC GTACGTCATT
  61 AACACAACAA CACAACTTTC TGCCTTCTTC CAGTCAGGAA TTGCGGCAAC AGATGACGAA
 121 TTGAATGCAC TTGCAAAAAA AGCGGGCGGC GGTAGCACTT TAAACATGCC GTACTGGAAT
 181 GACCTAGACG GAGATTCCCA AGTGTTGAAC GACACTGACG ACCTTGTACC GCAAAAAATC
 241 AACGCTGGAC AAGATAAAGC TGTCCTTATC CTTCGCGGTA ACGCTTGGAG TTCTCACGAT
 301 TTAGCGGCAA CACTTTCCGG TTCTGACCCA ATGCAGGCTA TCGGCTCCCG TGTAGCGGCA
 361 TACTGGGCGC GCGAAATGCA AAAGATTGTT TTCGCTGAAC TTGCAGGTGT GTTCTCTAAC
 421 GATGATATGA AAGACAACAA ACTCGATATC TCTGGAACGG CTGACGGTAT TTATTCAGCG
 481 GAAACTTTCG TTGATGCATC TTACAAGCTT GGAGATCATG AAAGCTTACT TACAGCTATC
 541 GGTATGCATT CTGCTACGAT GGCAAGCGCA GTTAAACAAG ACTTGATTGA GTTTGTCAAA
 601 GATTCCCAAA GTGGTATCCG TTTCCCGACA TACATGAATA AGCGTGTAAT CGTAGATGAT
 661 TCTATGCCAG TAGAAACGCT TGAAGATGGA ACTAAGGTAT TCACATCTTA CTTGTTCGGA
 721 GCTGGTGCTC TAGGATACGC AGAAGGACAA CCGGAAGTAC CAACAGAAAC AGCTTGCCCG
 781 AAATGCGAAC GTGCTGGTGA AATCGAAGGT ACCCCGTGCC CGGCTTGCTC CGGTAAAGGT
 841 GTTGATATTC TTATCAACCG TAAACACTTT GTTTTACACC CGCGCGGCGT AAAATTCACA
 901 GAAAACGCTA TGGCGGGAAC AACGCCTACG GACGAAGAAC TTGCTAACGG TGCGAACTGG
 961 CAACGCGTGT ACGACCCTAA GAAAATCCGT ATCGTTCAAT TCAAACACAG ACTACAAGCA
1021 TAA
```

(b)

```
        10         20         30         40         50         60
MAYTKISDVI VPELFNPYVI NTTTQLSAFF QSGIAATDDE LNALAKKAGG GSTLNMPYWN
        70         80         90        100        110        120
DLDGDSQVLN DTDDLVPQKI NAGQDKAVLI LRGNAWSSHD LAATLSGSDP MQAIGSRVAA

       130        140        150        160        170        180
YWAREMQKIV FAELAGVFSN DDMKDNKLDI SGTADGIYSA ETFVDASYKL GDHESLLTAI

       190        200        210        220        230        240
GMHSATMASA VKQDLIEFVK DSQSGIRFPT YMNKRVIVDD SMPVETLEDG TKVFTSYLFG

       250        260        270        280        290        300
AGALGYAEGQ PEVPTETACP KCERAGEIEG TPCPACSGKG VDILINRKHF VLHPRGVKFT

       310        320        330        340
ENAMAGTTPT DEELANGANW QRVYDPKKIR IVQFKHRLQA
```

Engineered protein G*13*P-A11-V34AT

(a), (b) Nucleotide and amino acid sequence of G*13*P-A11-V34AT. The region corresponding to AT is highlighted in brown. Areas in cyan correspond to the annealing sites for the primers.

Appendix 7. Primers for the engineered protein G*13*P-A11-V34AT

Template	Fragment Amplified	Designed Primers
G*13*P_FL	A: 1-777 (G*13*P)	1_F 5’ GG AAT TCT CAT ATG GCA TAC ACA AAA ATT TCA G 3’
		VI-R 5`TTC GCA TTT CGG GCA AGC TGT TTC TGT TGG TAC TTC 3`
Anti-TRAP	B: 766-849 (G*13*P -AT- G*13*P)	VII-F 5` GAA GTA CCA ACA GAA ACA GCT TGC CCG AAA TGC GAA 3`
		VIII-R 5’ ACG GTT GAT AAG AAT ATC AAC ACC TTT ACC GGA GCA 3’
G*13*P_FL	C: 838-1021 (G*13*P)	IX-F 5` TGC TCC GGT AAA GGT GTT GAT ATT CTT ATC AAC CGT 3’
		1_R 5’ CCA GAA CTC GAG TTA TGC TTG TAG TCT GTG TT 3’

Areas highlighted in brown correspond to the AT sequence. NdeI and XhoI restriction sites are highlighted in cyan and salmon, respectively.

Appendix 8. Sequencing results of G13P-V10-I35AT

>8928925.seq - ID: Heta_Col_1_T7-T7 on 2011/12/17-9:47:8 automatically edited with PhredPhrap, start with base no.: 42 Internal Params: Windowsize: 20, Goodqual: 19, Badqual: 10, Minseqlength: 50, nbadelimit: 1

TTTAAGAAGGAGATATACCATGGGCAGCAGCNATCNNNNNNANNNTCACAGCAGCGGCCTG
GTGCCGCGCGGCAGCCATATGGCATACACAAAAATTTCAGATGTTATCGTACCGGAGTTATT
TAACCCGTACGTCATTAACACAACAACACAACTTTCTGCCTTCTTCCAGTCAGGAATTGCGG
CAACAGATGACGAATTGAATGCACTTGCAAAAAAAGCGGGCGGCGGTAGCACTTTAAACAT
GCCGTACTGGAATGACCTAGACGGAGATTCCCAAGTGTTGAACGACACTGACGACCTTGTAC
CGCAAAAAATCAACGCTGGACNAGATAAAGCTGTCCTTATCCTTCGCGGTAACGCTTGGAGT
TCTCACGATTTAGCGGCAACACTTTCCGGTTCTGACCCAATGCAGGCTATCGGCTCCCGTGTA
GCGGCATACTGGGCGCGCGAAATGCAAAAGATTGTTTTCGCTGAACTTGCAGGTGTGTTCTC
TAACGNTGATATGAAAGACAACAAACTCGATATCTCTGGAACGGCTGACGGTATTTATTCAG
CGGAAACTTTCGTTGATGCATCTTACAAGCTTGGAGATCATGAAAGCTTACTTACAGCTATC
GGTATGCATTCTGCTACNATGGCAAGCGCAGTTAAACAAGACTTTGATTGAGTTTGTCAAAG
ATTCCCAAAGTGGTATCCGTTTTCCCGACATACATGAANTAAGCGTGTAATCGTAGATGANT
CTNATGCCANTAGAAAACGCTTGAAGATGGAACTAAGGTATTCNCATTCTTACTTTGTTCGG
GAGCTGNGCTCTAGGANTACGCAGAAANGACNAACCGGNAAGTCCCCAACAGAAAACAGTT
GCTTTGCCCNNAAATGCGAACGTGCTGGNTGAAATCGAAGGTNNCCCCGTGCCCGGCTT

>8912716.seq - ID: Heta_Col_1_pET-RP-pET-RP on 2011/12/14-19:49:39 automatically edited with PhredPhrap, start with base no.: 16 Internal Params: Windowsize: 20, Goodqual: 19, Badqual: 10, Minseqlength: 50, nbadelimit: 1

CNGGCTTTGTTACAGCCGGATCTCAGTGGTGGTGGTGGTGGTGCTCGAGTTATGCCCNNTAA
NNNGTGTTTGAATTGAACGATACGGATTTTCTTAGGGTCGTACACGCGTTGCCAGTTCGCAC
CGTTAGCAAGTTCTTCGTCCGTAGGCGTTGTTCCCGCCATAGCGTTTTCTGTGAATTTTACGC

CGCGCGGGTGTAAAACAAAGTGTTTACGGTTGATAAGAATATCGATAACACCTTTACCGGAG
CAAGCCGGGCACGGGGTACCTTCGATTTCACCAGCACGTTCGCATTTCGGGCAAGCAACTGT
TTCTGTTGGGACTTCCGGTTGTCCTTCTGCGTATCCTAGAGCACCAGCTCCGAACAAGTAAG
ATGTGAATACCTTAGTTCCATCTTCAAGCGTTTCTACTGGCATAGAATCATCTACGATTACAC
GCTTATTCATGTATGTCGGGAAACGGATACCACTTTGGGAATCTTTGACAAACTCAATCAAG
TCTTGTTTAACTGCGCTTGCCATCGTAGCAGAATGCATACCGATAGCTGTAAGTAAGCTTTCA
TGATCTCCAAGCTTGTAAGATGCATCAACGAAAGTTTCCGCTGAATAAATACCGTCAGCCGT
TCCAGAGATATCGAGTTTGTTGTCTTTCATATCATCGTTAGAGAACACACCTGCAAGTTCAGC
GAAAACAATCTTTTGCATTTCGCGCGCCCAGTATGCCGCTACACGGGAGCCGATAGCCTGCA
TTGGGTCAGAACCGGANAGTGTTGCCGCTAAATCCGTGAGAACTCCAAAGCGTTACCGCNN
AAGGATAANGGACAGNCTTTATCTTTGTCCNAGCGGTTGAATTTTTTTGCGGGTACAAAGGT
CNGTCA

Appendix 9. Sequencing results of G13P-A11-V34AT.

>8937685.seq - ID: Stigma_Col_1_T7-T7 on 2011/12/20-10:42:20 automatically edited with PhredPhrap, start with base no.: 35 Internal Params: Windowsize: 20, Goodqual: 19, Badqual: 10, Minseqlength: 50, nbadelimit: 1

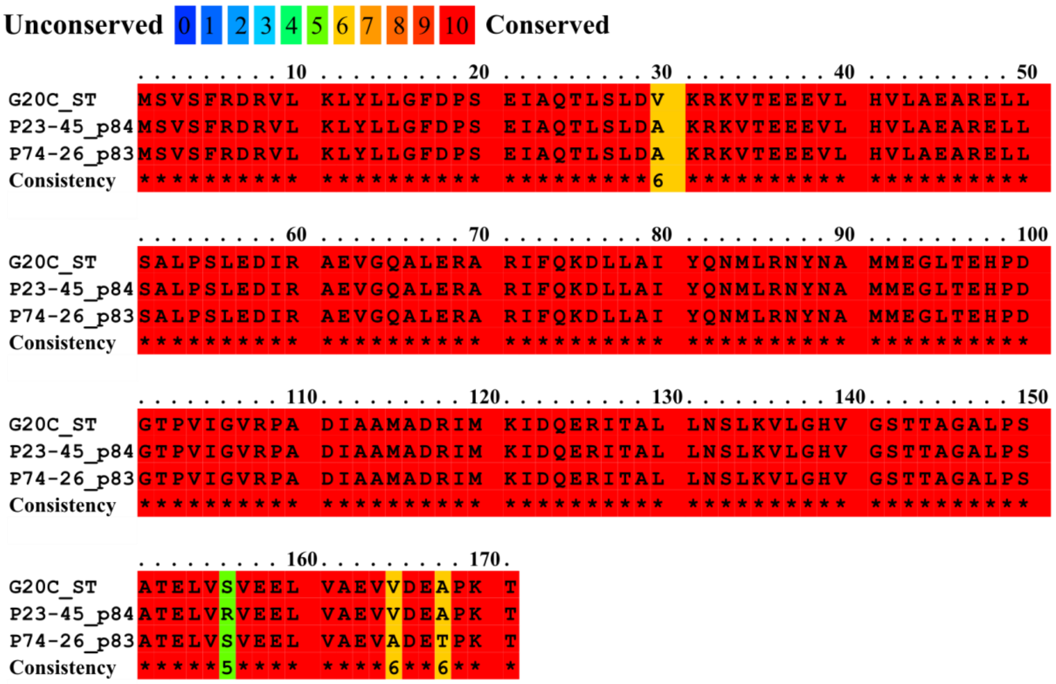
TTTGTTTAACTTTAAGAAGGNGNNNNACCATGGGCAGCAGCCATCNTCANNATCATCACAGC
AGCGGCCTGGTGCCGCGCGGCAGCCATATGGCATACACAAAAATTTCAGATGTTATCGTACC
GGAGTTATTTAACCCGTACGTCATTAACACAACAACACAACTTTCTGCCTTCTTCCAGTCAGG
AATTGCGGCAACAGATGACGAATTGAATGCACTTGCAAAAAAAGCGGGCGGCGGTAGCACT
TTAAACATGCCGTACTGGAATGACCTAGACGGAGATTCCCAAGTGTTGAACGACACTGACG
ACCTTGTACCGCAAAAAATCAACGCTGGACAAGATAAAGCTGTCCTTATCCTTCGCGGTAAC
GCTTGGAGTTCTCACGATTTANCGGCAACACTTTCCGGTTCNGANCCAATGCAGGCTATCGG
NTCNCGTGTAGCGGCATACTGGGCGCGCNAAATGCAAAAGATTGTTNTCNCTGAAC

>8912718.seq - ID: Stigma_Col_1_pET-RP-pET-RP on 2011/12/14-19:48:31 automatically edited with PhredPhrap, start with base no.: 53 Internal Params: Windowsize: 20, Goodqual: 19, Badqual: 10, Minseqlength: 50, nbadelimit: 1

GNTGGTGGTGGTGCTCGAGTTATNNNNNNANNNNGTGTTTGAATTGAACGATACGGATTTTC
TTAGGGTCGTACACGCGTTGCCAGTTCGCACCGTTAGCAAGTTCTTCGTCCGTAGGCGTTGTT
CCCGCCATAGCGTTTTCTGTGAATTTTACGCCGCGCGGGTGTAAAACAAAGTGTTTACGGTT
GATAAGAATATCAACACCTTTACCGGAGCAAGCCGGGNACGGGGTACCTTCGATTTCACCA
GCACGTTCGCATTTCGGGCAAGCTGTTTCTGTTGGTACTTCNGGTTGTCCTTCTGCGTATCCT
AGAGCACCA

Appendix 10. Sequences of G20C, P23-45 and P74-26 STs

a) Amino acid sequence



Unconserved 0 1 2 3 4 5 6 7 8 9 10 Conserved

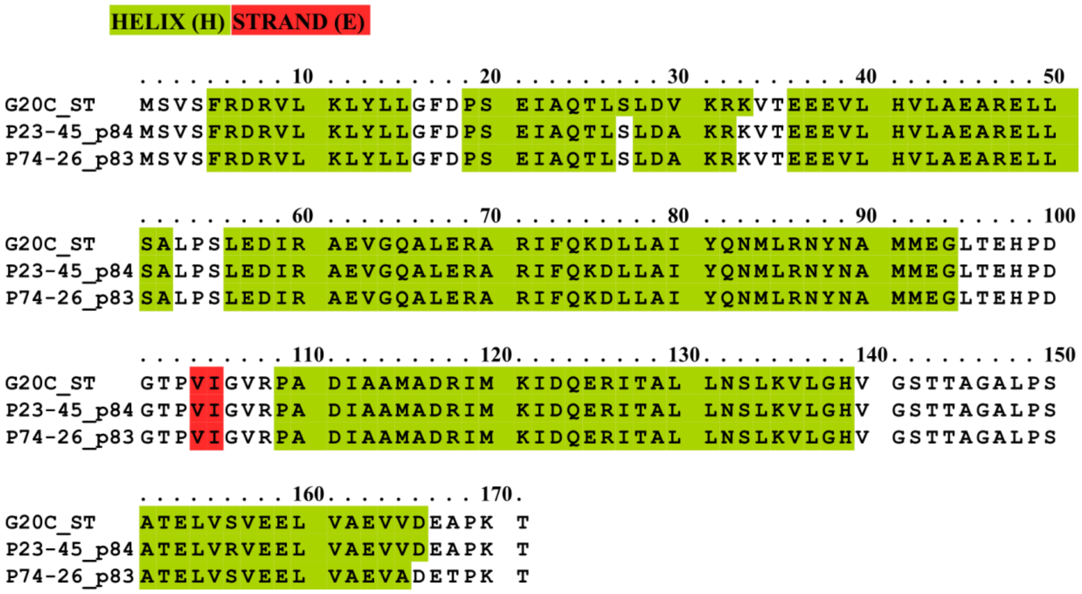
```
            .........10 .........20 .........30 .........40 .........50
G20C_ST     MSVSFRDRVL  KLYLLGFDPS  EIAQTLSLDV  KRKVTEEEVL  HVLAEARELL
P23-45_p84  MSVSFRDRVL  KLYLLGFDPS  EIAQTLSLDA  KRKVTEEEVL  HVLAEARELL
P74-26_p83  MSVSFRDRVL  KLYLLGFDPS  EIAQTLSLDA  KRKVTEEEVL  HVLAEARELL
Consistency **********  **********  *********6  **********  **********

            .........60 .........70 .........80 .........90 .........100
G20C_ST     SALPSLEDIR  AEVGQALERA  RIFQKDLLAI  YQNMLRNYNA  MMEGLTEHPD
P23-45_p84  SALPSLEDIR  AEVGQALERA  RIFQKDLLAI  YQNMLRNYNA  MMEGLTEHPD
P74-26_p83  SALPSLEDIR  AEVGQALERA  RIFQKDLLAI  YQNMLRNYNA  MMEGLTEHPD
Consistency **********  **********  **********  **********  **********

            .........110.........120.........130.........140.........150
G20C_ST     GTPVIGVRPA  DIAAMADRIM  KIDQERITAL  LNSLKVLGHV  GSTTAGALPS
P23-45_p84  GTPVIGVRPA  DIAAMADRIM  KIDQERITAL  LNSLKVLGHV  GSTTAGALPS
P74-26_p83  GTPVIGVRPA  DIAAMADRIM  KIDQERITAL  LNSLKVLGHV  GSTTAGALPS
Consistency **********  **********  **********  **********  **********

            .........160.........170.
G20C_ST     ATELVSVEEL  VAEVVDEAPK  T
P23-45_p84  ATELVRVEEL  VAEVVDEAPK  T
P74-26_p83  ATELVSVEEL  VAEVADETPK  T
Consistency *****5****  ****6**6**  *
```

b) Secondary structure prediction



HELIX (H) STRAND (E)

```
           .........10 .........20 .........30 .........40 .........50
G20C_ST    MSVSFRDRVL  KLYLLGFDPS  EIAQTLSLDV  KRKVTEEEVL  HVLAEARELL
P23-45_p84 MSVSFRDRVL  KLYLLGFDPS  EIAQTLSLDA  KRKVTEEEVL  HVLAEARELL
P74-26_p83 MSVSFRDRVL  KLYLLGFDPS  EIAQTLSLDA  KRKVTEEEVL  HVLAEARELL

           .........60 .........70 .........80 .........90 .........100
G20C_ST    SALPSLEDIR  AEVGQALERA  RIFQKDLLAI  YQNMLRNYNA  MMEGLTEHPD
P23-45_p84 SALPSLEDIR  AEVGQALERA  RIFQKDLLAI  YQNMLRNYNA  MMEGLTEHPD
P74-26_p83 SALPSLEDIR  AEVGQALERA  RIFQKDLLAI  YQNMLRNYNA  MMEGLTEHPD

           .........110.........120.........130.........140.........150
G20C_ST    GTPVIGVRPA  DIAAMADRIM  KIDQERITAL  LNSLKVLGHV  GSTTAGALPS
P23-45_p84 GTPVIGVRPA  DIAAMADRIM  KIDQERITAL  LNSLKVLGHV  GSTTAGALPS
P74-26_p83 GTPVIGVRPA  DIAAMADRIM  KIDQERITAL  LNSLKVLGHV  GSTTAGALPS

           .........160.........170.
G20C_ST    ATELVSVEEL  VAEVVDEAPK  T
P23-45_p84 ATELVRVEEL  VAEVVDEAPK  T
P74-26_p83 ATELVSVEEL  VAEVADETPK  T
```

c) Nucleotide sequence

```
G20C_ST         ATGAGCGTGAGTTTTAGGGACAGGGTGCTCAAGCTTTACCTGCTTGGCTTTGACCCTAGC 60
P23-45_p84      ATGAGCGTGAGTTTTAGGGATAGAGTGCTCAAGCTCTACCTGCTTGGCTTTGACCCTAGC 60
P74-26_p83      ATGAGCGTGAGTTTTAGGGACAGGGTGCTCAAGCTTTACCTGCTTGGCTTTGACCCTAGC 60
                ******************** ** *********** ************************

G20C_ST         GAGATTGCGCAGACCCTTAGCCTGGACGTCAAGCGCAAGGTTACAGAGGAGGAAGTTCTA 120
P23-45_p84      GAGATTGCGCAAACCCTTAGCCTGGACGCCAAGCGCAAGGTTACAGAGGAGGAAGTTCTA 120
P74-26_p83      GAGATTGCGCAAACCCTTAGCCTGGACGCCAAGCGCAAGGTTACAGAGGAGGAAGTTCTA 120
                *********** **************** *******************************

G20C_ST         CACGTCCTAGCCGAGGCTAGAGAGCTTCTTTCCGCCTTGCCTTCCCTTGAAGACATCCGG 180
P23-45_p84      CACGTCCTAGCCGAGGCTAGAGAGCTTCTTTCTGCCTTGCCTTCCCTTGAGGACATCCGG 180
P74-26_p83      CACGTCCTAGCCGAGGCTAGAGAGCTGCTTTCCGCCTTACCTTCCCTTGAGGACATACGG 180
                ************************** ***** ***** *********** ***** ***

G20C_ST         GCCGAGGTGGGTCAGGCTCTAGAGCGCGCCCGGATTTTCCAGAAAGACCTGCTAGCGATT 240
P23-45_p84      GCCGAGGTGGGTCAGGCCCTAGAGCGCGCCCGGATTTTCCAGAAAGACCTGCTAGCGATT 240
P74-26_p83      GCCGAGGTGGGTCAGGCCCTAGAGCGCGCCCGGATTTTCCAGAAAGACCTGCTAGCGATT 240
                **************** *******************************************

G20C_ST         TACCAGAACATGCTCCGCAACTACAACGCCATGATGGAAGGCTTGACCGAGCATCCAGAC 300
P23-45_p84      TACCAGAACATGCTCCGCAACTACAACGCCATGATGGAAGGCTTGACCGAGCATCCAGAC 300
P74-26_p83      TACCAAAACATGCTCCGCAACTACAACGCCATGATGGAAGGCTTGACCGAGCATCCAGAC 300
                ***** ******************************************************

G20C_ST         GGCACCCCGGTGATTGGCGTAAGACCGGCGGATATAGCCGCTATGGCCGACCGGATTATG 360
P23-45_p84      GGTACCCCGGTGATTGGCGTAAGACCGGCGGATATAGCCGCTATGGCCGACCGGATTATG 360
P74-26_p83      GGTACCCCGGTGATTGGCGTAAGACCGGCGGATATAGCCGCTATGGCCGACCGGATTATG 360
                ** *********************************************************

G20C_ST         AAGATTGACCAGGAGCGCATCACCGCTCTGCTCAATAGCCTCAAGGTACTAGGCCATGTC 420
P23-45_p84      AAGATTGACCAGGAGCGCATCACCGCTCTGCTCAATAGCCTCAAGGTGCTAGGCCATGTT 420
P74-26_p83      AAGATTGACCAGGAGCGCATCACCGCTCTGCTCAATAGCCTCAAGGTGCTAGGCCATGTC 420
                *********************************************** ***********

G20C_ST         GGGTCCACAACCGCCGGAGCTCTCCCCTCCGCTACAGAGCTAGTGAGCGTGGAGGAGCTG 480
P23-45_p84      GGGTCCACAACCGCCGGAGCTCTCCCCTCCGCTACAGAGCTAGTGCGCGTGGAGGAGCTG 480
P74-26_p83      GGGTCCACAACCGCCGGAGCTCTCCCCTCCGCTACAGAGCTAGTGAGCGTGGAGGAGTTA 480
                ********************************************* *********** *

G20C_ST         GTGGCGGAGGTGGTGGATGAAGCGCCTAAGACCTAG 516
P23-45_p84      GTGGCGGAGGTGGTGGATGAAGCGCCTAAGACCTAG 516
P74-26_p83      GTGGCGGAGGTGGCGGATGAAACGCCTAAGACCTAG 516
                ************* ******* **************
```

Appendix 10. Conservation of ST from thermophilic phages G20C, P23-45 and P75-26

Multiple sequence alignments to illustrate (a) amino acid sequence conservation, (b) secondary structure conservation and (c) nucleotide sequence conservation between G20C ST and ORF P23p84 and ORF P74p83 from phages P23-45 and P75-26. Figures a and b were prepared using the PRALINE server. Figure c was made using ClustalW.

Appendix 11. Derived Publications

Acta Crystallographica Section F
Structural Biology and Crystallization Communications
ISSN 1744-3091

The putative small terminase from the thermophilic dsDNA bacteriophage G20C is a nine-subunit oligomer

Juan Loredo-Varela,[a] Maria Chechik,[a] Vladimir M. Levdikov,[a] Ahmad Abd-El-Aziz,[a] Leonid Minakhin,[b] Konstantin Severinov,[b,c,d] Callum Smits[a]‡ and Alfred A. Antson[a]*

[a]York Structural Biology Laboratory, Department of Chemistry, University of York, York YO10 5DD, England, [b]Waksman Institute for Microbiology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA, [c]Department of Molecular Biology and Biochemistry, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA, and [d]Institutes of Molecular Genetics and Gene Biology, Russian Academy of Sciences, Moscow 119334, Russian Federation

‡ Current address: The Victor Chang Cardiac Research Institute, Darlinghurst, NSW 2010, Australia.

Correspondence e-mail: fred.antson@york.ac.uk

Received 3 May 2013
Accepted 19 June 2013

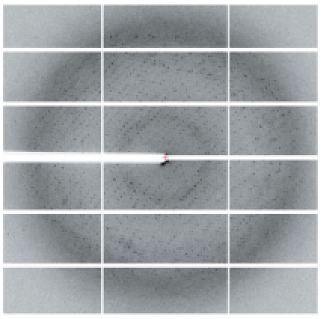
The assembly of double-stranded DNA bacteriophages is dependent on a small terminase protein that normally plays two important roles. Firstly, the small terminase protein specifically recognizes viral DNA and recruits the large terminase protein, which makes the initial cut in the dsDNA. Secondly, once the complex of the small terminase, the large terminase and the DNA has docked to the portal protein, and DNA translocation into a preformed empty procapsid has begun, the small terminase modulates the ATPase activity of the large terminase. Here, the putative small terminase protein from the thermostable bacteriophage G20C, which infects the Gram-negative eubacterium *Thermus thermophilus*, has been produced, purified and crystallized. Size-exclusion chromatography–multi-angle laser light scattering data indicate that the protein forms oligomers containing nine subunits. Crystals diffracting to 2.8 Å resolution have been obtained. These belonged to space group $P2_12_12_1$, with unit-cell parameters a = 94.31, b = 125.6, c = 162.8 Å. The self-rotation function and Matthews coefficient calculations are consistent with the presence of a nine-subunit oligomer in the asymmetric unit.

1. Introduction

During the assembly of double-stranded DNA bacteriophages, DNA is usually translocated into preformed procapsids by a molecular motor consisting of the small and large terminase proteins and the portal protein (Casjens, 2011; Feiss & Rao, 2012). The portal protein, a circular oligomer embedded in one of the fivefold symmetrical vertices of the icosahedral shell, contains a tunnel for DNA translocation. Initially, the small terminase specifically recognizes the bacteriophage genomic DNA and recruits the large terminase protein. Following DNA cutting by the large terminase, the complex of the two terminase proteins and DNA docks to the portal protein of an empty prophage. DNA translocation into the procapsid is fuelled by the ATPase activity of the large terminase protein (Sun *et al.*, 2008). During DNA translocation, the small terminase protein modulates the ATPase and nuclease activities of the large terminase protein (Gual *et al.*, 2000).

Three-dimensional structural data are available for oligomeric assemblies of small terminases from several viruses including Sf6, SF6, T4 and P22 (Zhao *et al.*, 2010; Büttner *et al.*, 2012; Sun *et al.*, 2012; Roy *et al.*, 2012). All were shown to assemble into ring-like structures composed of 8–12 subunits arranged symmetrically around a central axis. The main topological domains identified in the small terminase are (i) the C-terminal oligomerization domain, which establishes inter-subunit contacts around a central tunnel, and (ii) the exposed N-terminal domain, which in phages SPP1 and SF6 binds to the recognition *pac* site DNA (Büttner *et al.*, 2012). Interestingly, in phage P22 the DNA-binding function is attributed to a short segment at the C-terminus (Roy *et al.*, 2012).

Despite the availability of structural information, several questions concerning the mechanism by which the small terminase carries out its function remain to be answered. As mentioned above, there appear to be differences in the way that different terminases recognize the genomic DNA (Büttner *et al.*, 2012; de Beer *et al.*, 2002; Sun *et al.*, 2012; Roy *et al.*, 2012). Perhaps more importantly, there are



contradicting reports about the potential involvement of the central tunnel in DNA translocation (Roy *et al.*, 2012; Büttner *et al.*, 2012).

To answer some of these questions, we initiated structural and functional studies on the putative small terminase protein from the *Thermus thermophilus* bacteriophage G20C, which is a close relative of bacteriophages P23-45 and P74-26 (Minakhin *et al.*, 2008). Here, we report the production of recombinant protein in *Escherichia coli*, protein purification and crystallization. The results of size-exclusion chromatography coupled with multi-angle laser light scattering (SEC–MALLS) and crystal data indicate that the protein forms nine-subunit oligomers, like the small terminases found in bacteriophages SF6 and P22 (Büttner *et al.*, 2012; Roy *et al.*, 2012).

2. Materials and methods

2.1. Cloning

It was not possible to locate the small terminase gene based on sequence homology to small terminases from other viruses. However, as the small terminase is usually encoded by a gene immediately preceding the large terminase and portal protein genes, and because the corresponding G20C gene had an appropriate size, we decided to clone this gene of G20C. This gene corresponds to the ORF P23p84 (UniProtKB/TrEMBL A7XXB6) in the closely related phage P23-45 (Minakhin *et al.*, 2008). Forward and reverse primers containing the *Nde*I and *Hin*dIII restriction-site sequences, respectively, enclosing the full-length protein were designed as follows: forward, 5′-GGACAACATATGAGCGTGAGTTTTAGGGAC-3′; reverse, 5′-GGCAAGCTTCTAGGTCTTAGGCGCTTCATC-3′. The amplified segment was cloned into the pET28a vector (Novagen, Merck KGaA).

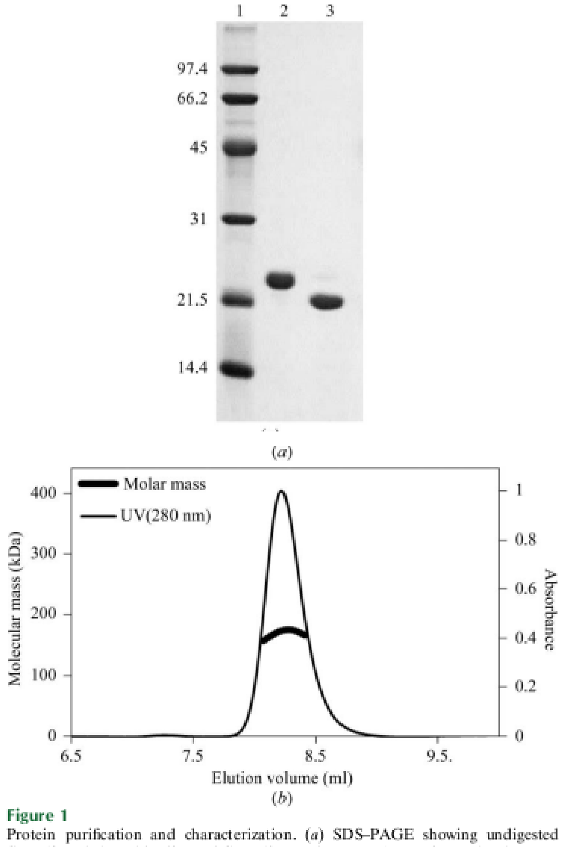


Figure 1
Protein purification and characterization. (*a*) SDS–PAGE showing undigested (lane 2) and thrombin-digested (lane 3) samples. Lane 1 contains molecular-mass markers (labelled in kDa). (*b*) SEC–MALLS analysis. The thin line corresponds to the absorbance at 280 nm. The thick line below the absorbance peak corresponds to the molecular weight calculated for the eluted protein.

2.2. Protein expression and purification

All chemicals were purchased from Sigma–Aldrich, unless stated otherwise. *E. coli* B834 cells (Novagen, Merck KGaA) were transformed with the recombinant DNA and grown at 310 K until the OD_{600} reached ~0.8. Protein expression was then induced with 1 m*M* IPTG at 289 K. Before sonication, the cells were lysed in a buffer consisting of 500 m*M* NaCl, 50 m*M* Tris pH 7.5, 20 m*M* imidazole, 100 µg ml^{-1} lysozyme, 0.7 µg ml^{-1} pepstatin A, 0.5 µg ml^{-1} leupeptin, 100 m*M* 4-(2-aminoethyl)benzenesulfonyl fluoride hydrochloride. The His-tagged protein was purified by Ni-affinity chromatography (ÄKTA, GE Healthcare) by binding the protein to nickel beads on a His-Trap column (GE Healthcare) and by further elution with an imidazole gradient. The binding and elution buffers consisted of 500 m*M* NaCl, 50 m*M* Tris pH 7.5 with 20 and 500 m*M* imidazole, respectively.

The His tag was cleaved by thrombin digestion (BD Biosciences) while the sample was dialysed against the binding buffer (no imidazole). One unit of thrombin per milligram of protein was used to digest the protein overnight. A second Ni-affinity chromatography was performed to separate cleaved protein from noncleaved protein, followed by size-exclusion chromatography in 250 m*M* NaCl, 20 m*M* Tris pH 7.5 using a Superdex 16/60 column (GE Healthcare).

2.3. Characterization of the oligomeric state by SEC–MALLS

Thrombin-digested protein was diluted to 4 mg ml^{-1} in 250 m*M* NaCl, 20 m*M* Tris pH 7.5 and loaded onto a BioSep SEC-s3000 gel-filtration column (Phenomenex) which was equilibrated with 250 m*M* NaCl, 20 m*M* Tris pH 7.5. Size-exclusion chromatography was carried out on a Shimadzu HPLC system with a flow rate of 0.5 ml min^{-1}. The elution was monitored at 280 nm using a SPD20A UV–Vis detector. Light-scattering data were recorded by a Dawn HELEOS II 18-angle light-scattering detector and the concentration of the eluting protein was measured using an inline Optilab rEX refractive-index monitor (Wyatt Technology). Data were analysed using the *ASTRA V* software package (Wyatt Technology).

2.4. Crystallization

The purified protein was concentrated to 21 mg ml^{-1} in a solution containing 175 m*M* NaCl and 10 m*M* Tris pH 7.5. Crystallization experiments using the Index screen (Hampton Research) were set up with a Mosquito nanolitre pipetting robot (TTP LabTech). Crystals grew within a few days from condition No. 80 of the Index screen in sitting drops at 293 K. These conditions were manually optimized in 24-well hanging-drop plates (Greiner Bio-One) and the obtained crystals were used as seeds for subsequent optimization experiments. A seed stock was produced using a tube with a seed bead (Hampton Research) and was stored in 100 µl mother liquor consisting of 0.4 *M* ammonium acetate, 0.1 *M* HEPES pH 7.5, 30%(*w*/*v*) PEG 3350. 0.5 µl of the seed stock was mixed with 1 µl protein solution and 1 µl mother liquor for the next round of optimization. The best crystals grew within one month using a reservoir solution consisting of 0.4 *M* ammonium acetate, 23%(*w*/*v*) PEG 3350, 0.1 *M* HEPES pH 7.5,

Acta Cryst. (2013). F**69**, 876–879 Loredo-Varela *et al.* • Putative small terminase **877**

9%(*v*/*v*) ethylene glycol. Crystals were tested in-house using an MSC MicroMax-007 HF rotating-anode X-ray generator (Rigaku) and a MAR345 detector (MAR Research).

2.5. X-ray data collection and processing

X-ray data were collected from a single cryocooled crystal on the I04 beamline at the Diamond Light Source, England. Data were collected at a wavelength of 0.9200 Å with a crystal-to-detector distance of 325.2 mm, a 0.2° crystal rotation per image and a total crystal rotation range of 180°. The data were indexed with *HKL*-2000 (Otwinowski & Minor, 1997) and processed with *XDS* (Kabsch, 2010) and *SHELX* beta (Sheldrick, 2010). The self-rotation function was calculated using *MOLREP* (Vagin & Teplyakov, 2010) with a resolution range of 10–3.0 Å and a radius of integration of 35 Å. Other crystallographic calculations were performed using the *CCP*4 suite of programs (Winn *et al.*, 2011).

2.6. Secondary-structure prediction

The secondary structure of the putative small terminase was predicted using *Jpred* (Cole *et al.*, 2008).

3. Results and discussion

3.1. Cloning, protein expression and purification

The recombinant protein, containing a His tag at the N-terminus with a thrombin protease cleavage site between the tag and the protein-coding region, was expressed in *E. coli* B834 cells at 289 K. The molecular weight of the expressed protein was 20 957 Da, or 19 074 Da after thrombin cleavage. Following thrombin digestion, size-exclusion chromatography produced a highly purified protein sample (Fig. 1*a*).

3.2. Oligomeric state determination by SEC–MALLS

SEC–MALLS analysis was performed to assess whether the protein forms oligomers containing multiple subunits, as observed for the small terminases of other phages. This experiment was performed with the final purified protein sample following thrombin digestion and the second Ni-affinity purification. The data revealed a homogeneous monodisperse protein preparation with an estimated molecular mass of ~170.9 kDa, corresponding to 8.9 subunits per oligomer (Fig. 1*b*). The data indicate that the putative small terminase protein forms nine-subunit oligomers, as observed for the small terminases of bacteriophages SF6 (Büttner *et al.*, 2012) and P22 (Roy *et al.*, 2012).

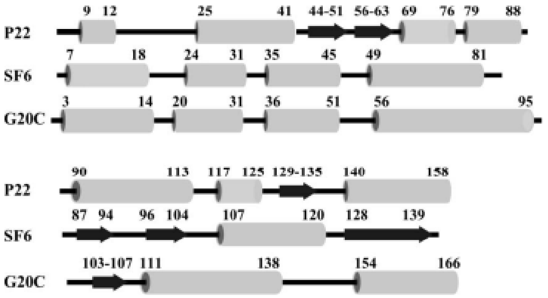


Figure 2
Secondary-structure alignment for small terminase proteins from bacteriophages SF6 and P22 and the putative G20C small terminase.

The predicted secondary structure of the G20C protein is consistent with the secondary structure of these two small terminases (Fig. 2). Interestingly, while the four N-terminal α-helices match the secondary structure observed in the SF6 small terminase, the α-helices at the C-terminus are more consistent with the secondary structure observed in the P22 protein.

3.3. Crystallization and X-ray data analysis

The best crystals were obtained by microseeding using 21 mg ml^{-1} protein solution containing 175 m*M* NaCl and 10 m*M* Tris pH 7.5 and

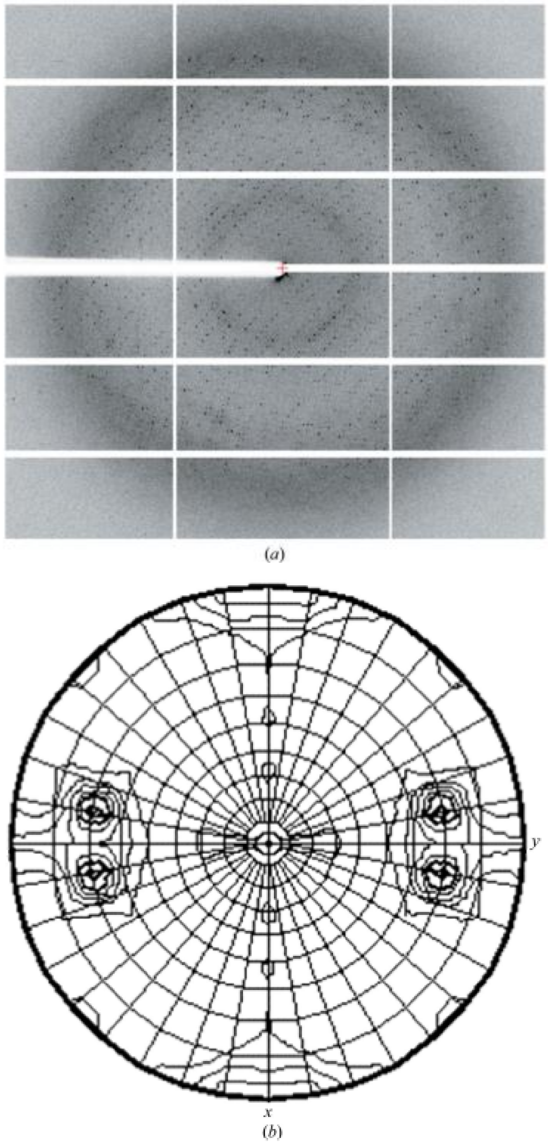


Figure 3
X-ray analysis. (*a*) Diffraction image. The resolution at the edge of the plate is 2.5 Å. (*b*) Stereographic projection (κ = 40° section) of the self-rotation function.

Table 1
X-ray data statistics.

Values in parentheses are for the outermost resolution shell.

X-ray source	I04, Diamond Light Source
Wavelength (Å)	0.92000
Temperature (K)	100
Space group	$P2_12_12_1$
Unit-cell parameters (Å)	a = 94.3, b = 125.6, c = 162.8
Resolution range (Å)	49.8–2.8 (2.89–2.80)
No. of unique reflections	48371 (4371)
$R_{\rm merge}$† (%)	9.6 (15.3)
Average $I/\sigma(I)$	16.3 (1.3)
Completeness (%)	99.9 (100)
Multiplicity	6.8 (6.1)

† $R_{\rm merge} = \sum_{hkl}\sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl}\sum_i I_i(hkl)$, where $I_i(hkl)$ is the intensity of the ith observation of reflection hkl, $\langle I(hkl)\rangle$ is the average value of the intensity, the sum $\sum_{hkl}$ is over all measured reflections and the sum $\sum_i$ is over i measurements of a reflection.

reservoir solution consisting of 0.4 *M* ammonium acetate, 23%(*w*/*v*) PEG 3350, 0.1 *M* HEPES pH 7.5, 9%(*v*/*v*) ethylene glycol. The synchrotron X-ray data from a crystal belonging to the orthorhombic space group $P2_12_12_1$, with unit-cell parameters a = 94.31, b = 125.6, c = 162.8 Å, extended to 2.8 Å resolution (Fig. 3*a*, Table 1).

The highest peaks in the self-rotation function calculated with *MOLREP* were 17% of the origin peak. These peaks are in the κ = 40° section, corresponding to ninefold rotational symmetry (Fig. 3*b*). The specific volume, corresponding to nine subunits per asymmetric unit, is 2.8 Å^3 Da^{-1}. This corresponds to a solvent content of 56.1% (Winn *et al.*, 2011; Matthews, 1968).

Although both SEC–MALLS and crystallographic data indicate nine-subunit oligomers, as observed for the P22 and SF6 small terminases (Roy *et al.*, 2012; Büttner *et al.*, 2012), structure determination by molecular replacement is not possible owing to a complete lack of sequence similarity. The next stage of this project will focus on experimental phasing.

4. Conclusions

The putative small terminase protein from the thermophilic bacteriophage G20C forms nine-subunit assemblies both in solution and in the crystal. The genomic context, the predicted secondary structure and the oligomeric state of the protein are consistent with this protein being the small terminase.

We thank Johan Turkenburg and Sam Hart for collecting the X-ray data. This project was supported by a CONACyT PhD fellowship to JL-V, the Wellcome Trust CIDCATS PhD programme 095024 (studentship to AAEA) and by Wellcome Trust fellowship 098230 to AAA.

References

Beer, T. de, Fang, J., Ortega, M., Yang, Q., Maes, L., Duffy, C., Berton, N., Sippy, J., Overduin, M., Feiss, M. & Catalano, C. E. (2002). *Mol. Cell*, **9**, 981–991.
Büttner, C. R., Chechik, M., Ortiz-Lombardiá, M., Smits, C., Ebong, I.-O., Chechik, V., Jeschke, G., Dykeman, E., Benini, S., Robinson, C. V., Alonso, J. C. & Antson, A. A. (2012). *Proc. Natl Acad. Sci. USA*, **109**, 811–816.
Casjens, S. R. (2011). *Nature Rev. Microbiol.* **9**, 647–657.
Cole, C., Barber, J. D. & Barton, G. J. (2008). *Nucleic Acids Res.* **36**, W197–W201.
Feiss, M. & Rao, V. B. (2012). *Adv. Exp. Med. Biol.* **726**, 489–509.
Gual, A., Camacho, A. G. & Alonso, J. C. (2000). *J. Biol. Chem.* **275**, 35311–35319.
Kabsch, W. (2010). *Acta Cryst.* D**66**, 125–132.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
Minakhin, L., Goel, M., Berdygulova, Z., Ramanculov, E., Florens, L., Glazko, G., Karamychev, V. N., Slesarev, A. I., Kozyavkin, S. A., Khromov, I., Ackermann, H. W., Washburn, M., Mushegian, A. & Severinov, K. (2008). *J. Mol. Biol.* **378**, 468–480.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Roy, A., Bhardwaj, A., Datta, P., Lander, G. C. & Cingolani, G. (2012). *Structure*, **20**, 1403–1413.
Sheldrick, G. M. (2010). *Acta Cryst.* D**66**, 479–485.
Sun, S., Gao, S., Kondabagil, K., Xiang, Y., Rossmann, M. G. & Rao, V. B. (2012). *Proc. Natl Acad. Sci. USA*, **109**, 817–822.
Sun, S., Kondabagil, K., Draper, B., Alam, T. I., Bowman, V. D., Zhang, Z., Hegde, S., Fokine, A., Rossmann, M. G. & Rao, V. B. (2008). *Cell*, **135**, 1251–1262.
Vagin, A. & Teplyakov, A. (2010). *Acta Cryst.* D**66**, 22–25.
Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.
Zhao, H., Finch, C. J., Sequeira, R. D., Johnson, B. A., Johnson, J. E., Casjens, S. R. & Tang, L. (2010). *Proc. Natl Acad. Sci. USA*, **107**, 1971–1976.

Abbreviations

A-domain	Axial domain
AEBSF	4-(2-Aminoethyl)-benzenesulfonyl fluoride hydrochloride
AEC	Anion exchange chromatography
Anti-TRAP	Anti-Tryptophan RNA-binding attenuation protein
ATP	Adenosine triphosphate
bp	base pairs
CCP4	Computational collaborative project 4
CTD	C-terminal domain
CV	Column volume
cryo-EM	Cryo-Electronic microscopy
DBD	DNA-binding domain
DNA	Deoxyribonuclease acid, ds for double stranded
EI	Electrospray ionisation
EMSA	Electrophoretic mobility shift assay
FL	Full length
FRET	Fluorescence resonance energy transfer
GST	Glutathione S-transferase
G*X*P	Gene *X* product
HEPES	2-[4-(2-hydroxyethyl) piperazin-1-yl] ethanesulfonic acid
IPTG	Isopropyl-β-D-thiogalacto-pyranoside
LB	Luria-Bertani medium
LT	Large terminase
MES	2-(N-morpholino) ethanesulfonic acid
MPD	2-methylpentane-2,4-diol
MR	Molecular Replacement
MW	Molecular weight
NAC	Nickel affinity chromatography
NMR	Nuclear magnetic resonance
NTD	N-terminal domain
OD	Oligomerisation domain
orf	open reading frames
PCR	Polymerase chain reaction
PDB	Protein data bank
P-domain	Peripheral domain
pET	Plasmid for expression by T7 RNA polymerase
POI	Protein of interest
Psi-BLAST	Position-specific iterated basic local alignment search tool
rpm	revolutions per minute
rmsd	Root mean square deviation
SDS-PAGE	Sodium dodecyl sulphate- polyacrylamide gel electrophoresis
SEC-MALLS	Size exclusion chromatography – Multiple-angle laser light scattering
SPR	Surface plasmon resonance
ST	Small terminase
WT	Wild type

References

1. Wommack, K.E. and Colwell, R.R. (2000) Virioplankton: Viruses in aquatic ecosystems. *Microbiol Mol Biol R*, **64**, 69-+.
2. Orlova, E.V. (2012) *Bacteriophages and Their Structural Organisation, Bacteriophages*. Intech, Rijeka, Croatia.
3. Riva, S., Polsinel.M and Falaschi, A. (1968) A New Phage of Bacillus Subtilis with Infectious DNA Having Separable Strands. *J. Mol. Biol.*, **35**, 347-&.
4. Alonso, J.C., Luder, G., Stiege, A.C., Chai, S., Weise, F. and Trautner, T.A. (1997) The complete nucleotide sequence and functional organization of Bacillus subtilis bacteriophage SPP1. *Gene*, **204**, 201-212.
5. Chai, S., Bravo, A., Luder, G., Nedlin, A., Trautner, T.A. and Alonso, J.C. (1992) Molecular analysis of the Bacillus subtilis bacteriophage SPP1 region encompassing genes 1 to 6. The products of gene 1 and gene 2 are required for pac cleavage. *J Mol Biol*, **224**, 87-102.
6. Yu, M.X., Slater, M.R. and Ackermann, H.W. (2006) Isolation and characterization of Thermus bacteriophages. *Arch Virol*, **151**, 663-679.
7. Minakhin, L., Goel, M., Berdygulova, Z., Ramanculov, E., Florens, L., Glazko, G., Karamychev, V.N., Slesarev, A.I., Kozyavkin, S.A., Khromov, I., Ackermann, H.-W., Washburn, M., Mushegian, A. and Severinov, K. (2008) Genome Comparison and Proteomic Characterization of Thermus thermophilus Bacteriophages P23-45 and P74-26: Siphoviruses with Triplex-forming Sequences and the Longest Known Tails. *J. Mol. Biol.*, **378**, 468-480.
8. Rao, V.B. and Feiss, M. (2008) The Bacteriophage DNA Packaging Motor. *Annu Rev Genet*, **42**, 647-681.
9. Casjens, S.R. (2011) The DNA-packaging nanomotor of tailed bacteriophages. *Nat Rev Micro*, **9**, 647-657.
10. Feiss, M. and Rao, V.B. (2012) The Bacteriophage DNA Packaging Machine. *Adv Exp Med Biol*, **726**, 489-509.
11. Lebedev, A.A., Krause, M.H., Isidro, A.L., Vagin, A.A., Orlova, E.V., Turner, J., Dodson, E.J., Tavares, P. and Antson, A.A. (2007) Structural framework for DNA translocation via the viral portal protein. *Embo J*, **26**, 1984-1994.
12. Oliveira, L., Tavares, P. and Alonso, J.C. (2013) Headful DNA packaging: bacteriophage SPP1 as a model system. *Virus research*, **173**, 247-259.
13. Gual, A., Camacho, A.G. and Alonso, J.C. (2000) Functional analysis of the terminase large subunit, G2P, of Bacillus subtilis bacteriophage SPP1. *J Biol Chem*, **275**, 35311-35319.
14. Roy, A., Bhardwaj, A., Datta, P., Lander, Gabriel C. and Cingolani, G. (2012) Small Terminase Couples Viral DNA Binding to Genome-Packaging ATPase Activity. *Structure*, **20**, 1403-1413.
15. Shen, X.D., Li, M., Zeng, Y.J., Hu, X.M., Tan, Y.L., Rao, X.C., Jin, X.L., Li, S., Zhu, J.M., Zhang, K.B. and Hu, F.Q. (2012) Functional identification of the DNA packaging terminase from Pseudomonas aeruginosa phage PaP3. *Arch Virol*, **157**, 2133-2141.
16. Buttner, C.R., Chechik, M., Ortiz-Lombardia, M., Smits, C., Ebong, I.O., Chechik, V., Jeschke, G., Dykeman, E., Benini, S., Robinson, C.V., Alonso, J.C. and Antson, A.A. (2012) Structural basis for DNA recognition and loading into a viral packaging motor. *P Natl Acad Sci USA*, **109**, 811-816.
17. Sun, S., Kondabagil, K., Draper, B., Alam, T.I., Bowman, V.D., Zhang, Z., Hegde, S., Fokine, A., Rossmann, M.G. and Rao, V.B. (2008) The Structure of the Phage T4 DNA Packaging Motor Suggests a Mechanism Dependent on Electrostatic Forces. *Cell*, **135**, 1251-1262.
18. McNicholas, S., Potterton, E., Wilson, K.S. and Noble, M.E. (2011) Presenting your structures: the CCP4mg molecular-graphics software. *Acta crystallographica. Section D, Biological crystallography*, **67**, 386-394.

19. Chai, S.H., Lurz, R. and Alonso, J.C. (1995) The Small-Subunit of the Terminase Enzyme of Bacillus-Subtilis Bacteriophage-Spp1 Forms a Specialized Nucleoprotein Complex with the Packaging Initiation Region. *J. Mol. Biol.*, **252**, 386-398.

20. Chai, S. and Alsonso, J.C. (1996) Distamycin-induced inhibition of formation of a nucleoprotein complex between the terminase small subunit G1P and the non-encapsidated end (pacL site) of Bacillus subtilis bacteriophage SPP1. *Nucleic Acids Res*, **24**, 282-288.

21. Chai, S.H., Kruft, V. and Alonso, J.C. (1994) Analysis of the Bacillus-Subtilis Bacteriophage-Spp1 and Bacteriophage-Sf6 Gene-1 Product - a Protein Involved in the Initiation of Headful Packaging. *Virology*, **202**, 930-939.

22. Smits, C., Chechik, M., Kovalevskiy, O.V., Shevtsov, M.B., Foster, A.W., Alonso, J.C. and Antson, A.A. (2009) Structural basis for the nuclease activity of a bacteriophage large terminase. *Embo Rep*, **10**, 592-598.

23. Oliveira, L., Cuervo, A. and Tavares, P. (2010) Direct Interaction of the Bacteriophage SPP1 Packaging ATPase with the Portal Protein. *J Biol Chem*, **285**, 7366-7373.

24. Leavitt, J.C., Gilcrease, E.B., Wilson, K. and Casjens, S.R. (2013) Function and horizontal transfer of the small terminase subunit of the tailed bacteriophage Sf6 DNA packaging nanomotor. *Virology*, **440**, 117-133.

25. Cornilleau, C., Atmane, N., Jacquet, E., Smits, C., Alonso, J.C., Tavares, P. and Oliveira, L. (2013) The nuclease domain of the SPP1 packaging motor coordinates DNA cleavage and encapsidation. *Nucleic Acids Res*, **41**, 340-354.

26. Tadokoro, T. and Kanaya, S. (2009) Ribonuclease H: molecular diversities, substrate binding domains, and catalytic mechanism of the prokaryotic enzymes. *Febs Journal*, **276**, 1482-1493.

27. Orlova, E.V., Gowen, B., Droge, A., Stiege, A., Weise, F., Lurz, R., van Heel, M. and Tavares, P. (2003) Structure of a viral DNA gatekeeper at 10 angstrom resolution by cryo-electron microscopy. *Embo J*, **22**, 1255-1262.

28. Dube, P., Tavares, P., Lurz, R. and Vanheel, M. (1993) The Portal Protein of Bacteriophage-Spp1 - a DNA Pump with 13-Fold Symmetry. *Embo J*, **12**, 1303-1309.

29. Becker, B., delaFuente, N., Gassel, M., Gunther, D., Tavares, P., Lurz, R., Trautner, T.A. and Alonso, J.C. (1997) Head morphogenesis genes of the Bacillus subtilis bacteriophage SPP1. *J. Mol. Biol.*, **268**, 822-839.

30. Droge, A., Santos, M.A., Stiege, A.C., Alonso, J.C., Lurz, R., Trautner, T.A. and Tavares, P. (2000) Shape and DNA packaging activity of bacteriophage SPP1 procapsid: Protein components and interactions during assembly. *J. Mol. Biol.*, **296**, 117-132.

31. Orlova, E.V., Dube, P., Beckmann, E., Zemlin, F., Lurz, R., Trautner, T.A., Tavares, P. and van Heel, M. (1999) Structure of the 13-fold symmetric portal protein of bacteriophage SPP1. *Nat Struct Biol*, **6**, 842-846.

32. Caspar, D.L. and Klug, A. (1962) Physical principles in the construction of regular viruses. *Cold Spring Harbor symposia on quantitative biology*, **27**, 1-24.

33. Crick, F.H.C. and Watson, J.D. (1956) Structure of Small Viruses. *Nature*, **177**, 473-475.

34. Prasad, B.V. and Schmid, M.F. (2012) Principles of virus structural organization. *Adv Exp Med Biol*, **726**, 17-47.

35. Johnson, J.E. (1996) Functional implications of protein-protein interactions in icosahedral viruses. *P Natl Acad Sci USA*, **93**, 27-33.

36. Johnson, J.E. and Speir, J.A. (1997) Quasi-equivalent viruses: A paradigm for protein assemblies. *J. Mol. Biol.*, **269**, 665-675.

37. Fokine, A., Leiman, P.G., Shneider, M.M., Ahvazi, B., Boeshans, K.M., Steven, A.C., Black, L.W., Mesyanzhinov, V.V. and Rossmann, M.G. (2005) Structural and functional similarities between the capsid proteins of bacteriophages T4 and HK97 point to a common ancestry. *P Natl Acad Sci USA*, **102**, 7163-7168.

38. Morais, M.C., Choi, K.H., Koti, J.S., Chipman, P.R., Anderson, D.L. and Rossmann, M.G. (2005) Conservation of the Capsid Structure in Tailed dsDNA Bacteriophages: the Pseudoatomic Structure of ϕ29. *Molecular Cell*, **18**, 149-159.

39. Wikoff, W.R., Liljas, L., Duda, R.L., Tsuruta, H., Hendrix, R.W. and Johnson, J.E. (2000) Topologically linked protein rings in the bacteriophage HK97 capsid. *Science*, **289**, 2129-2133.

40. Huang, R.K., Khayat, R., Lee, K.K., Gertsman, I., Duda, R.L., Hendrix, R.W. and Johnson, J.E. (2011) The Prohead-I Structure of Bacteriophage HK97: Implications for Scaffold-Mediated Control of Particle Assembly and Maturation. *J. Mol. Biol.*, **408**, 541-554.

41. White, H.E., Sherman, M.B., Brasiles, S., Jacquet, E., Seavers, P., Tavares, P. and Orlova, E.V. (2012) Capsid Structure and Its Stability at the Late Stages of Bacteriophage SPP1 Assembly. *J Virol*, **86**, 6768-6777.

42. Bamford, D.H., Grimes, J.M. and Stuart, D.I. (2005) What does structure tell us about virus evolution? *Curr Opin Struc Biol*, **15**, 655-663.

43. Parent, K.N., Khayat, R., Tu, L.H., Suhanovsky, M.M., Cortines, J.R., Teschke, C.M., Johnson, J.E. and Baker, T.S. (2010) P22 coat protein structures reveal a novel mechanism for capsid maturation: stability without auxiliary proteins or chemical crosslinks. *Structure*, **18**, 390-401.

44. Akita, F., Chong, K.T., Tanaka, H., Yamashita, E., Miyazaki, N., Nakaishi, Y., Suzuki, M., Namba, K., Ono, Y., Tsukihara, T. and Nakagawa, A. (2007) The crystal structure of a virus-like particle from the hyperthermophilic archaeon Pyrococcus furiosus provides insight into the evolution of viruses. *J Mol Biol*, **368**, 1469-1483.

45. Genomics., J.C.f.S. (2005) Crystal structure of a phage-related protein (BB3626) from *Bordetella bronchiseptica* RB50 at 2.05 A resolution.

46. Zhang, R., Hatzos, C., Abdullah, J., Joachimiak, A. (2008) Crystal structure of the putative capsid protein of prophage (E.coli CFT073). PDB: 3BQW.

47. Sutter, M., Boehringer, D., Gutmann, S., Guenther, S., Prangishvili, D., Loessner, M.J., Stetter, K.O., Weber-Ban, E. and Ban, N. (2008) Structural basis of enzyme encapsulation into a bacterial nanocompartment. *Nature Structural & Molecular Biology*, **15**, 939-947.

48. McHugh, C.A., Fontana, J., Nemecek, D., Cheng, N., Aksyuk, A.A., Heymann, J.B., Winkler, D.C., Lam, A.S., Wall, J.S., Steven, A.C. and Hoiczyk, E. (2014) A virus capsid-like nanocompartment that stores iron and protects bacteria from oxidative stress. *Embo J*, **33**, 1896-1911.

49. Chen, D.H., Baker, M.L., Hryc, C.F., DiMaio, F., Jakana, J., Wu, W.M., Dougherty, M., Haase-Pettingell, C., Schmid, M.F., Jiang, W., Baker, D., King, J.A. and Chiu, W. (2011) Structural basis for scaffolding-mediated assembly and maturation of a dsDNA virus. *P Natl Acad Sci USA*, **108**, 1355-1360.

50. Parent, K.N., Gilcrease, E.B., Casjens, S.R. and Baker, T.S. (2012) Structural evolution of the P22-like phages: comparison of Sf6 and P22 procapsid and virion architectures. *Virology*, **427**, 177-188.

51. Effantin, G., Boulanger, P., Neumann, E., Letellier, L. and Conway, J.F. (2006) Bacteriophage T5 structure reveals similarities with HK97 and T4 suggesting evolutionary relationships. *J. Mol. Biol.*, **361**, 993-1002.

52. Ionel, A., Velazquez-Muriel, J.A., Luque, D., Cuervo, A., Caston, J.R., Valpuesta, J.M., Martin-Benito, J. and Carrascosa, J.L. (2011) Molecular rearrangements involved in the capsid shell maturation of bacteriophage T7. *J Biol Chem*, **286**, 234-242.

53. Lander, G.C., Evilevitch, A., Jeembaeva, M., Potter, C.S., Carragher, B. and Johnson, J.E. (2008) Bacteriophage lambda stabilization by auxiliary protein gpD: Timing, location, and mechanism of attachment determined by cryo-EM. *Structure*, **16**, 1399-1406.

54. Liu, X.A., Zhang, Q.F., Murata, K., Baker, M.L., Sullivan, M.B., Fu, C., Dougherty, M.T., Schmid, M.F., Osburne, M.S., Chisholm, S.W. and Chiu, W. (2010) Structural changes in a marine podovirus associated with release of its genome into Prochlorococcus. *Nature Structural & Molecular Biology*, **17**, 830-U876.

55. Jiang, W., Baker, M.L., Jakana, J., Weigele, P.R., King, J. and Chiu, W. (2008) Backbone structure of the infectious epsilon15 virus capsid revealed by electron cryomicroscopy. *Nature*, **451**, 1130-1134.

56. Rizzo, A.A., Suhanovsky, M.M., Baker, M.L., Fraser, L.C.R., Jones, L.M., Rempel, D.L., Gross, M.L., Chiu, W., Alexandrescu, A.T. and Teschke, C.M. (2014) Multiple Functional Roles of the Accessory I-Domain of Bacteriophage P22 Coat Protein Revealed by NMR Structure and CryoEM Modeling. *Structure*, **22**, 830-841.

57. Poh, S.L., el Khaday, F., Berrier, C., Lurz, R., Melki, R. and Tavares, P. (2008) Oligomerization of the SPP1 scaffolding protein. *J. Mol. Biol.*, **378**, 551-564.

58. Vinga, I., Droge, A., Stiege, A.C., Lurz, R., Santos, M.A., Daugelavicius, R. and Tavares, P. (2006) The minor capsid protein gp7 of bacteriophage SPP1 is required for efficient infection of Bacillus subtilis. *Mol Microbiol*, **61**, 1609-1621.

59. Stiege, A.C., Isidro, A., Droge, A. and Tavares, P. (2003) Specific targeting of a DNA-binding protein to the SPP1 procapsid by interaction with the portal oligomer. *Mol Microbiol*, **49**, 1201-1212.

60. Plisson, C., White, H.E., Auzat, I., Zafarani, A., Sao-Jose, C., Lhuillier, S., Tavares, P. and Orlova, E.V. (2007) Structure of bacteriophage SPP1 tail reveals trigger for DNA ejection. *Embo J*, **26**, 3720-3728.

61. Lhuillier, S., Gallopin, M., Gilquin, B., Brasiles, S., Lancelot, N., Letellier, G., Gilles, M., Dethan, G., Orlova, E.V., Couprie, J., Tavares, P. and Zinn-Justin, S. (2009) Structure of bacteriophage SPP1 head-to-tail connection reveals mechanism for viral DNA gating. *P Natl Acad Sci USA*, **106**, 8507-8512.

62. Zairi, M., Stiege, A.C., Nhiri, N., Jacquet, E. and Tavares, P. (2014) The Collagen-like Protein gp12 is a Temperature Dependent Reversible Binder of SPP1 Viral Capsids. *J Biol Chem*.

63. Cherezov, V., Rosenbaum, D.M., Hanson, M.A., Rasmussen, S.G.F., Thian, F.S., Kobilka, T.S., Choi, H.-J., Kuhn, P., Weis, W.I., Kobilka, B.K. and Stevens, R.C. (2007) High-Resolution Crystal Structure of an Engineered Human β2-Adrenergic G Protein–Coupled Receptor. *Science*, **318**, 1258-1265.

64. Rosenbaum, D.M., Cherezov, V., Hanson, M.A., Rasmussen, S.G.F., Thian, F.S., Kobilka, T.S., Choi, H.-J., Yao, X.-J., Weis, W.I., Stevens, R.C. and Kobilka, B.K. (2007) GPCR Engineering Yields High-Resolution Structural Insights into β2-Adrenergic Receptor Function. *Science*, **318**, 1266-1273.

65. Weaver, L.H. and Matthews, B.W. (1987) Structure of Bacteriophage-T4 Lysozyme Refined at 1.7 a Resolution. *J. Mol. Biol.*, **193**, 189-199.

66. Granier, S., Kim, S., Shafer, A.M., Ratnala, V.R.P., Fung, J.J., Zare, R.N. and Kobilka, B. (2007) Structure and conformational changes in the C-terminal domain of the beta(2)-adrenoceptor - Insights from fluorescence resonance energy transfer studies. *J Biol Chem*, **282**, 13895-13905.

67. Kobilka, B.K., Kobilka, T.S., Daniel, K., Regan, J.W., Caron, M.G. and Lefkowitz, R.J. (1988) Chimeric Alpha-2-Adrenergic, Beta-2-Adrenergic Receptors - Delineation of Domains Involved in Effector Coupling and Ligand-Binding Specificity. *Science*, **240**, 1310-1316.

68. Zhao, H., Finch, C.J., Sequeira, R.D., Johnson, B.A., Johnson, J.E., Casjens, S.R. and Tang, L. (2010) Crystal structure of the DNA-recognition component of the bacterial virus Sf6 genome-packaging machine. *P Natl Acad Sci USA*, **107**, 1971-1976.

69. Sun, S., Gao, S., Kondabagil, K., Xiang, Y., Rossmann, M.G. and Rao, V.B. (2012) Structure and function of the small terminase component of the DNA packaging machine in T4-like bacteriophages. *P Natl Acad Sci USA*, **109**, 817-822.

70. Zhang, F., Joachimiak, G., Collart, F., Joachimiak, A. (2005) The crystal structure of a Phage protein (phBC6A51) from Bacillus cereus ATCC 14579. PDB 2AO9.

71. Ren, B., Pham, T.M., Surjadi, R., Robinson, C.P., Le, T.K., Chandry, P.S., Peat, T.S. and McKinstry, W.J. (2013) Expression, purification, crystallization and preliminary X-ray diffraction analysis of a lactococcal bacteriophage small terminase subunit. *Acta crystallographica. Section F, Structural biology and crystallization communications*, **69**, 275-279.

72. Roy, A., Bhardwaj, A. and Cingolani, G. (2011) Crystallization of the nonameric small terminase subunit of bacteriophage P22. *Acta Crystallogr F*, **67**, 104-110.

73. Němeček, D., Lander, G.C., Johnson, J.E., Casjens, S.R. and Thomas Jr, G.J. (2008) Assembly Architecture and DNA Binding of the Bacteriophage P22 Terminase Small Subunit. *J. Mol. Biol.*, **383**, 494-501.

74. Xiong, Y. and Sundaralingam, M. (2001) Protein-nucleic acid interaction: major groove recognition determinants. *Encyclopedia of Life Sciences (Macmlllan Reference Ltd, London)*, 1-8.

75. Sinden, R.R. (1994) *DNA bending, DNA structure and function*. Elsevier Science.

76. Wu, H.M. and Crothers, D.M. (1984) The locus of sequence-directed and protein-induced DNA bending. *Nature*, **308**, 509-513.

77. Trifonov, E.N. and Sussman, J.L. (1980) The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc Natl Acad Sci U S A*, **77**, 3816-3820.

78. Ulanovsky, L., Bodner, M., Trifonov, E.N. and Choder, M. (1986) Curved DNA: design, synthesis, and circularization. *Proc Natl Acad Sci U S A*, **83**, 862-866.

79. Privalov, P.L., Dragan, A.I. and Crane-Robinson, C. (2009) The cost of DNA bending. *Trends Biochem Sci*, **34**, 464-470.

80. Becker, A. and Murialdo, H. (1990) Bacteriophage-Lambda DNA - the Beginning of the End. *J Bacteriol*, **172**, 2819-2824.

81. Sternberg, N. and Coulby, J. (1987) Recognition and Cleavage of the Bacteriophage-P1 Packaging Site (Pac) .2. Functional Limits of Pac and Location of Pac Cleavage Termini. *J. Mol. Biol.*, **194**, 469-479.

82. Wu, H., Sampson, L., Parr, R. and Casjens, S. (2002) The DNA site utilized by bacteriophage P22 for initiation of DNA packaging. *Mol Microbiol*, **45**, 1631-1646.

83. Kumari, S.C.S. (2012) *Application of Therapeutic Phages in Medicine, Bacteriophages*. Intech, Rijeka, Croatia.

84. Carla M. Carvalho, S.B.S., Andrew M. Kropinski, Eugénio C. Ferreira & Joana Azeredo (2012) *Phages as Therapeutic Tools to Control Major Foodborne Pathogens: Campylobacter and Salmonella,Bacteriophages*. Intech, Rijeka, Croatia.

85. Zhu, B.G., Cai, G.F., Hall, E.O. and Freeman, G.J. (2007) In-Fusion (TM) assembly: seamless engineering of multidomain fusion proteins, modular vectors, and mutations. *Biotechniques*, **43**, 356-359.

86. Hamilton, M.D., Nuara, A.A., Gammon, D.B., Buller, R.M. and Evans, D.H. (2007) Duplex strand joining reactions catalyzed by vaccinia virus DNA polymerase. *Nucleic Acids Res*, **35**, 143-151.

87. Lopez, P.J., Marchand, I., Joyce, S.A. and Dreyfus, M. (1999) The C-terminal half of RNase E, which organizes the Escherichia coli degradosome, participates in mRNA degradation but not rRNA processing in vivo. *Mol Microbiol*, **33**, 188-199.

88. Studier, F.W. (1991) Use of Bacteriophage-T7 Lysozyme to Improve an Inducible T7 Expression System. *J. Mol. Biol.*, **219**, 37-44.

89. Studier, F.W. and Moffatt, B.A. (1986) Use of Bacteriophage-T7 Rna-Polymerase to Direct Selective High-Level Expression of Cloned Genes. *J. Mol. Biol.*, **189**, 113-130.

90. Bornhorst, J.A. and Falke, J.J. (2000) Purification of proteins using polyhistidine affinity tags. *Method Enzymol*, **326**, 245-254.

91. Porath, J., Carlsson, J., Olsson, I. and Belfrage, G. (1975) Metal Chelate Affinity Chromatography, a New Approach to Protein Fractionation. *Nature*, **258**, 598-599.

92. AB. (2010) Affinity Chromatography. Principles and Methods. Amersham Biosciences Ed.

93. Oganesyan, N., Kim, S.-H. and Kim, R. (2005) SOn-column Protein Refolding for Crystallization. *Journal of Structural and Functional Genomics*, **6**, 177-182.

94. Luft, J.R., Wolfley, J.R. and Snell, E.H. (2011) What's in a Drop? Correlating Observations and Outcomes to Guide Macromolecular Crystallization Experiments. *Cryst Growth Des*, **11**, 651-663.

95. HR. (2005) Crystallization. Research Tools. Hampton Research ed.

96. Bergfors, T. (2003) Seeds to crystals. *J Struct Biol*, **142**, 66-76.

97. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*, **28**, 235-242.

98. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res,* **25**, 3389-3402.

99. Cole, C., Barber, J.D. and Barton, G.J. (2008) The Jpred 3 secondary structure prediction server. *Nucleic Acids Res*, **36**, W197-W201.

100. Kabsch, W. and Sander, C. (1983) Dictionary of Protein Secondary Structure - Pattern-Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers*, **22**, 2577-2637.

101. Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195-202.

102. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. and Higgins, D.G. (2007) Clustal W and clustal X version 2.0. *Bioinformatics*, **23**, 2947-2948.

103. Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J. and Russell, R.B. (2003) Protein disorder prediction: Implications for structural proteomics. *Structure,* **11**, 1453-1459.

104. Stivala, A., Wybrow, M., Wirth, A., Whisstock, J.C. and Stuckey, P.J. (2011) Automatic generation of protein structure cartoons with Pro-origami. *Bioinformatics*, **27**, 3315-3316.

105. Krissinel, E. and Henrick, K. (2007) Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.*, **372**, 774-797.

106. Krissinel, E. and Henrick, K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D,* **60**, 2256-2268.

107. Kabsch, W. (2010) XDS. *Acta Crystallographica Section D,* **66**, 125-132.

108. Evans, P.R. (2011) An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta crystallographica. Section D, Biological crystallography*, **67**, 282-292.

109. Evans, P.R. and Murshudov, G.N. (2013) How good are my data and what is the resolution? *Acta Crystallogr D,* **69**, 1204-1214.

110. Sheldrick, G.M. (2010) Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. *Acta Crystallogr D*, **66**, 479-485.

111. McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C. and Read, R.J. (2007) Phaser crystallographic software. *J Appl Crystallogr,* **40**, 658-674.

112. Read, R.J. and McCoy, A.J. (2011) Using SAD data in Phaser. *Acta Crystallogr D,* **67**, 338-344.

113. Cowtan, K. (2010) Recent developments in classical density modification. *Acta crystallographica. Section D, Biological crystallography*, **66**, 470-478.

114. Cowtan, K. (2006) The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallographica Section D*, **62**, 1002-1011.

115. Murshudov, G.N., Skubak, P., Lebedev, A.A., Pannu, N.S., Steiner, R.A., Nicholls, R.A., Winn, M.D., Long, F. and Vagin, A.A. (2011) REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D,* **67**, 355-367.

116. Emsley, P. and Cowtan, K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr D,* **60**, 2126-2132.

117. Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G.W., McCoy, A., McNicholas, S.J., Murshudov, G.N., Pannu, N.S., Potterton, E.A., Powell, H.R., Read, R.J., Vagin, A. and Wilson, K.S. (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr D,* **67**, 235-242.

118. Shevtsov, M.B., Chen, Y.L., Gollnick, P. and Antson, A.A. (2005) Crystal structure of Bacillus subtilis anti-TRAP protein, an antagonist of TRAP/RNA interaction. *P Natl Acad Sci USA*, **102**, 17600-17605.

119. Matthews, B.W. (1968) Solvent Content of Protein Crystals. *J. Mol. Biol.*, **33**, 491-&.

120. Qin, L., Fokine, A., O'Donnell, E., Rao, V.B. and Rossmann, M.G. (2010) Structure of the Small Outer Capsid Protein, Soc: A Clamp for Stabilizing Capsids of T4-like Phages. *J. Mol. Biol.*, **395**, 728-741.

121. Veesler, D. and Cambillau, C. (2011) A Common Evolutionary Origin for Tailed-Bacteriophage Functional Modules and Bacterial Machineries. *Microbiol Mol Biol R*, **75**, 423-433.

122. Zhou, Z.H., Dougherty, M., Jakana, J., He, J., Rixon, F.J. and Chiu, W. (2000) Seeing the herpesvirus capsid at 8.5 angstrom. *Science*, **288**, 877-880.

123. Baker, M.L., Jiang, W., Rixon, F.J. and Chiu, W. (2005) Common ancestry of herpesviruses and tailed DNA bacteriophages. *J Virol*, **79**, 14967-14970.

124. Bowman, B.R., Baker, M.L., Rixon, F.J., Chiu, W. and Quiocho, F.A. (2003) Structure of the herpesvirus major capsid protein. *Embo J*, **22**, 757-765.

125. Hendrix, R.W. (1999) Evolution: The long evolutionary reach of viruses. *Curr Biol*, **9**, R914-R917.

126. Benson, S.D., Bamford, J.K.H., Bamford, D.H. and Burnett, R.M. (2002) The X-ray crystal structure of P3, the major coat protein of the lipid-containing bacteriophage PRD1, at 1.65 angstrom resolution. *Acta Crystallogr D*, **58**, 39-59.

127. Grimes, J.M., Burroughs, J.N., Gouet, P., Diprose, J.M., Malby, R., Zientara, S., Mertens, P.P.C. and Stuart, D.I. (1998) The atomic structure of the bluetongue virus core. *Nature*, **395**, 470-478.

128. Choi, H.K., Lee, S., Zhang, Y.P., McKinney, B.R., Wengler, G., Rossmann, M.G. and Kuhn, R.J. (1996) Structural analysis of Sindbis virus capsid mutants involving assembly and catalysis. *J. Mol. Biol.*, **262**, 151-167.

129. Valegard, K., Liljas, L., Fridborg, K. and Unge, T. (1990) The 3-Dimensional Structure of the Bacterial-Virus Ms2. *Nature*, **345**, 36-41.

130. Golmohammadi, R., Valegard, K., Fridborg, K. and Liljas, L. (1993) The Refined Structure of Bacteriophage-Ms2 at 2-Center-Dot-8-Angstrom Resolution. *J. Mol. Biol.*, **234**, 620-639.

131. Benson, S.D., Bamford, J.K.H., Bamford, D.H. and Burnett, R.M. (2004) Does common architecture reveal a viral lineage spanning all three domains of life? *Molecular Cell*, **16**, 673-685.

132. Hendrix, R.W., Smith, M.C.M., Burns, R.N., Ford, M.E. and Hatfull, G.F. (1999) Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. *P Natl Acad Sci USA*, **96**, 2192-2197.

133. Abrescia, N.G.A., Bamford, D.H., Grimes, J.M. and Stuart, D.I. (2012) Structure Unifies the Viral Universe. *Annual Review of Biochemistry, Vol 81*, **81**, 795-822.

134. Loredo-Varela, J., Chechik, M., Levdikov, V.M., Abd-El-Aziz, A., Minakhin, L., Severinov, K., Smits, C. and Antson, A.A. (2013) The putative small terminase from the thermophilic dsDNA bacteriophage G20C is a nine-subunit oligomer. *Acta Crystallographica Section F*, **69**, 876-879.

135. Benini, S., Chechik, M., Lombardia, M.O., Polier, S., Leech, A., Shevtsov, M.B. and Alonso, J.C. (2013) The 1.58 angstrom resolution structure of the DNA-binding domain of bacteriophage SF6 small terminase provides new hints on DNA binding. *Acta Crystallogr F*, **69**, 376-381.

136. Zhao, H., Kamau, Y.N., Christensen, T.E. and Tang, L. (2012) Structural and functional studies of the phage Sf6 terminase small subunit reveal a DNA-spooling device facilitated by structural plasticity. *J Mol Biol*, **423**, 413-426.

137. Brennan, R.G. and Matthews, B.W. (1989) The Helix-Turn-Helix DNA-Binding Motif. *J Biol Chem*, **264**, 1903-1906.

138. Dyson, H.J. and Wright, P.E. (2005) Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Bio*, **6**, 197-208.

139. Radhakrishnan, I., Perez-Alvarado, G.C., Parker, D., Dyson, H.J., Montminy, M.R. and Wright, P.E. (1997) Solution structure of the KIX domain of CBP bound to the transactivation domain of CREB: a model for activator:coactivator interactions. *Cell*, **91**, 741-752.

140. Shaywitz, A.J. and Greenberg, M.E. (1999) CREB: A stimulus-induced transcription factor activated by a diverse array of extracellular signals. *Annu Rev Biochem*, **68**, 821-861.

141. Crane-Robinson, C., Dragan, A.I. and Privalov, P.L. (2006) The extended arms of DNA-binding domains: a tale of tails. *Trends Biochem Sci*, **31**, 547-552.

142. Fraenkel, E. and Pabo, C.O. (1998) Comparison of X-ray and NMR structures for the Antennapedia homeodomain-DNA complex. *Nat Struct Biol*, **5**, 692-697.

143. Laughon, A., Boulet, A.M., Bermingham, J.R., Jr., Laymon, R.A. and Scott, M.P. (1986) Structure of transcripts from the homeotic Antennapedia gene of Drosophila melanogaster: two promoters control the major protein-coding region. *Molecular and cellular biology*, **6**, 4676-4689.

144. Casjens, S.R. and Thuman-Commike, P.A. (2011) Evolution of mosaically related tailed bacteriophage genomes seen through the lens of phage P22 virion assembly. *Virology*, **411**, 393-415.

145. Court, R., Chapman, L., Fairall, L. and Rhodes, D. (2005) How the human telomeric proteins TRF1 and TRF2 recognize telomeric DNA: a view from high-resolution crystal structures. *Embo Rep*, **6**, 39-45.

146. Lara-Gonzalez, S., Birktoft, J. J., Lawson, C.L. RNA polymerase alpha C-terminal domain (E. coli) and sigma region 4 (T. aq. mutant) bound to (UP,-35 element) DNA. PDB 3N97.

147. Ortega, M.E. and Catalano, C.E. (2006) Bacteriophage lambda gpNu 1 and Escherichia coli IHF proteins cooperatively bind and bend viral DNA: Implications for the assembly of a genome-packaging motor. *Biochemistry-Us*, **45**, 5180-5189.

148. Gual, A. and Alonso, J.C. (1998) Characterization of the small subunit of the terminase enzyme of the Bacillus subtilis bacteriophage SPP1. *Virology*, **242**, 279-287.

149. Camacho, A.G., Gual, A., Lurz, R., Tavares, P. and Alonso, J.C. (2003) Bacillus subtilis bacteriophage SPP1 DNA packaging motor requires terminase and portal proteins. *J Biol Chem*, **278**, 23251-23259.

150. Rao, V.B. and Black, L.W. (2010) Structure and assembly of bacteriophage T4 head. *Virology journal*, **7**.

151. Roy, A. and Cingolani, G. (2012) Structure of P22 Headful Packaging Nuclease. *J Biol Chem*, **287**, 28196-28205.

152. Zhao, H.Y., Christensen, T.E., Kamau, Y.N. and Tang, L. (2013) Structures of the phage Sf6 large terminase provide new insights into DNA translocation and cleavage. *P Natl Acad Sci USA*, **110**, 8075-8080.

153. Rao, V.B. and Black, L.W. (1988) Cloning, Overexpression and Purification of the Terminase Proteins Gp16 and Gp17 of Bacteriophage-T4 - Construction of a Defined Invitro DNA Packaging System Using Purified Terminase Proteins. *J. Mol. Biol.*, **200**, 475-488.

154. Kondabagil, K.R., Zhang, Z.H. and Rao, V.B. (2006) The DNA translocating ATPase of bacteriophage T4 packaging motor. *J. Mol. Biol.*, **363**, 486-499.

155. Poteete, A.R. and Botstein, D. (1979) Purification and Properties of Proteins Essential to DNA Encapsulation by Phage-P22. *Virology*, **95**, 565-573.

156. Maluf, N.K., Yang, Q. and Catalano, C.E. (2005) Self-association properties of the bacteriophage lambda terminase holoenzyme: Implications for the DNA packaging motor. *J. Mol. Biol.*, **347**, 523-542.

157. Chakravarty, S., Luger, K. (2009) Drosophila nucleosome structure. PDB 2NQB.