Structure and Function of Podovirus Sf6 Tail Complex

By

Lingfei Liang

B.E. Bioengineering, Beijing Institute of Technology, 2013

Submitted to the Department of Molecular Biosciences

and the Graduate Faculty of the University of Kansas

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

_____

Chair: Susan Egan

_____

Roberto N. De Guzman

_____

David Davido

_____

Mizuki Azuma

_____

Chris Fischer

Date Defended: 4/20/2018

The dissertation committee for Lingfei Liang certifies that this is the approved version of the following dissertation:

Structure and Function of Podovirus Sf6 Tail Complex

_____

Chair: Susan Egan

Date approved: 4/30/2018

Abstract

Sf6 is a double-stranded DNA (dsDNA) bacteriophage with a short, non-contractile tail. The tail is a sophisticated molecular machine made of 39 copies of four gene products, including the dodecameric tail adaptor gp7, the hexameric tail nozzle gp8, the trimeric tail needle gp9 and 6 copies of the trimeric tail spike gp14. It has been shown that the tail assembly occurs in a sequential manner. Here we report the high-resolution structure of the Sf6 tail adaptor protein gp7. Comparative structural studies reveal that during tail assembly the gp7 N-terminus undergoes structural rearrangement by repositioning two consecutive repeats of a conserved octad sequence motif, turning the molecule from the preassembly state to the postassembly state, which creates the binding site for the next tail component to attach to. These results provide a structural basis for a mechanism of sequence motifs repositioning by which the adaptor protein mediates the sequential assembly of the phage tail.

Tail nozzle gp8 is the following component attached to gp7 in the tail. It is highly conserved between Sf6 and P22, but the structure is not known yet. Here, we did Small-Angle X-ray Scattering (SAXS) analysis on gp8 monomer, showing a brick-shaped, globular protein with a small protrusion. Fitting of the SAXS model into the electron cryo-microscopy (cryoEM) map of the entire tail machine has aided in defining molecular boundaries between gp8 monomers and neighboring subunits of other tail components.

One of the important functions of the tail is to deliver viral DNA through host envelope to establish infection. Given the fact that the tail is too short to directly span bacterial envelope, it is assumed that during infection the short tail is extended by three DNA-injection proteins, gp11, gp12, and gp13, to drill through the three-layer envelope of the host cell to inject phage DNA into the host cytoplasm. We achieved the 3D EM reconstruction of gp12 decamer, revealing a tube-like assembly with a constricted channel presumably for dsDNA delivery. We then solved the X-ray structure of the gp12 N-terminal domain (gp12NTD) which, surprisingly, assembles into a undecamer in crystals. This 2.8Å gp12NTD structure represents the first high-resolution structure of tailed virus DNA-injection proteins. The gp12NTD molecule consists of eight α-helices, seven of which forms two helix bundles. Biochemical study suggests that the helix α8 is dispensable

for gp12 homo-oligomerization. Analysis on the tertiary structure and the locations of Gly and Pro residues, for the first time, provides experimental foundation for the assumption that internal proteins are partially unfolded when travelling through the narrow tail channel. We also show that P22-gp20 (Sf6-gp12 ortholog) NTD has a highly similar structure and it also assembles to a undecamer. By analyzing the structure characteristics and the conserved features between Sf6-gp12 and P22-gp20, we discussed the possible scenario of gp12/gp20 travelling through tail channel. The gp12CTD is monomeric in solution, and the C-terminal 27 residues are essential for gp12:gp13 interaction. The stoichiometry of gp12 and gp13 is likely to be 1:1. Similarly, gp20 binds with gp16 (counterpart of gp13). Our work sheds light on the roles of the two DNA-injection proteins (Sf6-gp12/P22-gp20 and Sf6-gp13/P22-gp16) in assembly of the extended tail for DNA delivery.

A high-resolution X-ray structure of the Non-Structural protein 1 N-terminal domain (NS1N) of Minute Virus of Mice (MVM) is also reported here as an effort to study DNA replication in parvovirus. MVM has a single-stranded DNA (ssDNA) genome with the two ends folding back to from double-stranded hetero-telomeres, providing origin of replication (Ori). NS1N binds to Ori to perform a series of functions including ssDNA nicking. The NS1N structure here shows potential sites for dsDNA binding, ssDNA binding and cleavage on a canonical fold of the histidine-hydrophobic-histidine superfamily of nucleases. Metal derivative crystal structures reveal the nickase active site with an architecture that allows highly versatile metal ligand binding. The structures support a unified mechanism of replication origin recognition for homotelomeric and heterotelomeric parvoviruses, mediated by a basic-residue-rich hairpin and an adjacent helix in the initiator proteins and by tandem tetranucleotide motifs in the replication origins.

Acknowledgements

This work would not have been possible without my previous mentor Dr. Liang Tang. He is a fantastic mentor who taught me far more than I could express both in science and in life. He has been guiding and supporting me for these five years. The knowledge I learned, the technique I built, and the work I did in Tang's lab prepares me well for my next stage of life.

I'm also grateful to Dr. Susan Egan for being willing to continue to mentor me and help me graduate. Susan is in my Committee and has been very supportive and helpful. She always offers professional ways to help me solve problems or, at least, tell me which direction I should go or which professor I can turn to.

I thank all the members of Tang lab and Egan lab that I have had the pleasure to work with during my time in MB. I am especially indebted to Dr. Haiyan Zhao who has taught me and trained me in the lab to grow into a real scientist.

All members of my dissertation committee, Dr. Susan Egan, Dr. Roberto De Guzman, Dr. David Davido, Dr. Mizuki Azuma, and Dr. Chris Fischer, have been very helpful throughout my graduate career. They listen to what I want to do; give suggestions; solve my confusions. I always feel very comfortable to turn to them for advices. I also want to thank the department and members of MB staff, especially John Connolly, Judi Harris, Linda Wiley, and Cynthia Rodriguez. Without their support, I would not be able to finish my graduate career so successfully.

Nobody has been more important to me in the pursuit of this degree in the US, away from home, than my family. My parents are the ones who I always share my happiness and progress with even though they know little about my research. Last but not least, I wish to give my most thank to my husband Congying. He is the one who will listen to my complaint or frustration or whatever at any time. He is the one who will encourage me and stand with me all the time.

It is a wonderful experience to study here at KU! This is the unique five years in my life!

Table of Contents

Chapter 1. Introduction

1.1 Introduction to podovirus Sf6

All viruses rely on cells for reproduction and all three domains of life can be infected by viruses. The infection process falls into several stages in general: attachment, entry, replication, assembly, and exit. Due to the huge diversity of viruses, these stages own greatly different details and mechanisms.

In the first part of this study, we explored the viral tail assembly and the genome entry process of a short-tailed double stranded DNA (dsDNA) bacteriophage Sf6. Tailed dsDNA bacteriophages constitute the order of *Caudovirales*. They are the most abundant biological entity in earth's biosphere, accounting for 96% of all known phages (1). Three families are defined by tail morphology. They are long-contractile tailed *Myoviridae*, long-noncontractile tailed *Siphoviridae*, and short tailed *Podoviridae*. The tail attached to the capsid is a unique, sophisticated macromolecular machine responsible for a variety of roles essential in viral life cycle, such as host cell recognition and viral DNA injection. Phage Sf6 is a P22-related podovirus (2) infecting *Shigella flexneri*. Like other P22-like phages, Sf6 possesses a short, non-contractile tail emanating from a unique pentameric vertex of the icosahedral capsid (3, 4). The tail is around 1.8 MDa comprised of 39 copies of four gene products, including the dodecameric tail adaptor gp7, the hexameric tail nozzle gp8, the trimeric tail needle gp9 and 6 copies of the trimeric tail spike gp14 (**Fig. 1-1, Table 1-1**). Structures and functions of the tail needle and tail spike proteins in Sf6 and P22 have been well characterized (5-8). The structure of tail adaptor in P22 was solved in complex with the portal protein (9). However, the structures of Sf6 tail adaptor and tail nozzles from both viruses are not known yet. P22 tail machine has been proved to be assembled in a sequential order (10). The successful assembly of the former tail protein onto virus particle is required for the assembly of the latter tail component (**Fig. 1-2**). Such sequential assembly behavior is also observed in families of *Siphoviridae* and *Myoviridae* when they assemble their tails (11). The molecular mechanism that mediates these sequential assembly remains to be determined.

Infection of Sf6 on *Shigella* cells is initiated by binding of the tail spikes to the primary receptor O

antigen on the cell surface (**Fig. 1-3**). The tail spike has an enzymatic activity that cleaves the polysaccharide receptors, pulling the viron down to the outer membrane surface. A second receptor, OmpA or OmpC, is then recognized most likely by the tail needle (3, 12, 13). This irreversible interaction triggers the release of tail needle and the injection of viral DNA. Compared with the long tailed caudoviruses, podoviruses have different mechanisms for genome delivery due to the limited length of the tail, meaning that they cannot directly traverse the three-layer cell envelope to inject DNA. It has been assumed that podoviral tail machines release internal proteins/DNA-injection proteins into cell envelope as a tail extension to drill through the three-layer envelope of the host cell and deliver phage DNA into the host cytoplasm (10, 14-19). The internal proteins of Sf6 and P22 are poorly understood since they are not well defined in the cryoEM reconstruction. The best scenario to characterize the internal proteins is probably in the tail extension where they assemble into a well-defined structure.

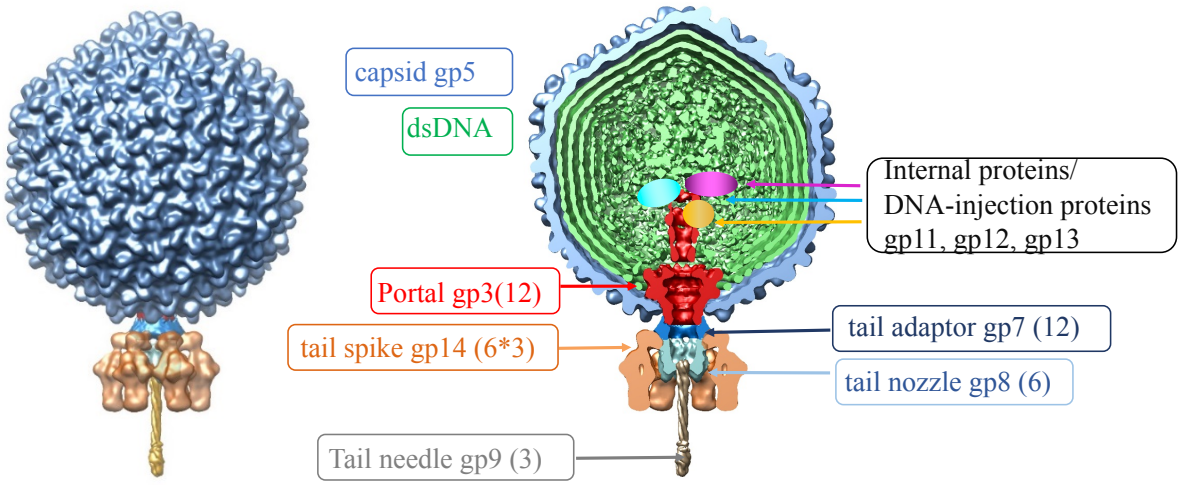


Fig. 1-1. The structure of P22 and Sf6 virion. The front view (left) and the cross-section (right) of the P22 virion cryoEM reconstruction (EMDB 1220) is segmented to highlight individual components, which are labeled with the name and copy number of Sf6 gene products.

Table 1-1. Sf6 and P22 virion assembly proteins

	Sf6 Protein	P22 ortholog	Function	Copy number in virion	% Identity
Capsid	gp5	gp5	Coat protein	415	14
	gp11	gp7	DNA injection	?	82
	gp12	gp20	DNA injection	?	33
	gp13	gp16	DNA injection	?	32
	gp3	gp1	Portal	12	30
Tail	gp7	gp4	Tail adaptor	12	35
	gp8	gp10	Tail nozzle	6	93
	gp9	gp26	Tail needle	3	35
	gp14	gp9	Tail spike	6x3	25

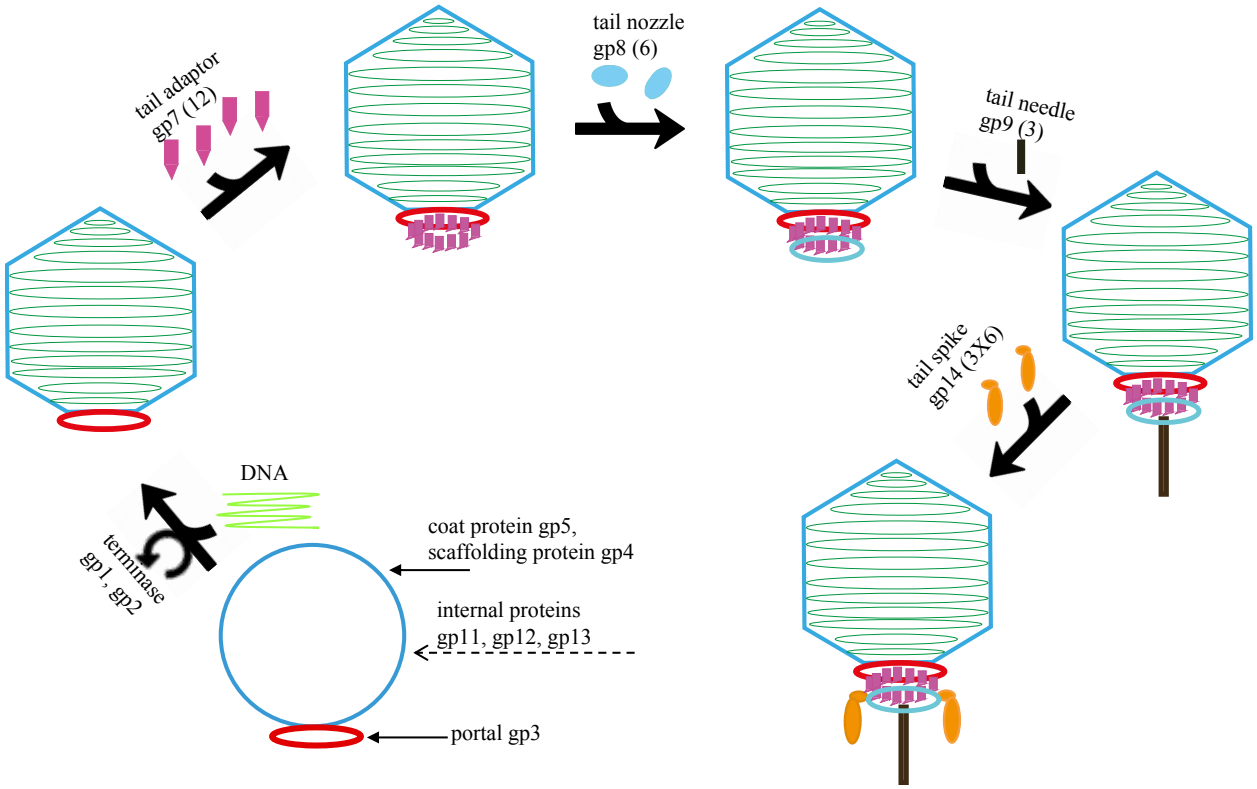


Fig. 1-2. Schematic diagram of Sf6 tail assembly. The tail adaptor gp7, tail nozzle gp8, tail needle gp9, and tail spike gp14 assembles onto the particle in a sequential manner.

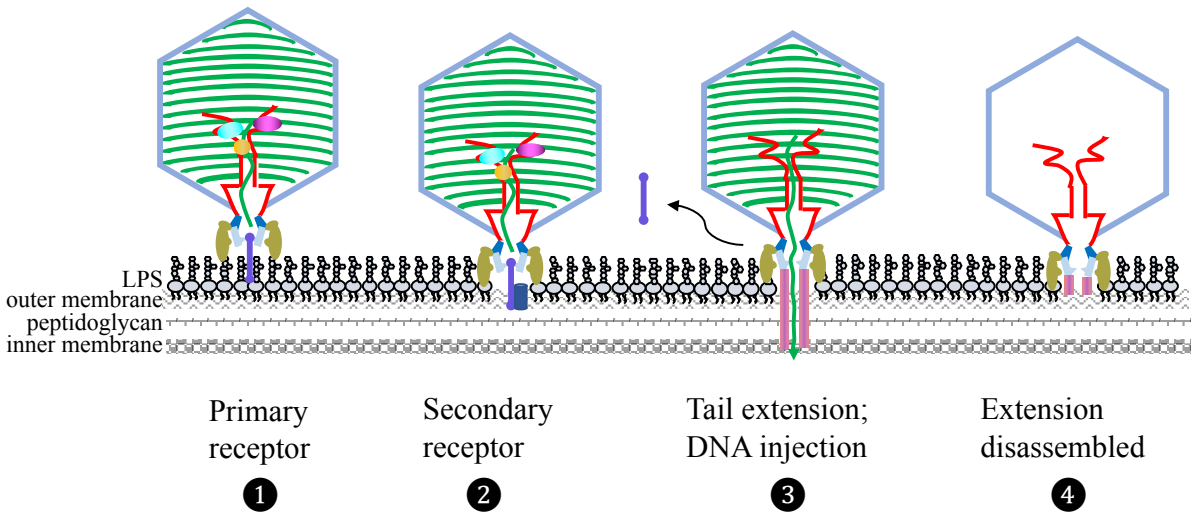


Fig. 1-3. Schematic diagram of Sf6 DNA delivery across a Gram-negative bacterial cell envelope.

1.2 Introduction to parvovirus MVM

Adapted from paper 'Tewary SK, *et al.* (2015) Structures of minute virus of mice replication initiator protein N-terminal domain: Insights into DNA nicking and origin binding. *Virology* 476:61-71.'

DNA replication is generally the next event after viral materials enter a host cell. In the second part of this study, we investigated the key role of the Non-Structural protein 1 (NS1) in Minute Virus of Mice (MVM) DNA replication.

MVM belongs to the genus *Protoparvovirus* in the *Parvoviridae,* a family of small isometric viruses containing a linear single-stranded DNA (ssDNA) genome (20). The best characterized strains are MVMp, a "prototype" strain that infects cells of fibroblast origin, and MVMi, an "immunosuppressive" strain that productively infects T lymphocytes in culture, but also shows specificity for endothelium and hepatic erythropoietic precursors when infecting neonatal mice (21). Like many rodent protoparvoviruses, in human cells MVM is oncoselective and oncolytic, infecting tumor cell lines while being non-pathogenic for normal cells (22), providing interesting therapeutic potential.

The MVM virion encapsidates a ssDNA genome of approximately 5 kb in a small protein capsid

of ~280Å in diameter with *T=1* icosahedral symmetry (23, 24). This genome contains two overlapping transcription units with P4 and P38 promoters (**Fig. 1-4**). Alternatively spliced mRNAs transcribed from the P4 promoter encode two major non-structural proteins, NS1 and NS2, that share a common 85 amino acid N-terminal domain (25). The viral DNA replication strategy, dubbed rolling hairpin replication (RHR) (26, 27), is a linear adaptation of the more widely employed rolling circle replication (RCR) mechanism (28) (**Fig. 1-5**). NS1 protein plays pivotal roles in initiating and directing viral DNA replication, as well as in viral DNA packaging and transcriptional activation of the viral promoters (24). Its N-terminal domain, the subject of the current study, contains overlapping site-specific double-stranded DNA (dsDNA) binding, ssDNA recognition, and origin-specific ssDNA nicking functions (29, 30) (**Fig. 1-6**). The linear ssDNA genome of MVM is flanked by short palindromic sequences that can fold into duplex hairpin telomeres, serving as an Origin of Replication (31). The two hairpins of MVM genome are disparate, differing in sequences and functions, providing two Origin of Replications, OriL and OriR. Within these origin sequences, NS1 binds site-specifically to 2-3 duplex reiterations of the tetranucleotide 5'-TGGT-3' (29, 30, 32). However, binding alone does not activate NS1's nicking function, which rather requires the cooperation of origin-specific cellular co-factors that use different mechanisms to further stabilize and orient NS1 in the nuclease complex (33-35). These allow NS1 to unwind proximal dsDNA, likely in an ATP-dependent manner, generating a region of ssDNA that encompasses the resolution site, which is subsequently nicked by the NS1 nuclease activity (36-40).

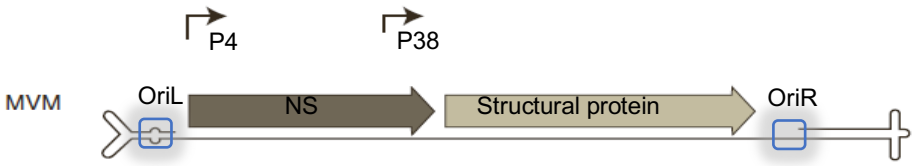


Fig. 1-4. The diagram of MVM genome. The terminal hairpin structures are magnified approximately 10-fold with respect to the intervening single-stranded region. The two promoters are indicated by arrows, and their major gene blocks by dark and light brown arrows indicating the N- to C-terminal direction. The two Ori are indicated by blue rectangles. (Adapted form Susan Cotmore and Peter Tattersall, 2013)

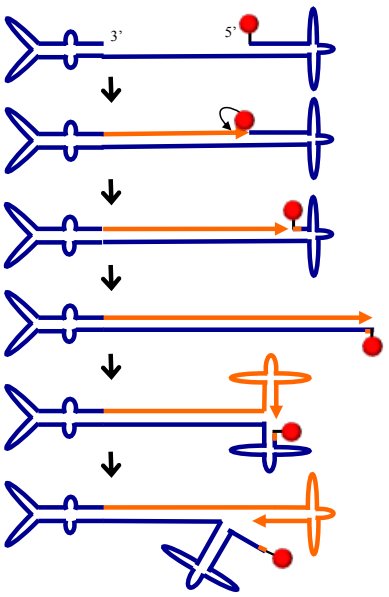


Fig. 1-5. MVM 'rolling hairpin replication' scheme. The genome is represented by a continuous line, blue for the original genome, orange for the progeny genome. The red sphere represents a NS1 molecule.

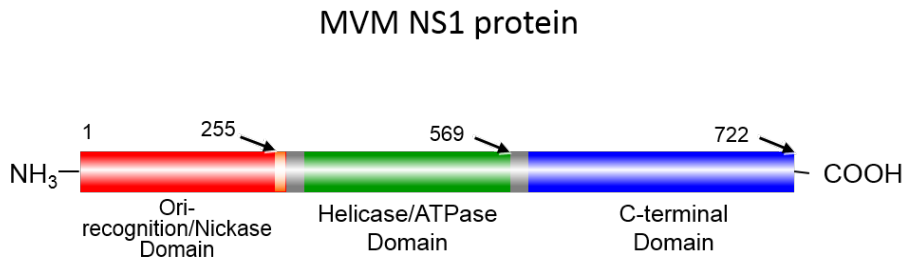


Fig. 1-6. Domain composition of MVM NS1. MVM NS1 is composed of three domains. The N-terminal domain plays a role in Ori-recognition and nickase activity. The middle domain has helicase and ATPase activity. The C-terminal domain is involved in recruiting cellular factors.

1.3 References

1. Fokine A & Rossmann MG (2014) Molecular architecture of tailed double-stranded DNA phages. *Bacteriophage* 4(1):e28281.
2. Casjens S, *et al.* (2004) The chromosome of Shigella flexneri bacteriophage Sf6: complete nucleotide sequence, genetic mosaicism, and DNA packaging. *Journal of molecular biology* 339(2):379-394.
3. Zhao H, Sequeira RD, Galeva NA, & Tang L (2011) The host outer membrane proteins OmpA and OmpC are associated with the Shigella phage Sf6 virion. *Virology* 409(2):319-327.
4. Parent KN, Gilcrease EB, Casjens SR, & Baker TS (2012) Structural evolution of the P22-like phages: comparison of Sf6 and P22 procapsid and virion architectures. *Virology* 427(2):177-188.
5. Bhardwaj A, Molineux IJ, Casjens SR, & Cingolani G (2011) Atomic structure of bacteriophage Sf6 tail needle knob. *The Journal of biological chemistry* 286(35):30867-30877.

6. Steinbacher S*, et al.* (1996) Crystal structure of phage P22 tailspike protein complexed with Salmonella sp. O-antigen receptors. *Proceedings of the National Academy of Sciences of the United States of America* 93(20):10584-10588.
7. Muller JJ*, et al.* (2008) An intersubunit active site between supercoiled parallel beta helices in the trimeric tailspike endorhamnosidase of Shigella flexneri Phage Sf6. *Structure* 16(5):766-775.
8. Olia AS, Casjens S, & Cingolani G (2007) Structure of phage P22 cell envelope-penetrating needle. *Nature structural & molecular biology* 14(12):1221-1226.
9. Olia AS, Prevelige PE, Jr., Johnson JE, & Cingolani G (2011) Three-dimensional structure of a viral genome-delivery portal vertex. *Nature structural & molecular biology* 18(5):597-603.
10. Israel V (1977) E proteins of bacteriophage P22. I. Identification and ejection from wild-type and defective particles. *Journal of virology* 23(1):91-97.
11. Yap ML, Mio K, Leiman PG, Kanamaru S, & Arisaka F (2010) The baseplate wedges of bacteriophage T4 spontaneously assemble into hubless baseplate-like structure in vitro. *Journal of molecular biology* 395(2):349-360.
12. Parent KN*, et al.* (2014) OmpA and OmpC are critical host factors for bacteriophage Sf6 entry in Shigella. *Mol Microbiol* 92(1):47-60.
13. Porcek NB & Parent KN (2015) Key residues of S. flexneri OmpA mediate infection by bacteriophage Sf6. *Journal of molecular biology* 427(10):1964-1976.
14. Perez GL, Huynh B, Slater M, & Maloy S (2009) Transport of phage P22 DNA across the cytoplasmic membrane. *Journal of bacteriology* 191(1):135-140.
15. Serwer P, Wright ET, Hakala KW, & Weintraub ST (2008) Evidence for bacteriophage T7 tail extension during DNA injection. *BMC research notes* 1:36.
16. Hu B, Margolin W, Molineux IJ, & Liu J (2013) The bacteriophage t7 virion undergoes extensive structural remodeling during infection. *Science* 339(6119):576-579.
17. Zhao H*, et al.* (2016) Structure of a Bacterial Virus DNA-Injection Protein Complex Reveals a Decameric Assembly with a Constricted Molecular Channel. *PloS one* 11(2):e0149337.
18. Casjens SR & Molineux IJ (2012) Short noncontractile tail machines: adsorption and DNA delivery by podoviruses. *Advances in experimental medicine and biology* 726:143-179.
19. Molineux IJ & Panja D (2013) Popping the cork: mechanisms of phage genome ejection. *Nature reviews. Microbiology* 11(3):194-204.
20. Cotmore SF*, et al.* (2014) The family Parvoviridae. *Archives of virology* 159(5):1239-1247.
21. Brownstein DG*, et al.* (1992) The pathogenesis of infection with minute virus of mice depends on expression of the small nonstructural protein NS2 and on the genotype of the allotropic determinants VP1 and VP2. *J Virol* 66(5):3118-3124.
22. Dupont F (2003) Risk assessment of the use of autonomous parvovirus-based vectors. *Current gene therapy* 3(6):567-582.
23. Agbandje-McKenna M, Llamas-Saiz AL, Wang F, Tattersall P, & Rossmann MG (1998) Functional implications of the structure of the murine parvovirus, minute virus of mice. *Structure* 6(11):1369-1381.
24. Cotmore SF & Tattersall P (2014) Parvoviruses: Small Does Not Mean Simple. *Annual Review of Virology* 1(1):null.
25. Pintel D, Dadachanji D, Astell CR, & Ward DC (1983) The genome of minute virus of mice, an autonomous parvovirus, encodes two overlapping transcription units. *Nucleic acids research* 11(4):1019-1038.
26. Cotmore SF & Tattersall P (2013) Parvovirus diversity and DNA damage responses. *Cold Spring Harbor perspectives in biology* 5(2).
27. Cotmore SF & Tattersall P (2005) A rolling-hairpin strategy: basic mechanisms of DNA replication in the parvoviruses. *Parvoviruses*, ed J. Kerr, S.F. Cotmore, M.E. Bloom, R.M. Linden and C.R. Parrish (Hodder Arnold, London, United kingdom), pp 171-181.
28. Kornberg A & Baker TA (1992) *DNA replication* (W.H. Freeman, New York) 2nd Ed pp xiv, 931 p.

29. Mouw M & Pintel DJ (1998) Amino acids 16-275 of minute virus of mice NS1 include a domain that specifically binds (ACCA)2-3-containing DNA. *Virology* 251(1):123-131.
30. Cotmore SF, Christensen J, Nuesch JP, & Tattersall P (1995) The NS1 polypeptide of the murine parvovirus minute virus of mice binds to DNA sequences containing the motif [ACCA]2-3. *Journal of virology* 69(3):1652-1660.
31. Kerr JR (2006) *Parvoviruses* (Hodder Arnold ; Distributed in the United States of America by Oxford University Press, London New York) pp xxviii, 598 p. , 524 p. of plates.
32. Cotmore SF, Gottlieb RL, & Tattersall P (2007) Replication initiator protein NS1 of the parvovirus minute virus of mice binds to modular divergent sites distributed throughout duplex viral DNA. *J Virol* 81(23):13015-13027.
33. Christensen J, Cotmore SF, & Tattersall P (2001) Minute virus of mice initiator protein NS1 and a host KDWK family transcription factor must form a precise ternary complex with origin DNA for nicking to occur. *Journal of virology* 75(15):7009-7017.
34. Cotmore SF, Christensen J, & Tattersall P (2000) Two widely spaced initiator binding sites create an HMG1-dependent parvovirus rolling-hairpin replication origin. *Journal of virology* 74(3):1332-1341.
35. Christensen J, Cotmore SF, & Tattersall P (1999) Two new members of the emerging KDWK family of combinatorial transcription modulators bind as a heterodimer to flexibly spaced PuCGPy half-sites. *Molecular and cellular biology* 19(11):7741-7750.
36. Nuesch JP, Cotmore SF, & Tattersall P (1995) Sequence motifs in the replicator protein of parvovirus MVM essential for nicking and covalent attachment to the viral origin: identification of the linking tyrosine. *Virology* 209(1):122-135.
37. Willwand K*, et al.* (1997) The minute virus of mice (MVM) nonstructural protein NS1 induces nicking of MVM DNA at a unique site of the right-end telomere in both hairpin and duplex conformations in vitro. *J Gen Virol* 78 (Pt 10):2647-2655.
38. Christensen J & Tattersall P (2002) Parvovirus initiator protein NS1 and RPA coordinate replication fork progression in a reconstituted DNA replication system. *Journal of virology* 76(13):6518-6531.
39. Wilson GM, Jindal HK, Yeung DE, Chen W, & Astell CR (1991) Expression of minute virus of mice major nonstructural protein in insect cells: purification and identification of ATPase and helicase activities. *Virology* 185(1):90-98.
40. Cotmore SF & Tattersall P (1989) A genome-linked copy of the NS-1 polypeptide is located on the outside of infectious parvovirus particles. *Journal of virology* 63(9):3902-3911.

Chapter 2: Structure of Sf6 Tail Adaptor Suggests Molecular Mechanism for Sequential Tail Assembly

Adapted from paper 'Liang L, Zhao H, An B, Tang L (2018) High-resolution structure of podovirus tail adaptor suggests repositioning of an octad motif that mediates the sequential tail assembly. *Proceedings of the National Academy of Sciences of the United States of America* 115(2):313-318.'

2.1 Introduction

The podovirus morphogenesis begins with the assembly of a procapsid, containing coat proteins, scaffolding proteins, and portal. Subsequently, the ~40kb viral genome is translocated into the procapsid through the central channel of the portal by DNA-packaging motor, resulting in densely packed DNA (1). The high pressure built by DNA packing in the capsid lattice triggers conformational change of portal protein, which subsequently releases the motor, terminating the DNA packaging (2-4). A set of tail proteins attach to the portal vertex to prevent the leakage of DNA. The tail adaptor, gp7 in the case of phage Sf6, is the first tail protein to polymerize on the portal, followed by the sequential incorporation of the tail nozzle, tail needle, and tail spikes (5). Eventually, the portal, tail adaptor, and tail nozzle make up a conduit that is to be utilized for phage DNA injection into host cells during infection. The tail needle protein blocks the end of this conduit, preventing DNA from leaking out. This needle probably also serves to penetrate part of host envelope and triggers the translocation of DNA-injection proteins as well as viral DNA (6). The last tail protein in the assembly is tail spike gp14 that has two domains. The tail spike head-binding domain (HBD) binds to the interface of the tail adaptor and the tail nozzle, and the tail spike receptor-binding domain (RBD) recognizes cell receptor and cleaves host O-antigen, pulling virions close to the host cell upon infection. Sequential assembly of tail components to the phage capsid has been recognized in several podoviruses such as T7 (7) and P22 (5, 8-10). This assembly mechanism differs from that in the other two families, *Myoviridae* and *Siphoviridae*, of which the phage head, tail, and tail fibers are assembled independently and subsequently join to form the complete virion (11). The sequential mechanism ensures

orderly assembly and avoids aberrant products from premature interaction between components, which may contribute to highly efficient, rapid phage assembly.

The podovirus sequential tail assembly has been studied for long (5). However, the underlying mechanisms remain to be understood. The X-ray structure of phage P22 portal:adaptor complex showed the dodecameric ring-like architecture of the tail adaptor (P22-gp4) that binds to the portal dodecamer with elongated C-terminal portions (12). This and the sub-nanometer resolution cryo-EM structures of the isolated P22 tail (13) and the P22 virion (14, 15) shed light on the assembly of the multiple components in the context of the tail and the virion. Here we report the high-resolution X-ray structure of phage Sf6 tail adaptor gp7 in its pre-assembly state. Comparison to the P22 tail adaptor structure reveals a conformational switch at the N-terminal segment upon tail assembly that ensures the subsequent attachment of the tail nozzle. Such conformational switch is enabled by repositioning of two consecutive repeats of a conserved sequence motif that are capable of binding to the same protein surface.

2.2 Methods

2.2.1 Production of Sf6-gp7

The DNA fragment encoding Sf6-gp7 (residues 1-124) was cloned into pET28b (Novagen) between Nco1/Xho1. The region after residue 124 was predicted as disordered by Phyre2 therefore removed. A His-tag followed the protein at the C-terminus. Protein overexpression was done by growth of E. coli strain B834(DE3) at 37°C in LB broth until OD_{600nm}=0.6, followed by induction at 30°C by adding IPTG to a final concentration of 1 mM. Cells were harvested after 3 h of growth and lysed on a French press in the resuspension buffer (20 mM Tris-HCl pH 8.5, 500 mM NaCl, 10 mM β-mercaptoethanol). The protein was purified with a Ni-NTA column (Qiagen) followed by gel filtration chromatography on a Hiload 16/60 Superdex 75 column (GE Healthcare) in the gel filtration buffer (20 mM Tris-HCl pH 8.5, 150 mM NaCl, 1 mM DTT, 1 mM EDTA), which showed an elution volume of 63.82 ml corresponding to a dimer. The eluted fractions were collected and concentrated to 13 mg/ml using a Millipore centricon (molecular weight cutoff 10 kDa) prior to crystallization.

2.2.2 Crystallization, X-ray data collection and structure determination

The purified Sf6-gp7 was crystallized with the hanging drop vapor diffusion method by mixing 1 µl of the protein solution with 1 µl of the well solution containing 15% (w/v) PEG8,000, 0.1M MES pH 6.0 and 0.28M $Ca(OAc)_2$. Crystals were dipped into well solution plus 15% and 30% ethylene glycol in succession prior to flash freezing in liquid nitrogen. The heavy atom derivative was obtained by soaking the native crystals in well solution with 5 mM $HgCl_2$ for 24h prior to flash freezing.

Sf6-gp7 native and heavy atom derivative X-ray data were collected at the Advanced Photon Source (APS). All data were indexed and integrated using XDS(16, 17), scaled using Aimless(18). The structure was determined by the single anomalous dispersion method using the mercury derivative data (**Table 2-1**). The experimental electron density map calculated with Autosol(19) in PHENIX (20) was of excellent quality, which allowed automated building of most of the model using PHENIX. Small loops of regions of residues 26-29 in chain A and 24-29, 96-101 in chain B were manually built using COOT (21). Structure refinement coupled with manual model building was performed with the programs PHENIX and COOT respectively.

2.2.3 Generation of gp7 dodecamer

Sf6 is a P22-like phage therefore has conserved architectures with P22 (22). The 9.4Å cryo-EM reconstruction of the P22 phage tail (EMD-5051) (13) and the 3.25Å x-ray structure of P22-gp4 dodecamer bound to the portal core (PDB-3LJ4) (12) have been determined, which offer the possibility to establish a Sf6-gp7 dodecamer model by symmetry operations. Two approaches were employed to achieve the Sf6-gp7 model. First of all, the P22-gp4 dodecamer was separated from the portal core and fitted into the corresponding volume of P22 tail cryo-EM map. The first approach was to superimpose one Sf6-gp7 monomer (peptide chain A was chosen) onto the P22-gp4 dodecamer, giving a RMSD of 1.269Å. While the second approach was to directly dock one Sf6-gp7 monomer onto P22 tail map; this led to slight downward movement of Sf6-gp7 compared with the one superimposed on P22-gp4. For both approaches, the P22-gp4 dodecamer symmetry was applied to Sf6-gp7 by using CCP4 (23) program lsqkab (24), giving gp4-based and map-based gp7 dodecamer models, respectively.

2.3 Results and Discussion

2.3.1 Overall structure

The Sf6-gp7 has 160 amino acid residues. We designed a truncated gp7 construct lacking 36 C-terminal residues which were predicted to be disordered. A C-terminal His-tag was added to the construct to facilitate the protein purification. The structure was determined at 1.78Å resolution (**Table 2-1**), with two molecules in the asymmetric unit related by a non-crystallographic two-fold symmetry (**Fig. 2-1A**). One magnesium ion and four calcium ions were found in the structure. The structure is composed of four α-helices, namely α1-α4, connected by three loops. Despite low sequence homology, this protein fold is common in tail adaptor proteins of podoviruses, siphoviruses, and probably myoviruses as well. Structural superimposition shows a Cα RMSD of 1.289Å between 53 atom pairs with podovirus P22 (gp4, PDB ID 3LJ4) (12), 0.884Å between 18 atom pairs with siphovirus HK97 (gp6, PDB ID 3JVO) (25), 1.252Å between 21 atom pairs with siphovirus SPP1 (gp15, PDB ID 2KBZ) (26), and 0.899Å between 21 atom pairs with a putative neck protein of a *Myoviridae* prophage found in the *Bacillus subtilis* genome (YqbG, PDB ID 1ZTS) (27) (**Fig. 2-2B)**. The highly-conserved portion is the two long helices α2 and α3 arranged in the antiparallel manner. The loop connecting helices α2 and α3, namely loop α2α3, spans over 26 residues (residues 51-76), and is well defined in the electron density map. Multiple H-bonds with the water molecules which are coordinated to the calcium ion between the two gp7 molecules help stabilize the loop α2α3 (**Fig. 2-3A**). According to the dodecameric assembly of the gp7 ortholog P22-gp4 (12), the dimeric arrangement of gp7 is not likely to be biologically relevant, meaning that in the phage tail the loop α2α3 is not stabilized in the same way as in the crystal. As discussed later in this manuscript, we fit the X-ray structure of gp7 to the cryo-EM map of the P22 tail (13). By fitting in the HBD of the neighboring tail spike at the meantime, we found that four loops of the HBD located within 7Å to the loop α2α3 of gp7 (**Fig. 2-3B**), suggesting the possible roles of these loops in stabilizing the loop α2α3. The C-terminal eight residues are highly extended and well defined in the electron density map for molecule A, but are disordered in molecule B (**Fig. 2-1B**).

Table 2-1. X-ray data collection and structure refinement statistics

Data collection	Native	Hg
Beamline	APS 23ID-D	APS 23ID-B
Wavelength (Å)	1.03324	1.00699
Resolution (Å)	29.6-1.78(1.81-1.78)*	50-2.66 (2.71-2.66)*
No. Measurements	235,103	154,347
Unique reflections	38,487 (1,936)*	11,922
Completeness (%)	98.9 (89.2)*	99.8
I/σ	17.5 (1.7)*	53.7 (9.96)*
R_{merge} (%)**	5.7 (67.9)*	9.6 (48)*
Space group	$P4_12_12$	$P4_12_12$
Unit cell (Å)	a=101.30, c=76.46	a=100.93, c=77.73

Structure refinement	
Resolution (Å)	29.55-1.78
R_{work}/R_{free}[a]	0.18/0.20
Number of atoms	
Protein/Water	1,838/230
B-factors	
Protein/Water	40.94/48.23
R.m.s deviations	
Bond lengths (Å)	0.007
Bond angles (°)	0.938
Ramachandran plot	
Most favored (%)	98.28
Allowed (%)	1.72
Disallowed (%)	0.00

*Values in the parentheses are for the outermost resolution shells.

**$Rmerge=\Sigma_{hkl}\Sigma_i|I_i(hkl)-\langle I(hkl)\rangle|/\Sigma_{hkl}\Sigma_i I_i(hkl)$, where $I_i(hkl)$ is the observed intensity of reflection hkl and $\langle I(hkl)\rangle$ is the averaged intensity of symmetry-equivalent measurements

[a]$R_{work}=\Sigma_{hkl}||F_{obs}|-|F_{calc}||/\Sigma_{hkl}|F_{obs}|$, where F_{obs} and F_{calc} are structure factors of the observed reflections and those calculated from the refined model, respectively. R_{free} has the same formula as R_{work}, except that it was calculated against a test set of the data that was not included in the refinement

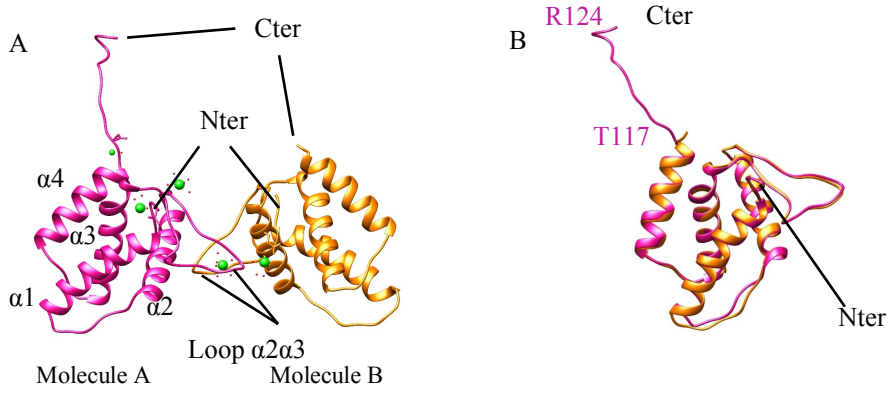


Fig. 2-1. The Sf6-gp7 is a globular protein composed of four alpha helices. *(A)* The ribbon diagram of the gp7 molecules A and B (colored in pink and orange) in the asymmetric unit. One magnesium (small green sphere) and four calcium (large green sphere) ions are found. *(B)* Superimposition of the gp7 molecules A and B shows the only significant difference at the C-terminal fragment (T117-R124).

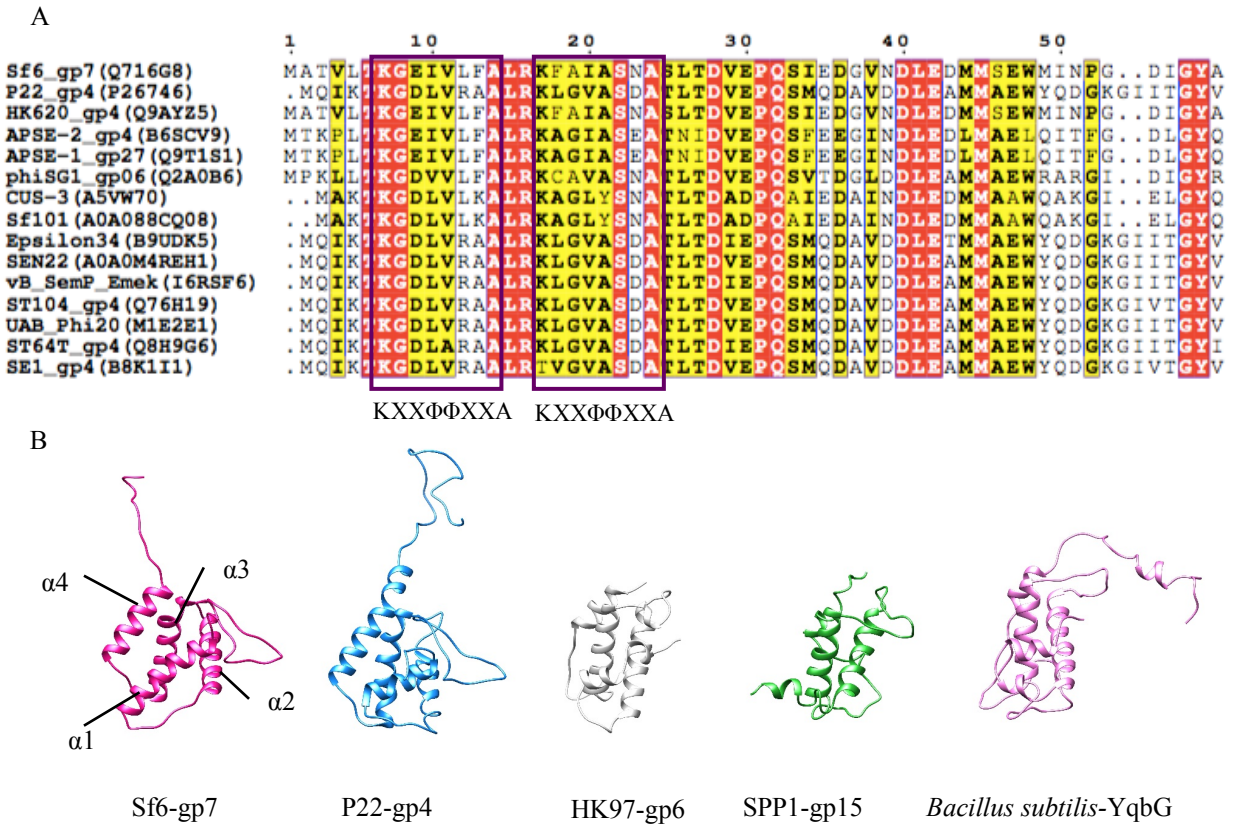


Fig. 2-2. The Sf6-gp7 possesses conserved octad-motif and overall protein fold. *(A)* Sequence alignment of the15 P22-like phage tail adaptor proteins with over 36% identity to gp7 shows that the octad motif is conserved within these polypeptides. *(B)* The Ribbon diagrams of podovirus tail adaptors Sf6-gp7 and P22-gp4, siphovirus tail adaptors HK97-gp6 and SPP1-gp15, putative myovirus neck protein YqbG. Superimposition of all five structures reveals a common, shared protein fold.

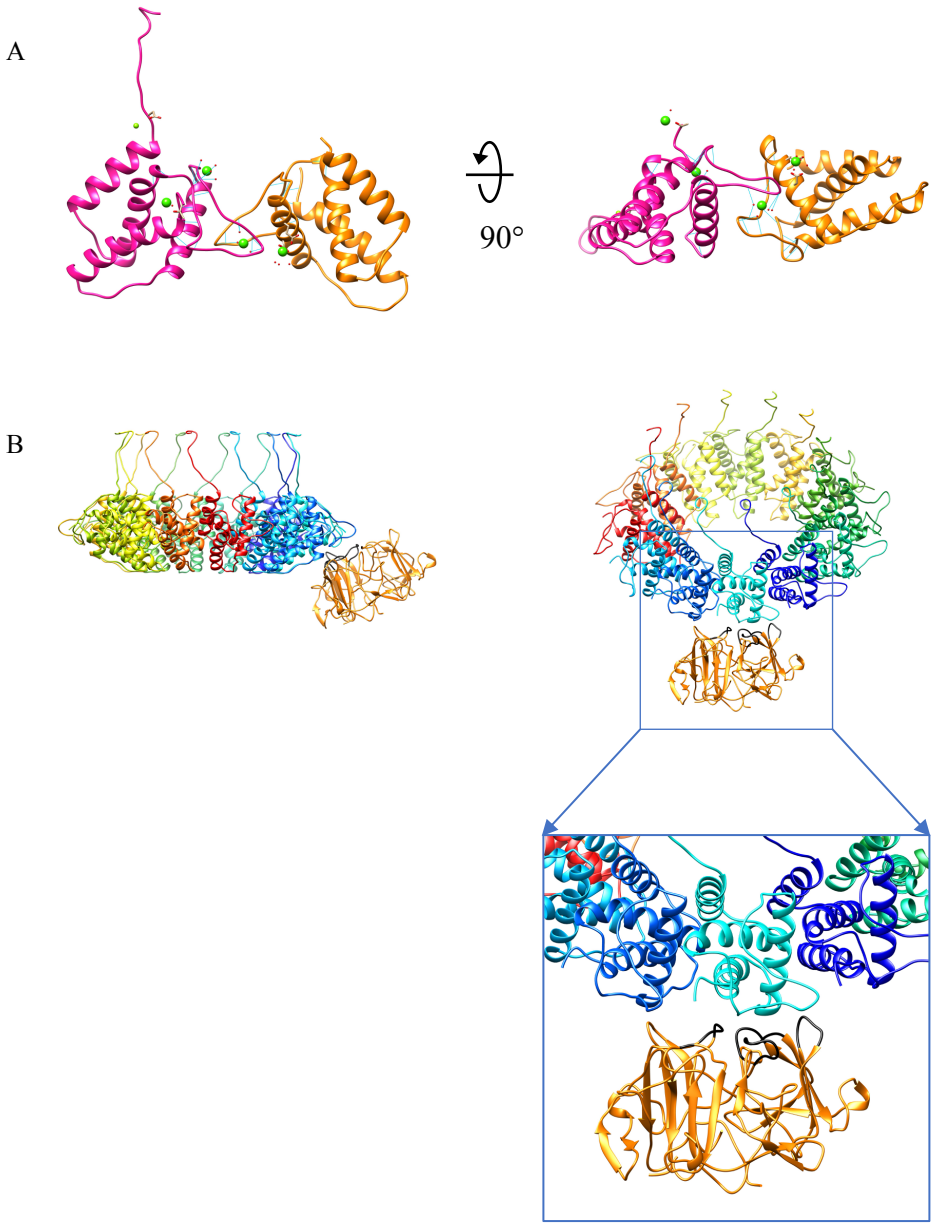


Fig. 2-3. Stabilization of the α2α3 loop. *(A)* In crystals, the α2α3 loop is stabilized by forming H-bonds (cyan lines) with water molecules (red spheres) that are coordinated to the calcium ion (green spheres) in between the two gp7 molecules. The side (left) and bottom (right) views are shown. *(B)* The relative positioning of the gp7 dodecamer and the HBD (orange) of a neighboring tail spike in the tail. The tilt view (right) is partially enlarged (bottom) to show the interface between the gp7 (cyan) and the HBD. The residues of the HBD located within 7Å to the α2α3 loop are colored in black, likely involved in stabilizing the α2α3 loop biologically.

2.3.2 A distinct conformation of the N-terminal portion

Sf6-gp7 and P22-gp4 share 37% identity and 56% similarity in amino acid sequences (**Fig. 2-4A**). We extract a gp4 monomer from the P22 portal-gp4 complex (12) and compare its structure with Sf6-gp7. The protein folds are highly conserved, while the N-terminal and C-terminal portions show significant differences (**Fig. 2-4B**). In the N-terminal segment, Sf6-gp7 has a long α1-helix and a nearly vertically oriented α1α2 loop, while P22-gp4 has the α1-helix half in length compared to that of Sf6-gp7 and an essentially horizontal α1α2 loop (**Fig. 2-4B**). The X-ray structure of Sf6-gp7 fits well into the isolated cryo-EM map of P22 tail (Fig. 1-5). However, the α1-helix and the α1α2 loop of Sf6-gp7 penetrate into the portion of the cryo-EM map corresponding to the tail nozzle (P22-gp10 or Sf6-gp8) which locates right underneath the adaptor in the tail complex, creating a steric clash (**Fig. 2-4C**). This indicates that the Sf6-gp7 α1-helix and the α1α2 loop must undergo conformational changes upon assembly from the monomers free in solution into the tail.

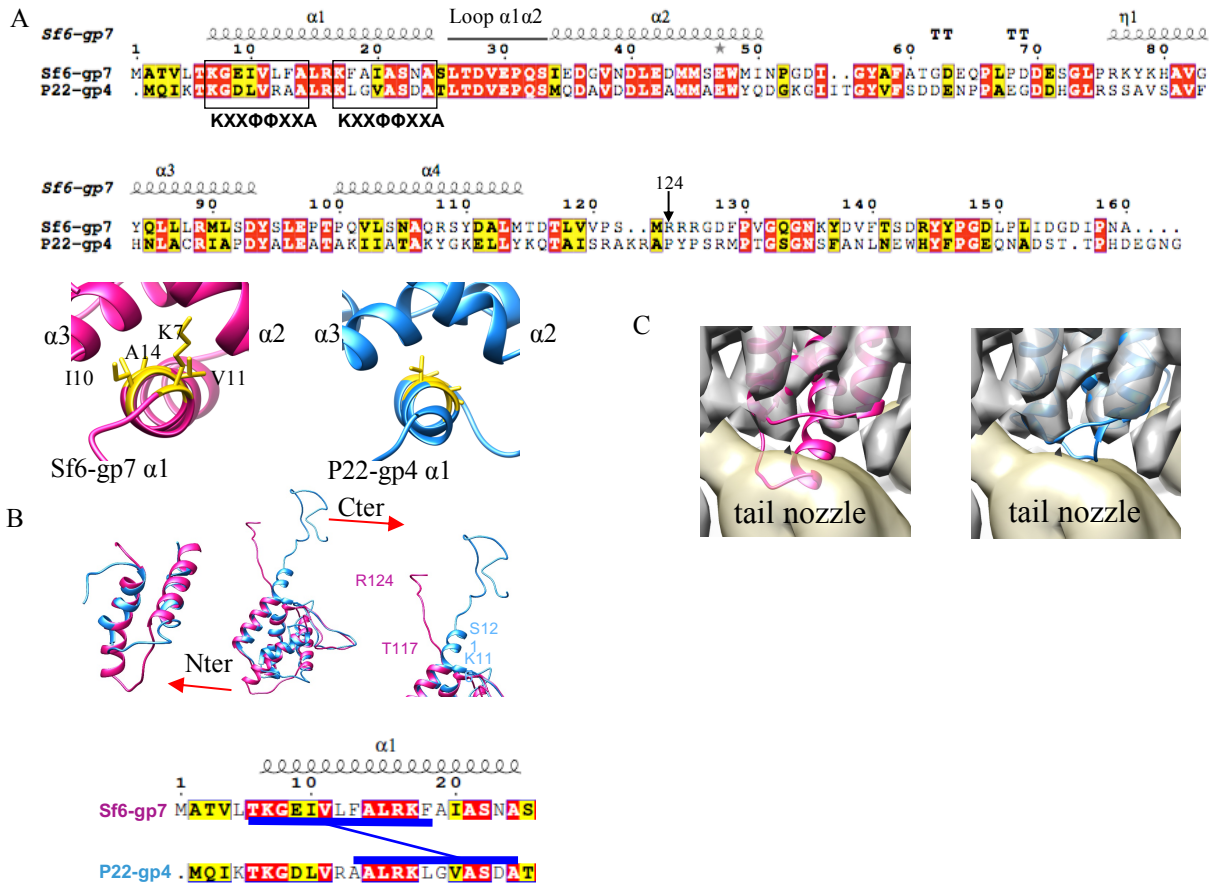


Fig. 2-4. The octad motif shifting in the α1-helix of tail adaptor. *(A)* The sequence alignment of Sf6-gp7 and P22-gp4 with 3D structure information of Sf6-gp7. The repeated octad motif in α1-helix is indicated with the black rectangles. The two images below are the projections of α1-helix in Sf6-gp7 (left) and P22-gp4 (right). The four conserved residues of the first gp7 octad motif and the second gp4 octad motif are colored in gold. *(B)* The superimposition of Sf6-gp7 (pink) and P22-gp4 (cyan) with N- and C-terminal segments enlarged. The sequences of α1-helix in Sf6-gp7 and P22-gp4 that are structurally superimposed are indicated by blue lines. *(C)* Sf6-gp7 (left) and P22-gp4 (right) are docked in the cryo-EM map of P22 tail. The Sf6-gp7, but not the P22-gp4, clashes with the EM volume of tail nozzle (map colored in yellow). The contour level of the map is 4.46.

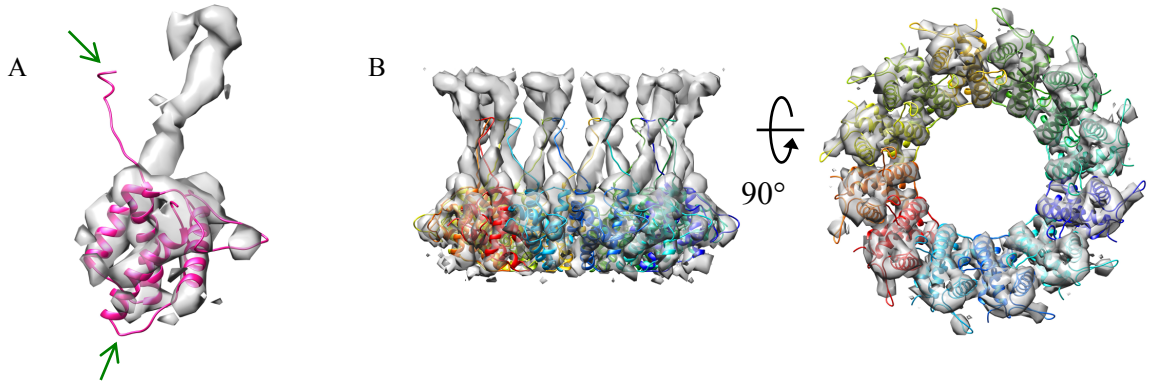


Fig. 2-5. Docking of the gp7 structure in the cryo-EM map of P22-gp4 ring segmented from the cryo-EM map of P22 tail. *(A)* The gp7 structure well fits the cryo-EM map except for the C-terminal segment and the loop α1α2 (indicated by arrows). *(B)* The gp7 pseudo-atomic dodecameric model fits the cryo-EM map of P22-gp4. The contour level of the map is 4.46.

Interestingly, structural superposition shows that the 1st half (residue number 6-18) of the Sf6-gp7 α1-helix overlays the P22-gp4 α1-helix (residue number 12-23) structurally (**Fig. 2-4B**), while sequence alignment shows that the 2nd half (residue number 14-24) of the Sf6-gp7 α1-helix aligns well with the P22-gp4 α1-helix (residue number 12-23) at the amino acid sequence level (**Fig. 2-4A**), indicating a correlation between the two structures. Notably, the amino acid sequences of tail adaptor loop α1α2 are highly conserved between phage Sf6 and P22. Moreover, the tail nozzle proteins in the two phages share as high as 93% sequence identity (22). Thus, it is reasonable to infer that the tail adaptor loop α1α2 may physically

interact with the essentially identical tail nozzle in phages Sf6 and P22, and the modes of interactions between the tail adaptor loop α1α2 and the tail nozzle are largely identical, which indicates consistent conformations for the loop α1α2 of Sf6-gp7 and P22-gp4 in the context of the tail. Given that Sf6-gp7 is crystallized in solution while P22-gp4 is in complex with the portal protein, these results together suggest that the distinct conformation of the N-terminal portion of Sf6-gp7 represents the pre-assembly state, whereas that observed in the P22 portal:gp4 complex structure represents the post-assembly state. While free in the infected cell cytoplasm after protein synthesis, the tail adaptor protein may adopt the conformation as observed in the present Sf6-gp7 X-ray structure, that is, its loop α1α2 and α1-helix adopts the vertical orientation. Upon assembly onto the portal ring, the tail adaptor undergoes conformational changes in its loop α1α2 and α1-helix, which move up so that the second half of the α1-helix takes place of the first half and the loop α1α2 adopts a horizontal orientation. Such conformational changes would eliminate the steric clash with the tail nozzle and create a specific binding site for the tail nozzle.

When transitioning from the pre-assembly state to the post-assembly state, the α1-helix repositions so that its second half takes the place of the first half in the pre-assembly state. Analysis of amino acid sequences reveals a sequence motif that supports such replacement. In Sf6-gp7, the amino acid region 7-24 (α1-helix) contains two repeats of an octad-motif KXXΦΦXXA (**Fig. 2-4A**), where K (position #1) is lysine; X is any residue; Φ (position #4 & 5) represents a hydrophobic residue; and A (position #8) is alanine. This motif is conserved in P22-gp4 (**Fig. 2-4A**) and many P22-like phages as well (**Fig. 2-2A**). In this motif, Lys, the two hydrophobic residues and Ala are located on the same side of the helix α1 and thus define a common surface (**Fig. 2-4A**). In the X-ray structures of Sf6-gp7 and the P22 portal:gp4 complex, this surface is the one that forms the interface between α1-helix with the rest of the molecule (mainly α2, α3-helices) (**Fig. 2-4A**). Thus, repositioning of the α1-helix during transition from the pre-assembly state to the post-assembly state is enabled by swapping of the two repeats of such octad sequence motif, which utilizes the essentially identical interface and would maintain essentially the same interactions between this helix to the binding interface on the molecule.

2.3.3 The C-terminal segment displays conformational flexibility

Superposition of molecules A and B gives an RMSD of 0.468 Å. The significant structural difference is at the C-terminal segment. The eight C-terminal residues in molecule B (residues T117-R124) show no electron density in the X-ray map and thus are disordered, while this region in molecule A forms a well-defined extended loop that is immobilized by interactions with adjacent molecules in the crystal (**Fig. 2-1B**). These data indicate conformational flexibility of Sf6-gp7 C-terminal segment. What's more, the conformation of the Sf6-gp7 C-terminal portion differs from that of P22-gp4 as seen in the P22 portal:gp4 complex structure. A part of this portion in P22-gp4 (residues K116-S121) is the C-terminal part of the α4-helix, while the remaining residues from R122 to the very C-terminus form an elongated loop that extensively interacts with the portal (**Fig. 2-4B**). It is likely that Sf6-gp7 C-terminal portion is flexible while free in solution prior to attachment to the portal. Upon binding to the portal, the C-terminal portion may adopt an extended conformation and make extensive contact with the portal akin to what was observed in the P22. The C-terminal portions of Sf6-gp7 and P22-gp4 both contain several glycine and proline residues that are conserved between the two phages and are well aligned (**Fig. 2-4A**), which may facilitate such conformational changes upon post-assembly with the portal. In fact, folding of unstructured segments upon binding to viral partners appears as a general strategy for the control of sequential tail assembly (26, 28, 29).

2.3.4 The bipolar electrostatic surface of the Sf6-gp7 monomer

Sf6-gp7 contains 19 negatively and 10 positively charged residues. While negatively charged residues spread over the molecule essentially uniformly, 7 out of 10 positively charged residues are concentrated on one face of the molecule. Calculation of the electrostatic potential using USCF Chimera(30) reveals a distinct bipolar feature, that is, a convex positively charged face on one side and a negatively charged face covering the rest of the molecule (**Fig. 2-6A**). During assembly of the ring-like gp7 dodecamer (see below) in the tail, the positively charged ridge of each monomer packs into the negatively charged groove of the adjacent subunit (**Fig. 2-6C**). This indicates that assembly of gp7 monomers into the ring-like dodecamer in the tail is mediated by electrostatic interactions. Each monomer mimics a molecular

"magnet" that joins the next one in a head-to-tail manner via interactions between its positively charged face and the negatively charged face of the next monomer. The similar bipolar feature is observed on gp7 orthologs P22-gp4 and HK97-gp6 (**Fig. 2-7** and **2-8**) (26, 31), suggesting that such charge:charge interactions between bipolar molecules may be a common mechanism for assembly of the adaptor protein monomers into ring-like oligomers in those phage tails.

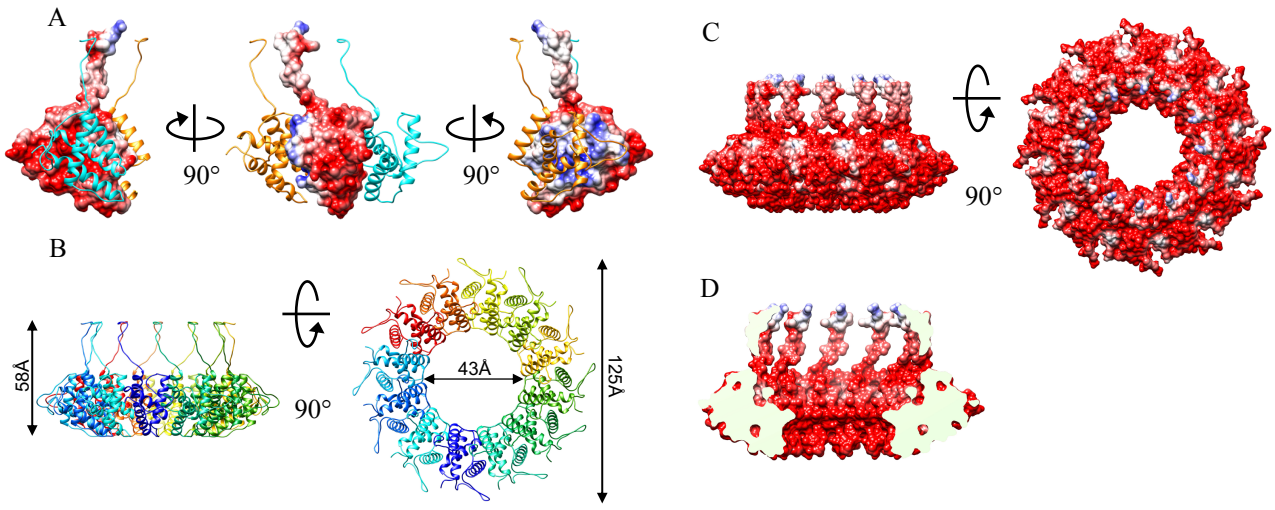


Fig. 2-6. The Sf6-gp7 monomer shows bipolar distribution of electric charges on its surface, whereas the dodecameric gp7 ring has the remarkably negative-charged surface. *(A)* Three neighboring gp7 monomers from the pseudo-atomic dodecameric model (shown in *(B)*) are shown as molecular surface colored according to electrostatic potential (middle) or ribbon (sides). The molecule is positively charged (blue) on one side, and negatively charged (red) on the other side. The positively charged ridge packs into the adjacent negatively charged groove during assembly. *(B)* Ribbon diagram of the pseudo-atomic model of gp7 dodecameric ring with chains colored in rainbow. *(C)* Surface representation of the gp7 dodecamric ring colored according to the electrostatic potential calculated with UCSF Chimera. *(D)* A cut-away view shows the negatively charged inner surface of gp7 dodecameric ring. For the color scheme of molecular surfaces, the blue corresponds to an electrostatic potential of +10kcal/(mol*e), and the red corresponds to an electrostatic potential of –10kcal/(mol*e).

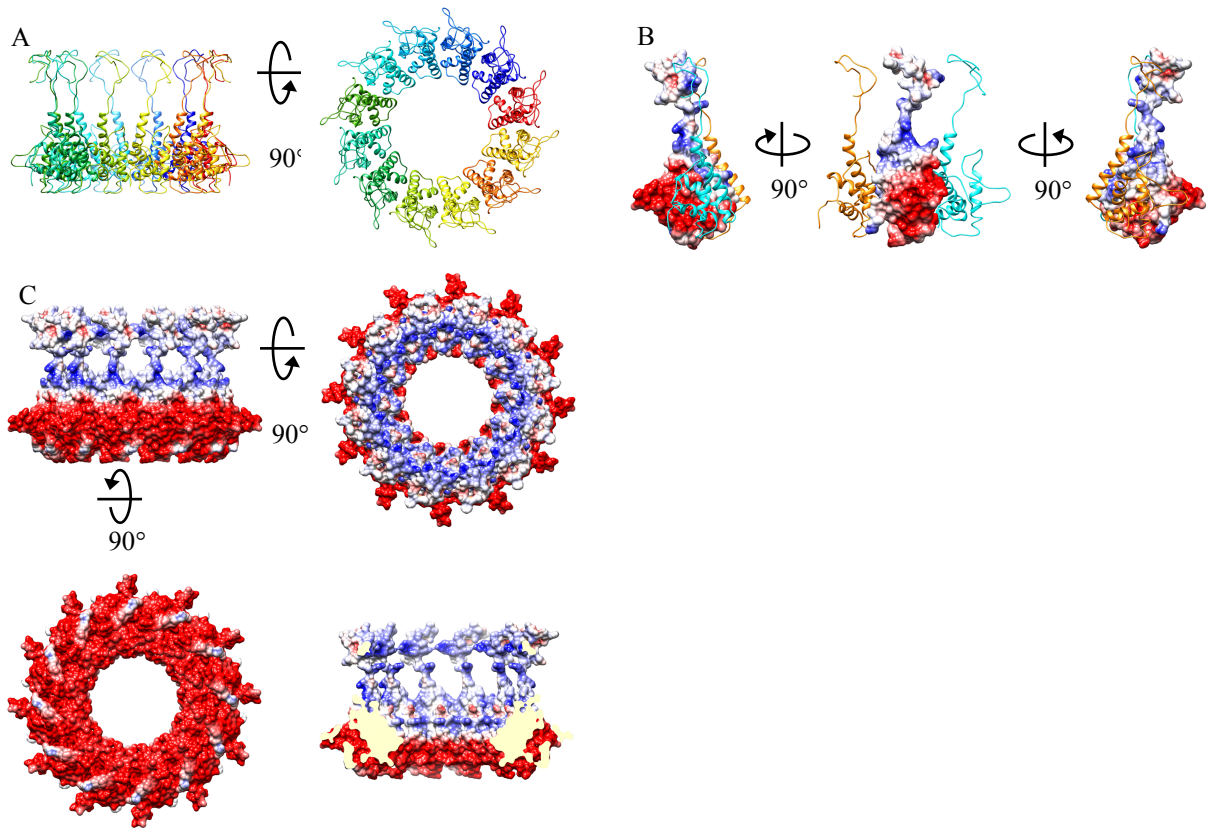


Fig. 2-7. The P22-gp4 shows bipolar distribution of electric charges on the surface; the P22-gp4 ring shows distinct positively and negatively charged surface. *(A)* Ribbon diagram of the P22-gp4 dodecameric ring with chains colored in rainbow. *(B)* Three neighboring P22-gp4 monomers from the dodecameric ring are shown as molecular surface colored according to the electrostatic potential (middle) or ribbon (sides). (*C)* Surface representation of the P22-gp4 dodecameric ring colored according to the electrostatic potential. The upper half of the ring shows positively charged surface, while the lower half shows negatively charged surface. Same color scheme of the electrostatic potential as that in **Fig.2-6**.

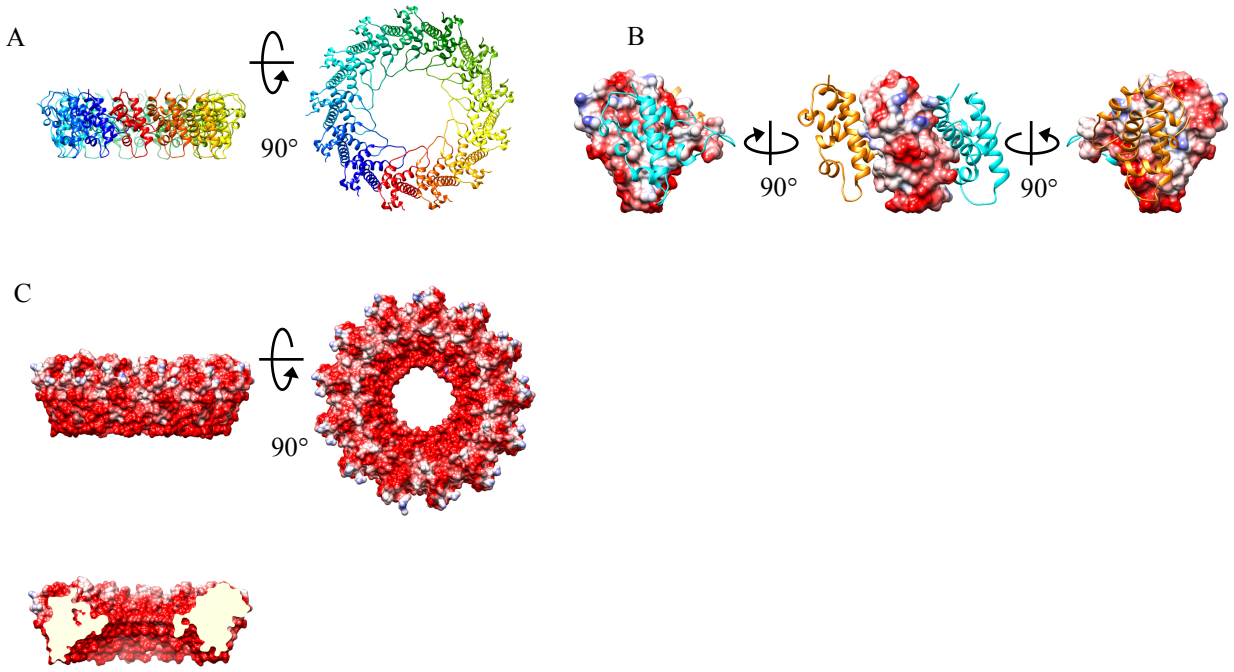


Fig. 2-8. The HK97-gp6 shows bipolar distribution of electric charges on the surface; the HK97-gp6 ring shows remarkably negatively charged surface. *(A)* Ribbon diagram of the HK97-gp6 dodecamer with chains colored in rainbow. *(B)* Three neighboring HK97-gp6 monomers from the dodecamer are shown as molecular surface colored according to electrostatic potential (middle) or ribbon (sides). Similar but not as significant as Sf6-gp7, one side of HK97-gp6 is positively charged (blue), and the other side is negatively charged (red). *(C)* Surface representation of the HK97-gp6 dodecamer colored according to electrostatic potential. The cut-away view shows the negatively charged inner surface of HK97-gp6 dodecamer. Same color scheme of the electrostatic potential as that in **Fig. 2-6**.

2.3.5 A model for the Sf6-gp7 dodecameric ring

The P22 portal:gp4 complex X-ray structure (12) and the sub-nanometer resolution cryo-EM map of P22 isolated tail (13) reveal how the P22-gp4 assembles with other components of the tail. Given the apparent sequence identity and common folds between tail component proteins, the overall assembly of the tail may be conserved between the two phages, which is also supported by the similarities of the tail

structures as shown in the cryo-EM asymmetric reconstructions of P22 (14) and Sf6 (22). The X-ray structure of the Sf6-gp7 fits well into the 9.4Å-resolution cryo-EM map of the P22 tail, except for the N-terminal and C-terminal portions where conformational changes occur as discussed above (**Fig. 2-5A**). The biological assembly of P22-gp4 is a ring-like dodecamer. To model the Sf6-gp7 dodecameric ring, we first docked one Sf6-gp7 monomer and the P22-gp4 dodecamer on the cryo-EM map of P22 tail. By applying the symmetry matrix of P22-gp4 dodecamer to Sf6-gp7, we could generate a pseudo-atomic model for the Sf6-gp7 dodecameric ring without significant clashes between monomers (**Fig. 2-6B**, **2-5B**). This assembly has a dimension of 125 Å by 58 Å, with an inner diameter of 43 Å.

The Sf6-gp7 monomer:monomer interface involves a complex network of salt bridges between the positively charged face of one monomer and the complementary negatively charged face of its adjacent monomer as described above, suggesting that electrostatic interactions plays a dominant role in stabilizing the Sf6-gp7 dodecamer. The Sf6-gp7 dodecameric model shows an overwhelmingly negatively charged surface, while the positively charged surfaces of monomers are buried. The ring assembly also shows a highly negatively charged internal surface of the channel. Such highly negatively charged internal surface of the channel has been described for several phage portals and tail adaptor proteins as a common characteristic and was thought to help to avoid non-specific DNA binding during DNA passage through these molecular channels (12, 25, 32). Hence, such a feature in Sf6-gp7 may facilitate translocation of phage dsDNA from inside of the phage into host cell cytoplasm.

The peripheral surface of the Sf6-gp7 dodecamer must provide binding sites for the tail spike and the tail nozzle. Thus, electrostatic interactions may play a major role in mediating assembly of the tail adaptor with the tail spike and tail nozzle. This may, at least partially, contribute to the ultra-stability of the tail, which is able to withstand heating to 60°C in 2M urea and in the presence of detergent in case of P22 (8). Additionally, the highly negatively charged surface of the tail adaptor ring assembly may be suitable for the need of switching from the terminase-bound state of the capsid at the end of DNA packaging process to the state for terminase dissociation and tail assembly. Interestingly, the highly negatively charged surface of the tail adaptor ring is also observed in P22-gp4, HK97-gp6 and Spp1-gp15 (**Fig. 2-7**, **2-8**, and **2-9**).

SPP1-gp15 shows some variance in that it contains a positively charged rim at the bottom (**Fig. 2-9C**). Thus, such a feature may be a common characteristic among those phages.

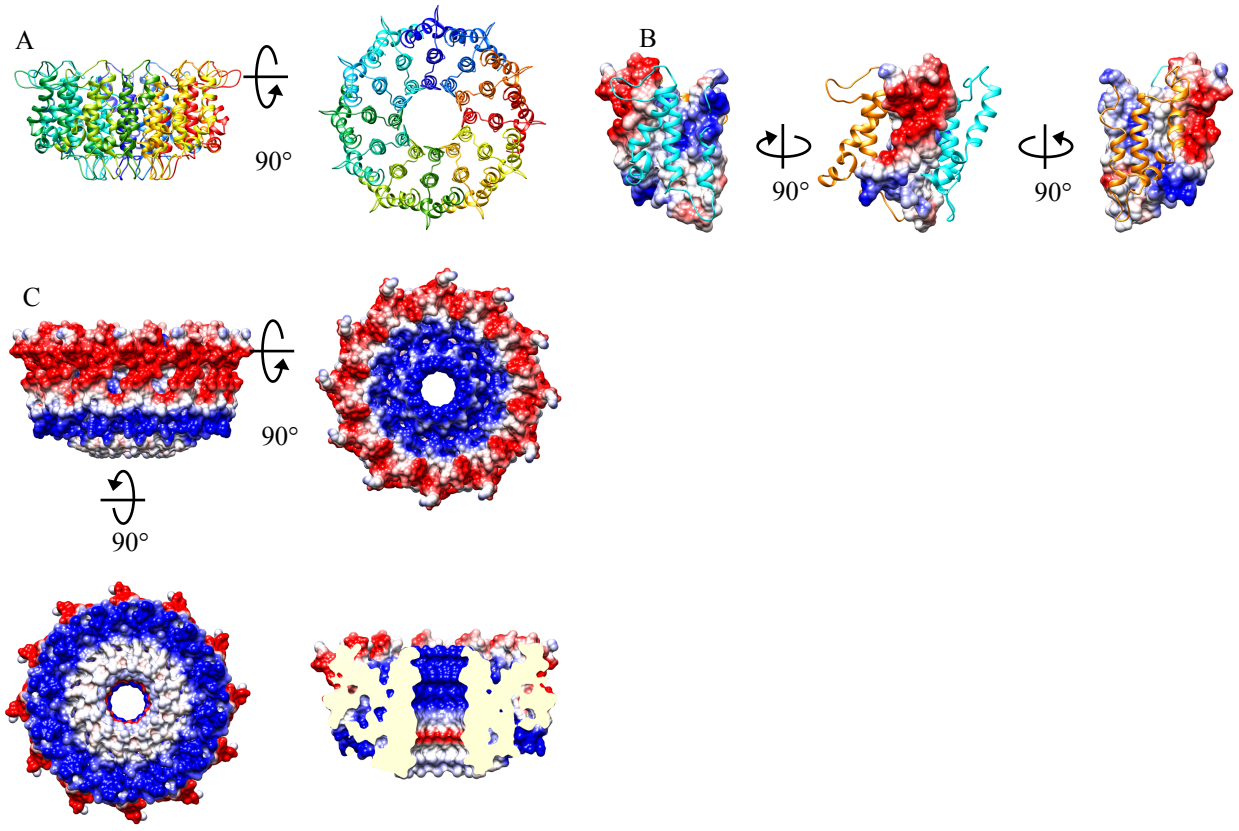


Fig. 2-9. The SPP1-gp15 shows bipolar distribution of electric charges on the surface; the SPP1-gp15 ring shows regularly distributed charged surface. *(A)* Ribbon diagram of the SPP1-gp15 dodecameric ring with chains colored in rainbow. *(B)* Three neighboring SPP1-gp15 monomers from the dodecameric ring are shown as molecular surface colored according to electrostatic potential (middle) or ribbon (sides). Differing from Sf6-gp7, the positively charged surface of SPP1-gp15 is not covered upon assembly. (*C)* Surface representation of the SPP1-gp15 dodecameric ring colored according to electrostatic potential. The upper half of the ring shows negative outer surface and positive inner surface, while the lower half shows opposite charge distribution. Same color scheme of the electrostatic potential as that in **Fig. 2-6**.

2.3.6 Implications for the gp7-mediated sequential assembly of the tail

The X-ray structure of Sf6 tail adaptor gp7 and the pseudo-atomic model of its dodecameric

assembly provide insight into the structural basis of the sequential assembly of the tails in P22-like phages. We propose that the Sf6-gp7 reported here and the P22-gp4 structure in the context of the portal:gp4 complex represent the pre- and post-assembly states of the tail adaptor, respectively. Our proposal is supported by the presence of the repeated octad motif KXXΦΦXXA in the α1-helix of Sf6-gp7 and P22-gp4. Comparison of the two X-ray structures shows that the 1st motif of Sf6-gp7 α1-helix is superimposed on the 2nd motif of the P22-gp4 α1-helix (**Fig. 2-4B**). Due to the distributions of the four conserved residues in the motif, such a positional replacement between the 1st and 2nd motifs in α1-helix still retains interactions of this α1-helix with the rest of the molecule. That is to say, upon assembly of the tail adaptor to the portal, the 2nd motif replaces the 1st motif, generating a different conformation for the highly-conserved loop α1α2. Such a conformational change in the loop α1α2 creates the binding site for the next tail component, i.e., the tail nozzle, and eliminates the steric clash between the tail adaptor and the tail nozzle as observed in fitting the Sf6-gp7 structure in the P22 tail cryo-EM map (**Fig. 2-4C**).

Based on these results, we propose a model for the sequential assembly of the phage tail in Sf6 (**Fig. 2-10**). In the infected cell cytoplasm, the newly synthesized tail adaptor protein molecules exist as monomers (33) that adopt the pre-assembly conformation as observed in the Sf6-gp7 X-ray structure, with the 1st octad motif occupying the interface with the protein core and a largely vertically oriented loop α1α2. The C-terminal portion may be flexible and unstructured. Upon completion of the DNA packaging process, the terminase dissociates from the capsid, and a tail adaptor subunit then binds to the portal dodecamer via its elongated C-terminal portion, which undergoes an induced fit from the flexible conformation to an extended and immobilized conformation that makes extensive contact with the portal. The binding probably triggers the upward repositioning of the second octad sequence motif in the α1-helix by ~10Å, which takes the place of the 1st motif, and this likely favors the binding of the next tail adaptor subunit as well as supported by that facts that P22-gp4 exists as monomer in solution (33) and binds to the portal ring in a highly cooperative manner (10). Accompanying the repositioning of the motif, the loop α1α2 moves away from protein core to a horizontal conformation, generating the binding site for the tail nozzle. This structural rearrangement allows assembly of Sf6-gp7 to the portal in the post-DNA-packaging capsid, followed by

attachment of the tail nozzle. The tail adaptor protein free in solution, which is in the pre-assembly conformation, would not be able to bind to the tail nozzle until it binds to the portal and undergoes the structural rearrangement at the α1-helix and the loop α1α2. After that, tail spike and tail needle will bind sequentially.

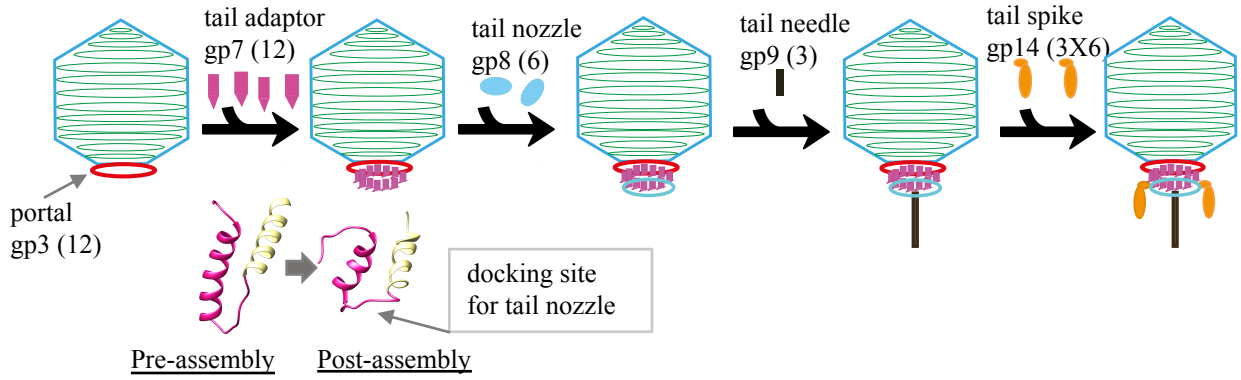


Fig. 2-10. The gp7-mediated phage Sf6 sequential tail assembly model.

Once the viral DNA (green line) is pumped into the procapsid to form the mature capsid (blue hexagon), the gp7 monomer (pink bullet) in solution that adopts the pre-assembly conformation begins to bind to the portal (red oval). The binding to the portal and/or to adjacent gp7 monomers induces the N-terminal segment (pink ribbon) to switch to the post-assembly conformation, allowing the attachment of the tail nozzle gp8 (light blue oval). The tail needle gp9 (brown line) and the tail spikes (gold) then attach to complete the tail assembly.

The sequential tail assembly of P22-like podovirus is enabled by a conserved conformational switch in the tail adaptor via replacement of two consecutive repeats of a sequence motif, which creates the binding site for tail nozzle only when it is time. Although *Siphoviridae* and *Myoviridae* tail and head assemblies occur independently, assembly of tails in those phages may also follow a sequential order as *Podoviridae* (34). For example, the baseplate wedge of the myovirus T4 is assembled sequentially, which is mediated by the induced conformational change caused by addition of new components (35). SPP1-gp15 has been shown to undergo conformational change during assembly to allow appropriate interaction between its helix α0 and loop α2α3 with the next tail protein gp16 (26, 31). It shares certain similarity with Sf6-gp7 in which the structural change allows the loop α1α2 to interact with gp8. The unique feature in Sf6-gp7 is that the

conformational change is achieved by repositioning of the two octad motifs. Such localized structural change may be more efficient and genetically stable compared with the large scale conformational change as in SPP1-gp15 because only several residues (the conserved residues of the octad motif) are essential. The tandem octad motif is highly conserved in P22-like phages, suggesting a conserved mechanism adopted by this group of phages. It will be interesting to see if the sequence motif repositioning mechanism is a recent evolutionary result adopted only among P22-like phages or it also exists in other Caudoviruses.

2.4 References

1. Earnshaw WC & Casjens SR (1980) DNA packaging by the double-stranded DNA bacteriophages. *Cell* 21(2):319-331.
2. Fokine A & Rossmann MG (2014) Molecular architecture of tailed double-stranded DNA phages. *Bacteriophage* 4(1):e28281.
3. Rao VB & Feiss M (2015) Mechanisms of DNA Packaging by Large Double-Stranded DNA Viruses. *Annual review of virology* 2(1):351-378.
4. Lokareddy RK*, et al.* (2017) Portal protein functions akin to a DNA-sensor that couples genome-packaging to icosahedral capsid maturation. *Nature communications* 8:14310.
5. Strauss H & King J (1984) Steps in the stabilization of newly packaged DNA during phage P22 morphogenesis. *J Mol Biol* 172(4):523-543.
6. Casjens SR & Molineux IJ (2012) Short noncontractile tail machines: adsorption and DNA delivery by podoviruses. *Advances in experimental medicine and biology* 726:143-179.
7. Cuervo A*, et al.* (2013) Structural characterization of the bacteriophage T7 tail machinery. *The Journal of biological chemistry* 288(36):26290-26299.
8. Tang L, Marion WR, Cingolani G, Prevelige PE, & Johnson JE (2005) Three-dimensional structure of the bacteriophage P22 tail machine. *The EMBO journal* 24(12):2087-2095.
9. Olia AS, Bhardwaj A, Joss L, Casjens S, & Cingolani G (2007) Role of gene 10 protein in the hierarchical assembly of the bacteriophage P22 portal vertex structure. *Biochemistry* 46(30):8776-8784.
10. Lorenzen K, Olia AS, Uetrecht C, Cingolani G, & Heck AJ (2008) Determination of stoichiometry and conformational changes in the first step of the P22 tail assembly. *Journal of molecular biology* 379(2):385-396.
11. Wood WB (1980) Bacteriophage T4 morphogenesis as a model for assembly of subcellular structure. *Q Rev Biol* 55(4):353-367.
12. Olia AS, Prevelige PE, Jr., Johnson JE, & Cingolani G (2011) Three-dimensional structure of a viral genome-delivery portal vertex. *Nature structural & molecular biology* 18(5):597-603.
13. Lander GC*, et al.* (2009) The P22 tail machine at subnanometer resolution reveals the architecture of an infection conduit. *Structure* 17(6):789-799.
14. Tang J*, et al.* (2011) Peering down the barrel of a bacteriophage portal: the genome packaging and release valve in p22. *Structure* 19(4):496-502.
15. Pintilie G, Chen DH, Haase-Pettingell CA, King JA, & Chiu W (2016) Resolution and Probabilistic Models of Components in CryoEM Maps of Mature P22 Bacteriophage. *Biophys J* 110(4):827-839.
16. Kabsch W (2010) Xds. *Acta crystallographica. Section D, Biological crystallography* 66(Pt 2):125-132.

17. Kabsch W (2010) Integration, scaling, space-group assignment and post-refinement. *Acta crystallographica. Section D, Biological crystallography* 66(Pt 2):133-144.
18. Evans PR & Murshudov GN (2013) How good are my data and what is the resolution? *Acta crystallographica. Section D, Biological crystallography* 69(Pt 7):1204-1214.
19. Terwilliger TC*, et al.* (2009) Decision-making in structure solution using Bayesian estimates of map quality: the PHENIX AutoSol wizard. *Acta crystallographica. Section D, Biological crystallography* 65(Pt 6):582-601.
20. Adams PD*, et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta crystallographica. Section D, Biological crystallography* 66(Pt 2):213-221.
21. Emsley P, Lohkamp B, Scott WG, & Cowtan K (2010) Features and development of Coot. *Acta crystallographica. Section D, Biological crystallography* 66(Pt 4):486-501.
22. Parent KN, Gilcrease EB, Casjens SR, & Baker TS (2012) Structural evolution of the P22-like phages: comparison of Sf6 and P22 procapsid and virion architectures. *Virology* 427(2):177-188.
23. Winn MD*, et al.* (2011) Overview of the CCP4 suite and current developments. *Acta crystallographica. Section D, Biological crystallography* 67(Pt 4):235-242.
24. Kabsch W (1976) A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. Sec. A* 32:922 - 923.
25. Cardarelli L*, et al.* (2010) The crystal structure of bacteriophage HK97 gp6: defining a large family of head-tail connector proteins. *Journal of molecular biology* 395(4):754-768.
26. Lhuillier S*, et al.* (2009) Structure of bacteriophage SPP1 head-to-tail connection reveals mechanism for viral DNA gating. *Proceedings of the National Academy of Sciences of the United States of America* 106(21):8507-8512.
27. Liu G*, et al.* (2006) NMR structure of protein yqbG from Bacillus subtilis reveals a novel alpha-helical protein fold. *Proteins* 62(1):288-291.
28. Maxwell KL, Yee AA, Arrowsmith CH, Gold M, & Davidson AR (2002) The solution structure of the bacteriophage lambda head-tail joining protein, gpFII. *Journal of molecular biology* 318(5):1395-1404.
29. Dyson HJ & Wright PE (2005) Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Bio* 6(3):197-208.
30. Pettersen EF*, et al.* (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* 25(13):1605-1612.
31. Chaban Y*, et al.* (2015) Structural rearrangements in the phage head-to-tail interface during assembly and infection. *Proceedings of the National Academy of Sciences of the United States of America* 112(22):7009-7014.
32. Cuervo A & Carrascosa JL (2012) Viral connectors for DNA encapsulation. *Curr Opin Biotechnol* 23(4):529-536.
33. Olia AS*, et al.* (2006) Binding-induced stabilization and assembly of the phage P22 tail accessory factor gp4. *Journal of molecular biology* 363(2):558-576.
34. Aksyuk AA & Rossmann MG (2011) Bacteriophage assembly. *Viruses* 3(3):172-203.
35. Yap ML, Mio K, Leiman PG, Kanamaru S, & Arisaka F (2010) The baseplate wedges of bacteriophage T4 spontaneously assemble into hubless baseplate-like structure in vitro. *Journal of molecular biology* 395(2):349-360.

Chapter 3: Structural Analysis of Sf6 Tail Nozzle gp8

3.1 Introduction

The *Shigella flexneri* phage Sf6 belongs to the *Podoviridae* family. It is closely related to phages P22, HK620 and ST64T (1). There have been significant studies on the tail architecture. The cryoEM reconstruction of the complete tail machine of phage P22 at the resolution of 9.4Å was reported (2). Although most of the structural proteins of Sf6 have amino acid sequences divergent from those of P22 (3), the overall mechanisms of virion morphogenesis and structural component folds are conserved for these phages (4). In Sf6, the crystal structures of the receptor-binding domain (RBD) of the tail spike (5) and the knob of the tail needle (6) were solved. The portal, the tail adaptor, the tail needle, the RBD and head-binding domain (HBD) of tail spikes in P22 were all solved (7-10), leaving the tail nozzle as the only tail protein for which the crystal structure is not obtained at all. According to the BLAST results of the protein amino acid sequences, this protein is highly conserved among five tail proteins (**Table 1-1, Fig. 3-1**) (93% identity with the P22 orthologs; in contrast to <=35% sequence identity for most other orthologous protein pairs), indicating that gp8 conducts highly conserved and vital functions.

As shown in the reconstruction, gp8 locates at the center of the tail machine, interacting with the adaptor, the tail spikes and the tail needle (2, 11-13) (**Fig. 1-1**). This spatial arrangement indicates that gp8 might serve to integrate the multiple tail components during the tail assembly. Gp8 is suggested to be indispensable during viral DNA transfer process. Since the Sf6 tail is too short to span over the three-layer envelope of its host cell, *Shigella flexneri*, like how myoviruses do, it is hypothesized that Sf6 injects three internal proteins to form an extended conduit in the host cell envelope for DNA delivery. This extended conduit is considered to be docked on tail protein gp8.

In this study, we overexpressed and purified the Sf6 gp8. The gp8 monomer bead model is obtained by small-angle X-ray scattering (SAXS) and is then fitted in the cryoEM map of the whole tail machine (2) to analyze protein-protein interfaces.

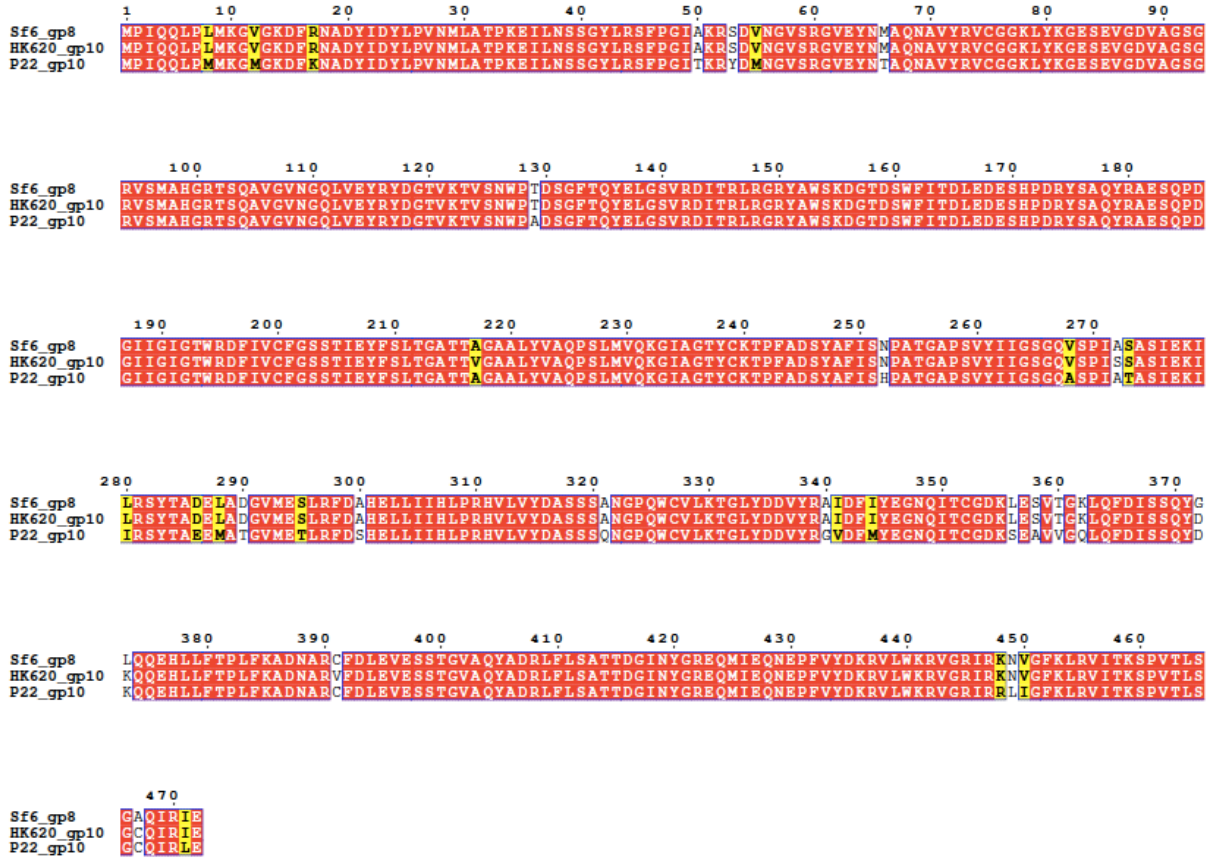


Fig. 3-1 Sequence alignment of gp8 orthologs from Sf6, HK620 and P22 shows that gp8 is highly conserved. Identical residues are colored in red and similar residues are in yellow.

3.2 Methods

3.2.1 Protein expression and purification

His-gp8

The gp8 DNA sequence from Sf6 was amplified by polymerase chain reaction (PCR) using forward and reverse primers containing the Nde1-EcoR1 restriction sites. The fragments were then cloned onto the pET28b vector using T4 ligase overnight at 16℃, placing the six-His tag at the N-terminus. The recombinant protein was expressed in *E.coli* strain B834(DE3) cells in LB broth with 30µg/ml kanamycin. After growth at 37℃ to an A600 of 0.6, IPTG was added to a final concentration of 1mM to induce protein expression. The culture was then incubated at 30℃ with shaking for 3 h. The cells were then collected,

resuspended in 30ml of lysis buffer (20mM Tris, pH 8.5, 500mM NaCl, 10mM β-mercaptoethanol), and lysed by French press. His-gp8 was then purified by immobilized metal affinity chromatography (IMAC) Ni-NTA. The elution was concentrated to 5ml by using a Millipore-Amicon concentrator with a molecular mass cutoff 10kDa before being applied to the size exclusion column Hiload 16/60 Superdex 200 prep grade (GE Healthcare) in gel-filtration buffer (20mM phosphate buffer pH 7.4, 300mM NaCl, 1mM EDTA, 1mM DTT). Fractions of his-gp8 monomer were pooled and concentrated. Typically, a liter of *E.coli* cells yielded ~12mg of purified his-gp8 monomer.

Tag-free gp8

Since his-gp8 did not crystallize under the searching conditions, tag-free gp8 was generated for crystallization trial. His-gp8 was first purified by Ni-NTA and concentrated to 2.5ml as described above. Then a PD-10 desalting column (GE Healthcare) was used to change the buffer of the protein to reaction buffer (20mM phosphate buffer pH 7.4, 300mM NaCl, 1mM DTT). The elution was treated with thrombin (0.5 NIH units thrombin per one milligram of protein) for 48hrs to cleave the His tag, after which the solution was applied to p-AMINOBENZAMIDINE-AGAROSE (pABA) beads and Nickel beads to remove thrombin and his tag/his-gp8, respectively. The flowthrough was collected and applied to the size exclusion column Hiload 16/60 Superdex 200 prep grade (GE Healthcare) in gel-filtration buffer (20mM phosphate buffer pH 7.4, 300mM NaCl, 1mM EDTA, 1mM DTT). Fractions of gp8 monomer were pooled and concentrated. Typically, a liter of *E.coli* cells yielded ~8mg of purified gp8 monomer.

MBP-gp8

The gp8 gene sequence was cloned on the pMAL-c5E vector with the restriction sites Nde1-EcoR1 to express maltose-binding protein (MBP) fusions. Six different lengths of polylinkers are designed. MBP-gp8_polylinker has the original polylinker; MBP-gp8_shortlinker has most part of the polylinker deleted, leaving five amino acid residues between MBP and gp8; MBP-gp8_delP has the five-residue-shorlinker but the second residue of gp8 proline is remove; MBP-gp8_delMP has the first residue methionine removed from MBP-gp8_delP; MBP-gp8_delMPI has the third residue isoleucine removed form MBP-gp8_delMP; MBP-gp8_del8 has the first eight residues of gp8 removed. All the MBP-gp8 fusions were overexpressed

in *E.coli* strain B834(DE3) cells in LB broth with 100μg/ml ampicillin supplemented with 0.2% glucose to repress the maltose genes on the chromosome of the E.coli host. After growth at 37℃ to an A600 of 0.6, IPTG was added to a final concentration of 0.3mM to induce protein expression. The culture was then incubated at 30℃ with shaking for 3 h. The cells were then collected, resuspended in 30ml of lysis buffer (20mM Tris pH 7.4, 200mM NaCl, 1mM EDTA, 1mM DTT), and lysed by French press. MBP-gp8 fusions were purified by amylose column followed by the size exclusion column Hiload 16/60 Superdex 200 prep grade (GE Healthcare) in gel-filtration buffer (20mM Tris pH 7.4, 150mM NaCl, 1mM EDTA, 1mM DTT). Fractions of MBP-gp8 monomer were pooled and concentrated for crystallization screening.

3.2.2 Solution X-ray scattering

The purified tag-free gp8 monomer was subject to the online SEC small-angle X-ray scattering (SEC-SAXS) at the Bio-SAXS beamline BL4-2 at the Stanford Synchrotron Radiation Lightsource (SSRL) using a Rayonix MX225-HE CCD detector (Rayonix, Evanston, IL) with a sample-to-detector distance of 1.7 m and a beam energy of 11 keV (wavelength $\lambda = 1.127$ Å) (14). Data collection and analysis for the SEC-SAXS were performed as previously described (15). Briefly, 20 μl protein sample at 4.7 mg/ml was applied onto a Superdex 200 PC 3.2/30 column equilibrated in the protein buffer with 5mM DTT. Eluate from the column was directly passed through a 1.5-mm-quartz capillary cell (Hampton Research, Aliso Viejo, CA) at 20°C in line with the X-ray beam. Scattering images were recorded with 1.5-second exposures every 5 seconds using the program Blu-Ice (16). The program SasTool (http://ssrl.slac.stanford.edu/~saxs/analysis/sastool.htm) was employed for data reduction including scaling, azimuthal integration, averaging and background subtraction. The first 100 images at the early part of the void volume were averaged and used as a buffer-scattering profile for the background subtraction. The scattering profiles were manually averaged after visual inspection in order to improve the signal-to-noise ratio. Guinier analysis was performed using the programs Primus (17) and AUTORG (18).

3.2.3 Fitting of the gp8 SAXS model into the cryoEM map

The 9.4Å cryoEM map of P22 tail machine (2) was downloaded from EMDB. X-ray structures of the portal and adaptor complex (7), the HBD (8) and the RBD (9) of the tail spikes, the tail needle (10)

were fitted into the map in the UCSF Chimera (19). The cryoEM map was enlarged by 5% using the mapman to better accommodate the portal and tail adaptor. Symmetry operations were applied to obtain hexamers of the HBD and the RBD trimmers. Surfaces of all x-ray structures were generated using the molmap so that the corresponding EM volume could be subtracted from the complete tail map. The remaining unassigned volume was then passed through a 20Å filter to generate a lower resolution map. This map was adjusted to a contour level of 0.344 resulting a volume of 399.8e^3 Å^3 which was of the calculated volume of gp8 hexamer. The beads model of gp8 monomer was fitted into this map computationally and manually and hexameric gp8 was generated based on the fitted monomer. The regions on the surrounding proteins that are within 7Å to gp8 was examined and the sequence alignment of these regions were analyzed among phage Sf6, P22 and HK620 by using ClustalW at EMBL-EBI and ESpript (20-23).

3.3 Results and Discussion

3.3.1 Expression and purification of gp8

The size exclusion chromatogram turned out to have three elusion peaks corresponding roughly to a monomer, dimer and hexamer, suggesting that gp8 exists in solution as a monomer, a dimer and a hexamer, regardless of which tag is used (**Fig. 3-2**). We incubated the monomer of his-gp8 at room temperature for three days and found that the monomer could associate into dimer and hexamer (**Fig. 3-2B**), indicating the continuous association and dissociation behavior of gp8.

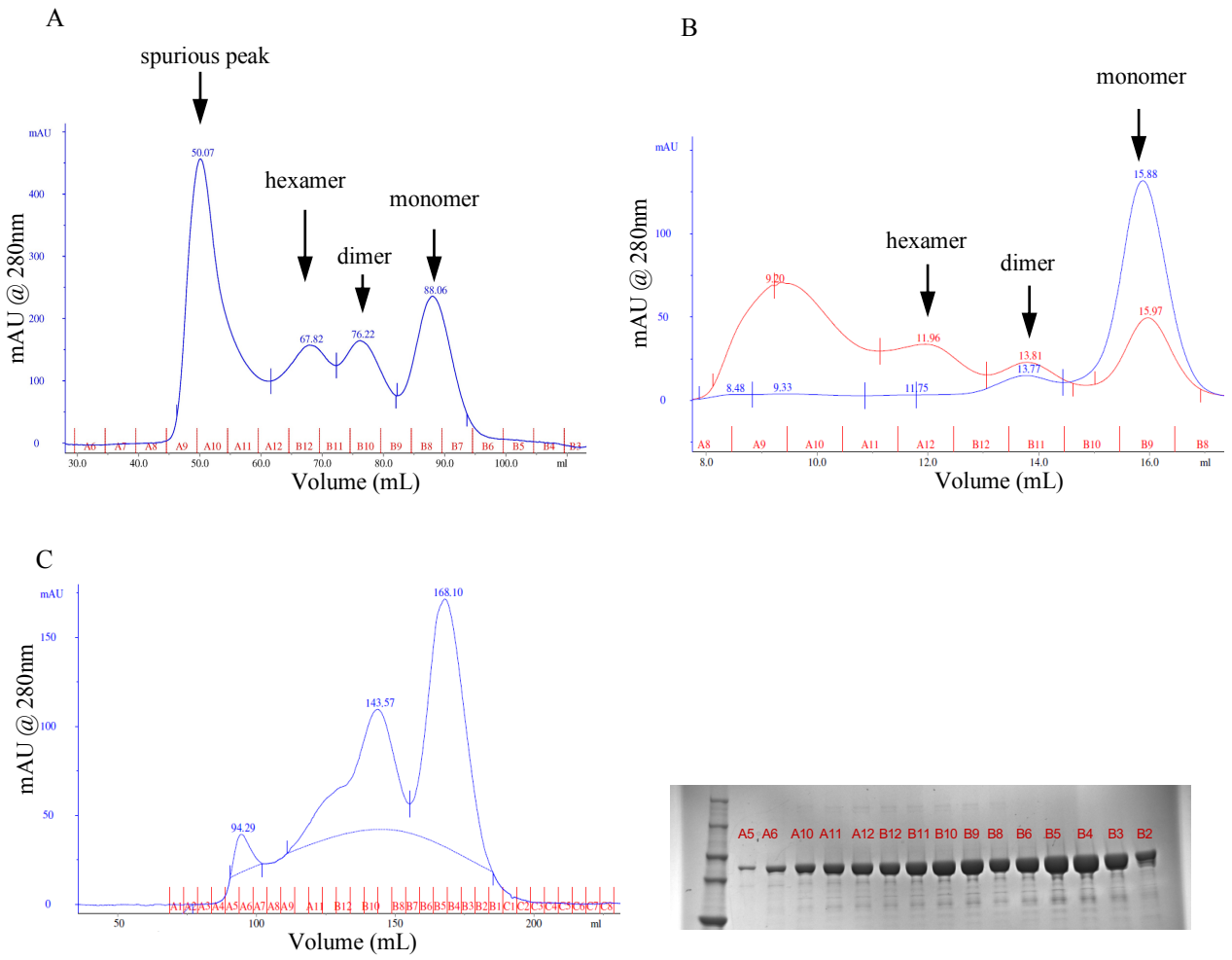


Fig. 3-2. Purification and oligomerization of Sf6 gp8. *(A)* The size exclusion chromatography (SEC) of his-gp8 shows three valid peaks corresponding to a monomer, a dimer and a hexamer, respectively. *(B)* Freshly purified and concentrated his-gp8 monomer exists mainly as monomer with a small amount of dimer (blue curve). After 3 days of incubation at room temperature, gp8 monomer associates into dimer, hexamer and other oligomers with higher molecular weight (red curve). *(C)* The SEC of MBP-gp8_shortlinker shows three similar peaks with that of his-gp8. The SDS-PAGE shows the proteins are in equilibrium among the three states without obvious gaps.

3.3.2 Crystallization trials fail to produce diffraction-quality crystals

His-gp8 was first tested using JBScreen, Wizard I&II, Hampton crystal screen HR2-110&112, however most drops generated heavy precipitate. In situ thrombin digestion by mixing 25uL of protein with 1uL of 100 units thrombin did not avoid precipitation. Similarly, tag-free gp8 also generated heavy precipitate. MBP-gp8 fusions were engineered with different linkers, including MBP-gp8_polylinker,

MBP-gp8_shortlinker, MBP-gp8_delP, MBP-gp8_delMP, MBP-gp8_delMPI, and MBP-gp8_del8 as explained in the methods. Among all these constructs, MBP-gp8_shortlinker was more promising by generating microcrystals in needle shape (**Fig. 3-3**). However, attempts to optimize conditions and add additives did not improve the crystals. The microcrystals did not diffract X-ray.

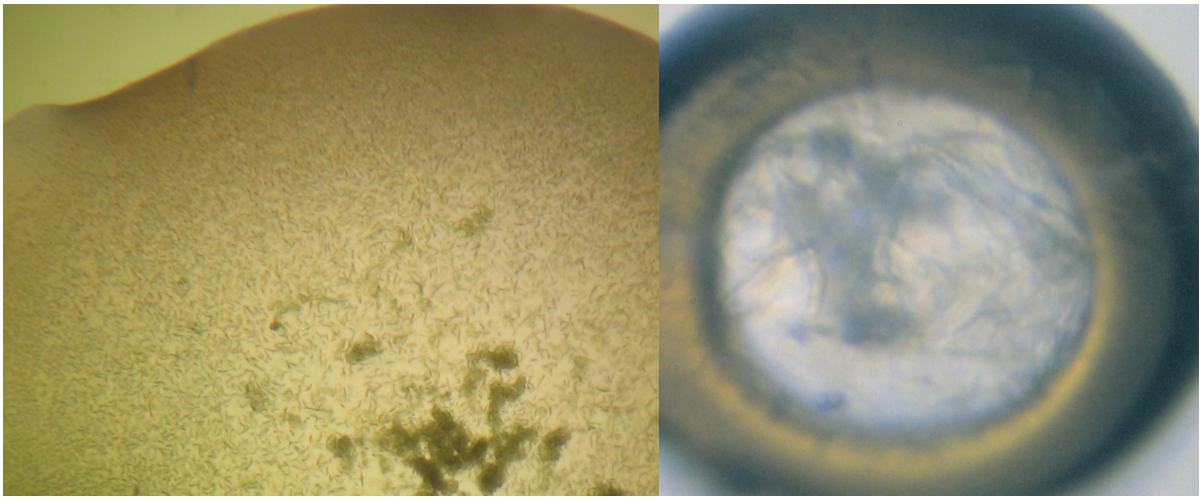


Fig. 3-3. MBP-gp8 fusion microcrystals (left) in hanging drops and (right) on loop.

3.3.3 SAXS analysis of the monomeric gp8 reveals a globular molecule

We applied monomeric tag-free gp8 to SAXS. The result unveiled a brick-shaped globular molecule with a protrusion on one side (**Fig. 3-4**).

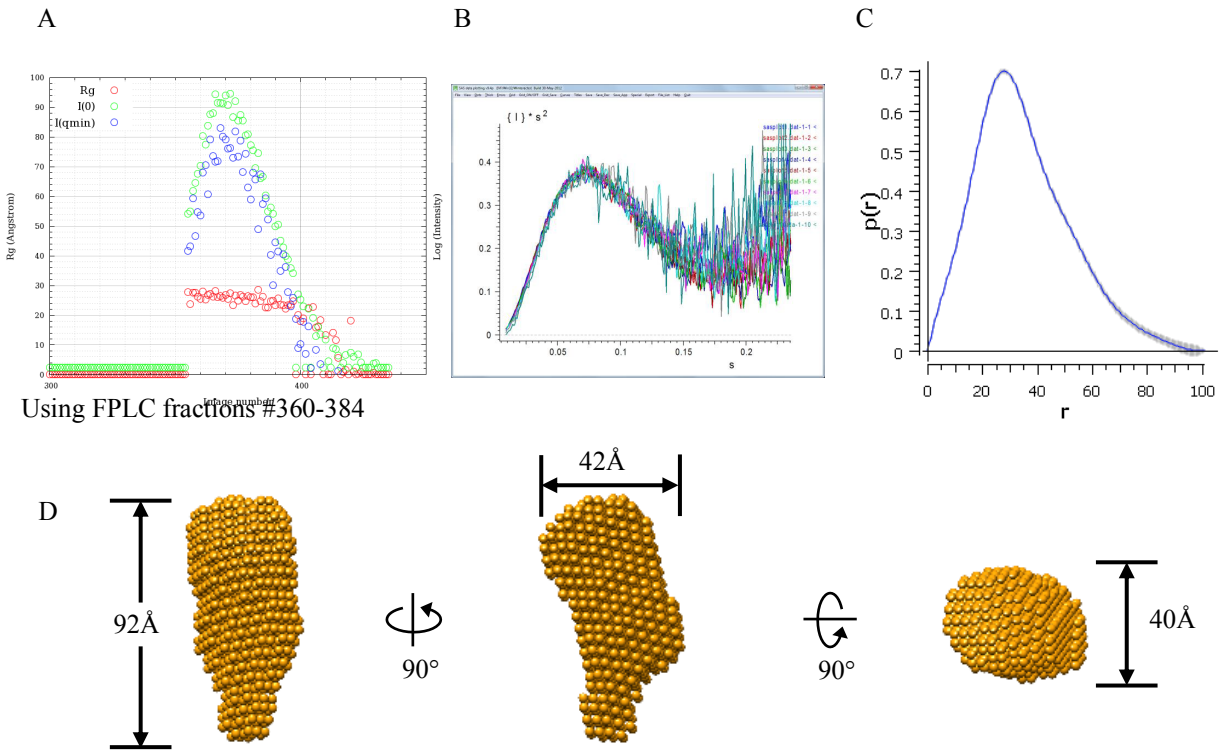


Fig. 3-4. The SAXS analysis of Sf6 gp8. *(A)* Rg/I plot from FPLC-SAXS. Fractions of the gp8 monomer peak in SEC elution profile (blue circles) was subject to SAXS data collection. *(B)* The Kratky plots of the SEC elution fractions. *(C)* The p(r) plot showing a Dmax of 101 Å. *(D)* The SAXS bead model viewed from the side, the front and the top respectively.

3.3.4 Fitting of the gp8 SAXS model into the cryoEM map

The cryoEM map of the complete tail machine (2)was enlarged by 5% so that the x-ray structures of the portal-adaptor complex, the HBD and the RBD of tail spikes, the tail needle can be covered better. However, twelve helices of the portal still deviated a little from the density map (**Fig. 3-5**). Since this part was involved in interactions with the capsid (12), we considered that there may be a small angle of twist at this region upon virion assembly.

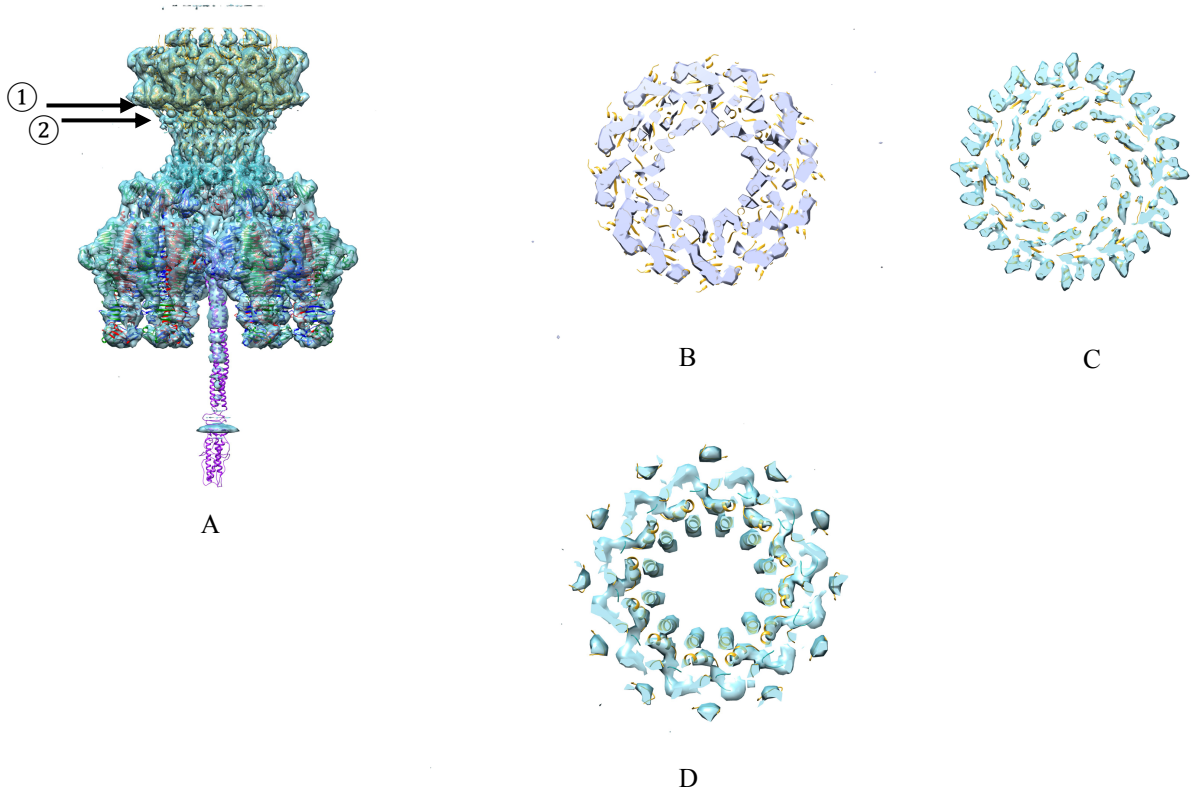


Fig. 3-5. The 5% enlarged cryoEM map fits the x-ray structures better with the region pointed by the arrow ① in *(A)*. The cross-sections at ① are highlighted as *(B)* (with the original map) and *(C)* (with the 5% enlarged map). There are twelve alpha-helices at the region ② deviates from the 5% enlarged density map as shown in *(D)*.

To examine the structural role of gp8 in the tail machine, after docking of the x-ray structures of the other four tail components, the unassigned volume of the complete tail machine was segmented out and the gp8 SAXS model was docked inside this volume (**Fig. 3-6A-D**). A good fit was observed, confirming the quality of the structural model obtained. The docking revealed versatile interactions between gp8 and the tail spikes, tail adaptor and tail needle. Furthermore, 6-fold symmetry was imposed to the SAXS model and the resulting hexameric model was also docked inside of the segmented cryoEM volume (**Fig. 3-6E&F**). This docking directly proved the *in vivo* hexameric state of gp8 at the first time and showed the boundaries between each gp8 monomer.

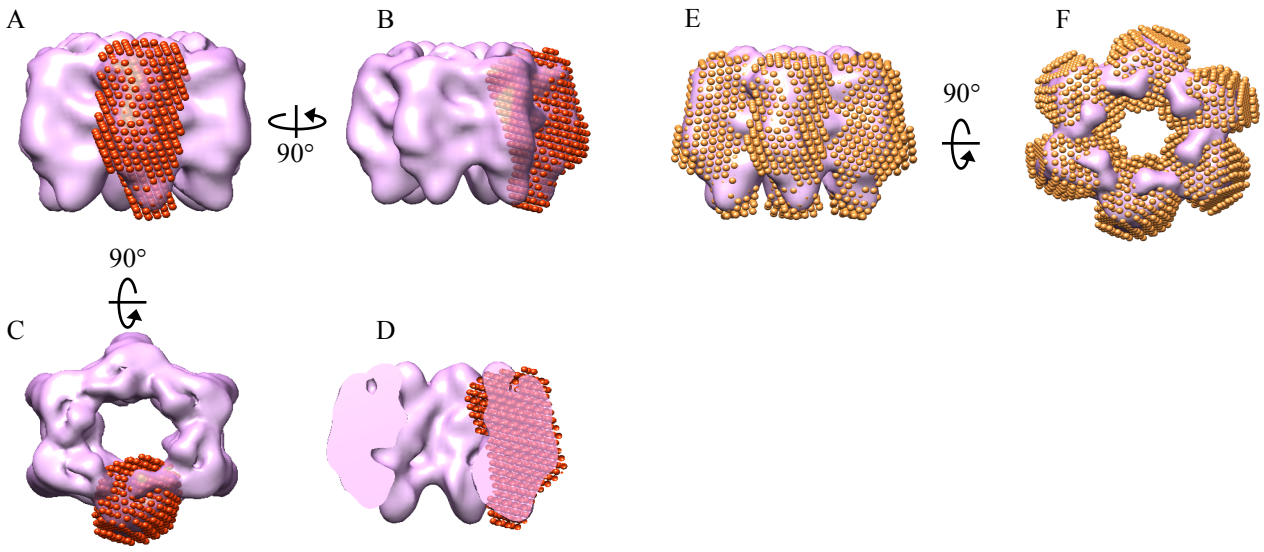


Fig. 3-6. Docking of the Sf6 gp8 SAXS model into the cryoEM map. *(A-D)* The bead model of gp8 monomer (red) is superposed on the segmented cryoEM map of the P22 tail nozzle hexamer (orchid). The map volume is adjusted to 399.8x10^{3}Å with the contour level of 0.849. The front, side, top and cutaway views are showed in *(A)*, *(B)*, *(C)* and *(D)*, respectively. *(E&F)* The hexameric bead model of gp8 monomer (gold) is generated by application of six-fold symmetry and is docked into the same cryoEM map as in panels *(A-D)*. The front and top views are showed in E and F, respectively.

3.3.5 Interactions between gp8 and the tail spike

The tail spike is homotrimeric and each monomer has two domains, the HBD and the RBD. The cryoEM reconstruction shows that gp8 closely interacts with the tail spike, mainly with the HBDs. As shown in the docking model, the tail spike hexamer is about 30° rotated around Z axis compared with gp8 hexamer (**Fig. 3-7**), which leads to the interlaced arrangement of these blocks. Each gp8 monomer interacts with two tail spikes on the left and right sides, respectively, acting like glue to help hold the six tail spikes together. When viewed closely, gp8 displays different modes of interaction with the HBDs and RBDs.

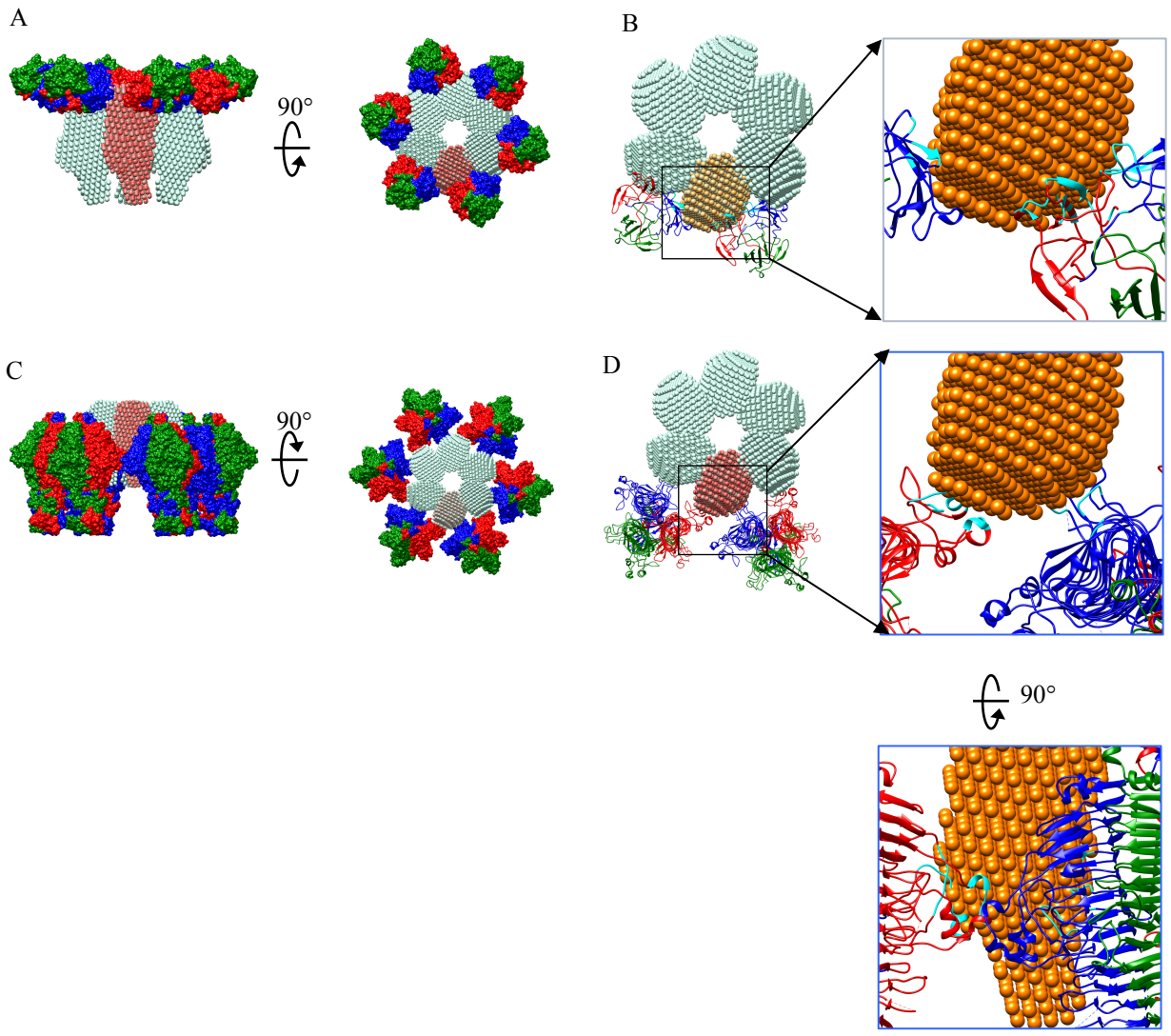


Fig. 3-7. Interaction of gp8 with the tail spike protein. *(A)* The gp8 hexamer (shown as hexmeric bead model) is located interlaced with respect to the HBDs of tail spikes (molecular surface, colored in red, green and blue by chains). Front and top views are shown on the left and right, respectively. *(B)* One gp8 monomer (bead model) binds two HBD trimers (ribbon, colored in in red, green and blue by chains). Regions on the HBDs interacting with gp8 are colored in cyan. *(C)* The gp8 hexamer (hexmeric bead model) is located interlaced with respect to the RBDs of tail spikes (molecular surface, colored in red, green and blue by chains). Front and top views are shown on the left and right, respectively. *(D)* One gp8 monomer (bead model) binds two RBD trimers (ribbon, colored in in red, green and blue by chains). Regions on the RBDs interacting with gp8 are colored in cyan.

One gp8 monomer interacts with two HBD trimers (**Fig. 3-7A**). In one of the HBD trimers, only one monomer is involved in interaction with gp8. The interactions occur on a beta sheet (amino acid

residues 59-62) and a small loop (amino acid residues 75-76) (**Fig. 3-7B**). In the other HDB trimer, two monomers interact with gp8, including a beta-turn-beta (amino acid residues 51-59) and a loop (amino acid residues 75-80) of the first monomer and a loop (amino acid residues 35-45) and a beta-turn-beta (amino acid residues 97-105) of the second monomer (**Fig. 3-7B**). All of these interacting regions are highly conserved among Sf6, P22 and HK620 except the last beta-turn-beta (**Fig. 3-8**).

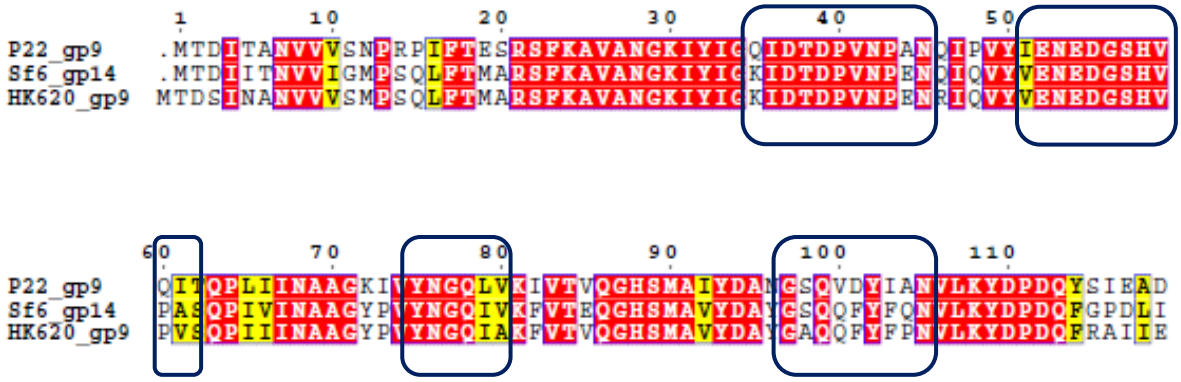


Fig. 3-8. Sequence alignment of HBDs of tail spike orthologs from P22, Sf6 and HK620. Identical amino acids are in red and similar amino acids are in yellow. Regions that are within 7Å to gp8 in the model are indicated with rounded rectangle

Similar to the situation of the HBD, one gp8 monomer interacts with two RBD trimers (**Fig. 3-7C**). However, only one RBD monomer is involved in interaction in each trimer (**Fig. 3-7D**). In one of the interacting RBD monomers, two loops (amino acid residues 311-314 and 368-372) extend closely to gp8. While in the other RBD, a loop (amino acid residues 200-205), a helix –loop (amino acid residues 218-228) and a helix (amino acid residues 248-253) are shown to be located within 7Å to gp8. It was shown that the three-fold axes of the HBD trimers and those of the RBD trimers formed 20°angles (11) and the flexible hinge between the HBD and the RBD was unlikely to support this confirmation. Gabriel C. Lander and his colleagues hypothesized that there should be gp8-tail spike interactions to keep the orientation of the RBDs (2). Our fitting of the gp8 SAXS model supports the hypothesis by giving out relatively concrete interacting regions between gp8 and the RBDs of the tail spikes.

3.3.6 Interactions between gp8 and tail adaptor

The docking of gp8 and the tail adaptor shows that each gp8 monomer aligns with a pair of tail adaptor monomers (**Fig. 3-9A**). The interactions occur between gp8 and a loop of which the amino acid sequence is highly conserved among different phages of the same morphology, such as Sf6, P22 and HK620 (**Fig. 3-10**). However, it seems that the pair of loops are not involved equally in the interaction. Amino acid residues 35-37 of one loop are located within 7Å to gp8 while the other loop has a few more residues involved, 35-40.

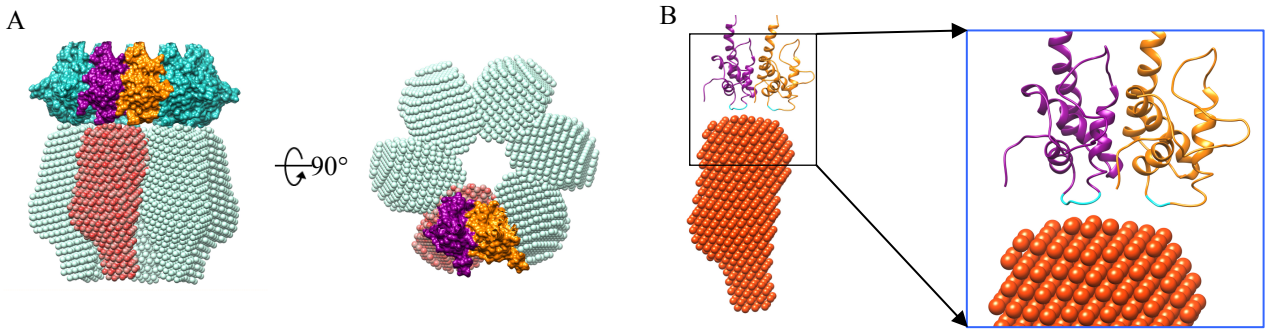


Fig. 3-9. Interaction of gp8 with the tail adaptor protein. *(A)* Symmetry mismatch between gp8 hexamer (shown as hexameric bead model) and the tail adaptor protein (molecular surface, colored in purple and gold for two chains, and cyan for the other ten chains). The 27 amino acid residues at the C-terminus of the tail adaptor are removed for clarity. Front and top views are shown on the left and right, respectively. *(B)* One gp8 monomer (bead model) interacts with two tail adaptor monomers (ribbon, colored in purple and gold) via two loops of the adaptor (cyan).

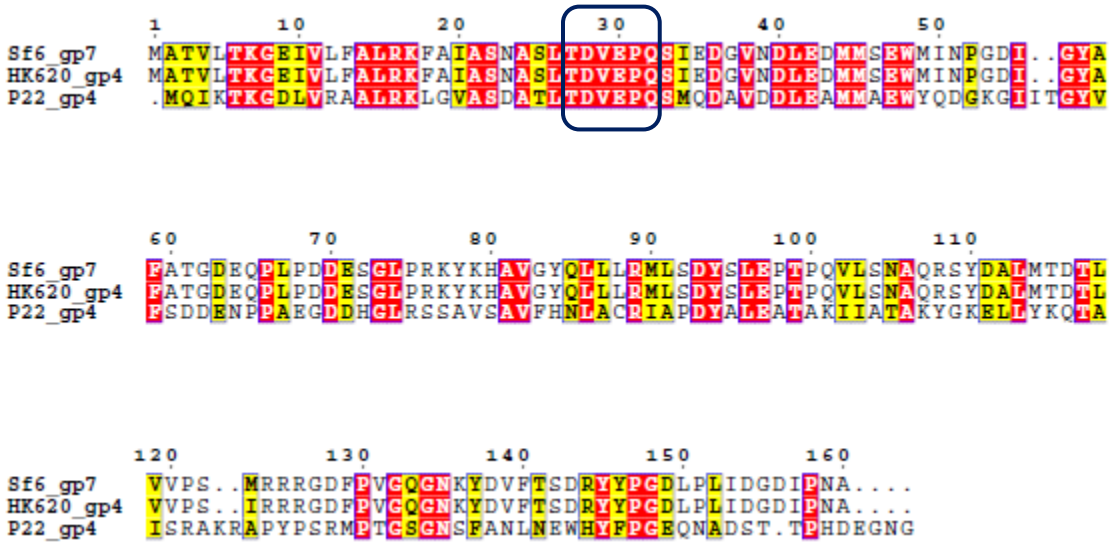


Fig. 3-10. Sequence alignment of the tail adaptor orthologs from Sf6, HK620 and P22. Identical amino acids are in red and similar amino acids are in yellow. Residues that are within 7Å to gp8 in the model are indicated with rounded rectangle.

It was shown previously that the extended C-terminus of the tail adaptor lies wedged between the capsid and portal (12, 24), which was crucial for the tail machine to attach to the capsid after DNA packaging. The 13 amino acid residues at the N-terminus were absent in the crystal structure (7). Since the N-terminus of the adaptor points at gp8, it is likely that the N-terminus extends to gp8, either wedges in the boundaries of gp8 monomers or interacts with other faces of gp8 to increase the stability of the tail machine.

3.3.7 Interactions between gp8 and tail needle

The x-ray structure of the tail needle (10) was fitted into the corresponding cryoEM volume of the tail machine. The needle is about 38 Å in width and 214Å in length (11), of which the 51Å portion at the N-terminus is inserted into the conduit formed by gp8. Gp8 is the only tail protein holding the needle (**Fig. 3-11**), mainly through the protuberance at the lower-middle of the inner cavity of gp8 hexamer. However, our SAXS model does not show this part of density, indicating the possible conformational difference between the isolated gp8 and the assembling gp8 in the tail. This may be due in part to the fact that the needle has to be released by gp8 at early moment of infection to allow the injection of pilot proteins and viral genome.

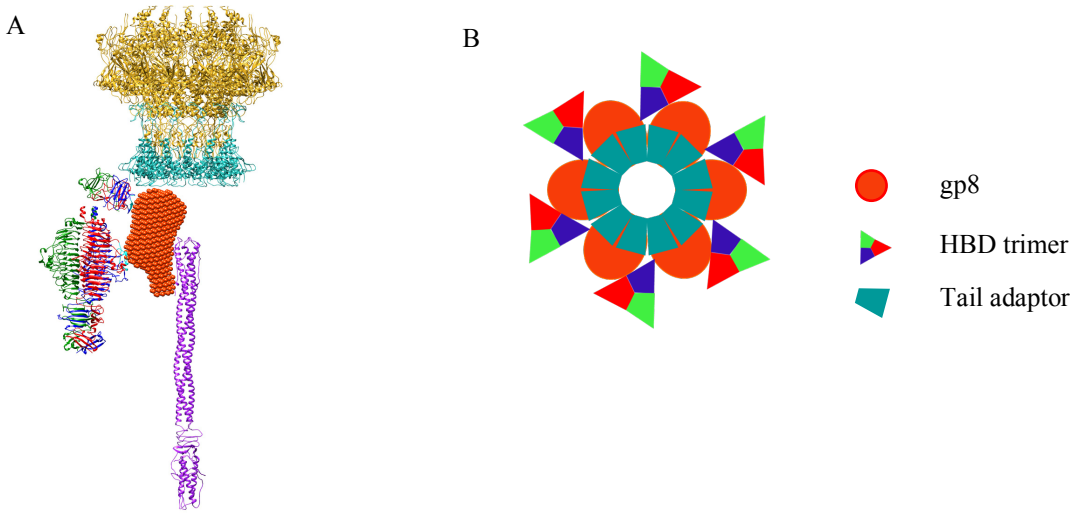


Fig. 3-11. A nexus of intermolecular interactions in the tail mediated by the tail nozzle gp8. *(A)* Available x-ray structures of the portal (gold), the tail adaptor (dark cyan), the tail spikes (RGB), the tail needle (purple) and the gp8 SAXS model (red) are docked into the cryoEM map of tail machine (not shown). *(B)* A schematic shows top view of the interaction nexus between gp8 and tail adaptor, gp8 and HBDs. Protein colors are in consistent with that in panel A.

3.3.8 Implications for roles of gp8 in assembly of the tail machine

We dock all of the five solved x-ray structures as well as the gp8 SAXS into the cryoEM map of the tail machine. Without showing the map, it is clear that gp8 is locating at such a prominent core position from where it integrates the portal-adaptor at the top, the tail spikes on its external surface and the needle on its internal surface (**Fig. 3-11A**). To address the central position of gp8 and its noteworthy role as a tail hub, we generate a top-view schematic containing gp8, the HBDs of tail spikes and the tail adaptor (**Fig. 3-11B**). With this novel recognition of gp8, we would like to further discuss the mode of tail assembly based on some previously proposed models. Strauss and King found that the tail adaptor was the first protein to be attached to the portal after DNA was pumped in, followed by the tail nozzle and needle and the assembly of tail spikes required both the tail adaptor and nozzle (25). This model was further described by Tang assisted by assignment of these proteins onto the cryoEM reconstruction of the tail machine (11). Olia and

his colleagues performed experiment conforming the sequential assembly of the tail adaptor and nozzle to the portal protein and indicating that, most likely, it was the tail nozzle hexamer instead of monomer that was assembled to the portal-adaptor complex (24, 26).

The structure arrangement that one gp8 monomer occupies two tail adaptor monomers enables us to infer the most possible behavior of gp8 during the tail assembly. We analyze that the situations when gp8 binds to the portal: tail adaptor complex one monomer by one monomer or one dimer by one dimer are almost impossible, because both of the two situations could easily result in odd number of tail adaptor monomers skipped. Since monomer, dimer and hexamer are the main forms of gp8 as we discussed, hexamer ends up being the most possible form to be assembled onto the portal: tail adaptor complex. This deduction supports the research which implied that only high oligomeric tail nozzle showed appreciable binding to the portal: adaptor complex (26).

Contributions: Haiyan Zhao and Liang Tang designed research; Lingfei Liang performed research; Lingfei Liang, Haiyan Zhao, and Liang Tang analyzed data. Thanks to Tsutomu Matsui and Thomas M. Weiss for performing the SAXS experiment.

3.4 References

1. Casjens SR & Thuman-Commike PA (2011) Evolution of mosaically related tailed bacteriophage genomes seen through the lens of phage P22 virion assembly. *Virology* 411(2):393-415.
2. Lander GC*, et al.* (2009) The P22 tail machine at subnanometer resolution reveals the architecture of an infection conduit. *Structure* 17(6):789-799.
3. Casjens S*, et al.* (2004) The chromosome of Shigella flexneri bacteriophage Sf6: complete nucleotide sequence, genetic mosaicism, and DNA packaging. *Journal of molecular biology* 339(2):379-394.
4. Parent KN, Gilcrease EB, Casjens SR, & Baker TS (2012) Structural evolution of the P22-like phages: comparison of Sf6 and P22 procapsid and virion architectures. *Virology* 427(2):177-188.
5. Muller JJ*, et al.* (2008) An intersubunit active site between supercoiled parallel beta helices in the trimeric tailspike endorhamnosidase of Shigella flexneri Phage Sf6. *Structure* 16(5):766-775.
6. Bhardwaj A, Molineux IJ, Casjens SR, & Cingolani G (2011) Atomic structure of bacteriophage Sf6 tail needle knob. *The Journal of biological chemistry* 286(35):30867-30877.
7. Olia AS, Prevelige PE, Jr., Johnson JE, & Cingolani G (2011) Three-dimensional structure of a viral genome-delivery portal vertex. *Nature structural & molecular biology* 18(5):597-603.
8. Steinbacher S*, et al.* (1997) Phage P22 tailspike protein: crystal structure of the head-binding domain at 2.3 A, fully refined structure of the endorhamnosidase at 1.56 A resolution, and the molecular basis of O-antigen recognition and cleavage. *Journal of molecular biology* 267(4):865-880.

9. Steinbacher S, *et al.* (1996) Crystal structure of phage P22 tailspike protein complexed with Salmonella sp. O-antigen receptors. *Proceedings of the National Academy of Sciences of the United States of America* 93(20):10584-10588.
10. Olia AS, Casjens S, & Cingolani G (2007) Structure of phage P22 cell envelope-penetrating needle. *Nature structural & molecular biology* 14(12):1221-1226.
11. Tang L, Marion WR, Cingolani G, Prevelige PE, & Johnson JE (2005) Three-dimensional structure of the bacteriophage P22 tail machine. *The EMBO journal* 24(12):2087-2095.
12. Tang J, *et al.* (2011) Peering down the barrel of a bacteriophage portal: the genome packaging and release valve in p22. *Structure* 19(4):496-502.
13. Lander GC, *et al.* (2006) The structure of an infectious P22 virion shows the signal for headful DNA packaging. *Science* 312(5781):1791-1795.
14. Martel A, Liu P, Weiss TM, Niebuhr M, & Tsuruta H (2012) An integrated high-throughput data acquisition system for biological solution X-ray scattering studies. *J Synchrotron Radiat* 19(Pt 3):431-434.
15. Matsui T, *et al.* (2014) Structural basis of the pH-dependent assembly of a botulinum neurotoxin complex. *J Mol Biol* 426(22):3773-3782.
16. McPhillips TM, *et al.* (2002) Blu-Ice and the Distributed Control System: software for data acquisition and instrument control at macromolecular crystallography beamlines. *J Synchrotron Radiat* 9(Pt 6):401-406.
17. Konarev PV, Volkov VV, Sokolova AV, Koch MHJ, & Svergun DI (2003) PRIMUS: a Windows PC-based system for small-angle scattering data analysis. *J Appl Crystallogr* 36:1277-1282.
18. Petoukhov MV, Konarev PV, Kikhney AG, & Svergun DI (2007) ATSAS 2.1 - towards automated and web-supported small-angle scattering data analysis. *J Appl Crystallogr* 40:S223-S228.
19. Pettersen EF, *et al.* (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* 25(13):1605-1612.
20. Robert X & Gouet P (2014) Deciphering key features in protein structures with the new ENDscript server. *Nucleic acids research* 42(Web Server issue):W320-324.
21. Larkin MA, *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21):2947-2948.
22. Goujon M, *et al.* (2010) A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic acids research* 38(Web Server issue):W695-699.
23. McWilliam H, *et al.* (2013) Analysis Tool Web Services from the EMBL-EBI. *Nucleic acids research* 41(Web Server issue):W597-600.
24. Olia AS, *et al.* (2006) Binding-induced stabilization and assembly of the phage P22 tail accessory factor gp4. *Journal of molecular biology* 363(2):558-576.
25. Strauss H & King J (1984) Steps in the stabilization of newly packaged DNA during phage P22 morphogenesis. *Journal of molecular biology* 172(4):523-543.
26. Olia AS, Bhardwaj A, Joss L, Casjens S, & Cingolani G (2007) Role of gene 10 protein in the hierarchical assembly of the bacteriophage P22 portal vertex structure. *Biochemistry* 46(30):8776-8784.

Chapter 4. Structure of Internal Protein gp12 Sheds Light on the Assembly of Podovirus DNA-Injection Apparatus

4.1 Introduction

Genetic material internalization is a key step for viral infection and propagation. Tailed double-stranded DNA (dsDNA) bacteriophages use a sophisticated tail complex to inject the dsDNA to host cytoplasm, leaving the empty capsid outside the cell wall. While myoviruses use a long contractile tail (1) and siphoviruses co-opt host membrane proteins (2) to overcome the barrier of cell envelope, podoviruses with a short tail use a more complex mechanism which involves in tail extension by internal proteins (3).

In 1977, Vance Israel first discovered that after the podovirus P22 adsorbed to *Salmonella* cells, three proteins gp7, gp20, and gp16 (now called internal proteins or DNA-injection proteins) were released from the virion (4). Phages with null mutants in any of these genes assemble normal-appearing virions that do not deliver their DNA properly (5, 6). Gp16 was proved to be essential for establishment of successful infection (7). Furthermore, gp16 partitioned into the extracted host cytoplasmic membrane and was able to mediate the active transport of P22 DNA across the membrane (8). Yan Jin et al. recently proved that, with the presence of receptors LPS and OmpA, P22 ejects the three DNA-injection proteins ahead of DNA ejection, suggesting a key role of these proteins in transferring P22 DNA during infection (9). Podovirus T7 is another prototype that has been extensively studied on its internal proteins (gp14, gp15, and gp16) and DNA injection. Similar with P22, the three T7 internal proteins are injected into host cells before DNA injection (10). Gp16 has a lytic transglycosylase motif at the N-terminal domain, essential for degrading the peptidoglycan under suboptimal conditions of growth (11). The gp16 C-terminal charged residues are required for viral genome entry and gp16 cooperates with gp15 in translocating DNA into cells (12). Gp15 and gp16 form a spiral ring complex *in vitro* (13). Gp15 binds to DNA and gp16 binds liposome, thus the gp15-gp16 complex binds both DNA and liposomes (13). Analysis on isolated outer and inner membranes of T7-infected cell suggested that gp14 associates with the outer membrane, whereas gp15 and gp16 form

a complex associated with both the outer and inner membranes (12). Hu and coworkers used T7 to infect *Escherichia coli* minicells, and observed an extended tail spanning the cell envelope using cryogenic-electron tomography (3), potentially formed by internal proteins. Similar elongated density spanning the host cell envelope was also observed in podovirus ε15 (14). However, both tail extensions were observed at relatively low resolution. To date, there is no available high-resolution structure of any of these internal proteins. Therefore, it is impossible to know the detailed structural characteristics and component organization of the extended tail.

The tail channel is presumably the only means of exit of the virus internal proteins and DNA. Podoviruses T7 and P22 have an average inner diameter of ~40Å throughout the portal and tail tube with a same constriction of 33Å at the portal (15-19). With such a constriction and the partial specific volume of 0.73 cm^3/g (20), we can calculate that a globular protein that could pass through should has a molecular weight of no more than 15.7kDa. This number is far smaller than all the known internal proteins of T7, P22, and Sf6, suggesting that these proteins must unfold in order to exit through the portal. A notable feature of these internal proteins is that they are all predicted to have high content of α-helices. Without any specific structural information, it was hypothesized that, like opening a carpenter's ruler where each segment of the ruler corresponds to an α-helix, these proteins open their tertiary structures to form a linear string of α-helices (21). This hypothesis well matches with the structural constriction of 33Å since the typical diameters of B-form DNA and α-helix are 20Å and 12Å respectively. The same constriction in T7 and P22 may represent the minimum diameter that accommodates one DNA duplex and one α-helix. Yet, experimental evidence is in need to support this hypothesis. When internal proteins arrive at the cell envelope, the α-helices should refold quickly into tertiary and quaternary structures. T7 internal proteins gp15 and gp16 have been observed to have relatively low thermal transition Tm and they can refold from partial unfolded states (13). Sf6 internal protein gp12 can also refold after urea denaturation (22).

Different podoviruses may have various forms of internal proteins inside the capsid. To our knowledge, T7, K1E, and P-SPP7 are the only podoviruses who have an easily distinguishable internal core

consisting of stacked rings of the three DNA-injection proteins (19, 23-25). The majority of podoviruses have undefined internal proteins, such as P22, Sf6, T4 (17, 26, 27). Recently Wu et al used Bubblegram Imaging (28) to study the location of the three proteins inside P22 virion and found that the three proteins were loosely clustered around the proximal end of the portal barrel (16), likely binding to the portal protein. For podoviruses like T7, the internal proteins must transit from the folded state inside the capsid to a partially unfolded state through the tail channel, then again to a folded state in the cell envelope. Whereas for other podoviruses, the internal proteins are likely partial unfolded inside the capsid. They may or may not undergo further unfolding to pass through the tail channel. But same with T7 internal proteins, they must refold rapidly in cell envelope for DNA delivery. To understand these states and processes, structures are of great significance.

Here we report the 2.8Å-resolution crystal structures of Sf6-gp12NTD and P22-gp20NTD, representing the first high-resolution structure of podovirus DNA-injection proteins. This structure serves as strong experimental support for the "carpenter's ruler" hypothesis, providing insight into the mechanism of protein folding and unfolding during the assembly of DNA-injection apparatus.

4.2 Methods

4.2.1 Protein expression and purification

Sf6-gp12NTDs and P22-gp20NTDs

Since gp12 was overexpressed in *E.coli* as inclusion bodies, we decided to work on the NTD and CTD separately. Sequence alignment between gp12 and the orthologs suggests that the two halves of gp12 have distinct conservation. Therefore, we cut the gp12 sequence in the middle after residue 211 which is the end of a predicted long alpha-helix to make the NTD and CTD construct. The two predicted transmembrane helices at the N-terminus were excluded from the NTD construct since they could largely decrease protein solubility. Three Sf6-gp12NTD constructs and their P22-gp20 orthologs were made for crystallization trials. They are Sf6-gp12(65-183)/P22-gp20(67-186), Sf6-gp12(65-211)/P22-gp20(67-214), and Sf6-gp12(65-221)/P22-gp20(67-224). The DNA fragment encoding each construct was cloned into

pET28b (Novagen) between *Nde1*/*Xho1* with an N-terminal 6xhis tag. Proteins were overexpressed in *E. coli* strain B834(DE3) with the induction of 1mM IPTG at 30°C for 3 hours. They were first purified by Ni affinity column and their homogeneity was analyzed by size exclusion chromatography (SEC) on Superdex 200. Sf6-gp12(65-183) and P22-gp20(67-186) showed a single peak corresponding to a monomer based on globular protein markers. Sf6-gp12(65-211/221) and P22-gp20(67-214/224) showed a large peak of oligomer and a much smaller peak of monomer, suggesting that these proteins spontaneously form a single species of oligomer in solution.

Sf6-gp12CTDs

The Sf6-gp12(211-431) and Sf6-gp12(211-404) were expressed and purified using the same protocol as above, except with SEC superdex 75. Both proteins showed a single peak as monomer.

Sf6-gp13, P22-gp20, and P22-gp16

The protocol of expression and purification is same as that of gp12-NTDs. P22-gp20 was partially soluble. The soluble portion in supernatant after centrifugation of cell lysate was purified and used in co-purification assay. Different with Sf6-gp12, P22-gp20 exists as monomer in solution. Sf6-gp13 and P22-gp16 are orthologs. Both proteins exist mainly as monomer in solution. The monomer SEC peak was pooled and used in protein-protein interaction assay.

4.2.2 Crystallization, X-ray data collection and crystallographic analysis

Sf6-gp12(65-211) was crystallized in 1.26M $(NH_4)_2SO_4$, 0.1 M MES pH 6.0. P22-gp20(67-214) was crystallized in 10% PEG 8,000, 0.1 M Sodium Acetate Trihydrate pH 5.0, 0.1 M $MgCl_2$, 0.4M NaCl. Selenomethionine gp20(67-214) was crystallized in 3% PEG 8,000, 0.1 M Sodium Acetate Trihydrate pH 4.0, 0.1 M $MgCl_2$. Data collection and processing was done with APS 23ID-D and SSRL 12-2. The structure of gp20(67-214) was solved by Single-wavelength Anomalous Diffraction (SAD) via CCP4-CRANK2(29). The poly-Alanine model of gp20(67-214) structure was then used as a template to solve gp12(65-211) structure via Phenix Autosol(30, 31). Both structures were refined using Phenix and Coot(32).

4.2.3 Negative staining EM

Protein sample (0.3 mg/ml for Sf6-gp12(65-211); 0.03mg/ml for P22-gp20(67-214); 9mg/ml for

Sf6 DNA-injection apparatus) was applied onto an EM grid for one minute, after which the excess moisture was wicked off with a piece of filter paper and the grid was rinsed in a protein buffer drop and a 1% Uranyl Acetate drop and dried. Lastly, the grid was soaked in a drop of 1% Uranyl Acetate for one minute before it was blotted dry. The samples were examined under the FEI Field Emission Transmission Electron Microscope (TEM) operating at 200kV.

4.2.4 Characterization of in vitro molecular interaction

All the proteins were purified with a N-terminal 6xhis tag. To study the protein-protein interactions using Ni-NTA, all the gp12 truncations had the tag retained while the tag on gp13 was removed by thrombin digestion. In the pull-down assay, a gp12 protein was incubated with gp13 for one hour, followed by Nickel affinity beads purification. The mixture, the wash fractions and eluent were loaded on a SDS-PAGE to check whether the no-tag gp13 was eluted with gp12 truncations or not. In the SEC co-purification assay, gp12-CTD and gp13 were incubated for 1hr before loaded on a size-exclusion column (Superdex 200 10/300GL). The chromatogram was then overlaid with those of pure gp12-CTD and gp13. A peak shift to left would suggest the formation of a complex. The SEC co-purification experiment for P22-gp20 and gp16 was conducted in a similar way.

4.2.5 In vivo assembly and isolation of phage Sf6 DNA-injection apparatus

M94 cells were cultured to OD_{600}=0.6 and pelleted. Cells were resuspended in 37°C PBS buffer containing 1mM $MgCl_2$. Purified Sf6 (protocol from (33)) was added with a MOI of 500. The culture was incubated at 100rpm 37°C for 15min, then cells were pelleted and resuspended in lysis buffer (20mM Tris-HCl pH 7.4, 150mM NaCl). The cells were treated with 0.2mg/ml lysozyme and 2mM EDTA (pH8). Spheroplasts were achieved by 45min incubation on ice and 15min in 37°C water bath, followed by lysis with 1mM PMSF and 1% LDAO (43.6mM). As the lysis near completion, 4mM $MgCl_2$ was added to active DNase for degradation of DNA. Additional 4mg DNaseI was gradually added to assist DNA degradation. Within 10min after addition of DNaseI, 10mM EDTA (pH8) was added to prevent re-aggregation of cell membrane and walls. Cell debris was removed. Phage particles were then collected by high-speed centrifugation at 26,892g, 4°C for 4 hours. The pellet was resuspended in lysis buffer containing 1mM

$MgCl_2$ and 1.4mM LDAO (10CMC) for overnight. Insoluble residue was removed by simple centrifugation and the prep was then examined with negative staining EM for phage particles associated with the DNA-injection apparatus.

4.3 Results and discussion

4.3.1 Sf6 assembles extended tail for DNA injection

To examine if Sf6 forms a similar tail extension as observed in T7 and ε15, we designed *in vivo* experiment to induce the assembly of the DNA-injection apparatus by allowing Sf6 to infect *Shigella flexneri* M94 cells. Sf6 genome entry normally occurs in a few minutes (faster than 10 minutes by time-lapse fluorescence microscopy (34)). We varied the incubation period of M94 with Sf6 from 5 to 25min and found that 15min-incubation resulted in most particles with extended tails examined by negative staining EM (nsEM).

Viruses were visualized at distinct stages of the DNA injection process—for example, (i) an intact phage before DNA injection, (ii) a part-empty particle with extended tail, and (iii) a ghost with long tail or (iv) short tail (the extended tail was disassembled) (**Fig. 4-1A**). Extended tails were apparent in the electron micrographs compared with the original tail. Six representative particles with outstanding extended tails are shown in **Fig. 4-1B**. These tails are at such a good contrast that we can directly measure their dimensions. They have an overall tail length of ~420Å with an inner diameter of ~32Å. The original length of Sf6 particle is about 200Å from the vertex of the capsid to the end of tail spikes. The central tail tube formed by the portal and two tail proteins gp7 and gp8 is shorter, about 120Å beyond the capsid vertex (17). The increased length of tail tube (~300Å) is comparable with those observed in T7 (450Å) and ε15 (200Å), and the inner diameter is close to the constriction of the portal ring (33Å) (16). Such tail extension could easily accommodate a DNA duplex.

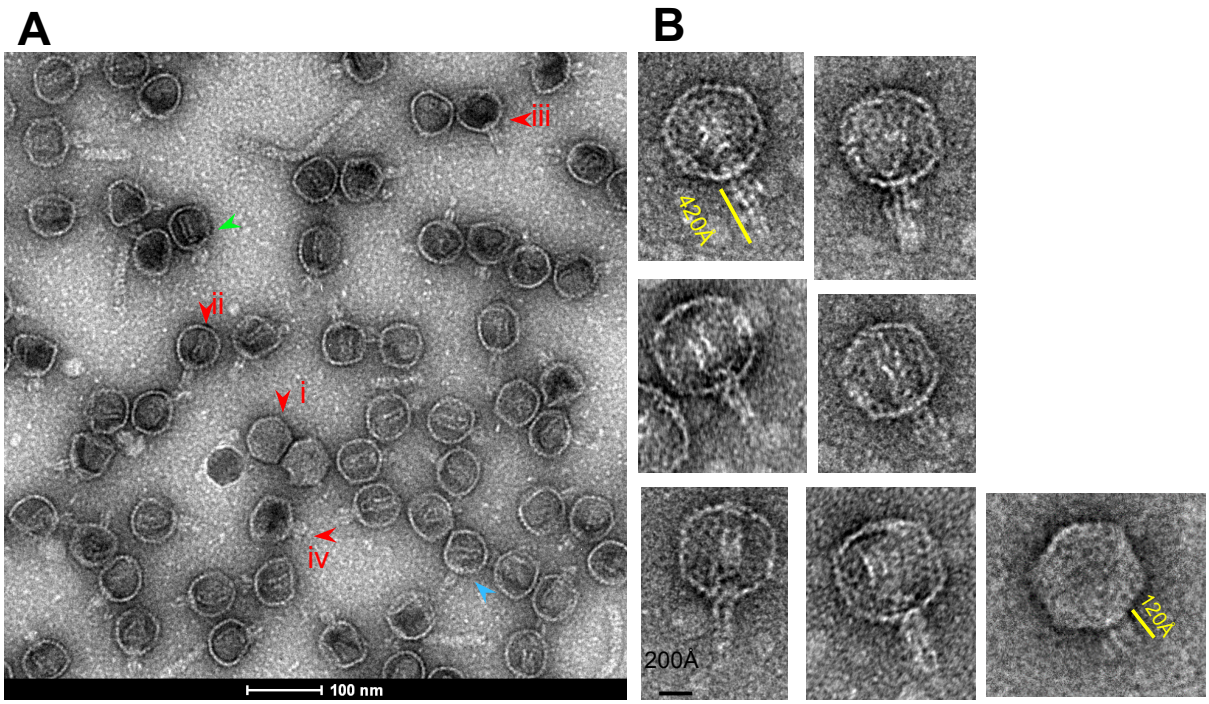
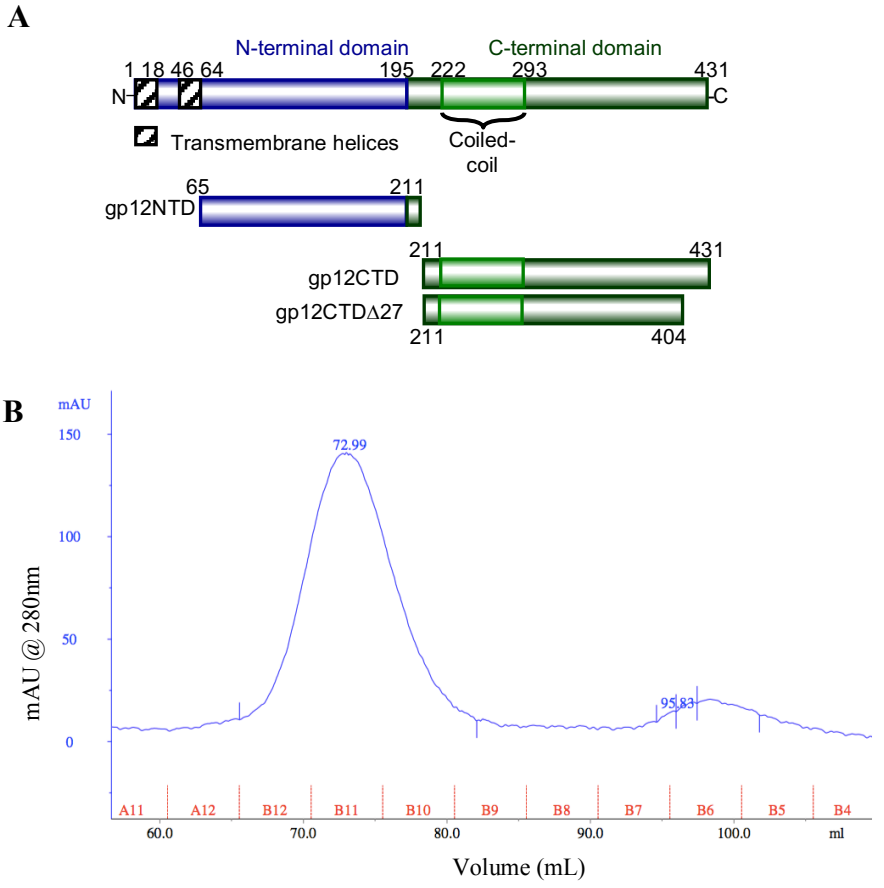


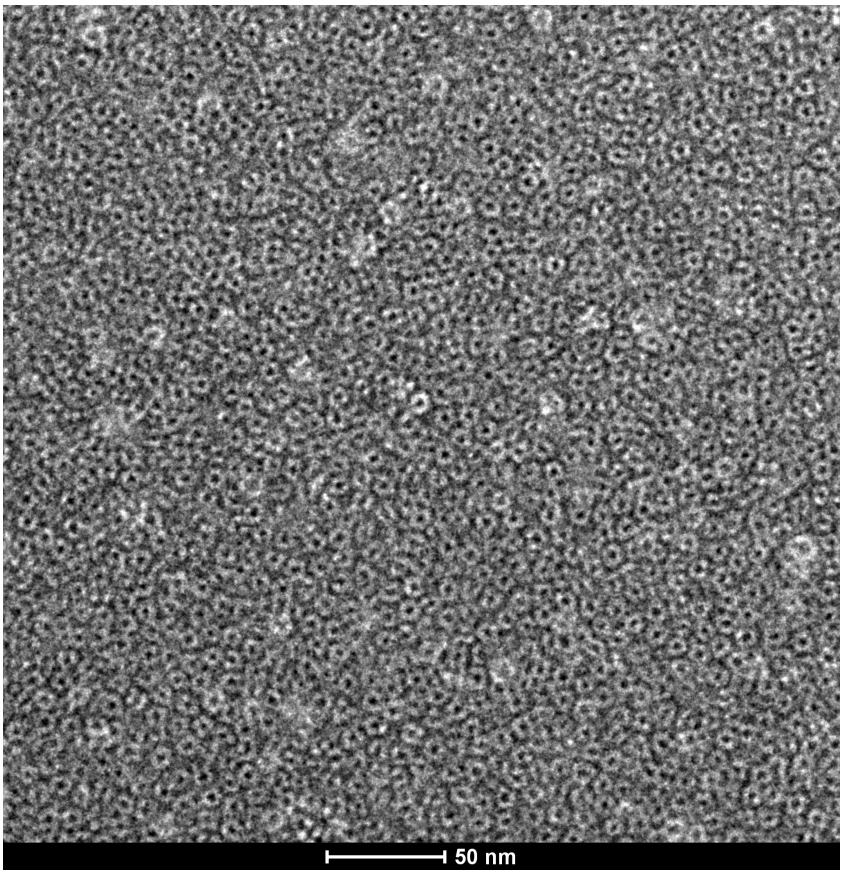
Fig. 4-1. Negative staining electron microscopic (nsEM) analysis of Sf6 DNA-injection apparatus. *(A)* An electron micrograph showing: (i) an intact phage before DNA injection; (ii) a part-empty particle with extended tail; (iii) a ghost with long tail; (iv) a ghost with short tail. Most *in situ* tubes are parallel to the tail (green arrow); few are at an angle (cyan arrow). *(B)* Six representative Sf6 particles with extended tail and a particle with its short tail before DNA injection

4.3.2 Gp12NTD/gp20NTD spontaneously assembles into a ring-like complex

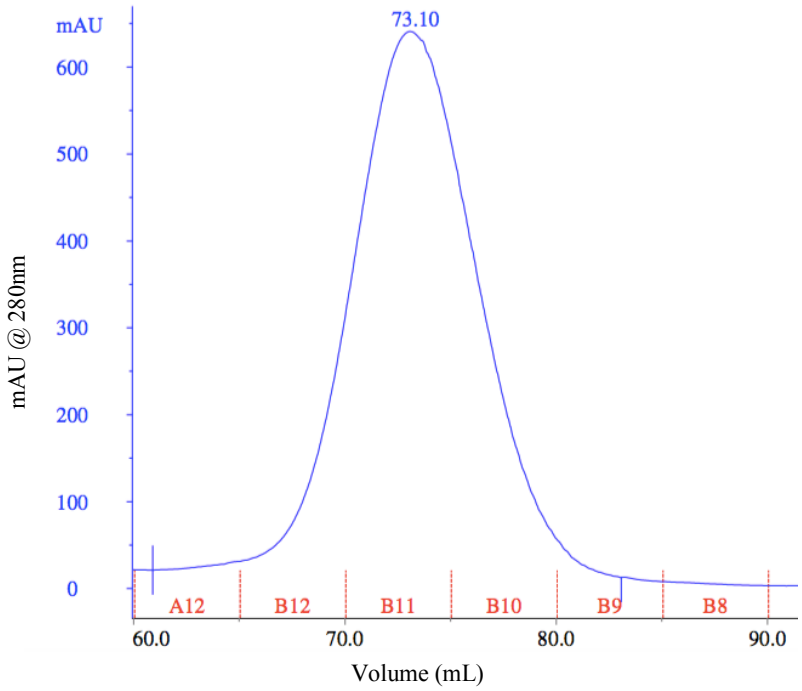
The Sf6-gp12 protein contains two domains as we learned from the 3D EM reconstruction (22). It has two predicted transmembrane helices at the N terminus and a coiled-coil structure in the middle of the sequence (**Fig. 4-2A**). We designed three constructs in the N-terminal domain without the transmembrane helices which could largely decrease protein solubility, gp12(65-183), gp12(65-211), and gp12(65-221). The gp12(65-211) (referred to as gp12NTD) and gp12(65-221) both form an oligomer in solution (**Fig. 4-2B**) while the short construct gp12(65-183) exists as a monomer regardless of concentration, suggesting that residues 184-211 are required for protein homo-oligomerization. The gp12NTD sample was examined with nsEM at the concentration of 0.3mg/ml. A remarkable ring-like structure with a central

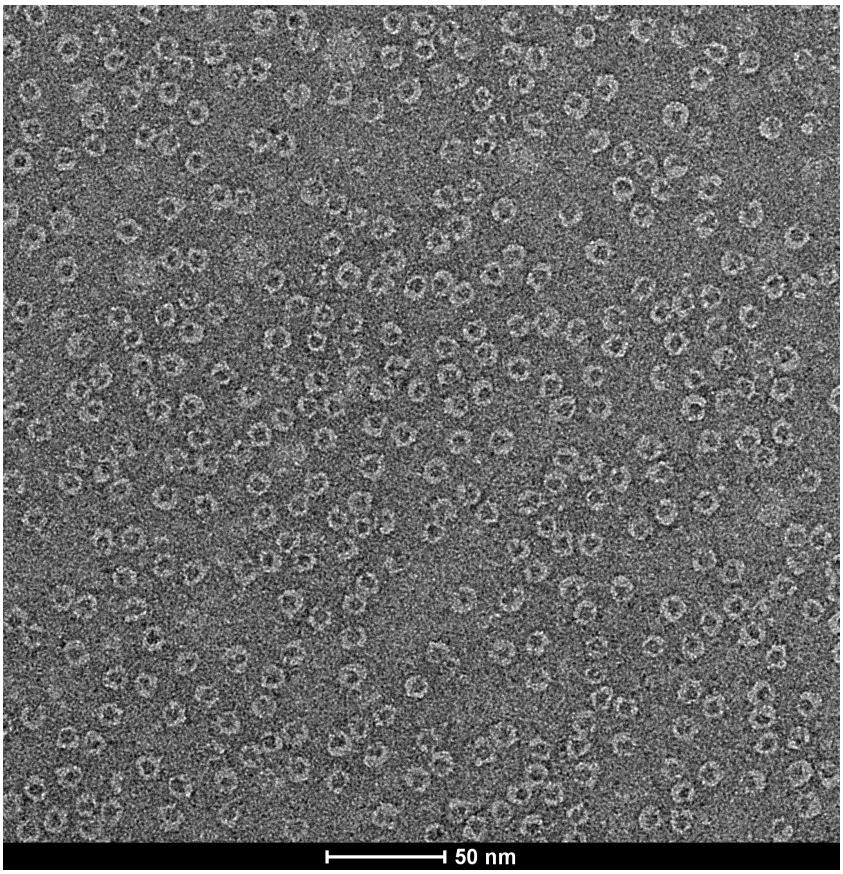
hole was observed as shown in **Fig. 4-2B**. The gp12 orthologue P22-gp20 was analyzed in the same way. Similarly, gp20NTD (residues 67-214) and gp20(67-224) (corresponding in sequence to gp12NTDand gp12(65-221) respectively) both form an oligomer in solution (**Fig. 4-2C**) while gp20(67-186) exists as a monomer. Gp20NTD forms a similar ring-like structure shown by nsEM (**Fig. 4-2C**). 2D classification and averaging analysis by EMAN2 (35) and RELION (36) resulted in classes predominantly representing the view down the ring axis of the particle. Close inspection revealed that this ring-like structure has two stacked concentric rings (**Fig. 4-2D** left). Roughly, the inner diameter of the whole ring is 36Å, and the outer diameter of the small and big ring is 75Å and 106Å respectively. By comparing the diameters of the P22-gp20NTD class-averages with those of the EM density of the full-length gp12 (**Fig. 4-2D** right), we found that the P22-gp20NTD resembles the structure of the gp12 crown domain. Considering the high sequence identity (42.9%) between gp12NTD and gp20NTD (**Fig. 4-5A**), we can safely assume that the gp12NTD corresponds to the crown domain and thus the CTD corresponds to the stem domain.




50 nm

C


mAU
600
500
400
300
200
100
0
mAU @ 280nm
73.10
A12
B12
B11
B10
B9
B8
60.0
70.0
80.0
90.0
Volume (mL)



D

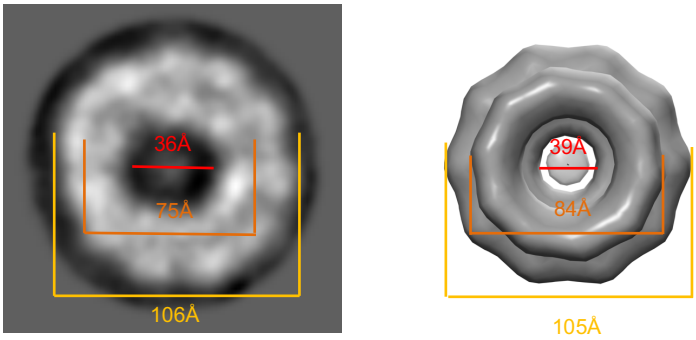


Fig. 4-2. Gp12NTD forms a ring assembly resembles the crown domain of the full-length EM reconstruction. *(A)* A schematic diagram of gp12 structure prediction and the three constructs. Gp12(65-183) and gp12(65-221) are not shown here. (*B)* Size exclusion chromatograph shows that most gp12NTD molecules exist as a oligomer in solution. The nsEM reveals that it forms a ring structure. *(C)* Gp20NTD also exist as an oligomer in solution. The nsEM reveals that it forms a ring structure. *(D)* The dimensions of gp20NTD (left, a representative 2D class averaging) suggest that the NTD may correspond to the crown domain. On the right is the top view of the crown domain of gp12 decamer

along the 10-fold axis (EMDB 6330).

4.3.3 The gp12NTD X-ray structure reveals a helical protein fold

To gain further insights into the structural properties and the intermolecular interactions that stabilize the oligomeric gp12, we tried to determine its X-ray structure. Complete diffraction data to 2.8Å resolution were obtained using the amino-terminal domain spanning residues 65-211 (gp12NTD). Unfortunately, experimental phasing with selenomethionine and heavy atom derivatives proved ineffective. Alternatively, gp20NTD structure was first phased by selenomethionine and was then used as the template to solve gp12NTD structure by molecular replacement (**Table 4-1**).

Table 4-1. X-ray data collection and structure refinement statistics of Sf6 gp12NTD and P22 gp20NTD

Data collection	**Sf6-gp12N**	**P22-gp20N**	
	Native	Native	Se-Met
Beamline	APS 23ID-D	SSRL 12-2	SSRL 12-2
Wavelength (Å)	0.9794	0.97934	0.97934
Resolution (Å)	78.82-2.84 (2.93-2.84)	39.32-2.85 (2.93-2.85)	39.22-3.01 (3.10-3.01)
No. Measurements	159,145 (13,645)	397,219 (28,068)	342,870(31,662)
Unique reflections	50,789 (4,366)	58,240 (4,358)	49,725 (4,575)
Completeness (%)	98.6 (99.0)	99.6 (96.6)	99.2 (99.3)
I/σ	9.6 (1.3)	11.3 (1.6)	10.6 (2.3)
R_{merge} (%)**	7.2 (82.8)	10.7 (172.4)	11.9 (90.9)
Space group	$P2_12_12$	$P4_2$	$P4_2$
Unit cell (Å)	*a*=159.63 *b*=159.65, *c*=85.40	*a*=165.23, *b*=165.23, *c*=92.88	*a*=165.77, *b*=165.77, *c*=92.44
Structure refinement			
Resolution (Å)	75.09-2.84	39.32-2.85	
R_{work}/R_{free}[a]	0.231/0.268	0.217/0.255	
Number of atoms			
Protein/Water	12,180/0	12,222/0	
B-factors			
Protein/Water	110.1/0	108.9/0	
R.m.s deviations			
Bond lengths (Å)	0.001	0.001	
Bond angles (°)	0.367	0.381	
Ramachandran plot			
Most favored (%)	98.72	98.86	
Allowed (%)	1.22	1.14	
Disallowed (%)	0.06	0	

*Values in the parentheses are for the outermost resolution shells.

**$R_{merge}=\Sigma_{hkl}\Sigma_i|I_i(hkl)-\langle I(hkl)\rangle|/\Sigma_{hkl}\Sigma_i I_i(hkl)$, where $I_i(hkl)$ is the observed intensity of reflection hkl and $\langle I(hkl)\rangle$ is the averaged intensity of symmetry-equivalent measurements

[a] $R_{work}=\Sigma_{hkl}||F_{obs}|-|F_{calc}||/\Sigma_{hkl}|F_{obs}|$, where F_{obs} and F_{calc} are structure factors of the observed reflections and those calculated from the refined model, respectively. R_{free} has the same formula as R_{work}, except that it was calculated against a test set of the data that was not included in the refinement

Gp12NTD crystallized in the space group $P2_12_12$, with eleven molecules in the crystallographic asymmetric unit related by a 11-fold symmetry in a mushroom-like shape (**Fig. 4-3A**). The undecamer has ~100Å external diameter and ~80Å height (**Fig. 4-3B**). The central channel constricts from about 42Å at the cap to about 23Å at the stem. The outer surface of the ring has remarkably negative electrostatic potential except at the opening of the cap, while the overall negatively charged surface of the interior is segmented by several positive residues, R107 and K108 at the N-terminus followed by K193 and R203 (**Fig. 4-3B**). Gp12NTD monomer consists of eight α-helices, forming a three-helix bundle and a four-helix bundle sharing the long helix α3 (**Fig. 4-3C**). The three long helices α1, α3, α8 and one short helix α6 share the same amino to carboxyl end orientation while all the other short helices have an opposite orientation. Likely due to molecular packing in crystals, ten out of eleven chains were less well defined at the loops connecting helices α3-α4 and α6-α7.

Gp20NTD has highly conserved protein fold with gp12NTD (**Fig. 4-3D-F** and **4-5A**). Superimposition of the two molecules results in a RMSD of 1.853Å between 144 CA atom pairs. Visual examination found the largest deviation at the loop between α6 and α7, which, in gp20NTD, tilts a little bit away from the α3 and α4 (**Fig. 4-3G**). Since this region in gp12NTD structure is not well defined and it does not interact with other regions in the undercameric assembly, we do not consider such deviation as a significant conformational difference between the two structures. Gp20NTD undecamer has similar dimensions with those of gp12NTD (**Fig. 4-3E**). The distribution of electrostatic potential of gp20NTD undecamer is similar with that of gp12NTD, but not as prominent (**Fig. 4-3E**).

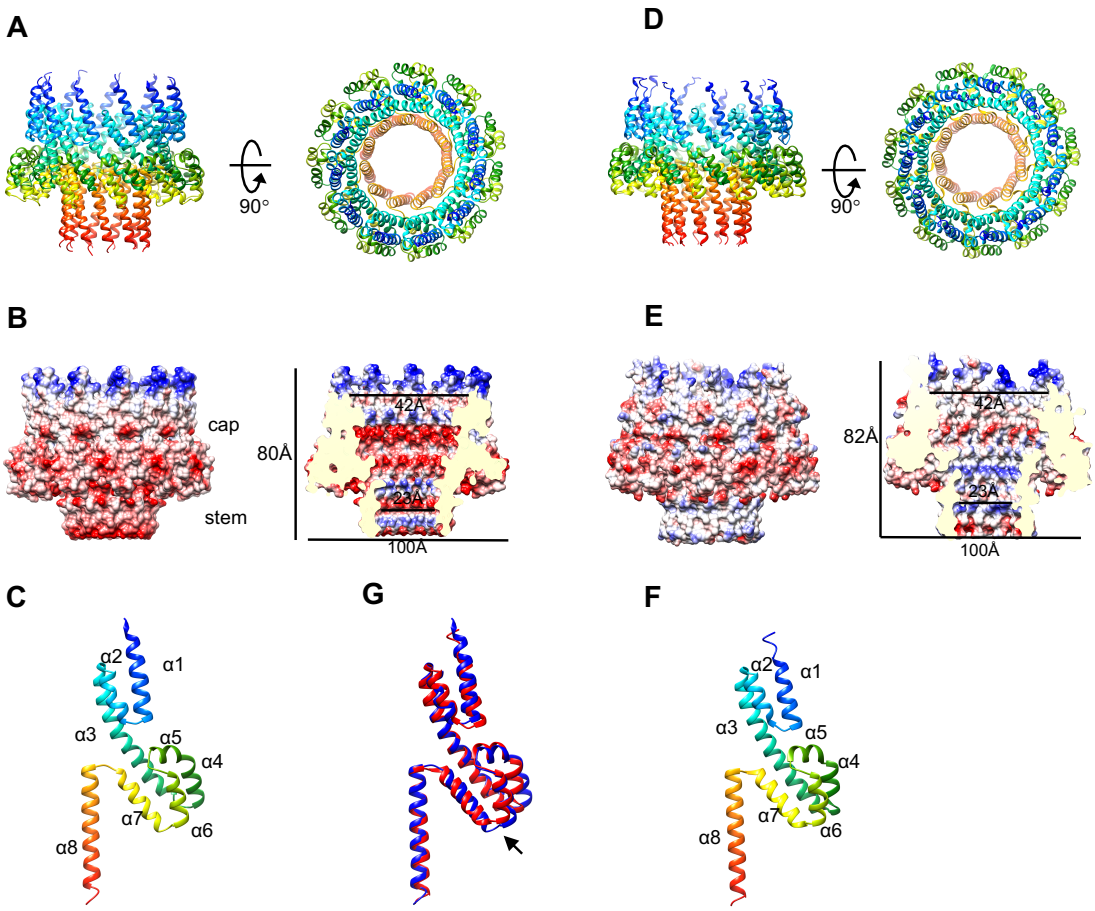


Fig. 4-3. The X-ray structures of gp12NTD (A-C) and gp20NTD (D-F).

(A, D) Ribbon diagram of gp12/gp20NTD undecamer with residues colored in rainbow from blue to red. In (A), one or both loops α3α4 and α6α7 are missing in chains other than chainA.

(B, E) Surface representation of undecameric gp12/gp20NTD colored according to electrostatic potential. Blue and red colours correspond to positive and negative potential of 10 kcal/(mol*e), respectively. In (B), this undecamer is generated by applying the 11-fold symmetry to gp12NTD chain A to make up missing loops in other chains.

(C, F) Gp12/gp20NTD monomer is shown in an orientation that the left is inside of tube, the right is outside of tube. Both molecules are composed of eight α-helices, colored with the same color scheme of the undecamer.

(G) Superimposition of gp12NTD (blue) and gp20NTD (red). The variation at the loop α6α7 is indicated with a black arrow.

4.3.4 Implications on how gp12/gp20 is translocated through the narrow tail tube

All of the three DNA-injection proteins are invisible in the asymmetric cryo-EM reconstruction of the entire P22 virion at 7.8Å resolution, suggesting that they are less well ordered (26).The portal-tail adaptor channel has an inner diameter ranging from 33Å to 42Å (16), while the gp12NTD/gp20NTD monomer has a dimension of 80Å *44Å *27Å. Thus, it could not pass through the channel as folded. What is more, it was shown that the P22 tail channel is partially occupied by dsDNA (26). While all of the DNA remained inside the capsid with osmotic suppression, the internal proteins can still be ejected (9). This suggests that the diameter of the protein structure going through the channel should be ~13Å at most, about the diameter of an α-helix. It has been hypothesized, as the "carpenter's ruler hypothesis", that T7 internal proteins convert to a linear conformation to go through the portal channel (21). Here with the structure of gp12NTD/gp20NTD, we would like to discuss, from the structural foundation, if it is possible that the molecule changes its conformation to a linear string of α-helices.

As mentioned above, the three long α-helices α1, α3, and α8 all have a consistent N to C orientation. Imagine that we hold the amino and carboxyl ends of the molecule (and treat each helix as a rigid body) and stretch it out along the orientation of the long helices until it becomes a linear string of α-helices, the three long α-helices do not change their orientation in this process while the short α-helices including α2, α4, α5, α7 will be flipped over to an opposite orientation (**Fig. 4-4**). In another word, when the N-/C-terminus goes into the portal, unfolding of the rest of the molecule could be achieved in a narrow space prior to the portal due to the fact that only short helices will turn around. Such feature matches well with the biological situation since, no matter in the capsid with densely packaged DNA or in the portal-tail channel, there is very limited space for protein unfolding.

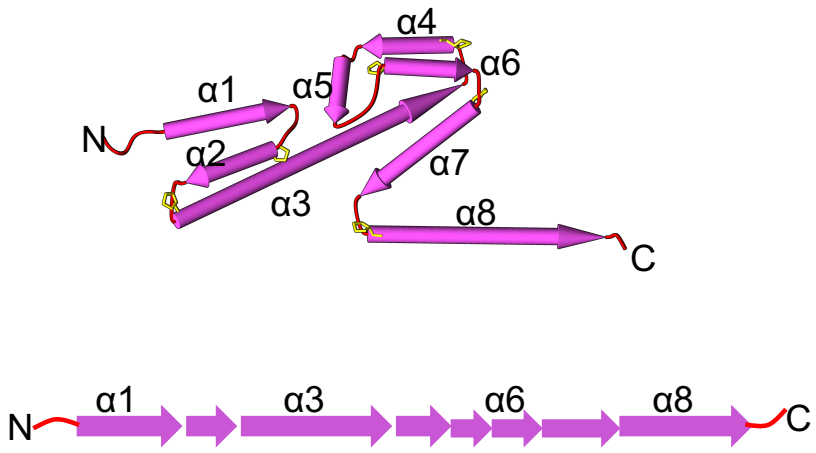
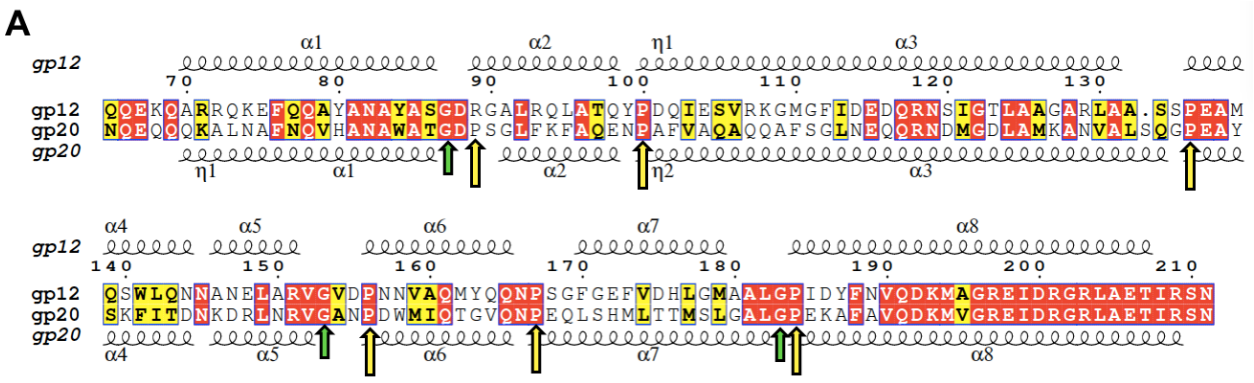


Fig. 4-4. A schematic diagram shows the partial unfolding process. Gp20NTD (90° rotated to left compared with **Fig. 4-3F**) is shown as pipes with the direction from the amino end to the carboxyl end. The six Pro residues are colored in yellow. The helices α1, α3, α6, and α8 do not change direction when unfolding.

In order to extend the molecule to a linear form, angles between successive helices have to be changed to ~180°. We found that all loops in the gp12NTD and gp20NTD structures have no more than four residues, suggesting that the angle/turn is restricted by the Φ-Ψ torsion angles of the backbone. For example, the angle between α1 and α2 is 12° in the folded structure. The Φ-Ψ torsion angles of the loop α1α2 backbone should allow a conversion from 12° to ~180°. Glycine has been known as the least sterically hindered amino acid as compared to others, thus a short loop with Gly should be more flexible than a loop with the same length. Three Gly residues (G87, G153, G183) are found in loops α1α2, α5α6, α7α8 of gp12NTD, and they are all conserved in gp20NTD (**Fig. 4-5**). Gp20NTD has an additional G137 in the α3α4 loop. Another amino acid that can introduce a turn in a peptide chain is proline. The *cis-trans*

interconversion of X-Pro peptide groups could switch a sharp turn to a mild one, thus making the structure more extended (X represents any amino acid residue prior to Pro). Strikingly, five Pro residues in gp12NTD are found nowhere else but at the first helical position (N1) of α3, α4, α6, α7, α8, all of which are conserved in gp20NTD (**Fig. 4-5**). Gp20NTD has an additional Pro at the N1 of helix α2. All of the X-Pro peptide groups are in the *cis* form, introducing a sharp turn between successive helices. Given that the *cis* and *trans* forms of X-Pro peptide groups are almost isoenergetic (37), and no favor is observed on one over the other (38), it is possible that, under certain circumstances, interconversion from *cis* to *trans* form occurs. Such conversion would promote the extension of the molecule to a linear conformation since the *trans* form of X-Pro peptide groups has an obtuse angle. The extra Pro in gp20NTD is at the N1 of α2 where gp12NTD has an Arg89 instead. However, the G87 in this loop as mentioned should be able to provide certain flexibility. No Pro is found between α4 and α5 in both gp12NTD and gp20NTD. These two helices are positioned in the structure with an obtuse angle of ~120°, thus may already be extended enough. Taken together, the Gly in loop regions and Pro on N1 positions provide structural basis for extending the gp12NTD/gp20NTD molecule to a linear form.



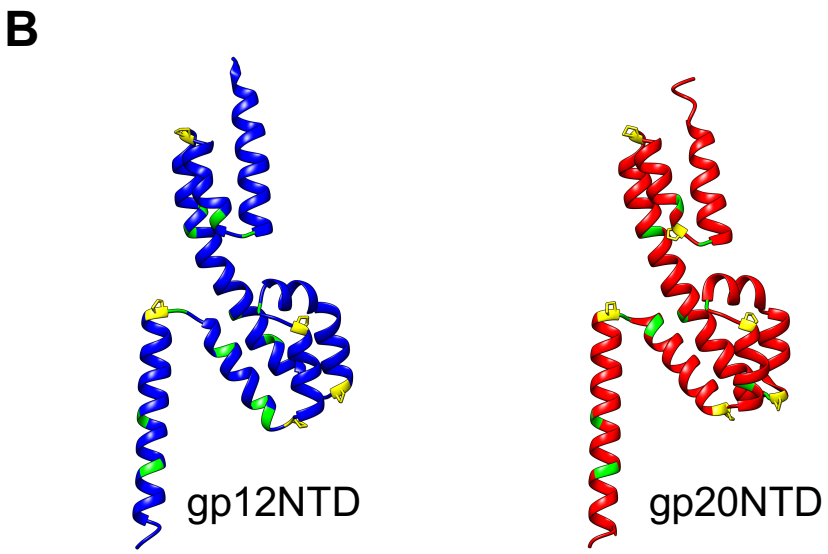


Fig. 4-5. Distribution of Pro and Gly in gp12NTD and gp20NTD.

(A) Sequence alignment of Sf6-gp12NTD and P22-gp20NTD. Secondary structures of gp12NTD and gp20NTD are shown on the top and bottom of the sequence alignment. Identical residues are in red while similar residues are in yellow. Pro residues and conserved Gly residues in loops are indicated with yellow and green arrows respectively. *(B)* Pro residues (yellow) locate only on the N1 positions. Gly residues (green) spread over the whole molecules, with three conserved Gly on the loops.

We then inspect carefully at the inter-helical intramolecular interactions. Seven and eight inter-helical hydrogen bonds are found in gp12NTD and gp20NTD respectively, resulting in only ~0.5 hydrogen bonds per closely located pair of helices. Therefore, hydrogen bonding is unlikely a major force that holds the tertiary structure. The two hydrophobic cores formed at the three-helix bundle (α1-α3) and the four-helix bundle (α3, α4, α6, α7) are more important for the tertiary structure (**Fig. 4-6**). For gp12NTD, the three-helix bundle forms a hydrophobic core mainly by F76, L92, L95, V106, M110, and I113. Residues in these positions in gp20NTD are all hydrophobic (**Fig. 4-5A**). Additionally, gp20NTD has L75, V81, F104 contributing to this hydrophobic core. The other hydrophobic core involves fourteen residues on same positions in gp12NTD and gp20NTD **(Fig.4-6&4-5A)**. Consistently, stability investigation of T7-gp15 with

the chaotrope guanidine hydrochloride (GuHCl) has indicated that hydrophobic interactions mainly contribute to the stability of gp15 (39). While it is not clear why and how the DNA-injection proteins unfold in the capsid or tail channel, a possible hypothesis is that the absence of water shell inside the capsid due to densely packaged DNA or hydrophobic interactions with other internal proteins stabilizes partially unfolded DNA-injection proteins. While T7-gp15 and gp16 both have relatively low thermal transition Tm (13), we find that gp12NTD also has a low Tm as 42°C, suggesting that the protein unfolding does not require much energy. When these proteins arrive at the aqueous environment in cell envelope, with no known chaperone proteins, hydrophobic collapse could potentially drive rapid protein folding.

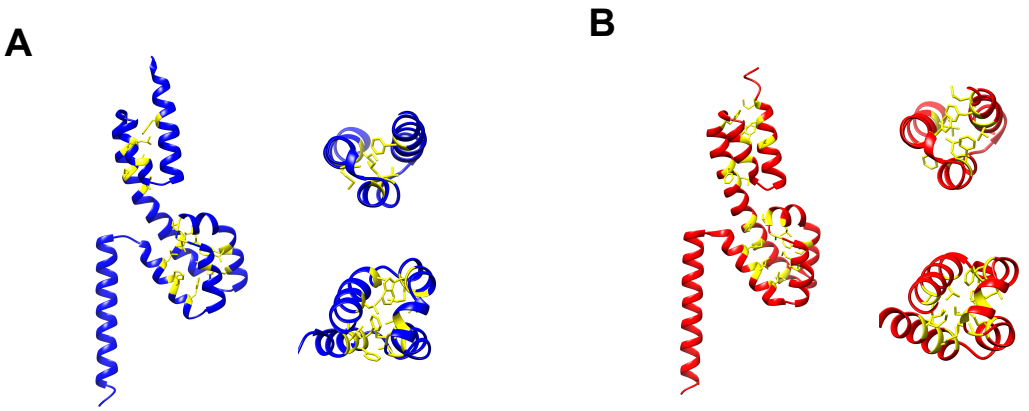


Fig. 4-6. Hydrophobic cores in gp12NTD (A) and gp20NTD (B). Hydrophobic residues contributing to the hydrophobic core are shown as atoms in yellow. Top views of the three-helix bundle and the four-helix bundle are shown on the right of the whole molecule.

The 3D EM reconstruction of full-length gp12 our lab published earlier (22) reveals the architecture of a molecular tube, suggesting gp12 as a component of the extended tail. The gp12 EM reconstruction is a decamer, while the X-ray structure of gp12NTD we determined here is an undecamer. In the EM study, class averages generated from reference-free 2D classification of gp12 full-length cryoEM images only showed 10-fold symmetry (22), suggesting the dominant percentage of decamer population, if other oligomers ever exist. The symmetry fold of the gp20NTD in the nsEM 2D class average cannot be determined due to the limited resolution, but it seems like to be more than 10 (**Fig. 4-2D** left). The two

structures differ in sequence at the N-terminal transmembrane helices and the C-terminal domain (CTD). The N-terminus could potentially play a role in oligomerization since there are four G/AXXXG/A motifs found in this position that are known to promote membrane protein oligomerization (40). On the other hand, the CTD probably does not affect oligomerization since we find that gp12CTD exists as a monomer in solution. In fact, oligomerization polymorphism has been observed in many ectopically expressed and assembled phage portals such as the portals of P22(41), T7(42), T3(43), SPP1(44), and also HSV-1(45). A modified purification protocol for P22 portal including heat shock and ultracentrifugation was proved efficient in producing homogenous dodecamer (46). In our case, purification protocol is likely an important factor leading to the different oligomerization states of gp12 since the full-length protein, but not the NTD, was purified through denature-refolding process. We also notice some different properties between full-length protein and the NTD. One example is that the solubility of the NTD was found to be highly dependent on salt concentration, protein tending to precipitate at NaCl concentration below 200mM. Directly fitting the undecamer into the EM map resulted in a correlation of 0.9283 (**Fig. 4-7)**. Except for the symmetry mismatch, visual inspection found two loops (loops connecting α1α2 and α6α7), a small fragment of helix α3, and the N termini being outside of the map. There was extra map volume at the N termini, consistent with our N-terminal truncation. We consider that although the oligomerization state is different in the EM and X-ray structures, the tertiary structure should be considerably similar.

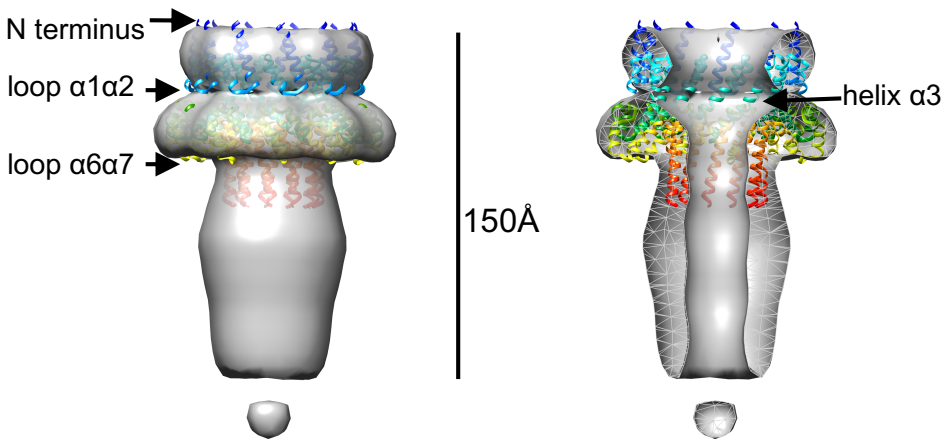


Fig. 4-7. Undecameric gp12NTD fitted in the decameric gp12 EM map. The regions that do not fit with the map is indicated by arrows, including loop α1α2, loop α6α7, helix α3, and the N terminus. Map: EMDB 6330.

Several features were observed when examining the interactions between gp12 monomers. First, each monomer is involved in extensive interactions with its two neighboring molecules, resulting in the burial of 35% (3,700Å^2) of the total solvent-accessible surface. Second, the gp12-gp12 interface contains a significant number of charged residues, and several of them are involved in hydrogen-bonding interactions and salt bridges (**Fig. 4-8**). In gp12NTD structure, there are seven hydrogen bonds at the cap, two hydrogen bonds and one salt bridges at the stem (helix α8) (**Fig. 4-8A**). Gp20NTD has eight hydrogen bonds at the cap, one hydrogen bonds and six salt bridges at the stem (**Fig. 4-8B**). While all these hydrogen bonds should stabilize the neighboring monomers, the ones at the stem may be more important because we learned from protein purification process that residues 184-211 are required for gp12 protein oligomerization. Gp12(65-183) exists as a monomer in solution while gp12NTD forms an oligomer. This is also true for gp20NTD. We think that the hydrogen bonds and salt bridges at the stem may first orient the two helices α8 in a parallel pattern so that homo-oligomerization could occur correctly.

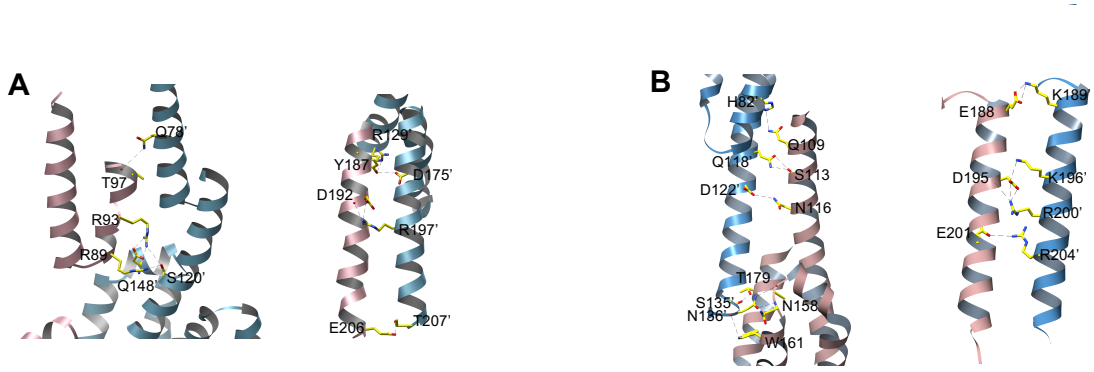


Fig. 4-8. Intermolecular interactions at the cap (left) and stem (right). The two neighboring monomers are colored blue and pink. The residues involved in hydrogen-bonding interactions are highlighted with dotted lines depicting potential hydrogen bonds.

(A) Gp12NTD has five hydrogen bonds at the cap, two at the connection. One salt bridge and two hydrogen bonds at the stem may orient the neighboring α8 helices in the process of oligomerization.

(B) Gp20NTD has four hydrogen bonds at the cap, four at the connection. Six salt bridges and one hydrogen bond at the stem may orient the neighboring α8 helices in the process of oligomerization.

Overall, we proved that Sf6 forms an extended tail for DNA injection. Gp12 is very likely one component of the tail extension. The conserved gp12NTD and gp20NTD structures provide valuable information on understanding how the internal proteins travel through the narrow tail tube.

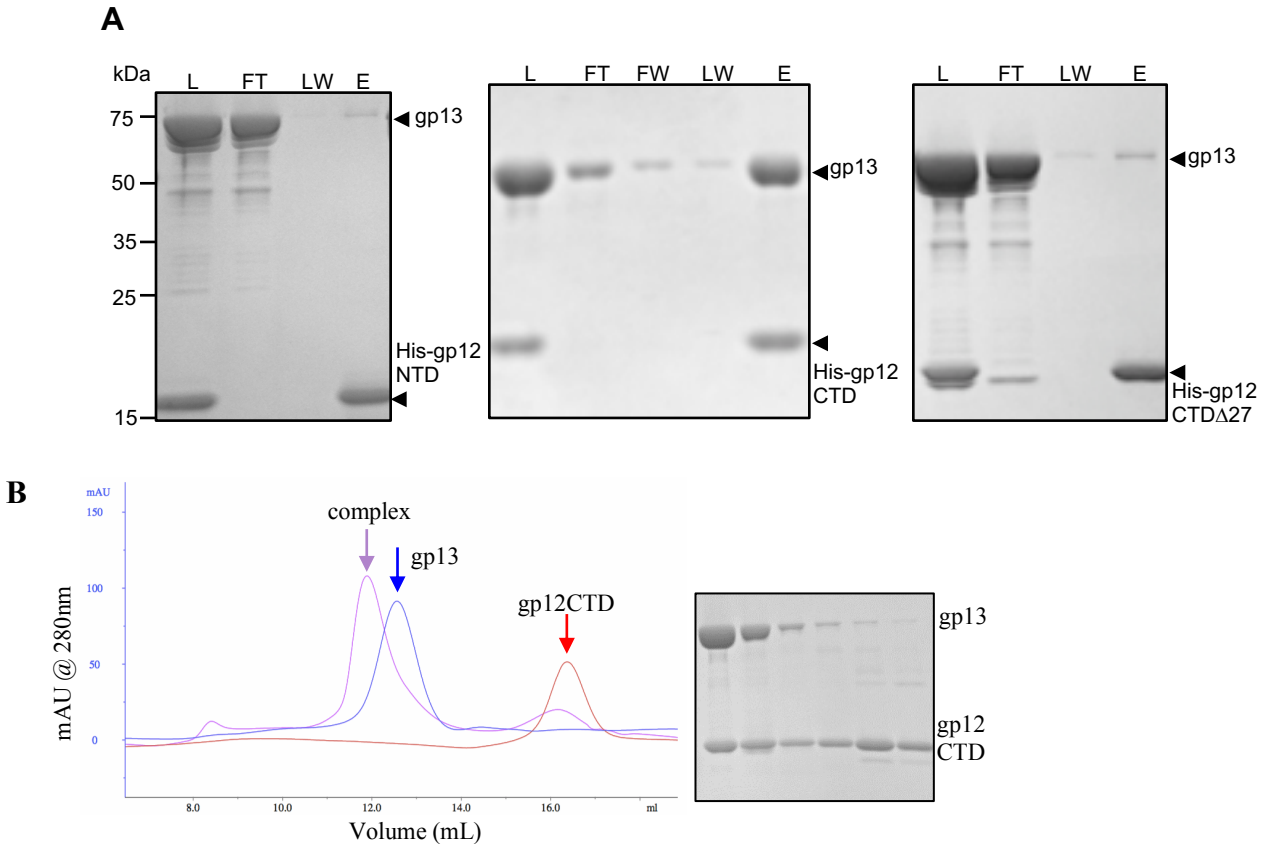
4.3.5 The gp12 interacts with gp13 by the 27 C-terminal residues

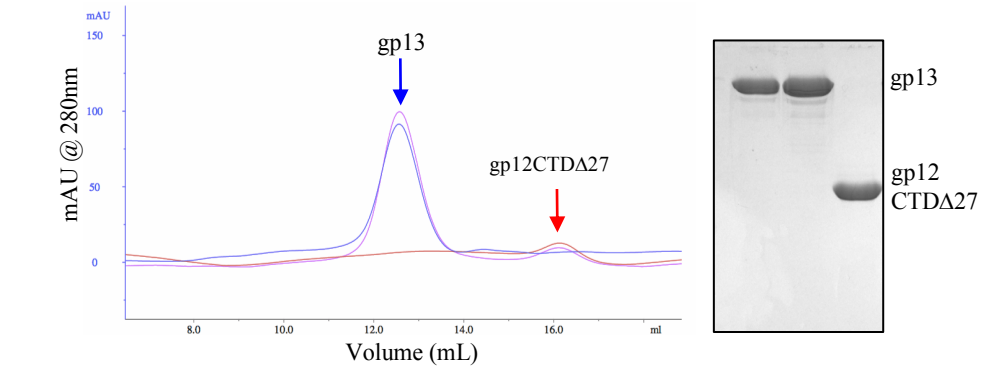
The 3D EM reconstruction of gp12 revealed that the decameric assembly was 150Å in length and might not be long enough to span the three-layer envelope of the Gram-negative *Shigella* (22). The extended tail observed in T7 was 450Å (3). Accordingly, it is likely that gp11 and/or gp13 forms a similar tube assembly attaching to gp12 to traverse the envelope.

To assess the interactions with gp13, his-tagged gp12NTD, gp12CTD (residues 211-431), and gp12CTDΔ27(residues 211-404) (**Fig. 4-2A**) were incubated with gp13 individually. The mixture was examined by both pull-down assay and SEC co-purification assay. In the pull-down assay, gp12NTD failed to pull down gp13 since gp13 was in the flow-through and washing fractions but not in the elution fraction

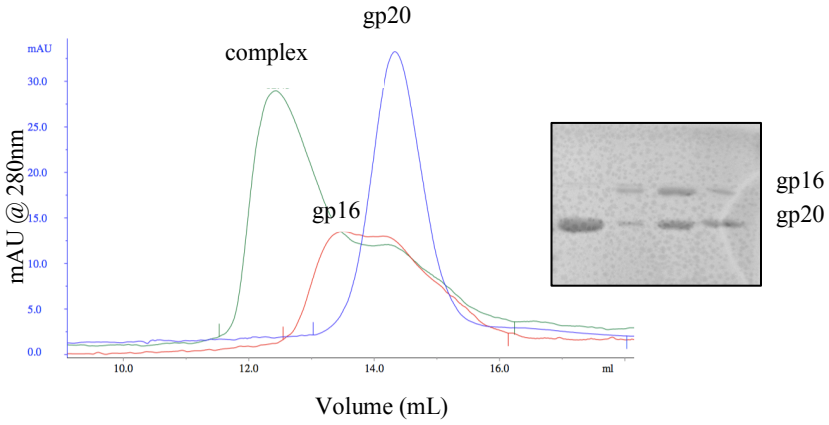
(**Fig. 4-9A** left). Instead, gp12CTD was eluted with significant amount of gp13 in the pull-down assay (**Fig. 4-9A** middle) and it appeared in a new peak of large molecular weight with gp13 in the co-purification assay (**Fig. 4-9B** top). Furthermore, this interaction was eliminated in the gp12CTDΔ27 as shown by both assays (**Fig. 4-9A** right**, Fig. 4-9B** bottom). Taken together, gp12 interacts with gp13 through its CTD, and the 27 C-terminal residues are required for such interaction. More refined truncations can be designed to further investigate the roles of specific residues for binding gp13.

Interaction between P22-gp20 and P22-gp16 was tested with SEC co-purification. Both gp20 and gp16 showed a single peak of ~73 and ~49kDa on SEC, respectively. Given their theoretic molecular weight of 64 and 50kDa, the two proteins are thus most likely monomeric. A complex of gp16-gp20 was detected by incubating the two proteins at room temperature for 1 hour and subsequent analysis by SEC (**Fig. 4-9C**).





C



D

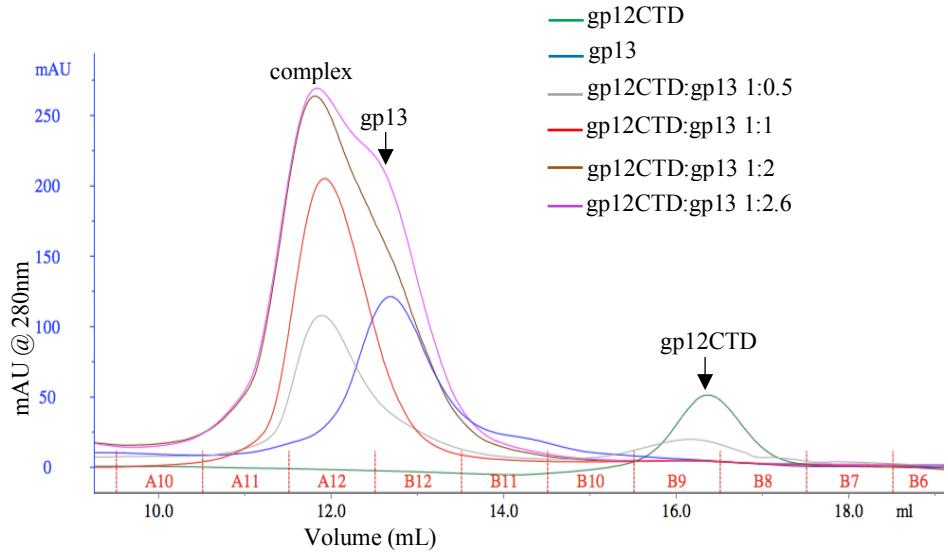


Fig. 4-9. The gp12 27 C-terminal residues are essential for interaction with gp13.

(A) His-tagged gp12NTD (left), gp12CTD (middle), and gp12CTDD27 (right) are used to pull down gp13 from solution individually. Only gp12CTD pulls down significant amount of gp13. L: loading; FT: flow-through; FW: first wash; LW: last wash; E: elution.

(B) Gp12CTD and gp12CTDD27 are incubated with gp13 individually and loaded on Superdex 200 10/300GL.

Gp12CTD, but not gp12CTDD27, forms a new peak with gp13. Red: gp12CTD/ gp12CTDD27; Blue: gp13; Purple: mixture of gp12CTD/ gp12CTDD27 and gp13.

(C) Complex formation of P22-gp20 and P22-gp16 *in vitro*. After mixed, P22-gp20 and gp16 are eluted in a new SEC peak. SDS-PAGE proves the co-localization of the two proteins in the new peak. Blue: P22-gp20; Red: P22-gp16; Green: mixture of P22-gp20 and gp16.

(D) Gp12CTD and gp13 are mixed at various molar ratios. At the molar ratio of 1:1, essentially all proteins elute in a single peak.

The stoichiometry between gp12CTD and gp13 was studied by SEC. The pure gp12CTD and gp13 were loaded separately on the SEC to find out their elution volumes. We then fixed the loading amount of gp12CTD and titrated it with gp13 at the molecular ratio of 1:0.5 to 1:2.6. In all titrations, a large-molecular-weight-peak of about 220kDa where both proteins located was found (**Fig. 4-9D**). When $M_{gp12CTD}$:M_{gp13} equaled to 1:0.5, there was excess gp12CTD eluting off at its normal elution volume. When the amount of gp13 increased to $M_{gp12CTD}$:M_{gp13} of 1:1, all gp12CTD moved to the large-molecular-weight-peak. We conclude that the stoichiometry between gp12 and gp13 is likely to be 1 to 1.

4.3.6 A working model for the assembly of Sf6 DNA-injection apparatus

In the gp12NTD/gp20NTD structure, we saw many inward-facing long side chains pointing to the N termini as observed in the H protein of tail-less bacteriophage ΦX174 (47). Such orientation (like an outside-in harpoon as described in the paper) and the scattered positive charges may serve to prevent DNA from moving backward, suggesting that the DNA is transported from the C to N termini. This is consistent with our assumption, partially based on the prediction of two predicted N-terminal transmembrane helices, that the gp12 N-terminus is located at the inner membrane and the C terminus is in periplasmic space.

The gp12 EM reconstruction is ~150Å in length, suggesting that gp12 alone is likely not long enough to span the entire cell envelope, more proteins may lengthen the channel. Gp13 exits mainly as a monomer in solution and it seems to be elongated since gp13 passes down the size exclusion column much

faster than a globular protein of the same molecular weight. Pull-down and co-purification of gp12 truncations with gp13 suggests that the gp12 CTD, but not NTD, interacts with gp13, and the stoichiometry is most likely to be 1:1. The 27 C-terminal residues are essential for such interaction. If, as we proposed, gp12 locates the cap at the inner membrane and extends the stem domain towards peptidoglycan, gp13 would mostly be in the periplasmic space. Gp13 does not spontaneously form oligomer in solution. We tried to use full-length gp12 decamer to induce gp13 oligomerization, expecting a larger tubular structure. However, both proteins precipitated out of solution immediately when they were mixed.

Take all together, we propose a working model for assembly of Sf6 DNA-injection apparatus (**Fig. 4-10**). After DNA injection is triggered, the three internal proteins are ejected prior to DNA. The way that these proteins travel through the narrow tail tube is to be partially unfolded, but with the helical structures. Once they arrive at cell envelope and get fully hydrated, the proteins can refold probably driven by hydrophobic interactions. We do not yet know the location of gp11 and gp13. But we propose that gp12 decameric tube attaches to the inner membrane with the N-terminus, while the C-terminus in the periplasmic space binds with gp13.

Cryo-EM is revolutionizing the structural biology in recent years. It is tempting to study the complete DNA-injection apparatus with cryo-EM which could elucidate how the DNA-injection proteins assemble. Meanwhile, there may be difficulty in synchronize the ejection of the proteins and achieve homogenized sample since P22-like phages inject DNA more rapidly than T7 (48), and the channel may be closed or degraded after DNA injection to ensure cell integrity (14, 49). Alternatively, liquid cell TEM may be applied to image the dynamic process.

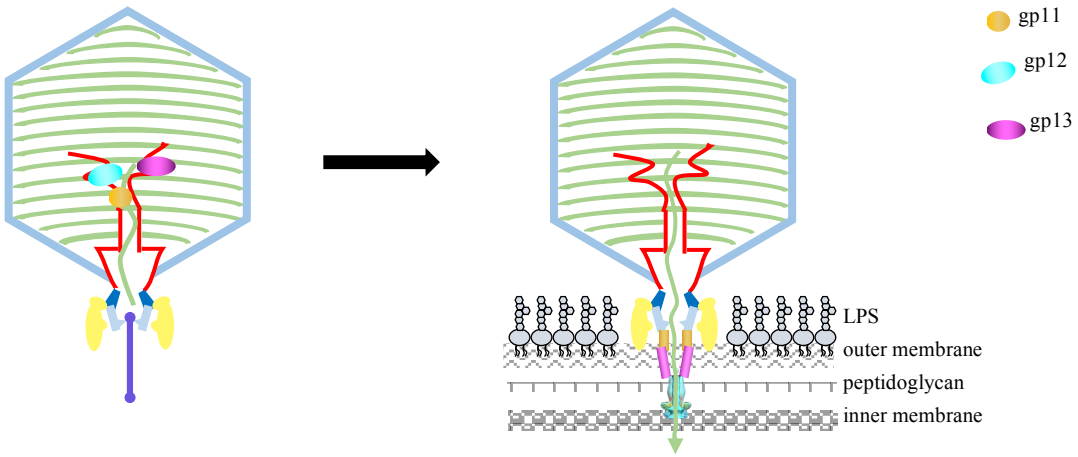


Fig. 4-10 A working model of Sf6 DNA-injection apparatus. Once DNA injection is triggered, the internal proteins are first delivered through the tail channel at a partially unfolded state as helical segments, followed by protein refolding and assembly in cell envelope to form a tail extension. The gp12 forms a decameric tube with its N-terminus docked on the inner membrane, C-terminus in the periplasmic space interacting with gp13.

4.4 Reference

1. Leiman PG & Shneider MM (2012) Contractile tail machines of bacteriophages. *Advances in experimental medicine and biology* 726:93-114.
2. Esquinas-Rychen M & Erni B (2001) Facilitation of bacteriophage lambda DNA injection by inner membrane proteins of the bacterial phosphoenol-pyruvate: carbohydrate phosphotransferase system (PTS). *J Mol Microbiol Biotechnol* 3(3):361-370.
3. Hu B, Margolin W, Molineux IJ, & Liu J (2013) The bacteriophage t7 virion undergoes extensive structural remodeling during infection. *Science* 339(6119):576-579.
4. Israel V (1977) E proteins of bacteriophage P22. I. Identification and ejection from wild-type and defective particles. *Journal of virology* 23(1):91-97.
5. Botstein D, Waddell CH, & King J (1973) Mechanism of head assembly and DNA encapsulation in Salmonella phage p22. I. Genes, proteins, structures and DNA maturation. *Journal of molecular biology* 80(4):669-695.
6. King J, Lenk EV, & Botstein D (1973) Mechanism of head assembly and DNA encapsulation in Salmonella phage P22. II. Morphogenetic pathway. *Journal of molecular biology* 80(4):697-731.
7. Hoffman B & Levine M (1975) Bacteriophage P22 virion protein which performs an essential early function. II. Characterization of the gene 16 function. *Journal of virology* 16(6):1547-1559.
8. Perez GL, Huynh B, Slater M, & Maloy S (2009) Transport of phage P22 DNA across the cytoplasmic membrane. *Journal of bacteriology* 191(1):135-140.
9. Jin Y*, et al.* (2015) Bacteriophage P22 ejects all of its internal proteins before its genome. *Virology* 485:128-134.
10. Serwer P, Wright ET, Hakala KW, & Weintraub ST (2008) Evidence for bacteriophage T7 tail extension during DNA injection. *BMC research notes* 1:36.
11. Moak M & Molineux IJ (2000) Role of the Gp16 lytic transglycosylase motif in bacteriophage T7 virions at the initiation of infection. *Mol Microbiol* 37(2):345-355.

12. Chang CY, Kemp P, & Molineux IJ (2010) Gp15 and gp16 cooperate in translocating bacteriophage T7 DNA into the infected cell. *Virology* 398(2):176-186.
13. Lupo D*, et al.* (2015) The T7 ejection nanomachine components gp15-gp16 form a spiral ring complex that binds DNA and a lipid membrane. *Virology* 486:263-271.
14. Chang JT*, et al.* (2010) Visualizing the structural changes of bacteriophage Epsilon15 and its Salmonella host during infection. *Journal of molecular biology* 402(4):731-740.
15. Lander GC*, et al.* (2006) The structure of an infectious P22 virion shows the signal for headful DNA packaging. *Science* 312(5781):1791-1795.
16. Olia AS, Prevelige PE, Jr., Johnson JE, & Cingolani G (2011) Three-dimensional structure of a viral genome-delivery portal vertex. *Nature structural & molecular biology* 18(5):597-603.
17. Parent KN, Gilcrease EB, Casjens SR, & Baker TS (2012) Structural evolution of the P22-like phages: comparison of Sf6 and P22 procapsid and virion architectures. *Virology* 427(2):177-188.
18. Agirrezabala X*, et al.* (2005) Structure of the connector of bacteriophage T7 at 8A resolution: structural homologies of a basic component of a DNA translocating machinery. *Journal of molecular biology* 347(5):895-902.
19. Agirrezabala X*, et al.* (2005) Maturation of phage T7 involves structural modification of both shell and inner core components. *The EMBO journal* 24(21):3820-3829.
20. Erickson HP (2009) Size and shape of protein molecules at the nanometer level determined by sedimentation, gel filtration, and electron microscopy. *Biol Proced Online* 11:32-51.
21. Molineux IJ (2001) No syringes please, ejection of phage T7 DNA from the virion is enzyme driven. *Mol Microbiol* 40(1):1-8.
22. Zhao H*, et al.* (2016) Structure of a Bacterial Virus DNA-Injection Protein Complex Reveals a Decameric Assembly with a Constricted Molecular Channel. *PloS one* 11(2):e0149337.
23. Guo F*, et al.* (2013) Visualization of uncorrelated, tandem symmetry mismatches in the internal genome packaging apparatus of bacteriophage T7. *Proc Natl Acad Sci U S A* 110(17):6811-6816.
24. Leiman PG*, et al.* (2007) The structures of bacteriophages K1E and K1-5 explain processive degradation of polysaccharide capsules and evolution of new host specificities. *Journal of molecular biology* 371(3):836-849.
25. Liu X*, et al.* (2010) Structural changes in a marine podovirus associated with release of its genome into Prochlorococcus. *Nature structural & molecular biology* 17(7):830-836.
26. Tang J*, et al.* (2011) Peering down the barrel of a bacteriophage portal: the genome packaging and release valve in p22. *Structure* 19(4):496-502.
27. Fokine A*, et al.* (2004) Molecular architecture of the prolate head of bacteriophage T4. *Proceedings of the National Academy of Sciences of the United States of America* 101(16):6003-6008.
28. Wu W*, et al.* (2016) Localization of the Houdinisome (Ejection Proteins) inside the Bacteriophage P22 Virion by Bubblegram Imaging. *MBio* 7(4).
29. Winn MD*, et al.* (2011) Overview of the CCP4 suite and current developments. *Acta crystallographica. Section D, Biological crystallography* 67(Pt 4):235-242.
30. Adams PD*, et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta crystallographica. Section D, Biological crystallography* 66(Pt 2):213-221.
31. Terwilliger TC*, et al.* (2009) Decision-making in structure solution using Bayesian estimates of map quality: the PHENIX AutoSol wizard. *Acta crystallographica. Section D, Biological crystallography* 65(Pt 6):582-601.
32. Emsley P & Cowtan K (2004) Coot: model-building tools for molecular graphics. *Acta crystallographica. Section D, Biological crystallography* 60(Pt 12 Pt 1):2126-2132.
33. Zhao H, Sequeira RD, Galeva NA, & Tang L (2011) The host outer membrane proteins OmpA and OmpC are associated with the Shigella phage Sf6 virion. *Virology* 409(2):319-327.
34. Parent KN*, et al.* (2014) OmpA and OmpC are critical host factors for bacteriophage Sf6 entry in Shigella. *Mol Microbiol* 92(1):47-60.

35. Tang G, *et al.* (2007) EMAN2: an extensible image processing suite for electron microscopy. *Journal of structural biology* 157(1):38-46.
36. Scheres SH (2012) RELION: implementation of a Bayesian approach to cryo-EM structure determination. *Journal of structural biology* 180(3):519-530.
37. Macknight WJ (1976) Macromolecules. *Science* 191(4233):1260.
38. Wedemeyer WJ, Welker E, & Scheraga HA (2002) Proline cis-trans isomerization and protein folding. *Biochemistry* 41(50):14637-14644.
39. Leptihn S, Gottschalk J, & Kuhn A (2016) T7 ejectosome assembly: A story unfolds. *Bacteriophage* 6(1):e1128513.
40. Russ WP & Engelman DM (2000) The GxxxG motif: a framework for transmembrane helix-helix association. *Journal of molecular biology* 296(3):911-919.
41. Cingolani G, Moore SD, Prevelige PE, Jr., & Johnson JE (2002) Preliminary crystallographic analysis of the bacteriophage P22 portal protein. *Journal of structural biology* 139(1):46-54.
42. Kocsis E, Cerritelli ME, Trus BL, Cheng N, & Steven AC (1995) Improved methods for determination of rotational symmetries in macromolecules. *Ultramicroscopy* 60(2):219-228.
43. Valpuesta JM, *et al.* (2000) Structural analysis of the bacteriophage T3 head-to-tail connector. *Journal of structural biology* 131(2):146-155.
44. Dube P, Tavares P, Lurz R, & van Heel M (1993) The portal protein of bacteriophage SPP1: a DNA pump with 13-fold symmetry. *The EMBO journal* 12(4):1303-1309.
45. Trus BL, *et al.* (2004) Structure and polymorphism of the UL6 portal protein of herpes simplex virus type 1. *Journal of virology* 78(22):12668-12671.
46. Lorenzen K, Olia AS, Uetrecht C, Cingolani G, & Heck AJ (2008) Determination of stoichiometry and conformational changes in the first step of the P22 tail assembly. *Journal of molecular biology* 379(2):385-396.
47. Sun L, *et al.* (2014) Icosahedral bacteriophage PhiX174 forms a tail for DNA transport during infection. *Nature* 505(7483):432-435.
48. Rhoades M & Thomas CA, Jr. (1968) The P22 bacteriophage DNA molecule. II. Circular intracellular forms. *Journal of molecular biology* 37(1):41-61.
49. Gonzalez-Garcia VA, *et al.* (2015) Conformational changes leading to T7 DNA delivery upon interaction with the bacterial receptor. *The Journal of biological chemistry* 290(16):10038-10044.

Chapter 5: Structure of MVM NS1N provides insight into DNA nicking and origin binding

Adapted from paper 'Tewary SK, *et al.* (2015) Structures of minute virus of mice replication initiator protein N-terminal domain: Insights into DNA nicking and origin binding. *Virology* 476:61-71.'

5.1 Introduction

Minute Virus of Mice (MVM) belongs to the family of *Parvoviridae* which is characterized by small icosahedral, non-enveloped capsid containing a linear single-stranded DNA (ssDNA) genome of about 5kb, with short unique terminal palindromic sequences that fold back on themselves to form hairpin duplexes. MVM is a common infection in laboratory mice, typically with no clinical signs of infection. There are two variant forms of MVM: one is the prototype MVMp which infects cells of fibroblast origin; the other is MVMi, an immunosuppression strain infecting T lymphocyte. Autonomous parvoviruses are inherently oncoselective, having demonstrated enhanced fitness and toxicity in many rodent and human cancer cell lines, compared to their normal counterparts, under the same growth conditions in vitro, although there is considerable variability in tropism among different Parvoviruses and among different cell lines (1-4), providing interesting therapeutic potential.

Over 95% of MVM virions contain a negative-sense strand of DNA. At the end of the genome are two different hairpin telomeres which harboring viral disparate viral replication origins OriR and OriL, when expressed in duplex replicative-form DNA. The ability of the telomeres to unfold and refold during genome replication are critical to the unique parvoviral 'rolling-hairpin replication (RHR)' strategy, in which DNA synthesis is alternatively primed by the 3' end of one telomere and then the other.

MVM genome contains two promoters, the non-structural gene promoter P4 and the capsid gene promoter P38. Alternatively spliced mRNAs transcribed from the P4 promoter encode two major non-structural proteins NS1 and NS2 which share a common N-terminal segment of 85 amino acids. NS1 is an 83-kDa mainly nuclear phosphoprotein, the only viral nonstructural protein required in all cell types. With

three major functional domains, it is involved in multiple processes necessary for virus propagation, such as DNA amplification, particle assembly and packaging, and cytotoxicity.

Once the MVM virus in the nucleus, it must await the G1/S transition. Synthesis of the complementary strand creates functional duplex DNA promoters P4 and P38 which subsequently transcribe and translate to get proteins. NS1 binds site-specifically to duplex DNA at repetitions of the tetranucleotide sequence 5'-TGGT-3', which are found at multiple locations throughout the genome (5). Site specific binding requires assembly of NS1 molecules into multimers, which occurs in the presence of ATP (6). Phosphorylation at T435 and S437 by protein kinase C lambda activate the helicase function of NS1 (7, 8), allowing the unwinding of dsDNA to generate a region of ssDNA that encompasses the nick site which is subsequently nicked by the NS1 nickase activity (9-13) assisted by some cellular co-factors (14-16). After nicking DNA and becoming covalently esterified to the newly created 5' end of DNA, NS1 is believed to function as the replicative 3' to 5' helicase, hydrolyzing ATP to provide the energy needed to unwind double-stranded DNA (9). The C-terminus of NS1 has a transcriptional activation domain which functions to upregulate the P38 promoter when NS1 binds to its trans-activating region (TAR) (6, 17).

MVM NS1 is composed of three domains. In this study, we put our focus on the N-terminal nickase domain (NS1N). Orthologs of the MVM NS1 nickase domain in AAV ($Rep^{68/78}$) and HBoV were structurally characterized (18, 19). The AAV Rep nickase domain was shown structurally to bind to the five tetranucleotide motifs as well as a stem-loop structure in the AAV Ori (20). The MVM NS1 N-terminal domain (NS1N) shares low sequence identities of 17.8% and 18.7% with that of HBoV and AAV, respectively. Here we report the crystal structure of the N-terminal nickase domain (NS1N) and the derivatives with metal ligands.

5.2 Methods

5.2.1 Overexpression and purification of NS1N

The DNA fragment encoding MVM NS1N (residues 1–255; protein_id="AAA67109.1") was

cloned into pET28b (Novagen) between *Nde1/BamH1* and a stop codon was introduced after Asp255. The N-terminally His-tagged protein was over-expressed in E. coli strain B834(DE3). The protein was purified with a Ni-NTA column (Qiagen) followed by gel filtration chromatography on a Hiload 16/60 Superdex 75 column (GE Healthcare), which showed an elution volume of 64.8ml corresponding to a monomer. The eluted fractions were collected in tubes containing the gel filtration buffer (20 mM Tris–HCl pH 8.0, 150 mM NaCl, 1 mM DTT, 1 mM EDTA and 5% glycerol) supplemented with a final concentration of 100mM L(+)Arginine (ACROS Organics) to prevent rapid protein precipitation and promote single crystal growth. The protein was concentrated to approximately 12 mg/ml using a Millipore centricon (molecular weight cutoff 10kDa) prior to crystallization.

5.2.2 Electrophoretic Mobility Shift Assay (EMSA) of NS1N

On Left-Hand End (LHE) of MVM DNA there is a repeated 5'-TGGT-3' motif which is supposed to be one of the NS1 binding sites. 23bp downstream of the NS1 binding site is the potential nick site. We design three different sizes of DNA oligomers to characterize the binding of NS1 to the viral genomic DNA. All of the DNA oligomers start from 3bp prior to the binding site and stop before the nick site (-3~15bp), right at the nick site (-3~23bp) or after the nick site (-3~27bp). The forward and reverse DNA fragments are mixed and heated at 95 ºC for 10min in the annealing buffer (10mM Tris 7.5, 5mM KCl, 2mM $MgCl_2$) followed by slow cooling down at room temperature. Then the annealed DNA are incubated with NS1N protein at the molecular ratio of 1:0, 1:1, 1:3, 1:6, 1:12 for 1hr at room temperature. The samples are then loaded on TBE gels for DNA analysis.

5.2.3 NS1N crystallization, X-ray data collection, and structure determination

The purified MVM NS1N was crystallized with the hanging drop vapor diffusion method by mixing 1 μl protein solution with 1 μl of the well solution containing 2.8 M sodium formate and 100 mM sodium acetate trihydrate pH 5.4. Heavy atom derivatives were obtained by soaking the native crystals in 3.5 M sodium acetate solution with 10 mM HgCl2 or Pb(C2H3O2)2 for 5–7 min prior to flash freezing. MVM NS1N native and heavy atom derivative X-ray data were collected at the Advanced Photon Source (APS) and Stanford Syn- chrotron Radiation Lightsource (SSRL), respectively. Data were processed with

xia2 (21, 22). The structure was determined by the multiple isomorphous replacement method using the mercury and lead derivative data with SOLVE in PHENIX (23). Small loops of regions of residues 72–81 and 191–195 were manually built using COOT (24). The coordinates and reflection data have been deposited with RCSB Protein Data Bank with the accession codes of 4PP4.

Crystals containing magnesium were obtained by soaking native crystals in a 10mM $MgCl_2$ solution for 5min. X-ray data were collected at APS. Structures were solved with the molecular replacement method using the native structure as the search model, and structure refinement and model building were performed with PHENIX and COOT, respectively. The coordinates and reflection data have been deposited with RCSB Protein Data Bank with the accession codes of 4R94.

5.3 Results and Discussion

5.3.1 The reiterated 5'-TGGT-3' motif binds to NS1N; the following bases until #23 enhances binding

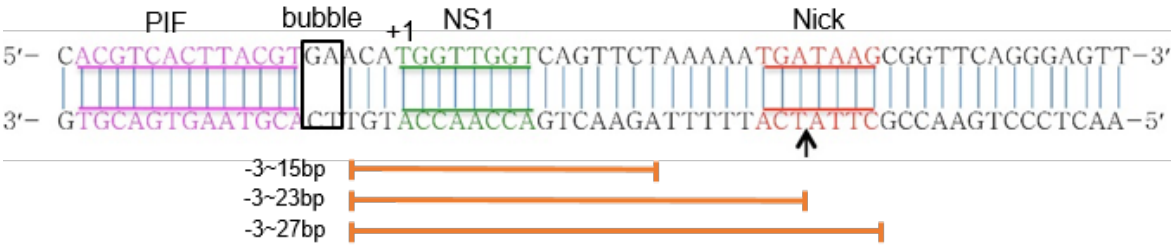
As a preliminary test, excess NS1N was mixed and incubated with dsDNA (dsDNA -3~15bp, dsDNA -3~23bp, dsDNA -3~27bp) or hetero-strand DNA (hDNA -3~15bp+-3~23bp, hDNA -3~15bp+-3~27bp) (**Fig. 5-1A**) at a molar ration of 12 to 1. The TBE gel of EMSA result showed a nucleoprotein complexes band in every reaction, suggesting that all these DNA could bind to NS1N (**Fig. 5-1B**). Notably, dsDNA -3~23bp resulted in a much stronger band compared with dsDNA -3~15bp, indicating that base pairs 16-23 enhanced such interaction.

Another EMSA was then designed to assess the approximate stoichiometry of DNA to NS1N. NS1N was mixed with the five types of DNA molecules at different molar ratios from 0:1 to 12:1 (**Fig. 5-1C**). A non-specific dsDNA was included as a negative control. As the results, the negative control DNA barely showed a nucleoprotein complex bands at the ratio of 6:1 and 12:1. The other three dsDNA all showed a brighter nucleoprotein complex band at lower protein to DNA ratio. Consistently, the dsDNA -3~23bp showed a brighter nucleoprotein band compared with that of the dsDNA -3~15bp. The dsDNA -3~27bp did not show significant brighter nucleoprotein band than the dsDNA -3~23bp. In the case of

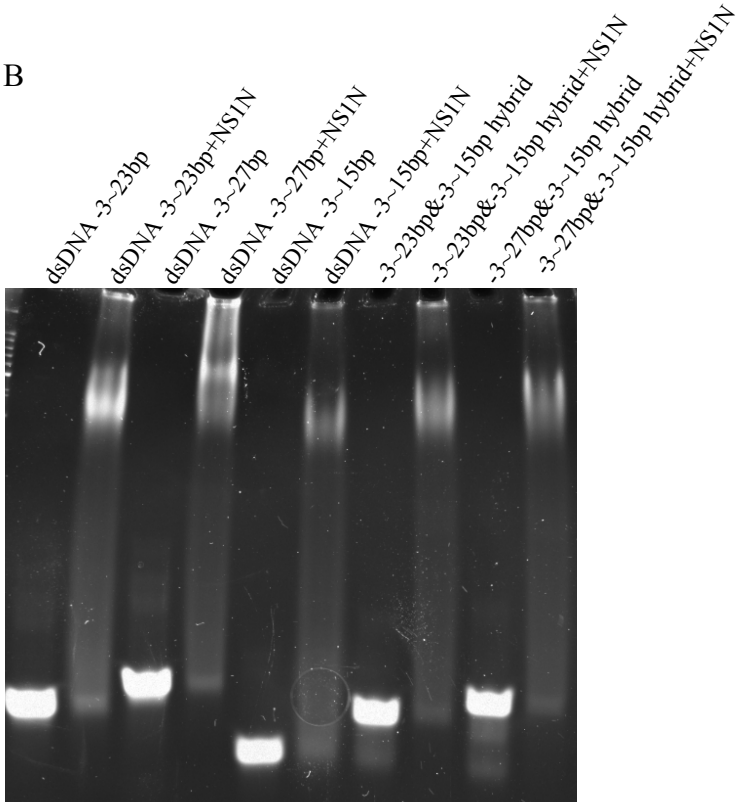
dsDNA -3~23bp, the first nucleoprotein complex band appeared at the ratio of 1:1 and the band density increased as the increasing amount of protein in the reaction.

Taken together, the 5'-TGGT-3' motif binds to NS1N and this interaction is significantly strengthened by the following DNA sequence until #23bp. The sequence at the nick site does not seem to

A NS1N binding. We conclude that more than one NS1 molecule binds to the DNA due to the fact that the nucleoprotein complex band becomes brighter as the increasing of protein:DNA molar ratio.



B



C

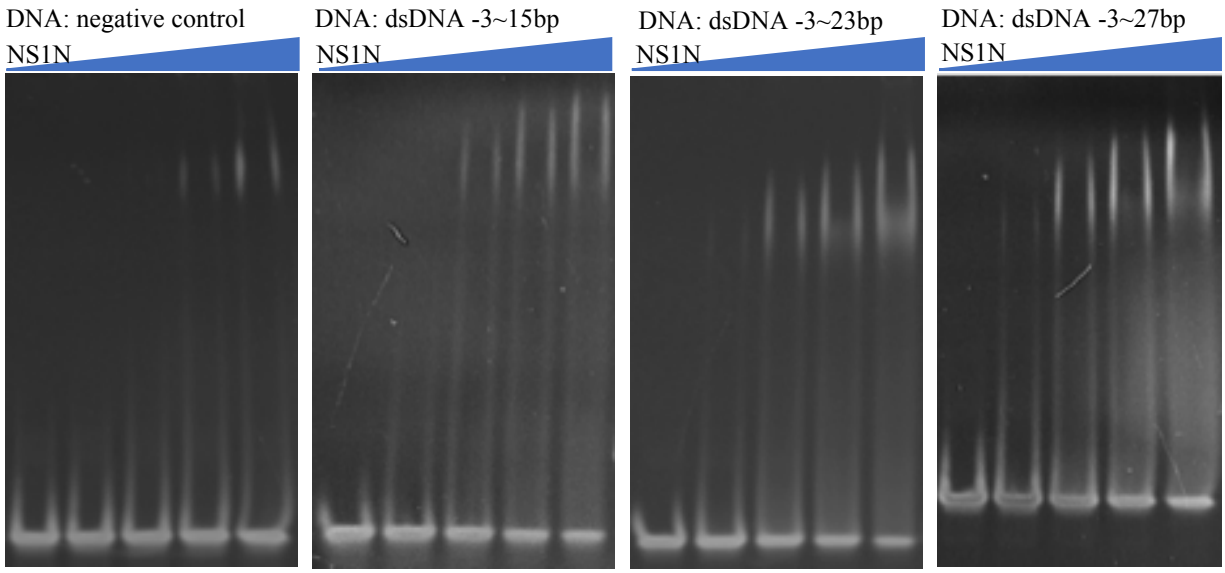


Fig. 5-1. NS1N and DNA interaction. *(A)* The MVM OriL dsDNA contains the parvovirus initiation factor (PIF) binding site, the NS1 binding site, and a nick site indicated with the black arrow. Three DNA oligos designed to test interaction with NS1N are indicated with orange lines. *(B)* TBE gel of the EMSA result on DNA fragments and NS1N at a molecular ratio of 1 to 12. *(C)* Assessment of the interaction of different dsDNA with increasing NS1N concentration.

5.3.2 Conserved nickase active site among the subfamily *Parvovirinae*.

The NS1N structure was solved with residues 6-253 at 1.45 Å resolution (**Table 5-1**). The crystallographic asymmetric unit contains one NS1N molecule. It consists of a centrally placed five-stranded antiparallel β-sheet (β6/β1/β5/β4/β8) flanked by four α-helices on one side (α2/α3/α4/α5) and three helices on the other side (α1/α6/α7) (**Fig. 5-2A**). The central β-sheet forms a cleft that embraces the nickase active site, which is surrounded by helix α6, a small loop between β4 and β5, and the loop L10.

Table 5-1. X-ray data collection and structure refinement statistics for the native crystal and heavy atom derivatives

Data collection	Native	Pb	Hg	Mg^{2+}
Beamline	APS 23ID-D	SSRL 7-1	SSRL 7-1	APS 23IDB
Wavelength (Å)	0. 97949	0. 97946	0. 97946	**0**.97934
Resolution (Å)	33.00-1.45 (1.49-1.45)*	34.5-1.56 (1.56-1.60)*	34.4-1.56 (1.56-1.60)*	50.00-1.67 (1.70-1.67)*
No. Measurements	142,612	183,404	148,706	151,693
Unique reflections	51,499 (3,649)*	41,626 (2,901)*	41,677 (2,794)*	33,535 (1,610)*
Completeness (%)	98.6 (95.8)*	99.0 (95)*	99.0 (91.3)*	97.2 (95.9)*
I/σ	20.4 (2.0)*	16.3 (2.3)*	16 (2.3)*	37.1 (3.1)*
R_{merge} (%)**	2.7 (35.2)*	5.1 (38.2)*	5.1 (51.3)*	3.6 (45.3)*
Space group	$P2_12_12$	$P2_12_12$	$P2_12_12$	$P2_12_12$
Unit cell (Å)	a=56.92, b=121.61, c=41.81	a=56.99, b=121.24, c=41.81	a=57.05, b=121.29, c=41.73	a=57.02, b=121.34, c=41.58
Structure refinement				
Resolution (Å)	33.71-1.45			19.68-1.67
R_{work}/R_{free} [a]	0.17/0.19			0.16/0.19
Number of atoms				
Protein	2,074			2.074
Water	242			253
B-factors				
Protein	25.8			22.9
Water	32.9			44.5
R.m,s deviations				
Bond lengths (Å)	0.009			0.010
Bond angles (°)	1.151			1.190
Ramachandran plot (%)				
Most favored regions	97.56			97.97
Allowed regions	2.44			2.03
Disallowed regions	0.00			0.00

*Values in the parentheses are for the outermost resolution shells.

**Rmerge=$\Sigma_{hkl}\Sigma_i|I_i(hkl)-\langle I(hkl)\rangle|/\Sigma_{hkl}\Sigma_i I_i(hkl)$, where $I_i(hkl)$ is the observed intensity of reflection hkl and $\langle I(hkl)\rangle$ is the averaged intensity of symmetry-equivalent measurements

[a] $R_{work}=\Sigma_{hkl}||F_{obs}|-|F_{calc}||/\Sigma_{hkl}|F_{obs}|$, where F_{obs} and F_{calc} are structure factors of the observed reflections and those calculated from the refined model, respectively. R_{free} has the same formula as R_{work}, except that it was calculated against a test set of the data that was not included in the refinement

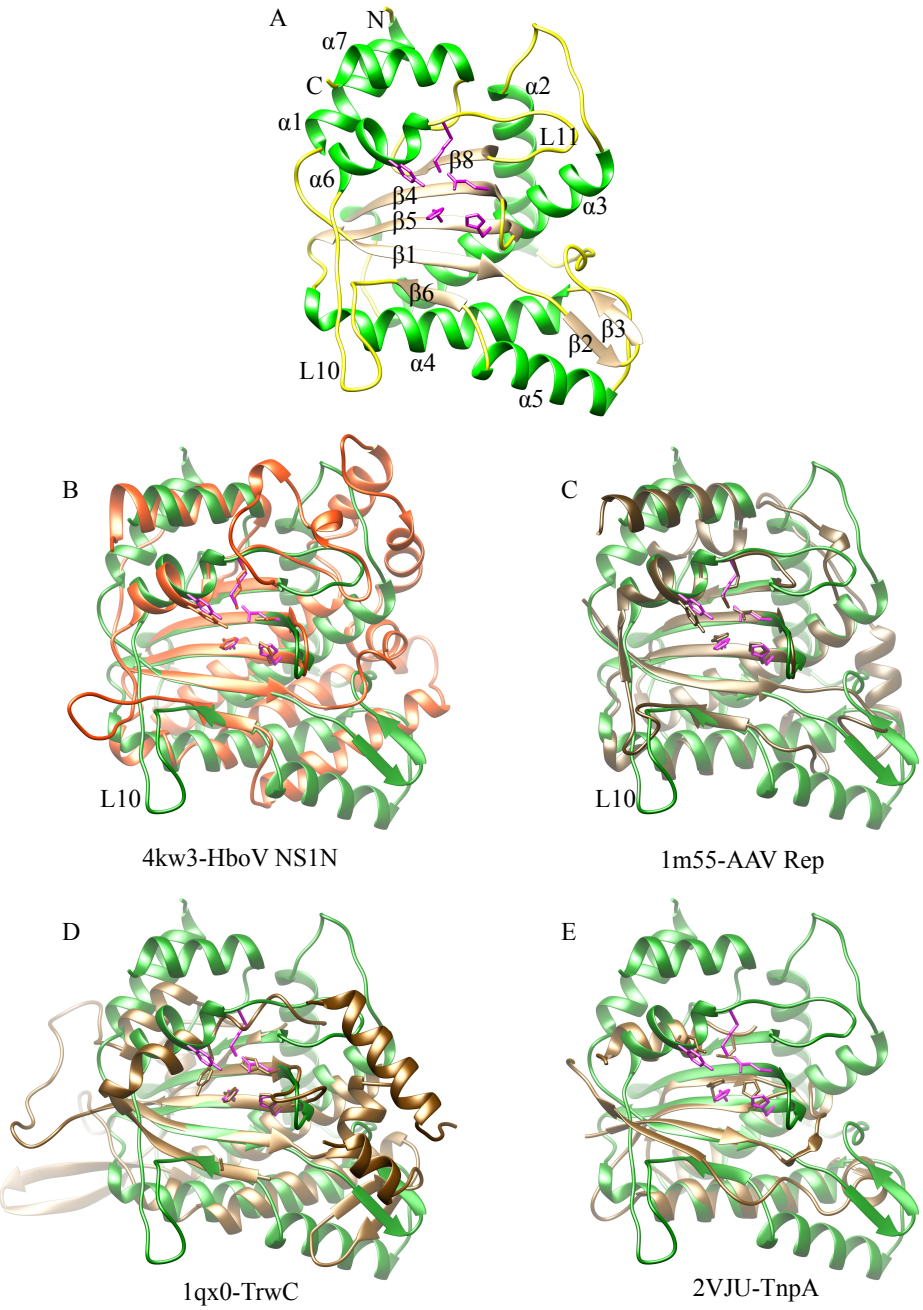


Fig. 5-2. *(A)* The overall structure of MVM NS1N. The α helices, β strands and loops are in green, brown and yellow, respectively. Side chains of the nickase active site residues are shown as stick models in magenta. The amino- and carboxyl-termini are indicated with N and C respectively. The secondary structural elements are labeled. *(B-E)* superimposition of MVM NS1N (green) with HBoV NS1N (orange), AAV Rep (brown), TrwC (brown) and TnpA (brown) respectively. The nickase active site residues are shown as stick models.

Structure superposition with the HboV NS1 (18) and AAV Rep (19) reveals highly conserved left and central motifs but less conserved right motif (**Fig. 5-2B and C**), suggesting that the left and central

motifs are involved in the conserved function-nickase while the right one is likely involved in virus-specific functions. The nickase active site of MVM NS1N contains two histidine residues (His127/His129) separated by a spacer residue Cys128 (**Fig. 5-2A**). The two His residues are approached by Glu119, Lys 214, and the catalytic residue Tyr210 (12). The structure-based sequence alignment of MVM NS1N with AAV Rep and HBoV NS1N shows strict conservation of these active site residues (**Fig. 5-3**), reflecting the structural and functional unity of the active site architecture in the *Parvoviridae* family. The core fold comprised of central β-sheet and bilateral helices bears apparent similarity to bacterial conjugative relaxases TrwC (25) and TraI (26) and transposase TnpA (27, 28), which are all members of the HUH-superfamily nucleases (**Fig. 5-2D and E**). The histidine-hydrophobic-histidine (HUH) motif and the catalytic tyrosine residue, comprises the characteristic structural elements that are found to be ubiquitous among the HUH-superfamily endonucleases that cleave and/or ligate ssDNA and play roles in DNA replication in some bacterial and eukaryotic viruses, bacterial conjugation and transposition (29).

Nevertheless, the MVM NS1N nickase active site exhibits differences from those of other HUH-superfamily endonucleases. A third His residue is frequently observed in conjugative relaxases TrwC and TraI, while in MVM NS1N this position is occupied by Glu119, which is invariable among parvoviruses (**Fig. 5-3**). In transposase TnpA, this position is a Gln residue. Such a difference may result in differential preferences and binding affinity for metal ligands. Several HUH endonucleases, such as gpA in bacteriophage phiX174 (30), the A protein in bacteriophage P2 (31), conjugative relaxases TrwC (25) and TraI (26), contain two Tyr residues in the active site, which are functionally essential and are believed to alternate in the cleavage and ligation reactions. MVM NS1N contains only a single active Tyr residue (12), which superposes well with Tyr18 in TrwC and Tyr16 in TraI. The lack of a second catalytic tyrosine residue in MVM NS1 supports the idea that no joining reaction is needed in NS1-directed DNA replication in MVM.

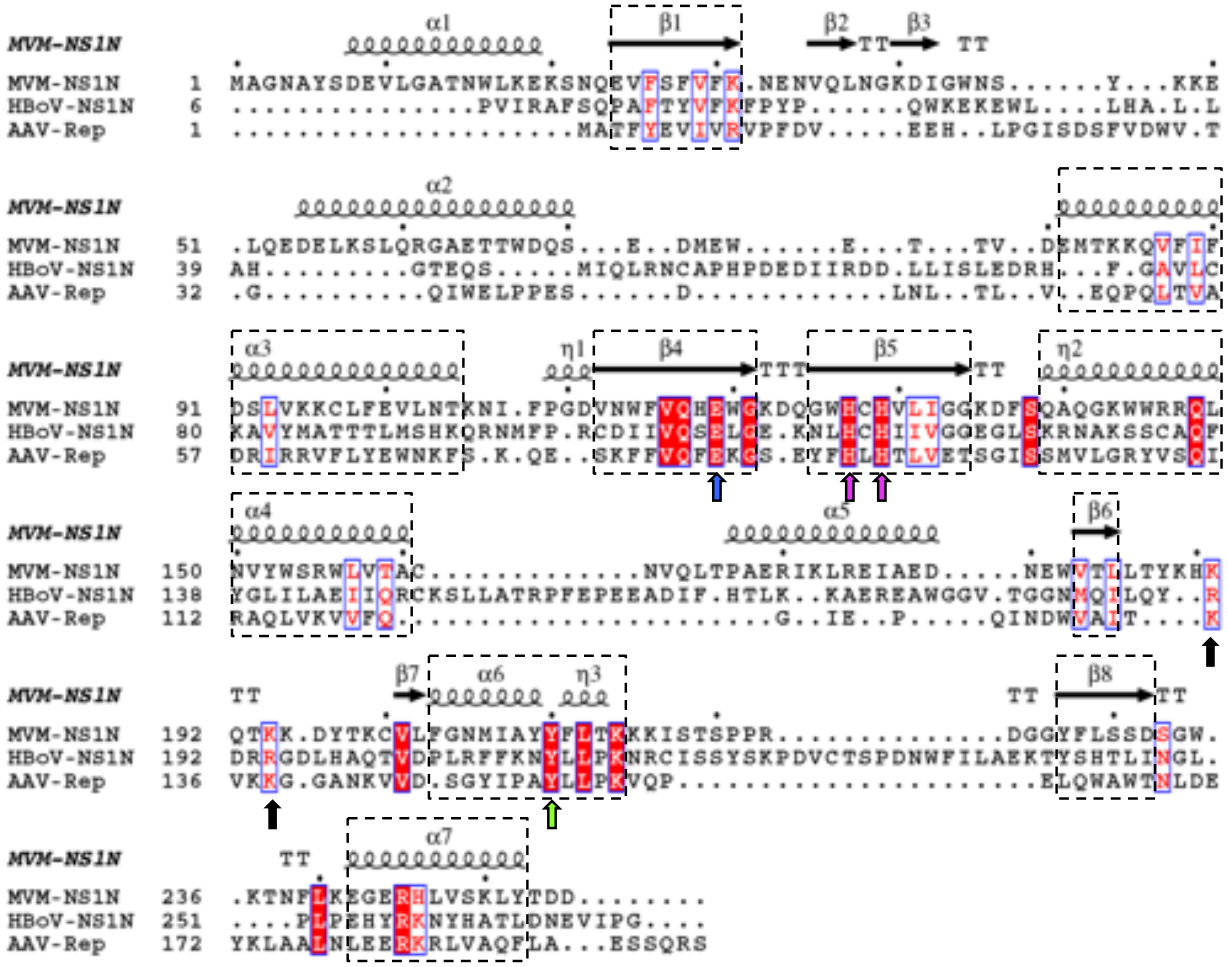


Fig. 5-3. Structure-based sequence alignment of the parvovirus N-terminal nuclease domains. The secondary structural elements based on the structure of MVM NS1N are shown above the sequences. α, α-helices; β, β-strands; T, turn; η, 3_{10} helices. The conserved core folds among MVM NS1N, AAV Rep and HBoV NS1N are highlighted in dashed rectangles. The conserved active site residues His117/His119, Tyr210 and Glu119 are indicated with arrows in magenta, green and blue, respectively. The Ori-binding loop of MVM NS1N exhibits a KXXKK motif (black arrows), corresponding to KXKK in AAV Rep and RXRR in HBoV NS1, where X is any amino acid residue.

5.3.3 The MVM NS1 nickase active site can coordinate Mg^{2+}

It has been known that divalent metal ions are required for DNA cleavage activities for the HUH-superfamily of endonucleases, although it remains an open question what metal ion is physiologically used. For example, it was reported that DNA cleavage activity of the relaxase TrwC, an HUH-superfamily endonuclease, can be activated by a variety of divalent metal ions including Mg^{2+}, Mn^{2+}, Ca^{2+}, Zn^{2+}, Cu^{2+} and Ni^{2+} in in vitro assays (32). DNA nicking of AAV Rep is metal-dependent (33). Mn^{2+} but not Mg^{2+} supported DNA nicking by AAV Rep (19, 34). Mn^{2+} and Zn^{2+} but not Mg^{2+} bound to

the AAV Rep nickase active site (19), but Mg^{2+} was observed in the nickase active site of AAV Rep complexed with a DNA stem-loop harboring the secondary Rep-binding element (20). Different with the case of AAV Rep, Mg^{2+} was shown to support DNA nicking in MVM NS1 (11, 12). We took the crystallographic approach to gain insights into metal ligand binding in MVM NS1N. The MVM NS1N native structure does not display any metal ion in the active site (**Fig. 5-4A**). The MVM NS1N crystals were soaked with 100mM $MgCl_2$ and the structure was determined and refined (**Table 5-1, Fig. 5-4B**). It shows excellent electron density in the active sites for Mg^{2+}. The Mg^{2+} is coordinated with the epsilon nitrogen atoms of His127 and His129, a side chain carboxyl oxygen atom of Glu119, and three water molecules in a nearly perfect six-member, octahedral configuration.

In addition to this, our lab also observed the binding of other metal ions including Mn^{2+}, Ni^{2+}, Co^{2+}, Cu^{2+}, and Zn^{2+}, suggesting that the MVM NS1 nickase active site is highly versatile in metal ligand binding. It remains to be determined which metal ion supports the nickase activity of MVM NS1.

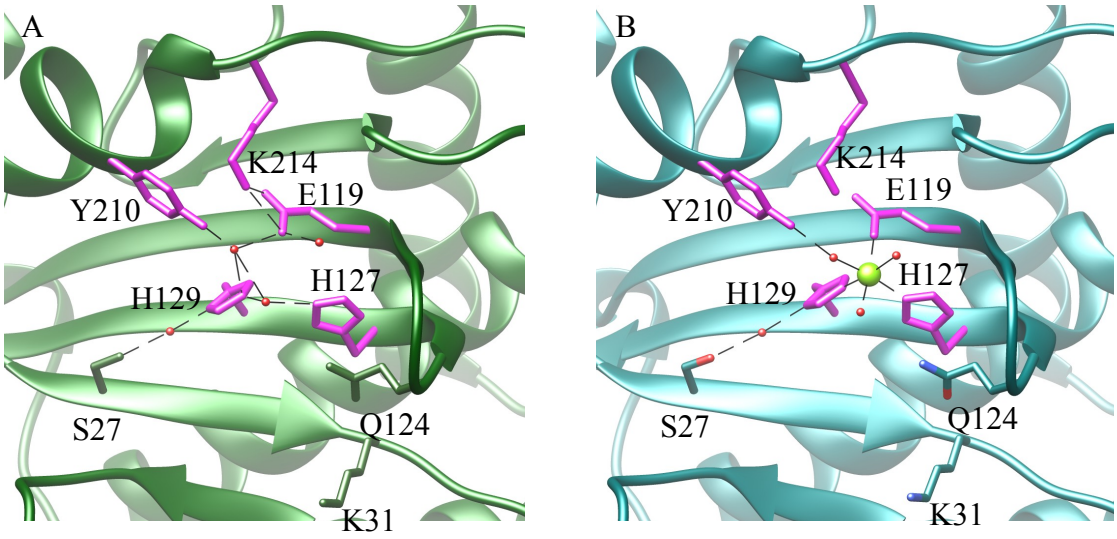


Fig. 5-4. *(A)* The H-bond network in the nickase active site of the native structure. The active site residues are shown as stick models in magenta and labeled. Residues Ser27, Gln124 and Lys31 are shown as stick models in green. Water molecules, red spheres. H-bond, dashed lines. *(B)* The conformation of the nickase active site bound with the six-member, octahedrally coordinated Mg^{2+} (green sphere).

5.3.4 Insight into ssDNA nicking and dsDNA binding.

Soaking with ssDNA molecules did not generate an X-ray structure of NS1N complexed with ssDNA. To gain insight into the binding and cleavage of the ssDNA substrate, the MVM NS1N structure was superimposed with that of the Y16F mutant of the conjugative plasmid F factor TraI, a HUH endonuclease, in complex with a ssDNA molecule that extends beyond its nicking site (26). Superimposition of the two structures resulted in a root mean square deviation (RMSD) of 2.64 Å for 104 Cα atoms. The MVM NS1N residues His127 and His129 align well with His157 and His159 of the TraI HUH motif, respectively, while Glu119 in MVM NS1N superposes with the third histidine residue (His146) in the TraI histidine triad (**Fig. 5-5A**). The MVM NS1N Tyr210 occupies a same position as the TraI catalytic residue Tyr16 (mutated into Phe in the structure). The metal ions in the MVM NS1N structures fit well with the Mg^{2+} ion in the TraI active site.

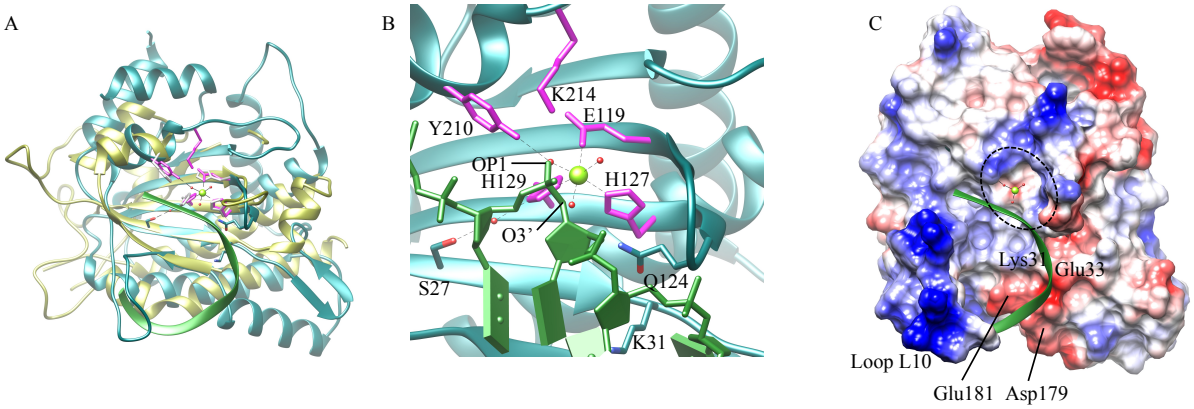


Fig. 5-5. A model for binding of MVM NS1N to the ssDNA substrate. *(A)* Superposition of MVM NS1N (cyan) with TnpA (yellow; 2a0i). The nickase active site residues are shown as stick models in magenta for MVM NS1N and yellow for TnpA. The bound Mg^{2+} in MVM NS1N is shown as a green sphere. The backbone of the bound ssDNA in TnpA is shown as a ribbon diagram in green. *(B)* Closeup of the MVM NS1N nickase active site docked with the ssDNA (green) from TnpA. The three water molecules coordinated with the Mg^{2+} are shown as red spheres. Notice the OP1 and O3' atoms of ssDNA (indicated) nearly overlap with two Mg^{2+}-coordinating water molecules respectively. *(C)* Electrostatic potential surface of MVM NS1N with docked ssDNA (green ribbon diagram) showing how the DNA aligns with the nickase active site (dashed ellipsoid) and approaches the catalytically

essential metal ion (green sphere). Residues Lys31, Glu33, Asp179 and Glu181 as well as the loop L10 are indicated.

The superimposition shows that the ssDNA substrate docks onto the MVM NS1N active site via a curved path, circumventing the loop L10, crossing β6 and β1, and extending into the active site (**Fig. 5-5A**). Interestingly, two water molecules coordinating with Mg^{2+} in MVM NS1N occupy nearly the same positions as oxygen atoms O3′ and OP1 in a backbone phosphate in the docked DNA from TraI (**Fig. 5-5B**), and the phosphate is positioned between the catalytic Tyr210 and the bound Mg^{2+}, with a distance of 2.69 and 2.38 Å from the oxygen atom OP1 of the DNA backbone phosphate to the side chain hydroxyl oxygen of Tyr210 and the Mg^{2+}, respectively. The oxygen atom O3' from the DNA backbone is at a distance of 2.51 Å from the Mg^{2+}. The phosphorous atom of the DNA backbone phosphate lies within 3.35 Å from the side chain hydroxyl oxygen of Tyr210. Such a conformation of the MVM NS1N active site together with the docked DNA may mimic the catalytic intermediate state of MVM NS1N poised for cleavage of the scissile phosphate bond of ssDNA.

It has been known that NS1 binds to the viral telomere to direct viral DNA replication during infection (35), suggesting that the initial binding site for NS1 in the Ori region in viral DNA is of double-stranded nature. Analysis of the electrostatic potential surface of MVM NS1N shows a highly basic region formed by the loop L10 (residues 186–202) and the N-terminal proximity of helix α4, which is rich in positively charged residues and can potentially serve to bind to DNA (**Fig. 5-6A**). To gain insight into the detailed protein:DNA interaction in Ori-recognition, the MVM NS1N structure is superimposed onto the structure of AAV Rep complexed with DNA, resulting in an RMSD of 2.42 Å for 159Cα atoms (**Fig. 5-2C**). The loop L10 is considerably longer than the corresponding loops in AAV Rep and HBoV NS1. To avoid clashes between the lengthier loop L10 and DNA, the model was manually adjusted by moving the DNA slightly. This results in a plausible model for DNA binding of MVM NS1, in which the loop L10 inserts into the DNA major groove, and the N-terminal proximity of helix α4 interacts with the DNA minor groove, akin to those observed in AAV Rep (20).

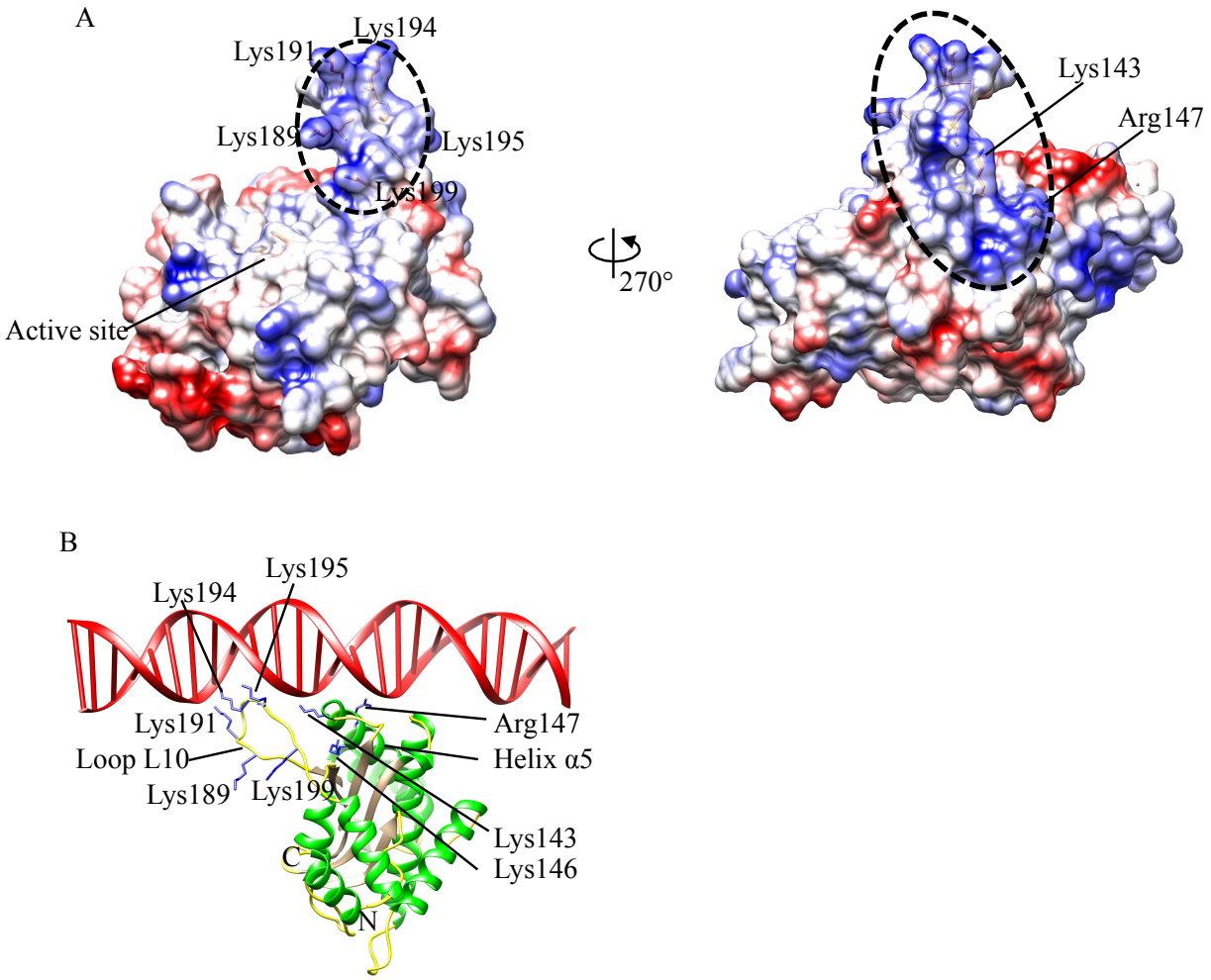


Fig. 5-6. An Ori-binding model of MVM NS1N. *(A)* Electrostatic potential surface of MVM NS1N. Positively charged residues in the putative Ori-binding site (dashed ellipsoid) are indicated. *(B)* The MVM NS1N structure with docked dsDNA.

The MVM NS1N loop L10 contains three basic residues Lys191, Lys194 and Lys195, corresponding to Lys135, Lys137 and Lys138 of AAV Rep, which may interact with DNA by inserting side chains into the DNA major groove as in AAV Rep (**Fig. 5-6B**) (20). These residues in MVM NS1 form a KXXXK motif, showing consistence with the KXKK/RXRR motif in AAV Rep and HBoV NS1 respectively (**Fig. 5-3**) but meanwhile indicating variations in details of interactions with DNA. The loop L10 contains two lysine residues (Lys189 and Lys199), which may also be involved in DNA interaction

(**Fig. 5-6B**). In addition to the differences in lengths, the loop L10 in MVM NS1 adopts a conformation different from those in HBoV and AAV Rep (**Fig. 5-2B** and **C**). This suggests that these loops bear structural flexibility that allows conformational changes necessary for DNA binding. This is consistent with the weak electron density for the residues 191-193 located at the tip of the loop in MVM NS1N and the fact that the loops of the two molecules in the crystallographic asymmetric unit in the HBoV NS1N X-ray structure adopt different conformations (18). Several positively charged residues near the N-terminal end of helix α4, Lys143, Arg146 and Arg147, are proximal to the DNA phosphate backbones, providing additional charge-neutralizing interaction with DNA. The presence of positively charged residues around the N-terminal end of helix α4 is a conserved feature in AAV Rep and HBoV NS1, albeit the exact locations of those residues vary (**Fig. 5-3**).

The conserved pattern of protein:DNA interactions as well as variations in details among MVM, AAV and HBoV are in agreement with similarities and differences in the sequences and structures of the NS1-bound DNA motifs in those parvoviruses. The Rep-binding site in the AAV Ori is five consecutive tetranucleotide motifs (20). In MVM, the OriL and OriR also contain tetranucleotide motifs 5'-TGGT-3' (36), although the repeating number differs from that in AAV. The GC-rich NS1-binding site in the Ori region of B19V can be viewed as five consecutive repeats of the tetranucleotide motif 5'-GCCG-3' (37, 38). While the exact DNA sequences of those motifs differ, the presence of consecutive tetranucleotide motifs appears to be a conserved feature among AAV, MVM and B19V, reflecting a conserved mode for NS1/Rep:DNA interactions.

These results support a unified mechanism for recognition of replication origins by NS1/Rep for homo-telomeric parvoviruses such as AAV and B19V and hetero-telomeric parvoviruses such as MVM and HBoV, which is mediated by a basic-residue-rich surface loop or hairpin and an adjacent helix in NS1/Rep proteins and tandem tetranucleotide motifs in viral DNA replication origin.

5.3.5 X-ray structure of NS1(6-264) suggests an unstructured region.

T278 is a consensus Protein Kinase C (PKC) phosphorylation site, located within the essential self-association motif (amino acids 261-278) (39). It is hypothesized that phosphorylation at this site

could influence NS1 oligomerization. A T278A mutant supports this hypothesis by showing completely abolished NS1's ability to co-transport cytoplasmic mutants of NS1 into the nucleus (7). Since glutamic acid mimics the phosphorylated threonine, we decide to make a NS1(1-285) T278D mutant through site-directed mutation to see if this promotes protein oligomerization. However, the protein remains as monomeric regardless of mutation. The X-ray structure did not show a defined 278 residue (**Fig. 5-6**), suggesting that it is in a loop region. Within the NS1 self-association motif, residues 261-266 and 276-278 are critical for interaction (39). The NS1(1-285)T278D structure shows an unstructured region from residue 256 till 264, stabilized by hydrogen bonding with the ß2 of an adjacent molecule in crystal (**Fig. 5-7**). This structure implies that the NS1 self-association may mediated by a loop region.

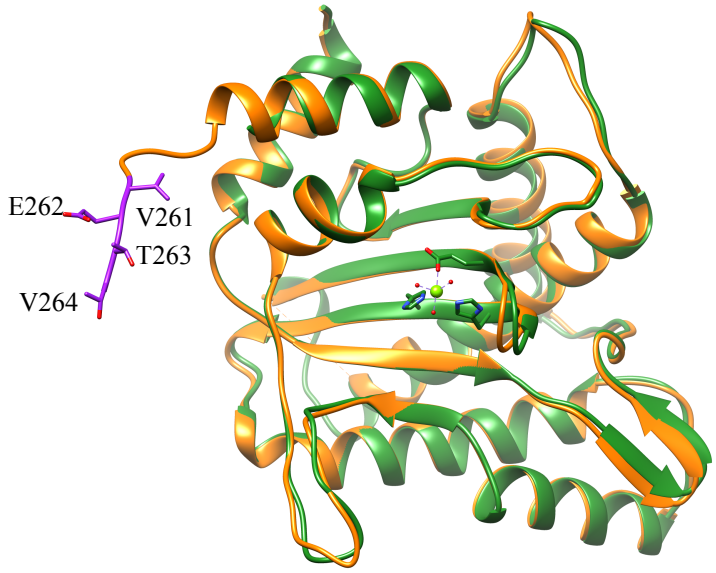


Fig. 5-7. The residues 261-264 are extended in the NS1(1-285)T278D construct. The NS1(1-255) (green) is superposed on the NS1(1-285)T278D structure (orange). Residues 261-264 are in purple and labelled.

Contributions: The research was directed by Haiyan Zhao and Liang Tang. The native X-ray structure of NS1N and the Mn^{2+}, Ni^{2+}, Co^{2+}, Cu^{2+}, and Zn^{2+} derivatives were solved and analyzed by Sunil K. Tewary. Lingfei performed the EMSA; solved the Mg^{2+} coordinated NS1N structure and the NS1(6-264) structure.

5.4 References

1. Legrand C, Mousset S, Salome N, & Rommelaere J (1992) Cooperation of oncogenes in cell transformation and sensitization to killing by the parvovirus minute virus of mice. *J Gen Virol* 73 (Pt 8):2003-2009.
2. Mousset S, Ouadrhiri Y, Caillet-Fauquet P, & Rommelaere J (1994) The cytotoxicity of the autonomous parvovirus minute virus of mice nonstructural proteins in FR3T3 rat cells depends on oncogene expression. *Journal of virology* 68(10):6446-6453.
3. Dupont F*, et al.* (2000) Tumor-selective gene transduction and cell killing with an oncotropic autonomous parvovirus-based vector. *Gene Ther* 7(9):790-796.
4. Cornelis JJ*, et al.* (2004) Cancer gene therapy through autonomous parvovirus--mediated gene transfer. *Curr Gene Ther* 4(3):249-261.
5. Cotmore SF, Christensen J, Nuesch JP, & Tattersall P (1995) The NS1 polypeptide of the murine parvovirus minute virus of mice binds to DNA sequences containing the motif [ACCA]2-3. *Journal of virology* 69(3):1652-1660.
6. Christensen J, Cotmore SF, & Tattersall P (1995) Minute virus of mice transcriptional activator protein NS1 binds directly to the transactivation region of the viral P38 promoter in a strictly ATP-dependent manner. *Journal of virology* 69(9):5422-5430.
7. Corbau R, Duverger V, Rommelaere J, & Nuesch JP (2000) Regulation of MVM NS1 by protein kinase C: impact of mutagenesis at consensus phosphorylation sites on replicative functions and cytopathic effects. *Virology* 278(1):151-167.
8. Nuesch JP, Christensen J, & Rommelaere J (2001) Initiation of minute virus of mice DNA replication is regulated at the level of origin unwinding by atypical protein kinase C phosphorylation of NS1. *Journal of virology* 75(13):5730-5739.
9. Christensen J & Tattersall P (2002) Parvovirus initiator protein NS1 and RPA coordinate replication fork progression in a reconstituted DNA replication system. *Journal of virology* 76(13):6518-6531.
10. Cotmore SF & Tattersall P (1989) A genome-linked copy of the NS-1 polypeptide is located on the outside of infectious parvovirus particles. *Journal of virology* 63(9):3902-3911.
11. Nuesch JP, Cotmore SF, & Tattersall P (1995) Sequence motifs in the replicator protein of parvovirus MVM essential for nicking and covalent attachment to the viral origin: identification of the linking tyrosine. *Virology* 209(1):122-135.
12. Willwand K*, et al.* (1997) The minute virus of mice (MVM) nonstructural protein NS1 induces nicking of MVM DNA at a unique site of the right-end telomere in both hairpin and duplex conformations in vitro. *J Gen Virol* 78 (Pt 10):2647-2655.
13. Wilson GM, Jindal HK, Yeung DE, Chen W, & Astell CR (1991) Expression of minute virus of mice major nonstructural protein in insect cells: purification and identification of ATPase and helicase activities. *Virology* 185(1):90-98.
14. Christensen J, Cotmore SF, & Tattersall P (1999) Two new members of the emerging KDWK family of combinatorial transcription modulators bind as a heterodimer to flexibly spaced PuCGPy half-sites. *Molecular and cellular biology* 19(11):7741-7750.
15. Cotmore SF, Christensen J, & Tattersall P (2000) Two widely spaced initiator binding sites create an HMG1-dependent parvovirus rolling-hairpin replication origin. *Journal of virology* 74(3):1332-1341.
16. Christensen J, Cotmore SF, & Tattersall P (2001) Minute virus of mice initiator protein NS1 and a host KDWK family transcription factor must form a precise ternary complex with origin DNA for nicking to occur. *Journal of virology* 75(15):7009-7017.
17. Rhode SL, 3rd & Richard SM (1987) Characterization of the trans-activation-responsive element of the parvovirus H-1 P38 promoter. *Journal of virology* 61(9):2807-2815.

18. Tewary SK, Zhao H, Shen W, Qiu J, & Tang L (2013) Structure of the NS1 protein N-terminal origin recognition/nickase domain from the emerging human bocavirus. *Journal of virology* 87(21):11487-11493.
19. Hickman AB, Ronning DR, Kotin RM, & Dyda F (2002) Structural unity among viral origin binding proteins: crystal structure of the nuclease domain of adeno-associated virus Rep. *Molecular cell* 10(2):327-337.
20. Hickman AB, Ronning DR, Perez ZN, Kotin RM, & Dyda F (2004) The nuclease domain of adeno-associated virus rep coordinates replication initiation using two distinct DNA recognition interfaces. *Molecular cell* 13(3):403-414.
21. Kabsch W (2010) Xds. *Acta crystallographica. Section D, Biological crystallography* 66(Pt 2):125-132.
22. Winter G, Lobley CM, & Prince SM (2013) Decision making in xia2. *Acta crystallographica. Section D, Biological crystallography* 69(Pt 7):1260-1273.
23. Adams PD*, et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta crystallographica. Section D, Biological crystallography* 66(Pt 2):213-221.
24. Emsley P, Lohkamp B, Scott WG, & Cowtan K (2010) Features and development of Coot. *Acta crystallographica. Section D, Biological crystallography* 66(Pt 4):486-501.
25. Guasch A*, et al.* (2003) Recognition and processing of the origin of transfer DNA by conjugative relaxase TrwC. *Nat Struct Biol* 10(12):1002-1010.
26. Larkin C*, et al.* (2005) Inter- and intramolecular determinants of the specificity of single-stranded DNA binding and cleavage by the F factor relaxase. *Structure* 13(10):1533-1544.
27. Barabas O*, et al.* (2008) Mechanism of IS200/IS605 family DNA transposases: activation and transposon-directed target site selection. *Cell* 132(2):208-220.
28. Ronning DR*, et al.* (2005) Active site sharing and subterminal hairpin recognition in a new class of DNA transposases. *Molecular cell* 20(1):143-154.
29. Chandler M*, et al.* (2013) Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. *Nature reviews. Microbiology* 11(8):525-538.
30. Hanai R & Wang JC (1993) The mechanism of sequence-specific DNA cleavage and strand transfer by phi X174 gene A* protein. *The Journal of biological chemistry* 268(32):23830-23836.
31. Odegrip R & Haggard-Ljungquist E (2001) The two active-site tyrosine residues of the a protein play non-equivalent roles during initiation of rolling circle replication of bacteriophage p2. *Journal of molecular biology* 308(2):147-163.
32. Boer R*, et al.* (2006) Unveiling the molecular mechanism of a conjugative relaxase: The structure of TrwC complexed with a 27-mer DNA comprising the recognition hairpin and the cleavage site. *Journal of molecular biology* 358(3):857-869.
33. Davis MD, Wu J, & Owens RA (2000) Mutational analysis of adeno-associated virus type 2 Rep68 protein endonuclease activity on partially single-stranded substrates. *Journal of virology* 74(6):2936-2942.
34. Yoon M*, et al.* (2001) Amino-terminal domain exchange redirects origin-specific interactions of adeno-associated virus rep78 in vitro. *Journal of virology* 75(7):3230-3239.
35. Li L, Cotmore SF, & Tattersall P (2013) Parvoviral left-end hairpin ears are essential during infection for establishing a functional intranuclear transcription template and for efficient progeny genome encapsidation. *Journal of virology* 87(19):10501-10514.
36. Cotmore SF & Tattersall P (2005) A rolling-hairpin strategy: basic mechanisms of DNA replication in the parvoviruses. *Parvoviruses*, ed J. Kerr, S.F. Cotmore, M.E. Bloom, R.M. Linden and C.R. Parrish (Hodder Arnold, London, United kingdom), pp 171-181.
37. Tewary SK, Zhao H, Deng X, Qiu J, & Tang L (2014) The human parvovirus B19 non-structural protein 1 N-terminal domain specifically binds to the origin of replication in the viral DNA. *Virology* 449:297-303.

38. Guan W, Wong S, Zhi N, & Qiu J (2009) The genome of human parvovirus b19 can replicate in nonpermissive cells with the help of adenovirus genes and produces infectious virus. *J Virol* 83(18):9541-9553.
39. Pujol A*, et al.* (1997) Inhibition of parvovirus minute virus of mice replication by a peptide involved in the oligomerization of nonstructural protein NS1. *Journal of virology* 71(10):7393-7403.

Chapter 6. Overall Discussion and Future Directions

6.1 Podovirus Sf6 tail machine

Podovirus tail machine is relatively simple in composition compared with those long tails. Of the four tail proteins, three have high-resolution X-ray structures in both Sf6 and P22. Tail nozzle Sf6-gp8/P22-gp10 is the only one that has never been structurally described. Nevertheless, dynamic processes such as tail assembly and structure rearrangement during DNA injection are poorly understood, likely due to the small size compared with the capsid.

Genetic studies suggest that podovirus P22 assembles tail machine in a sequential manner (1). Our comparative analysis on Sf6-gp7 and P22-gp4 X-ray structures provides a molecular mechanism that mediates the sequential attachment of tail nozzle by repositioning a conserved sequence motif to create a new docking site (**Chapter 2**). Sequential assembly mediated by structure conformational change is common in many cases. One example is that myovirus SPP1 tail protein gp15 undergoes large conformational change during assembly to allow appropriate interactions with the next tail protein gp16. In comparison, Sf6-gp7 is unique with the repositioning of the octad sequence motif. Although we listed several strong evidences to support this proposal, direct evidence, such as gp7 structure in a post-assembly state, will be more convincing. This can in principle be achieved by determining a high-resolution structure of Sf6 tail machine through cryoEM. Crystallization of Sf6 portal and tail adaptor complex as that of P22 is unlikely because we find that Sf6 portal does not form homo-oligomer in solution as the P22 portal does. We tried to mutate the key residues in the gp7 first motif to drive the second motif to replace the first one. However, the protein is very sensitive to such mutations and forms inclusion bodies when overexpressed in *E.coli* cells.

Another piece of information we can confirm by a high-resolution structure of Sf6 tail machine is the interaction interface between gp7 and gp8. In Chapter 3, we obtained a low-resolution beads model of gp8 monomer by SAXS. Fitting the gp8 monomer into cryoEM map of P22 tail suggests possible interfaces between gp8 and other tail components. We find that, in gp7, the loop 1 is the only region involved in

interaction with gp8. However, this is at relatively low resolution, we expect to see how the interface looks like, especially how one gp8 molecule interacts with two same loops.

6.2 Podovirus Sf6 DNA-injection apparatus

Given the fact that the P22 portal-tail adaptor channel has a constriction of 33Å (2) and that all internal proteins can be ejected prior to DNA (3), it is hypothesized that the internal proteins have to be unfolded. The gp12NTD/gp20NTD monomeric structure turns out to have a dimension of 80Å*44Å*27Å, which surely cannot go through the tail channel as folded. The architecture of the helices and positions of Gly and Pro residues both indicate that gp12/gp20 can potentially unfold to a linear string of helices. The other two internal proteins gp11 and gp13 are also predicted by Phyre2 (4) to be helical only. We found that the melting temperature of gp12 is as low as 42℃, suggesting that only a small amount of energy is required for protein unfolding. The factors during infection process that could facilitate partial unfolding and refolding includes pressure and chemical environment. However, it is pointed out that protein unfolding occurs only at high pressure of several kilobars (5, 6). Physical pressure inside virus capsid, tens of atmospheres, will not break tertiary structure. Instead, the dehydration environment in the capsid is probably the dominant factor that stabilizes unfolded internal proteins (7). To package the complete genome, some hydration layers must be stripped from the DNA and internal proteins. Hence, the absence of hydration shell around the protein therefore leads to exposure of hydrophobic cores within the capsid. When proteins are ejected into cell envelope, full hydration allows hydrophobic collapse and protein refolding. This assumption is consistent with the gp12NTD structure in which there are very few inter-helical hydrogen bonds. The major interactions that hold the tertiary structure are the two hydrophobic cores in the three-helical bundle and the four-helical bundle respectively. Such hydrophobic interactions are thus the key for protein partially unfolding and folding.

An open question is how internal proteins coordinate with each other and with DNA to be ejected. Since internal proteins locate around the barrel of the portal, close to the center of capsid (8), whereas the end of viral DNA is in the lumen of tail nozzle gp8, much closer to the proximal opening of tail channel.

What force contributes to retaining DNA at place but allowing internal proteins to be ejected? We would consider the internal proteins are ejected in a specific order. This might be resolved if the locations of these proteins in cell envelope are determined. We have established well-developed protocol for production of Sf6 particles with extended tail. High-resolution structure of the extended tail determined by cryoEM or cryoET would have critical significance on revealing the podovirus DNA injection mechanism.

6.3 Structures and functions of MVM NS1

MVM NS1 is an 83kDa protein with three domains. Each domain is integrated with several functions. The N-terminal domain recognizes Ori dsDNA and have site- and strand-specific nickase activity (9). The middle domain binds and hydrolyzes ATP, using the energy to unwind dsDNA (10, 11). The C-terminal domain is mainly involved in interaction with cellular factors. Additionally, there are some other motifs for i.e. nuclear localization signal, oligomerization, trans-activation, and phosphorylation.

We solved the X-ray structure of the N-terminal domain, revealing a conserved catalytic fold with His-hydrophobic-His motif. By superposing with orthologs, we came up with a MVM NS1N ssDNA and dsDNA binding model. The model is likely to be true given the conservation of proteins and the matching between protein and DNA. However, we suggest that it is worthwhile to determine the structure of NS1 binding with ssDNA and dsDNA. One possible approach is to soak crystals with small segments of ssDNA or dsDNA. It largely depends on the size of solvent channel whether DNA strand can go to certain binding sites. Another way is to pre-mix protein and DNA to obtain complex for crystallization. However, it can be difficult to get pure and homogeneous sample.

6.4 References

1. Israel V (1977) E proteins of bacteriophage P22. I. Identification and ejection from wild-type and defective particles. *Journal of virology* 23(1):91-97.
2. Tang J*, et al.* (2011) Peering down the barrel of a bacteriophage portal: the genome packaging and release valve in p22. *Structure* 19(4):496-502.
3. Jin Y*, et al.* (2015) Bacteriophage P22 ejects all of its internal proteins before its genome. *Virology* 485:128-134.
4. Kelley LA, Mezulis S, Yates CM, Wass MN, & Sternberg MJ (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols* 10(6):845-858.

5. Hillson N, Onuchic JN, & Garcia AE (1999) Pressure-induced protein-folding/unfolding kinetics. *Proceedings of the National Academy of Sciences of the United States of America* 96(26):14848-14853.
6. Zhang M & Wu Y (2011) Pressure-induced structural and hydration changes of proteins in aqueous solutions. *Anal Sci* 27(11):1139-1142.
7. Leptihn S, Gottschalk J, & Kuhn A (2016) T7 ejectosome assembly: A story unfolds. *Bacteriophage* 6(1):e1128513.
8. Wu W*, et al.* (2016) Localization of the Houdinisome (Ejection Proteins) inside the Bacteriophage P22 Virion by Bubblegram Imaging. *MBio* 7(4).
9. Nuesch JP, Cotmore SF, & Tattersall P (1995) Sequence motifs in the replicator protein of parvovirus MVM essential for nicking and covalent attachment to the viral origin: identification of the linking tyrosine. *Virology* 209(1):122-135.
10. Wilson GM, Jindal HK, Yeung DE, Chen W, & Astell CR (1991) Expression of minute virus of mice major nonstructural protein in insect cells: purification and identification of ATPase and helicase activities. *Virology* 185(1):90-98.
11. Christensen J, Pedersen M, Aasted B, & Alexandersen S (1995) Purification and characterization of the major nonstructural protein (NS-1) of Aleutian mink disease parvovirus. *Journal of virology* 69(3):1802-1809.