Insights and structural comparison of portal proteins in bacteriophages

Morgan Sperratore, Brittany Grenyer, Melinda Harrison*

Cabrini University, 610 King of Prussia Road, Radnor, PA 19087, USA

Received: February 8, 2019; **accepted:** March 29, 2019.

A bacteriophage is a virus that is able to infect a bacterium host. Catdawg is a mycobacteriophage belonging to Cluster O and the siphoviridae family. The portal protein of Catdawg forms a hole, or portal, that enables DNA passage during packaging and ejection. It also forms the junction between the phage head (capsid) and the tail proteins. The structure of many phage portal proteins is unknown and understanding structural information would be valuable to the phage community. In this study, a homology model and GEPARD Plot of Catdawg's portal protein was generated. The predicted homology model was validated using Procheck and then compared to known portal protein crystal structures. The results confirm that Catdawg's portal protein does not contain a similar conformation as the known crystal structures. The shapes of the known protein capsids were icosahedral and Catdawg's portal protein contains a prolate shape which may contribute to why the structures were different.

Keywords: Bacteriophages; portal protein; homology model.

Financial support: MH wished to thank Cabrini University for the generous support of this research project.

***Corresponding author:** Melinda Harrison, Cabrini University, 610 King of Prussia Road, Radnor, PA 19087, USA. E-mail: Melinda.a.harrison@cabrini.edu.

Introduction

Bacteriophages (phages) are viruses that infect bacteria and one of the oldest and most ubiquitous organisms on earth [1]. Phages are studied due to their therapeutic use in fighting bacterial infections and as an alternative to drug resistance bacteria [1].

Phages were first found in the 1910s and were discovered to be bacteria killers by Fredrick William Twort and Felix d'Herelle [2]. Twort described the appearance of the plaques, the clear area after phage infection and lysis of the cell, as having a "glassy transformation" [3]. This discovery of phages motivated scientists to further investigate them. In the 1940's, researchers used phages to fight Salmonella in chickens and *Clostridium* perfringens in victims who had gas gangrene [3]. Phage infections produce progeny phages and eventually result in the death of the host bacterium. The progeny phages infect neighboring cells until the bacteria are no longer available [3]. There are two life cycles of the phage; lytic and lysogenic [3]. The lytic cycle occurs when the phage infects a host bacterium. The progeny phages are assembled and burst out of the host bacterium. The cycle will continue when the progeny phages land on a neighboring bacteria cell [3]. The other phage cycle is the lysogenic or temperate cycle. Phages that go through the lysogenic cycle have transposases that allows the phage to insert its DNA into the DNA of the host bacterium. The DNA will remain in the host DNA until the conditions are optimal for the assembly of progeny phage. Once the conditions are optimal for replication of the phage, the DNA will excise

out of the host DNA and progeny phage will be assembled [3].

Phages contain 3 main structures: the head, tail, and tail fibers (Figure 1) [4]. The head can be icosahedral or prolate in shape. An icosahedral head or capsid has 12 vertices and 20 identical face [3]. A prolate head contains an elongated head which are few in numbers [5]. The tail is a hollow shaft that is connected to the capsid and the tail fibers. The shaft is helical in shape and allows for the DNA in the capsid to flow through the tail during adsorption. The tail tube punctures the cell and allows for the DNA to enter the cell [3]. There are three types of morphologies of the tail: podovirus, myovirus, and siphovirus (Figure 2) [6]. The podovirus tail morphology is short and stubby. The tail morphology of the myovirus is a contractile tail with a wide sheath around the tail. Siphoviruses are phages with long, non-contractile tail morphology [6, 7, 8]. The tails are essential in the life cycle of the phages [9].

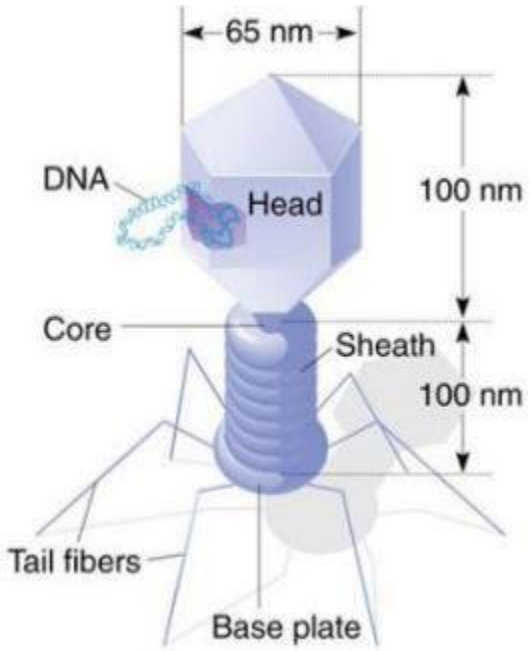

Figure 1. The main structural parts of the phage [4].

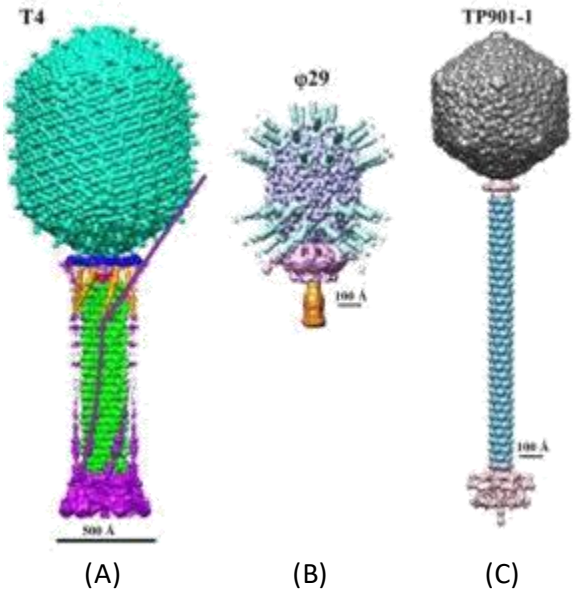

Figure 2. Phage tail morphologies. The different tail morphologies are *myovirus* (A), *podovirus* (B), and *siphovirus* (C). The *myovirus* tail is thicker than the other morphologies. The *podovirus* has the shortest tail of the three and the *siphovirus* has the longest tail [6].

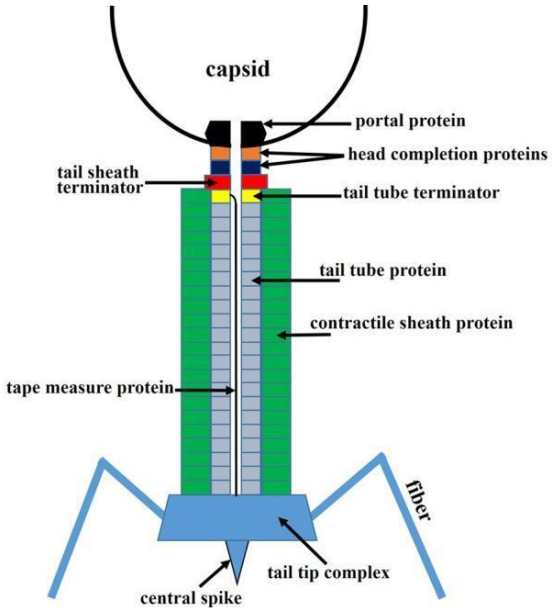

Figure 3. The main structural parts of the general *myovirus* phage [6].

The portal protein forms a hole, or portal, that enables DNA passage during packaging and ejection. It also forms the junction between the phage head (capsid) and the tail proteins. (Figure 3) [10]. Portal proteins have 4 similar regions:

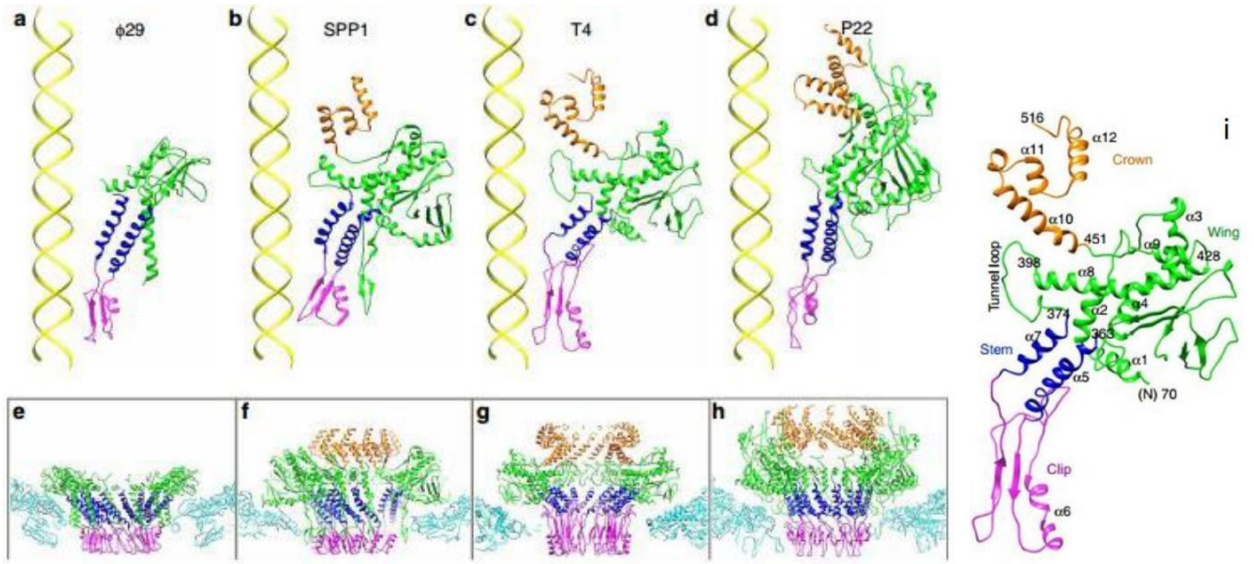

Figure 4. Known crystal structures of portal proteins. Phages that have crystal structure for the portal protein are ϕ29 (*Bacillus* phage) (a and e), SPP1 (*Bacillus subtilis* phage (b and f), T4 (*E. coli*) (c and g), and P22 (*Salmonella*) (d and h). The top row of portal proteins is just one of the 12 identical subunits, where the bottom row is the complete arrangement of all of the 12 identical subunits. The portal protein subunit broken down into each of the parts that make up to subunit: crown, wing, stem, and the clip (i) [16].

crown, wing, stem, and clip. Differences are visible mainly in the crown and wing regions [6]. Another regulatory element that is paired with the portal protein at the base of the head near the neck of the phage is the terminase. The portal protein and terminase work hand-in-hand when the DNA is being packaged into the capsid of the phage. The portal protein helps with the packaging of the DNA and the terminase feeds DNA into the prohead, then cleaves the DNA when the concatemer is fully packaged into the capsid [6, 11]. Concatemers are long, connected sequences of phage DNA that is replicated through rolling circle replication and the cos sites are found at the beginning and end of the complete sequence. The cos sites are the areas in which DNA are cleaved during DNA packaging. When the DNA is finished being packed into the capsid of the phage, the terminase will cut the recognition site in order to package DNA into the next phage capsid [12].

Collections of Phages have been isolated from many different bacteria and their nucleotide sequence studied [13]. Previous studies have shown that phages with similar nucleotide sequence and gene content are grouped into the same cluster designated as "Cluster A, B, C, *etc*." [13, 14]. Some of the clusters can be further divided by nucleotide relatedness into subclusters "A1; B1; *etc*.". Additional studies have indicated that phages swap genes; delete genes and slightly alter their gene sequence [13, 15].

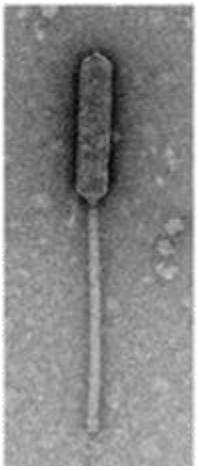

Figure 5. Electromicrograph picture of the mycobacteriophage Catdawg with a prolate head and a long non-contractile tail [5].

Most phage proteins structures have not been determined. To date, 16.6% (1,729 out of 10,420) of *Mycobacterium* phages archived on

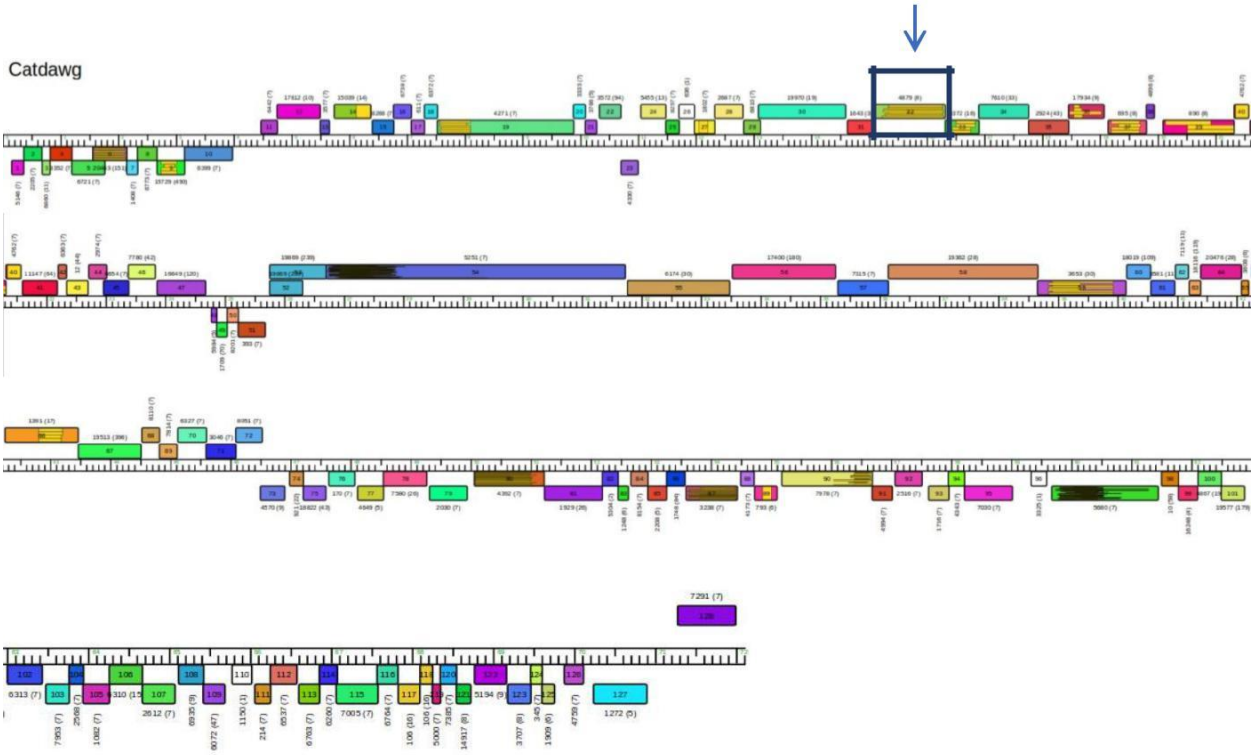

Figure 6. Phamerator map of phage Catdawg. The gene of interest, gene 32, indicated with an arrow [5].

PhagesDB have been sequenced but there is very little structural information on these sequenced proteins [16]. Structural studies were conducted on the lysin B protein in *Mycobacterium smegmatis* phage D29 and several protein structures were determined in the *Mycobacterium tuberculosis* phage H37Rv [17, 18]. The *Mycobacterium abscessus* subsp. bolletii phage portal protein was analyzed, and based on sequence analysis, found to be almost identical to *Corynebacterium diphtheriae* phage (HK97) portal protein [18]. Crystal structures for the portal protein in phages ϕ29 (Bacillus phage), SPP1 (*Bacillus subtilis* phage), T4 (*E. coli*), and P22 (Salmonella) (Figure 4) have been determined [19]. Two additional portal proteins crystal structures were also isolated from phage HK97 and phage G20c [20, 21].

The work described herein involves research with a lytic phage named Catdawg. Catdawg was isolated from *Mycobacterium smegmatis* mc²155 at Cabrini College in 2011 by Sarah Carzo (Figure 5). Phage Catdawg's portal protein is 1,230 bps long, reads in the forward direct, and contains 409 amino acids [16]. The start site for its portal protein was confirmed by Cresawn *et al.*, 2015 via mass spectrometry [5]. The theoretical isoelectric point (pI) is 5.24 and the molecular weight is 45,471.63 Dalton (Da) predicted through exPasy [22]. The Catdawg genome contains 128 genes and is 72,108 bps in size (Figure 6) [16]. Catdawg is an O cluster siphovirus phage and some of its members include phages Dylan, Corndog, Firecracker, Smooch, YungJamal, and Zakhe101 [5, 16]. We seek to determine if the known portal proteins structure will be similar or different from Catdawg portal protein. We also seek the structure of the *Mycobacterium smegmatis* mc²155 phage Catdawg portal protein. We believe the structure may have similar folds to other known portal protein structures which have the same tail morphology.

Materials and methods

Homology modeling of Catdawg portal protein GP32

The program PHYRE2 [23] was used to generate a predicted homology model of GP32, the portal protein. The FASTA sequence of the portal protein was obtained from phagesDB.org and uploaded into the program. The program searched the known crystal structures that contained similarity to the uploaded sequence. Phyre2 uses the alignment of hidden Markov models via HHsearch to significantly improve accuracy of alignment and detection rate [23]. After the generation of a predicted structure, the Protein Data Bank three-dimensional model was uploaded to Procheck, a validation program [24]. Procheck is a program used to determine the accuracy of the predicted model compared to other models [24].

Comparison of known portal protein structures

The sequences of the six known portal proteins were analyzed using the protein software, PyMOL [25, 26], in order to analyze the similarities and differences of the folding of the protein. The bioinformatics program, PyMOL, superimposed the known crystal structures to the portal protein homology model made previously. The root-mean-square deviation (RMSD) of atomic positions is the measure of the average distance between the atoms (usually the backbone atoms) of superimposed proteins [27]. The RMSD values were generated after the alignment of the structures. The percent coverage was calculated between the atoms that were aligned vs. the number of amino acids contained in the homology modeled portal protein. Genome Pair Rapid Dotter (GEPARD) [28] was used to align the protein sequences in a dot plot viewer.

Results

Homology modeling of portal protein

The homology model was generated for the portal protein through the online server PHYRE2 [23]. The modeling resulted with 78% coverage of the targeted protein with a 90% confidence. The program generated our homology model using the known crystal structure of HK97, which is a phage from *Corynebacterium diphtheria* (Figure 7A) [18]. The confidence represents the probability that the match between the Catdawg portal protein and the HK97 phage protein is a true homology. A confidence >90% indicates confidence that your protein adopts the overall fold and the core is modelled at high accuracy [23].

Comparison of known portal protein structures

The six known portal protein crystal structures are from phages ϕ29, SPP1, T4, P22, HK97, and G20c, which were compared to the portal protein from Catdawg using PyMOL to superimpose the structures for better visual comparison (Figure 7, Table 1) [21, 23, 29, 30, 31, 32].

GEPARD dot plot

A GEPARD dot plot [28] was also generated to compare the known portal protein sequences against each other as well as the portal proteins in Phages Corndog and Catdawg (Figure 8). The GEPARD dot plot is used to visualize the similarities and differences between sequences. A “dot” is placed on the plot where the sequences intersect with the same amino acid. The similarities on the GEPARD dot plot were found only between the portal proteins of Catdawg and Corndog portal proteins. The sequence identity was low between the known portal protein structures and that of Catdawg with a range of 10-15%. The GEPARD plot also demonstrated the low sequence homology between the known portal crystal structures.

Discussion

Homology model of the Catdawg portal protein

A homology model of the portal protein from Catdawg was generated using PHYRE2 and was modelled based on the known HK97 protein structure from phage *Corynebacterium*

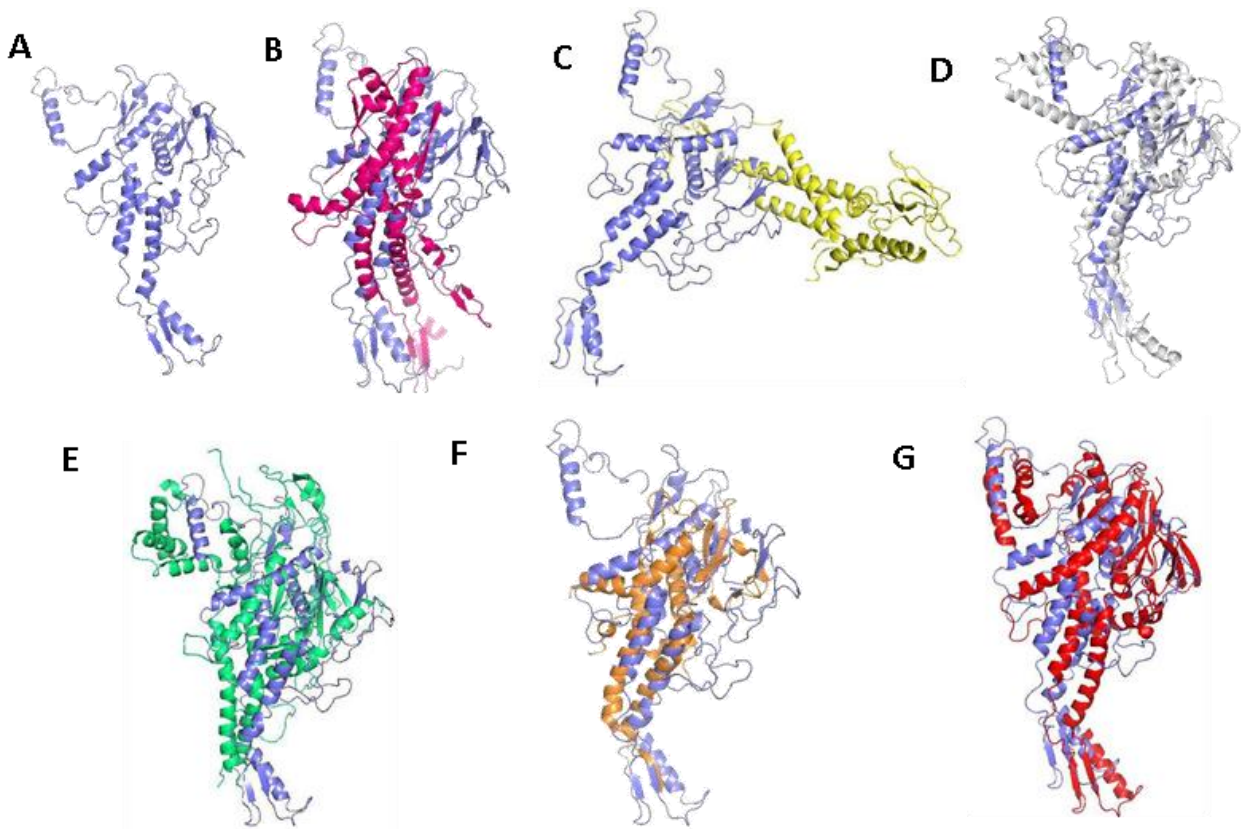

Figure 7. (A) Phyre2 Homology Model of the portal protein from phage Catdawg. The model generated by Phyre2 was aligned to *Corynebacterium diphtheriae* phage, HK97 [23]. **(B)** Alignment of the Catdawg Portal Protein (blue) and SPP1 Portal Protein (pink) in PyMOL. The RMSD value was 32.569. **(C)** Alignment of the Catdawg Portal Protein (blue) and ϕ29 Portal Protein (yellow) in PyMOL. The RMSD value was 18.243. **(D)** Alignment of the Catdawg Portal Protein (blue) and T4 Portal Protein (grey) in PyMOL. The RMSD value was 14.412. **(E)** Alignment of the Catdawg Portal Protein (blue) and P22 Portal Protein (green) in PyMOL. The RMSD value was 28.408. **(F)** Alignment of the Catdawg Portal Protein (blue) and HK97 Portal Protein (orange) in PyMOL. The RMSD value was 13.978. **(G)** Alignment of the Catdawg Portal Protein (blue) and G20c Portal Protein (red) in PyMOL. The RMSD value was 19 .687.

Table 1. Comparison of the known portal protein structures to the Catdawg homology model [32].

Known Portal Protein Structures	PDB ID	Compare to homology model of Portal Protein from Catdawg	RMSD (Å)	Percent Coverage	Overall Charge
Phage ϕ29	1H5W	Phage Catdawg	18.243	13%	-7.921
Phage SPP1	2JES_W	Phage Catdawg	32.569	40%	-45.625
Phage T4	3JA7_D	Phage Catdawg	14.412	37%	-14.539
Phage P22	3LJ5.K	Phage Catdawg	28.408	55%	-36.977
Phage HK97	3KDR_B	Phage Catdawg	13.978	30%	-14.38
Phage G20C	4ZJN.B	Phage Catdawg	19.687	55%	-10.175

diphtheriae [18]. The model covers 78% of our portal protein but excludes the beginning of the protein which correlates to the crown region of the protein [23] and is also the portion of the protein at the base of the capsid. Based on our homology model results for the Catdawg portal protein, if protein structural studies were attempted, we recommend removing the crown

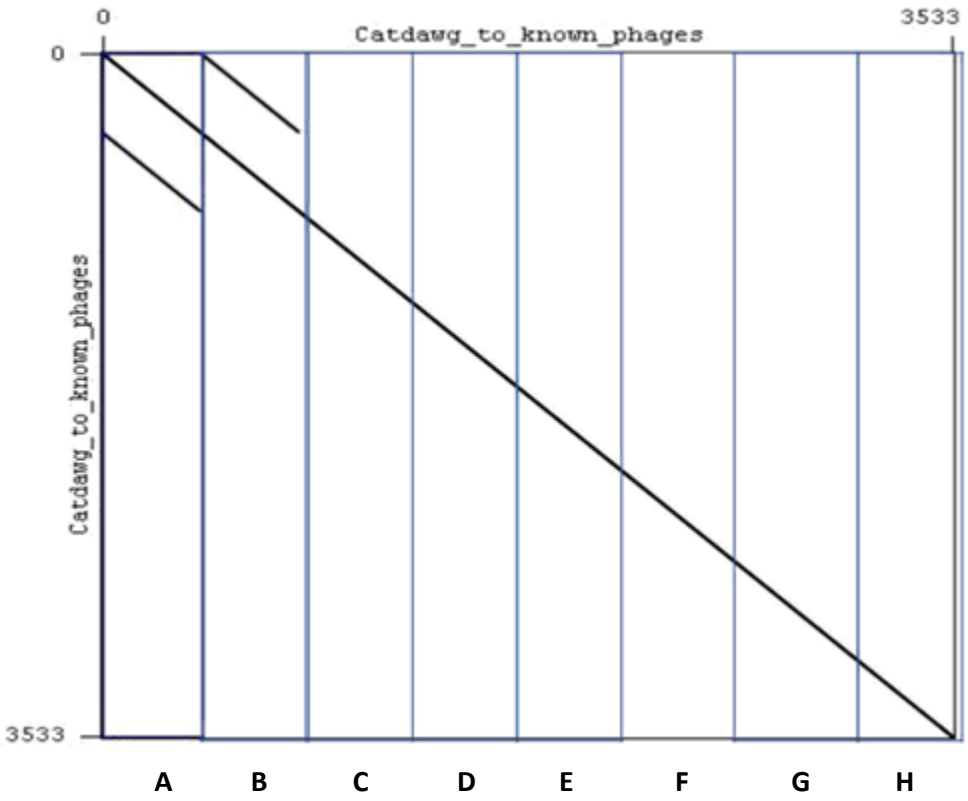

Figure 8. GEPARD comparison between Catdawg, Corndog, and the known structures of portal proteins. Catdawg (A), Corndog (B), G20c (C), ϕ29 (D), SPP1 (E), P22(F), T4 (G), and HK97 (H) protein sequences are compared to each other. Similarities were only observed bet ween Cluster O phages Catdawg (A) and Corndog (B).

region because it is unstructured. Additionally, the unstructured beginning region of the homology model could be attributed to a difference in capsid shape. The known portal protein structure (ϕ29) contains an icosahedral capsid and Catdawg's capsid contains a prolate capsid. The different shape of the capsid could change the structure of the portal protein.

The Catdawg portal protein was observed to have low identity with all of the known structures with a range from 10-15%. This low identity could mean that the 6 other phages are not similar to one another as they would be in a cluster like Catdawg and Corndog contained in cluster O. The portal protein sequence of phages Catdawg and Corndog was compared and found to be 100% identical; data not shown [32]. This data confirms the similarities between phages in the same cluster.

GEPARD dot plot Catdawg's portal protein

The GEPARD analysis was used to visualize the relationship between the 6 known portal proteins, the Catdawg portal protein, and the Corndog portal protein. The GEPARD dot plot compares each of the portal proteins to all of the other portal proteins as well as itself. The diagonal line present on the GEPARD dot plot was indicative of the comparison of each of the portal proteins to themselves. A "dot" is placed where there is similarity between the portal proteins. Besides the standard diagonal line, the only dots present were between Catdawg and Corndog, which was utilized as a positive control since they are already very similar and are confirmed to have 100% sequence identity (see above). The known portal protein structure sequences did not have any prominent dots indicating similarity between them as well as Catdawg.

PyMOL sequence alignments

The PyMOL data, with the alignments of sequence and structure for the known structures, along with our portal protein structures are presented. The RMSD value is a value that is used by crystallographers to conduct analysis on the protein. The alignments comparing the Catdawg portal protein to the other known structures were greater than 12 Å, which is not desirable for crystallography. The desired RMSD value for alignments is less than 2.00 Å for crystallography studies to indicate that it is a good structural alignment as well as sequence alignment. The control comparison when Catdawg portal protein aligned with itself exhibited an RMSD value of less than 2.00 Å [data not shown]. The appearance of the known structures resembled the Catdawg portal protein. The known structure alignment to the Catdawg portal protein was found largely in the stem and clip area of the common portal protein. All of the known phages that were compared to the Catdawg portal protein contained an icosahedral capsid; the Catdawg portal protein contains a prolate capsid. The change in capsid shape may explain why the crown and parts of the wing region were not fully aligned with the Catdawg portal protein. In a prolate head, the portal protein has a slimmer appearance than a portal protein of an icosahedral head, through physical appearance. The known structure that was mostly aligned to the Catdawg portal protein was the HK97 structure.

Research significance

Phages are abundant and may be useful as an alternative therapy for anti-biotic resistance bacteria; disease therapy and food safety [1]. Little is known about the structure of many phage proteins. We focused on the *Mycobacterium smegmatis* portal protein in phage Catdawg. This research was significant for the phage community since the portal protein is part of the DNA packaging process and not many crystal studies have been completed with *Mycobacterium* phages. The Catdawg portal protein is negatively charged which explains why the DNA does not bind when the DNA is fed through the octamer or 13-mer complete portal protein.

The understanding of the portal proteins shape and function could aid in understanding the portal proteins of other prolate capsid phages. We hope our study of phage portal proteins will bring awareness and lead to more structural studies in order to understand the differences we observed.

Acknowledgements

The authors would like to acknowledge Dr. Anne Coleman and Dr. Krystle Mclaughlin, for their support and guidance while researching and writing this paper.

References

1. Sulakvelidze A. 2011. Safety by Nature: Potential Bacteriophage Applications. Microbe, 6(3): 122-126.
2. Clokie M, Millard A, Letarov A, Heaphy S. 2015. Phages in nature. Bacteriophage, 1(1): 31-45.
3. Feiner R, Argov T, Rabinovich L, Sigal N, Borovok I, Herskovits A. 2015. A new perspective on lysogeny: prophages as active regulatory switches of bacteria. Nature Reviews Microbiology, 13(10): 641-650.
4. Russell P. 2010. Igenetics: DNA the genetic material. San Francisco: Pearson Education, Inc.
5. Cresawn SG, Pope WH, Jacobs-Sera D, Bowman CA, Russell DA, Dedrick RM, *et al*. 2015. Comparative genomics of cluster of mycobacteriophages. PloS one, 10(3): e0118725.
6. Fokine A and Rossmann M. 2014. Molecular architecture of tailed double-stranded DNA phages. Bacteriophage, 4.
7. Ghosh N, McKillop TJ, Jowitt TA, Howard M, Davies H, Holmes DF, *et al*. 2012. Collagen-like proteins in pathogenic *E. coli* strains. PloSone, 7(6): e37872.
8. McGrath S, Neve H, Seegers J, Eijlander R, Vegge C, Brøndsted L, *et al*. 2006. Anatomy of a lactococcal phage tail. J. of bacteriology, 188(11): 3972-3982.
9. Bo H, Margolin W, Molineux I, Liu J. 2013. The bacteriophage T7 virion undergoes extensive structural remodeling during infection. Science, 339(6119): 576-579.
10. Duda R, Martincic K, Hendrix RW. (1995). Genetic basis of bacteriophage HK97 prohead assembly. Journal of Molecular Biology, 247(4):636-647.
11. Camacho A, Gual A, Lurz R, Tarvares P, Alonso J. 2003. Bacillus subtilis bacteriophage spp1 DNA packaging motor requires terminase and portal proteins. J. of Biological Chemistry, 278: 23251-23259.
12. Nicholl D. 2008. An introduction to genetic engineering. New York: Cambridge University Press.
13. Pope WH, Bowman CA, Russell DA, Jacobs-Sera D, Asai DJ, Cresawn SG, Jacobs Jr WR, Hendrix RW, Lawrence JG, Hatfull G. 2015. Science Education Alliance Phage Hunters Advancing

Genomics and Evolutionary Science, Phage Hunters Integrating Research and Education, Mycobacterial Genetics Course: Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. elife 4.
14. Hatfull G, Jacobs-Sera D, Lawrence J, Pope W, Russell D, Ko C, *et al*. 2010. Comparative genomic analysis of 60 mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. J. of molecular biology, 397(1): 119-143.
15. Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, *et al*. 2003. Origins of highly mosaic mycobacteriophage genomes. Cell, 113: 171-182.
16. Phagesdb. 2019. Mycobacterium phage catdawg. Retrieved from: http://phagesdb.org/phages/Catdawg/
17. Payne K, Sun Q, Sacchettini J, Hatfull G. 2009. Mycobacteriophage lysin B is a novel mycolylarabinogalactan esterase. Mol. Microbiol, 73(3):367-381.
18. Sassi M, Bebeacua C, Drancourt M, Cambillau C. 2013. The first structure of a mycobacteriophage, the *mycobacterium abscessus* subsp. bolletii phage Araucaria. J. of virology, 87(14): 8099-8109.
19. Sun L, Zhang X, Gao S, Rao P, Padilla-Sanchez V, Chen Z, Sun S, *et al*. 2015. Cryo-em structure of the bacteriophage t4 portal protein assembly at near-atomic resolution. Nature Communications, 6. PDB ID: 3JA7_D.
20. Shi W, Ostrov D, Gerchman S, Graziano V, Kycia H, Studier B, Almo S, Burley S. 2009. New York Structural GenomiX Research Consortium: The crystal structure of a hk97 family phage portal protein from corynebacterium diphtheriae to 2.9A PDB ID: 3KDR_B.
21. Shi W, Ostrov D, Gerchman S, Graziano V, Kycia H, Studier B, Almo S, Burley S. 2015. New York Structural GenomiX Research Consortium: Crystal structure of the bacteriophage g20c portal protein. PDB ID: 4ZJN.B.
22. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. 2003. ExPASy: the proteomics server for in-depth protein knowledge and analysis Nucleic Acids Research, 31(13): 3784-3788.
23. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nature Protocols, 10(6): 845-858.
24. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. 1993. PROCHECK - a program to check the stereochemical quality of protein structures. J. App. Cryst., 26: 283-291.
25. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.
26. The AxPyMOL Molecular Graphics Plugin for PowerPoint, Version 2.0 Schrödinger, LLC.
27. Reva B, Finkelstein AV, Skolnick J. 1998. What is the probability of a chance prediction of a protein structure with a rmsd of 6 Å? Folding and Design, 3(2): 141-147.
28. Krumsiek J, Arnold R, Rattei T. 2007. GEPARD: A rapid and sensitive tool for creating dot plots on genome scale. Bioinformatics, 23(8): 1026-1028.
29. Lebedev A, Krause M, Isidro A, Vagin A, Orlova E, *et al*. 2007. Structural framework for DNA translocation via the viral portal protein. The EMBO Journal, 26(7): 1984-1994.
30. Simpson AA, Leiman PG, Tao Y, He Y, Badasso MO, Jardine PJ, *et al*. 2001. Structure determination of the head–tail connector of bacteriophage ϕ29. Acta Crystallographica Section D: Biological Crystallography, 57(9): 1260-1269.
31. Olia AS, Prevelige Jr PE, Johnson JE, Cingolani G. 2011. Three-dimensional structure of a viral genome-delivery portal vertex. Nature structural & molecular biology, 18(5): 597-603.
32. The Worldwide Protein Data Bank. Retrieved February 4, 2019 from https://www.wwpdb.org
33. BLASTp. 2019. Retrieved from: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&BLAST_SPEC=blast2seq&DATABASE=n/a&QUERY=&SUBJECTS=