Research Into Discovery of New Bacteriophages For Water Treatment In Qatar

Ahmed Warraich, Supervised by Dr. Valentin Ilyin, Computational Biology, Carnegie Mellon University Qatar

Abstract:

Phages are viruses that have evolved to target one or several similarly related bacterial strains. As bacteria develop drug resistance, humans have turned to phages as an alternative form of antibacterial treatment. We need to identify phages that are effective at destroying hazardous bacteria. In this research, an unknown phage genome has been isolated and analyzed to learn about its characteristics and potentially identify the bacteria it targets and lyses. The process begins from genomic reads files isolated in the lab and assembled using gene sequencing software. Afterward, careful analysis of the genome and the use of public genetic databases helped identify potential proteins within the genome. So far, 50,000 of 85,000 base pairs have been analyzed and their proteins have been identified 4 of which with their specific functions. The 4 proteins were: T4 Phage Tail Protein (used for attaching and penetrating the cell membrane), Thymidylate Synthase (used for nucleotide production), T7 Leading Strand DNA Polymerase (part of a larger protein unit designed to copy DNA effectively), and the Portal Protein (acts as a foundation in which to assemble the phage's components into one cohesive unit). This phage shares many proteins with E. Coli phages, Salmonella Phages, as well as the hosts' cells E. Coli, and Salmonella, leading to the conclusion that this phage must either be an E. Coli phage or a Salmonella Phage.

Method:

Turning reads into a viable genome: Prior to this project, a separate group took samples from human waste and isolated several phage genomes. These genomes were copied and cut up in varying pieces, that a machine read and transcribed into a file of those reads. The beginning of this project was taking these reads and putting them into genome assembly programs. Two were used in the project, Mira and spADEs. The Mira program gave inconsistent results (two segments of genome versus one from spADEs) the spADEs result was used for further investigation. It is important to note that comparing both genomes resulted in a match of 99.99%, allowing for the dismissal of one genome as they are practically identical.

```
>NODE_1_length_88237_cov_16.768863
AAAAAATCTCTGCTAATGAGTTTGGTAAGATTGACTACAGTAAGATTCCTTCTCTTGCTG
CTGCACGCTACCAAAAACTCTTTAACCGCAAAGATGGAGAACGCTACAAAGCTTACATAG
AGTCGTTATCAAAAGGTGAGGCCAAGATTAACGCTGGTGCTGTTTACCCATACGATGTGA
TTAAGTCTATCAAACATGGTAATGCAGATGTTGCCAATGAGCAGTGGAAAGCACTACCAA
ACTGGATGGCAGAAGGTGAGAACATCTTGTGTATGACTGATGTTTCAAGCTCAATGTCTT
GGGTGAATCTTGGTTCAATCACTGCCCTTGATATTGGTGTATCACTTGCTTTGTATGTAG
CAGAACGCAATACTGGTTGCTTTAAGAATGAGTTAATGGTTTACTCAACAAACCCTCACT
TCATCGAACTGAGTGGTGATTTACGAAACCGTCATCGTCAGGTGATGCGTCATGTTGAGT
ATGGTTCAACTAACTTACAAGCAGCTTTTGACCGTATTCTTGAGACAGGTAAGAGAAACA
```

Figure 1: A portion of the spADEs produced genome. It has 88,000 base pairs, and is one segment (or contig as it's commonly called)

Annotating genome: The contig was taken and put it into DNA master, a program used to annotate genomes. The initial attempt was to have the program automatically annotate the genome. Auto annotate did not identify potential genes so the work shifted to manual annotation of 50,000 base pairs. Manual process involves looking for potential gene starts and looking up their protein products (that is, what proteins their codons coded for). The product was taken to the NCBI database and a protein BLAST was completed to identify the similarities between the BLAST results and the gene before determining if it was a valid protein identity or not. This was repeated along the entire 50,000 base pair segment. The goal was to find as many proteins as possible, where there were sections that the BLAST could not identify were labeled as suspicious or a potential protein for future testing. This is because bacteriophages have a very dense genome in order to maximize their productivity.

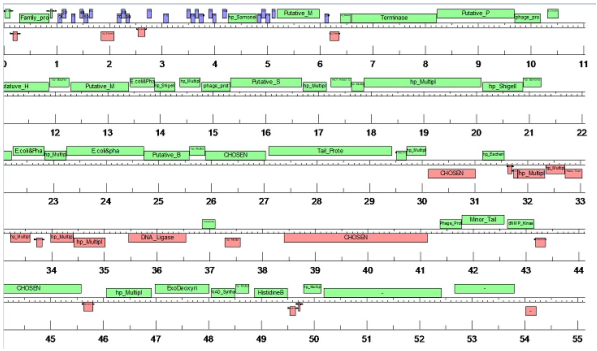

Figure 2: Annotated genome up to 50,000 base pairs. Green boxes indicate a gene on the "top" strand of DNA, red boxes indicate a gene on the "bottom" strand of DNA.

Deep dive into four identified proteins of the unknown phage to known proteins:

T4 Phage Tail Protein: This unknown protein is similar to the T4 Phage Tail protein, specifically one of the tail fiber proteins (either long or short), and depending on its identity, the fibre can carry out different functions. The long tail protein is made for holding onto the cell membrane, while the short tail protein is for plunging the needle into the cell to insert genetic material. (Bartual, 2010)

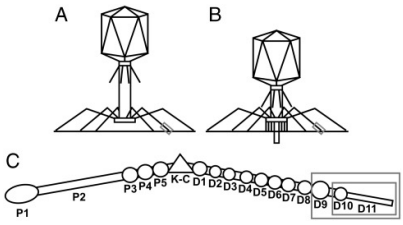

Figure 3: Diagram of Long and Short Tail Fibers of T4 Bacteriophage

Thymidylate Synthase: The function of thymidylate synthase in a cell is to convert dUMP into dTMP, a precursor to DNA nucleotides. This benefits the phage as it can make its host create more nucleotides for it to copy the phage's genetic material which leads to increased efficiency in producing copies of the phage. This gene originally came from E. Coli, however it makes sense that it's in the phage's genome as periodically phages going through the lysogenic cycle will incorporate their host's genes into theirs. (JJ Medicine, 2018)

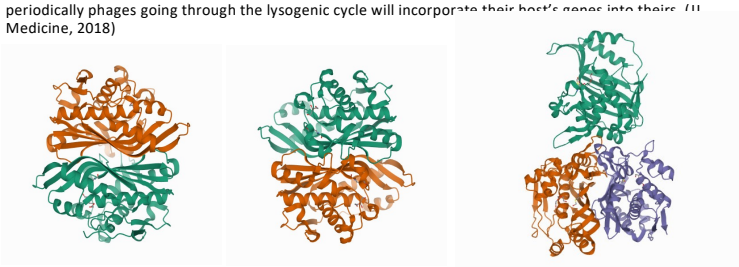

Figure 4: 3 X-Ray crystallography images of different conformations of the Thymidylate Synthase protein. These images specially come from (Mayclin, Et. Al., 2017).

T7 Leading Strand DNA Polymerase: This protein is part of a larger DNA polymerase-helicase complex. This protein is commonly found in pages, and its purpose is to help increase rates of DNA replication. Usually when DNA is to be replicated, it needs a helicase molecule to unzip it before finding DNA polymerase to help create the new strands. The issue this presents is that its slow since it's up to chance that the genetic material runs into these proteins in that order. Eukaryotic cells bypass this by sectioning these proteins and genetic materials into a separate section in the cell, the nucleus, so that they are in closer proximity to each other this speeding up replication. Phages however, overcome this by creating a larger unit where both DNA polymerase and helicase are together so that the genetic material can be unzipped and copied at the same time. This greatly speeds up the time it takes for phages to replicate/reproduce. (Gao, Et. Al.)

Figure 5: X-Ray crystallography image of T7 Leading Strand DNA Polymerase.

Portal Protein: The protein itself is very crucial in the construction of the the phage as it serves as a foundation for the head and the tail portion to be built upon. The portal protein also resembles a camera aperture and slides open when the needle of the phage has pierced into the body of the host in order to let the genetic material of the phage insert the bacteria and begin replication. (Williams, Et. Al., 2015)

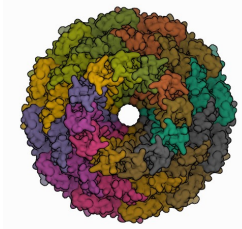

Figure 6: X-Ray crystallography image of Portal Protein

Conclusions and Future Goals: The unknown phage under study is either an E. Coli. Phage or a Salmonella Phage due to fact that it contains a large number of genes that are similar to other E. Coli and Salmonella phages. It is more likely that this unknown phage is an E. Coli phage since it has more matching genes present in the genome analyzed so far. Additional research is needed, such as filling out the rest of the genome and completing the BLAST for all 88,000 base pairs. Eventually the author wishes test out potential hypotheses for the function of certain identified genes and submit findings to NCBI so that the phage can be added to the collective library/database stored on those servers.

Acknowledgement: I'd like to acknowledge and thank Prof. Valentin Ilyin for guiding me through this project, I hope to continue this research and work on many projects in the future. I would also like to acknowledge the work that Prof. Annette Vincent and the students from the previous class did, without them these samples would not be available for my use.

References:

Bartual, S. G., Otero, J. M., Garcia-Doval, C., Llamas-Saiz, A. L., Kahn, R., Fox, G. C., & van Raaij, M. J. (2010). Structure of the bacteriophage T4 long tail fiber receptor-binding tip. *Proceedings of the National Academy of Sciences*, *107*(47), 20287–20292. https://doi.org/10.1073/pnas.1011218107

Crystal structure of the bacteriophage G20C portal protein. (n.d.).

Gao, Y., Cui, Y., Fox, T., Lin, S., Wang, H., de Val, N., Zhou, Z. H., & Yang, W. (2019). Structures and operating principles of the replisome. *Science*, *363*(6429), eaav7003. https://doi.org/10.1126/science.aav7003

JJ Medicine. (2018, July 6). *One carbon metabolism | tetrahydrofolate and the folate cycle*. https://www.youtube.com/watch?v=JmWrtzoe9pU

Mayclin, S. J., Delker , S. L., Lorimer, D. D., & Edwards, T. E. (n.d.). *Crystal structure of thymidylate synthase from Elizabethkingia anophelis NUHP1*. https://doi.org/http://dx.doi.org/10.2210/pdb6auj/pdb

Prevelige, P. E., & Cortines, J. R. (2018). Phage assembly and the special role of the portal protein. *Current Opinion in Virology*, *31*, 66–73. https://doi.org/10.1016/j.coviro.2018.09.004

Williams, L. S., Turkenburg, J. P., Levdikov, V. M., Minakhin, L., Servinov, K., & Antson, A. A. (2015). Crystal Structure of the Bacteriophage G20C Portal Protein. *RCSB Protein Data Bank*. https://doi.org/10.2210/pdb4zjn/pdb