Complete Genomes of Two Novel Active Prophages Discovered by Bioinformatics Methods from High-Throughput Sequencing Data

Xiangcheng Xie¹, Qiang Sun³, Xiangke Liao¹, Yigang Tong³ and Shaoliang Peng²*

¹College of Computer, National University of Defense Technology, Changsha 410073, China, 109 De-Ya Road, Changsha 410073, China.
²College of Computer Science and Electronic Engineering & National Supercomputer Centre in Changsha, Hunan University, 252 Lu-Shan South Street, Changsha 410073, China.
³State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, 20 Dong-Da Street, Beijing 100071, China.
Email: pengshaoliang@nudt.edu.cn

Abstract. Bacteriophages constitute key gene transfer agents in many bacteria. Specifically, they may confer gene mobility to Staphylococcus warneri that usually dwells in the skin of human and animals. In this study, we first extracted samples of bacteria from foot skin of foot ulcers sufferers and develop these bacteria. Then, our bioinformatic methods for prophage predicting were applied to the genome sequences of Staphylococcus warneri which are assembly from high-throughput sequencing data. Two novel sequences predicted to encode complete prophages were found in the bacterial genome. Using the inducing agent mitomycin C, we produced phage suspensions. These indeed encompassed the progeny of the two identified prophages (denoted IME1367_01 and IME1367_02). From high-throughput sequencing data of the suspensions, we obtained the two complete sequence of phage IME1367_01 and IME1367_02. By electron microscopic, we observed these two different phages. The complete genomes of IME1367_01 and IME1367_02 are submitted to NCBI: “https://www.ncbi.nlm.nih.gov/nuccore/KY653124.1”&“https://www.ncbi.nlm.nih.gov/nuccore/KY653121.1”. It will enrich our understanding of the lysogenic phages of Staphylococcus warneri.

1. Introduction
Viruses that infect bacteria - bacteriophages (phages) - play significant roles in the evolution of bacteria, at both the individual and community levels. Bacteriophages can typically be divided into two groups, lytic and temperate. Lytic phage infect propagate within and then lyse their host bacterial cells as part of their life cycle, while temperate phages may exist benignly within the DNA of their bacterial host. Temperate phages can physically integrate into one of the native replicons (plasmid or chromosome) of their preferred bacterial host. As agents of horizontal gene transfer (HGT), temperate phages can enhance the fitness of their host cells in the form of lysogenic conversion and/moron genes, for instance by providing so-called auxiliary metabolic genes (AMGs) as well as virulence or pathogenicity traits [1]. Moreover, phages function in the biological ‘warfare’ among neighboring bacterial cells and can modulate the formation of bacterial biofilms at the population level [2].
Temperate phages integrate their nucleic acid into host bacterial genomes in a process termed lysogeny genetic material of a bacteriophage in host chromosome, which is referred to as prophage,
can be passed to daughter cells during cell division. Prophages may be induced under certain circumstances (such as UV [3-4], hydrogen peroxide or mitomycin C treatment [5]), which activates a lytic cycle [6]. Thus, prophages are viewed as "dangerous molecular time bombs". DNA could be transferred among different hosts during phage infection through either the lytic cycle or the lysogenic cycle. Previous studies showed that more than 50% of marine isolates harbor prophages, which play a significant role in shaping the phenotypic traits of their hosts.

Lysogenic phage plays an important role in bacteria virulence and multi-drug resistance acquisition. When lysogenic phage inserts into bacterial genome, it has an important impact on host bacterial pathogenicity and/or multi-drug resistance. Taking EHEC O104:H4 as an example, the main toxic genes were encoded by prophage [7]. Prophage is the product of lysogenic phage integrated into the bacterial chromosome [8]. Lysogenic phages often participate in the host adaptation to the environment and contribute to the host metabolism [9]. Recently lytic phages are extensively studied for the potential use as antibacterial agents against multiple drug resistant bacteria. For the safe use of lytic phages, careful examination to exclude lysogenic phages is critical since lysogenic phages are usually encoded by the host and contaminated the lytic phage preparations without notice. Better understanding of lysogenic phages will help to understand the pathogenicity of bacteria and its particular metabolic pathways, as well as produce safe lytic phage products.

2. Materials and Methods

2.1. Active Prophage Predicting

After assembling, prophages are usually within the contigs, which could be predicted by prediction tools. But sometimes, some prophages occupy the whole contig and shows characteristic that these contigs themselves are cyclized. These prophages are active prophages. In order to better understand this phenomenon, we start with the mechanism of lysogenic phages’ integration and removal.

![Figure 1. Integration and excision of lysogenic phages.](image-url)

Figure 1. Integration and excision of lysogenic phages.

After invading hosts, lysogenic phages’ attP site reassortment with bacterial attB site, which forms attL and attR on the ends of prophages. At this point, phages enter into the lysogenic cycle. Under induction, phages conduct reverse reaction of integration with the action of phages, which excise prophage’ DNA from bacterial genome encoding factors and hosts’ factors. At this moment, phages enter into lytic cycle and start with producing offspring.
Prophages can be considered as clusters of phage-like genes within a bacterial genome. In the history of evolution, most of the integrated prophages lost their ability to be induced due to deletions or critical mutations of some important phage genes, and resulted in a lot of defective prophage remnants in the bacterial genomes. Recently integrated prophages remain functional and can be induced UV light or some chemicals, but the incidence of bacteria carrying functional prophages is not high (according to our experience, it is about 10~20%). In order to get as much prophage regions as possible, we used both clusters of phage-like genes (hypothetic phage genes) and functional phage protein genes (like genes encoding capsid, terminase, tail fiber, lysin, holin) to identify possible ranges within a bacterial genome contigs. To identify clusters of phage-like genes, we designed program which used the general DBSCAN algorithm which performs quite well. DBSCAN takes two parameters: the cluster size n and a distance e. The parameter n defines the minimal number of phage-like genes required to form a prophage cluster and e is the maximal spatial distance between two neighbor genes within the same cluster. In our case, the spatial distance between two genes is just the number of ORFs between them. Empirically, we set n to be 6, since prophages generally have more than five proteins. The value of e was set to 5. To identify functional protein genes, we compare bacterial genome to database of phage’s protein and find out all functional protein genes. The regions found by clusters and functional protein may overlap. After merging the overlapped regions, we get possible prophage regions of a bacterial genome. Those contigs that contains clusters are possible prophages.

The appearance of contig of phages that self loop and contains clusters of phage-like genes indicates that part of such phages are enter into lytic cycle. In the high-throughput sequencing data, the reads that across both ends of a contig come from the lysogenic phages under proliferating status. At the same time, the contig of this phage also connects with the bacterial contigs. The reads that links phage’s contig and bacterial contig come from phages in the states of integration. In order to reveal the connecting relationship of contigs intuitively, we use ContigScape [10] which is a visualization plugin developed on the strength of Cytoscape [11]. ContigScape can read the file, 454Contigs.ace, which is created by Newbler while assembling raw data. This file is parsed and graphical displayed in CytoScape.

In our case, we import 454Contigs.ace into ContigScape and find two contigs that are in line with the characteristics of lysogenic phages. As shown in figure.2, the blue dumbbell-like graphs represent contigs. The numbers in the circle at the both ends represent the serial numbers of contigs. "S" and
“E” respectively represent the start and end of the contigs. The green line represent reads that connect contigs. The figure on the line show the number of reads. The contig10 is the active prophage we predicted. As we can see, there are 18 reads across the start to the end and make this contig self-loop. The contig21 and contig1 in its each end are the contigs of bacterial genome, and there are respectively 45 and 50 reads connected with contig10. According to the connecting relationship displayed in figure.2, we can grep reads across the each end of contig10, make it self-loop and get the phage’s whole sequence of unbound state. We can also grep reads to fill up the gaps between contig10 and contig21, contig1. So we get prophage integrating in bacterrial genome and its surrounding sequences. We called this prophage IME1367_01.

![Figure 2. Connecting relationship of contigs in ContigScape](image)

In the same way, we get another active prophage in figure.3. The contig16 is the active prophage we predicted. As we can see, there are 305 reads across the start to the end and make this contig self-loop. The contig8 and contig12 in its each end are the contigs of bacterial genome, and there are respectively 50 and 46 reads connected with contig16. According to the connecting relationship displayed in figure.3, we can grep reads across the each end of contig16, make it self-loop and get the phage’s whole sequence of unbound state. We can also grep reads to fill up the gaps between contig16 and contig8, contig12. So we get prophage integrating in bacterrial genome and its surrounding sequences. We called this prophage IME1367_02.

Analyzing this two prophages’ information of sequencing depth can also help us understand active prophage. The file, 454ContigGraph.txt generate by Newbler, record average sequencing depth of all contigs. We obtain the average sequencing depth of six contigs in figure.1 and figure.2. The results is in table.1 and table.2.

| Table 1. Average sequencing depth of three contigs of IME1367_01. |
|---------------------------------------------------------------|
| The number of conig | Contig21 | Contig10 | Contig1 |
| Average sequencing depth | 43.2 | 62.1 | 42.1 |

| Table 2. Average sequencing depth of three contigs of IME1367_02. |
|---------------------------------------------------------------|
| The number of conig | Contig12 | Contig16 | Contig8 |
| Average sequencing depth | 39.2 | 69.7 | 38.5 |
Both in table.1 and table.2, the two contigs of bacterial genome shows similar average sequencing depth. But the active prophages shows higher average sequencing depth than bacterial genome obviously. In the premise that the influence of GC bias on the sequencing depth could be ignored, we can estimate the situation of active prophages. As for IME1367_01, on average the number of lysogenic phages induced on one bacteria is 0.46. As for IME1367_02, on average the number of lysogenic phages induced on one bacteria is 0.19.

2.2. Active Prophage Induction

In order to verify that our prophages are actually active, we induced the stain for free lysogenic phages. Prophages may be induced under certain circumstances (such as UV, The radiation or mitomycin C treatment), which activates a lytic cycle. Given the finding of the IME1367_01 and IME1367_02 encoding sequence in the Staphylococcus warneri’s genome, cultures of this organism were screened for the presence of virions, using induction with different levels of MMC. Following addition of MMC, prophages had been induced to excise from the host genome, resulting in production of enhanced levels of phage progeny. We extract supernate and sequencing it, because there is no bacteria in supernate. Then we assemble raw data by Newbler 3.0. In our result of assembling, there are two contigs, which are the same with the two active phages’ contigs above. That means our prediction about active prophages are right.

2.3. Active Prophage Induction

We annotated the two complete induced sequences and shows results in figure.4 and figure.5. As we can see in figure.4, prophage IME1367_01 has 88 CDS, and there are some functional protein, such as integrase, tail, lysis. As we can see in figure.5, prophage IME1367_02 has 34 CDS, and there are some functional protein, such as integrase and terminase.

![Figure 4. Result of 1367_01’s annotation](image-url)
2.4. Phages’ Morphology under Electron Microscopic

Crude induced lysate was filtered with 0.22-µm-pore-size filter and centrifuged to pellet the cell derivates, then stored in -20 °C for one night prior imaging. As we can see in figure 6, Image(a) shows a typical Siphoviridae family Image(b) shows a Myoviridae family, the image(b) also shows induced phages has lost its capsid structure.

![Image(a)](image1.png) ![Image(b)](image2.png)

**Figure 6.** Transmission electron microscopy photograph of lysogenic phages induced from Staphylococcus warneri strain.
3. Conclusion
This paper conducts a method based on bioinformatics to predict the active prophages from bacterial high throughput sequencing data. This method is intuitive. We find two active prophages in Staphylococcus warneri’s genome. Moreover, we successfully induced the two prophages from the host and annotated the two sequences to find some phage’s functional protein, which shows obvious characteristics of lysogenic phages. There was no report about the finding of lysogenic phages on Staphylococcus warneri. Our methods may find some enzymes and pave the way for the future antibiotic.

4. Acknowledgments
This work was supported by National Key R&D Program of China 2017YFB0202602, 2017YFC1311003, 2016YFC1302500, 2016YFB0200400, 2017YFB0202104; NSFC Grants 61772543, 61625202, 61272056; The Funds of State Key Laboratory of Chemo/Biosensing and Chemometrics; the Fundamental Research Funds for the Central Universities; and Guangdong Provincial Department of Science and Technology under grant No. 2016B090918122.

5. References
[1] Brüssow H, Canchaya C, Hardt W D 2004 Microbiol Mol Biol R 68 560-602
[2] Secor P R, Sweere J M, Michaels L A, Malkovskiy A V 2015 Cell Host & Microbe 18 549-559.
[3] Lamont I, Brumby A M and Egan J B 1989 Proc Natl Acad Sci U S A 86 5492-96
[4] Sylwia Bloch, Bożena NejmanFaleńczyk, Gracja Topka, Aleksandra Dydecka, Katarzyna Licznerska and Magdalena Narajczyk 2015 Toxins 7 3727-39
[5] Fortier L and Moineau S. 2007 Appl & Environ Microb 73 7358-66.
[6] Chen F, Wang K, Stewart J, Belas R 2006 Appl Environ Microbiol 72 4995-5001.
[7] Hamelin K, Bruant G, Elshaarawi A, Hill S, Edge T A and Fairbrother J 1986 Anal Chim Acta 186(2) 349-50
[8] L Barksdale A and Arden S B 1974 Annu Rev Microbiol 28 265-99
[9] Stevens R H, Zuanazzi D, Mb D A M 2013 Microb Pathog 54(1) 54-59
[10] Tang B, Qi W, Yang M 2013 Bmc Genomics 14 289
[11] Shannon P, Markiel A, Ozier O 2003 Genome Research 13(11) 2498-504