

***In silico* GENOMIC STUDY OF PHAGE  
SPECIFIC TO *Pseudomonas aeruginosa* USM AR2**

**NURUL ASHIQIN BINTI MOHAMAD ZAWAWI**

**UNIVERSITI SAINS MALAYSIA**

**2018**

***In silico* GENOMIC STUDY OF PHAGE SPECIFIC  
TO *Pseudomonas aeruginosa* USM AR2**

by

**NURUL ASHIQIN BINTI MOHAMAD ZAWAWI**

**Thesis submitted in fulfillment of the requirements  
for the degree of  
Master of Science**

**August 2018**

## **ACKNOWLEDGEMENT**

In the name of God, the Most Gracious, the Most Merciful

First and foremost, I would like to express my gratitude to Allah S.W.T, the Most Gracious and the Most Merciful for His blessings and for giving me strength, thoughts, good health that allow me to complete the study. In general, this work cannot be successfully completed without continuous assistance and cooperation from many individuals. Specifically, I would like to thank my main supervisor, Associate Professor Dr. Yahya Mat Arip for his valuable comments, advice, supports, patience and guidance throughout my journey to finish the study. Special thanks to my beloved parents, siblings, members of Rhaudatul Mahabbah, Kak Liyana and Kak Wan for their unrelenting support and prayers to ensure I will keep on persisting and pursuing my dreams. In addition, I would love to extend my heartfelt appreciation to my friends and other related persons who have, in one way or another, offered me invaluable assistance and advice. To who has contributed to my study but not listed here, I owe you my heartfelt appreciation. Thank you.

## TABLE OF CONTENTS

<b>ACKNOWLEDGEMENT</b>	<b>ii</b>
<b>TABLE OF CONTENTS</b>	<b>iii</b>
<b>LIST OF TABLES</b>	<b>vii</b>
<b>LIST OF FIGURES</b>	<b>ix</b>
<b>LIST OF ABBREVIATIONS</b>	<b>xi</b>
<b>ABSTRAK</b>	<b>xiv</b>
<b>ABSTRACT</b>	<b>xvi</b>
<b>CHAPTER 1 - INTRODUCTION</b>	<b>1</b>
<b>CHAPTER 2 - LITERATURE REVIEW</b>	<b>3</b>
2.1 Discovery of phage	3
2.2 The nature of phage	3
2.3 Phages are everywhere	6
2.4 Classification of phage	7
2.5 Phage genomic	11
2.4.1 Sequences diversity	11
2.4.2 Mosaic genome architecture	12
2.4.3 Overlapping gene	13
2.6 Host	15
2.6.1 <i>Pseudomonas aeruginosa</i>	15
2.7 Rhamnolipid	16
2.8 <i>Pseudomonas</i> phage	25
2.8.1 <i>Pseudomonas</i> phage SM1	26
2.9 Bioinformatics	27
2.9.1 Genome sequencing and assembly	27

2.9.2	Genome annotation	27
<b>CHAPTER 3 - METHODOLOGY</b>		<b>29</b>
3.1	Outline of study	29
3.2	Materials	30
3.3	Media and buffers	31
3.3.1	Bacteria media	31
3.3.2	Bacteria media containing ampicillin	31
3.3.3	Preparation of 1% agarose gel	32
3.3.4	Buffers and working solutions	32
3.4	Sodium dodecyl sulfate polyacrylamide gel electrophoresis	33
3.4.1	Sample preparation for SDS-PAGE	34
3.4.2	Preparation of SDS-PAGE	34
3.4.3	SDS-PAGE	35
3.5	Bacterial strain	36
3.5.1	Bacterial host of <i>Pseudomonas</i> phage SM1	36
3.5.2	Bacterial glycerol stock	37
3.6	Plasmid cloning vector	37
3.7	Bacteriophage enrichment	39
3.8	Plaque assay	39
3.9	DNA extraction of <i>Pseudomonas</i> phage SM1	40
3.10	Genome sequencing and annotation of <i>Pseudomonas</i> phage SM1	41
3.11	Cloning of pET32-Capsid containing scaffold gene	42
3.12	Treatment of T4 DNA polymerase and annealing of pET32- Capsid containing scaffold gene into Ek/LIC expression vector	44

3.13	Transformation of pET32-Capsid containing scaffold into expression host	45
3.14	PCR colony	46
3.15	Plasmid isolation	47
3.16	Plasmid sequencing	48
3.17	Expression of recombinant capsid containing scaffold protein	48
3.18	Sequencing of the recombinant capsid containing protein	49
3.19	Transmission electron microscope of capsid containing scaffold protein	50
<b>CHAPTER 4 - RESULTS</b>		<b>51</b>
4.1	Bioinformatics analyses	51
4.1.1	General overview of <i>Pseudomonas</i> phage SM1 genome	51
4.1.2	Open reading frames (ORF) prediction and genome annotation of <i>Pseudomonas</i> phage SM1	54
4.1.3	Analysis of predicted proteins of <i>Pseudomonas</i> phage SM1	63
4.1.4	Genomic organization of <i>Pseudomonas</i> phage SM1	69
4.2	Cloning capsid containing scaffold gene into expression vector pET32 Ek/LIC	71
4.3	Plasmid sequencing	74
4.4	Expression of capsid containing scaffold protein	79
4.5	Transmission electron microscope of capsid containing scaffold protein	82
4.6	Peptide sequencing	84
<b>CHAPTER 5 - DISSCUSSIONS</b>		<b>86</b>

<b>CHAPTER 6 - CONCLUSIONS AND RECOMMENDATIONS</b>	<b>102</b>
<b>REFERENCES</b>	<b>103</b>
<b>APPENDICES</b>	
<b>LIST OF PRESENTATIONS</b>	

## LIST OF TABLES

		<b>Page</b>
Table 2.1	Bradley's phage classification scheme	9
Table 2.2	Main type of biosurfactants	20
Table 2.3	Application of biosurfactants in bioremediation of contaminated sites	21
Table 2.4	Patents on biosurfactants produced by microorganisms	22
Table 2.5	Applications of rhamnolipid	23
Table 3.1	Materials and suppliers	30
Table 3.2	Recipes for bacteria media	31
Table 3.3	Compositions of buffers and working solutions	32
Table 3.4	Buffers and solutions for Sodium Dodecyl Sulfate (SDS-PAGE)	33
Table 3.5	Compositions of SDS-PAGE	35
Table 3.6	Bacterium strain	36
Table 3.7	pET vector features	37
Table 3.8	Specific primers	43
Table 3.9	PCR recipes	43
Table 3.10	PCR cycling parameter	43
Table 3.11	T4 DNA Polymerase treatment on insert	45
Table 3.12	Components of annealing reaction	45
Table 3.13	PCR recipes	46
Table 3.14	PCR cycling parameter	46
Table 4.1	Whole genome sequence alignment of <i>Pseudomonas</i> phage SM1 against NCBI nucleotide databases	53
Table 4.2	Whole genome annotation of <i>Pseudomonas</i> phage SM1	55





## LIST OF FIGURES

		<b>Page</b>
Figure 2.1	Phage replication cycle	5
Figure 2.2	Structure of surfactant	19
Figure 3.1	Flowchart summarizing the work on <i>Pseudomonas</i> phage SM1	29
Figure 3.2	Circular map and sequence reference points of the pET32 Ek/LIC	38
Figure 3.3	Flowchart summarized procedure of whole genome sequencing and annotation of <i>Pseudomonas</i> phage SM1	41
Figure 3.4	Summary of the cloning capsid containing scaffold gene <i>Pseudomonas</i> phage SM1 into the Ek/LIC vector	42
Figure 4.1	Genomic organization of <i>Pseudomonas</i> phage SM1	70
Figure 4.2	Agarose gel electrophoresis of PCR capsid containing scaffold gene	72
Figure 4.3	Agarose gel electrophoresis of colony PCR on colonies 1-4 carrying pET32-Capsid containing scaffold plasmid	73
Figure 4.4 (A)	Sequence alignment of colony 1 with capsid containing scaffold gene of <i>Pseudomonas</i> phage SM1	75
Figure 4.4 (B)	Sequence alignment of colony 2 with capsid containing scaffold gene of <i>Pseudomonas</i> phage SM1	76
Figure 4.4 (C)	Sequence alignment of colony 3 with capsid containing scaffold gene of <i>Pseudomonas</i> phage SM1	77
Figure 4.4 (D)	Sequence alignment of colony 4 with capsid containing scaffold gene of <i>Pseudomonas</i> phage SM1	78

Figure 4.5	SDS-PAGE of recombinant capsid containing scaffold protein expression induced by IPTG	80
Figure 4.6	SDS-PAGE of purified expressed capsid containing scaffold protein	81
Figure 4.7	Electron microscope image of capsid containing scaffold protein of <i>Pseudomonas</i> phage SM1	83
Figure 4.8	Identity of the recombinant capsid containing scaffold protein was determined by MALDI-TOF	85
Figure 5.1	Genome organization of HIV-1 virus by reading frames	92
Figure 5.2	Translation from the bicistronic mRNA	93
Figure 5.3	Visualization of ORF 1 <i>Pseudomonas</i> phage SM1	94
Figure 5.4	The head assembly pathway in morphogenesis of the bacteriophage T4 virion	96
Figure 5.5	Prediction cleavage sites of serine protease in the capsid containing scaffold protein sequences	97

## LIST OF ABBREVIATIONS

APS	Ammonium persulfate
aa	Amino acid
BLAST	Basic Alignment Search Tool
C <sub>2</sub> H <sub>9</sub> NaO	Sodium acetate trihydrate
bp	Base pair
dATP	Deoxyadenosine triphosphate
dH <sub>2</sub> O	Distilled water
ddH <sub>2</sub> O	Double distilled water
DNA	Deoxyribonucleic acid
DNAse	Deoxyribonucleic acid enzyme
dNTP	Deoxynucleotide triphosphate
dsDNA	Double stranded deoxyribonucleic acid
DTT	Dithiothreitol
<i>E.coli</i>	<i>Escherichia coli</i>
EB	Elution buffer
EDTA	Ethylenediaminetetraacetic acid
g	Gram
HCl	Hydrochloric acid
ICTV	International Committee On Taxonomy Viruses
IPTG	Isopropyl-beta-D-thiogalactopyranoside
KH <sub>2</sub> PO <sub>4</sub>	Potassium dihydrogenphosphate,
kbp	Kilobase pair
KCl	Potassium chloride

kDa	Kilodalton
LB	Luria Bertani
LIC	Ligation independent cloning
M	Molar
Mg	Milligram
mg/mL	Milligram/mililitre
mL	Mililitre
mM	Milimolar
NaCl	Sodium chloride
Na <sub>2</sub> HPO <sub>4</sub>	Sodium phosphate
NaOH	Sodium hydroxide
NCBI	National Center For Biotechnological Information
Nm	Nanometer
NR	DNA clean up buffer
NW	Column wash buffer N
OD	Optical density
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PFU	Plaque forming unit
pH	Potential of hydrogen
RNA	Ribonucleic acid
RNAse	Ribonucleic acid enzyme
rpm	Revolutions per minute
SDS	Sodium dedocyl sulphate
SDS-PAGE	Sodium dedocyl sulfat polyacrylamide gel electrophoresis

ssDNA	Single stranded deoxyribonucleic acid
SOC	Super Optimal broth with Catabolite repression
TBE	Tris- borate EDTA
TEM	Transmission electron microscope
TEMED	Tetramethylethylenediamine
Tris base	Tris (hydroxymethyl)- aminomethane
Tris-HCl	Tris hydrochloric acid
ug	Microgram
ug/mL	Microgram/microliter
uL	Microliter
uM	Micrometer
U/uL	Atomic mass unit/ microliter
V	Voltage
v/v	Volume/volume
w/v	Weight/ volume
°C	Degree celcius
%	Percentage

# **KAJIAN GENOM *In siliko* FAJ KHUSUS KEPADA *Pseudomonas aeruginosa***

## **USM AR2**

### **ABSTRAK**

Peningkatan kemajuan dalam teknologi pengklonan dan penjujukan telah menghasilkan koleksi genom bakteriofaj atau faj yang besar. Penjujukan dan anotasi genom faj yang lengkap menyumbang kepada penghasilan maklumat penting dalam penemuan kepelbagaian populasi faj. Analisis *in siliko* genom memudahkan pengenalpastian gen dan protein virus dengan lebih baik. Hasil dapatan kajian penjujukan lengkap ke atas genom *Pseudomonas* faj SM1 mendapati asid dioksiribonukelik (DNA) *Pseudomonas* faj SM1 terdiri daripada 93 191 bp dengan kandungan G+C sebanyak 55.2 %. Analisis keseluruhan genom mendedahkan sejumlah seratus dua puluh sembilan kerangka bacaan (ORF) telah dikenalpasti dalam genom *Pseudomonas* faj SM1. Keseluruhan ORFs telah ditemui tersusun diatas unting positif melainkan satu ORF sahaja. Berdasarkan beberapa analisis perisian, 60 ORFs tidak menunjukkan sebarang persamaan kepada protein yang telah dikenalpasti dalam pengkalan data, justeru dianggap sebagai penemuan unik kepada faj ini. Selain itu, 18 ORFs yang lain ditemui berpadanan dengan protein berfungsi dan 51 ORFs mengkodkan protein hipotesis. Penjajaran genom *Pseudomonas* faj SM1 dengan genom *Pseudomonas* faj lengkap yang tersimpan dalam pengkalan data NCBI (Pusat Kebangsaan Untuk Maklumat Bioteknologi) menunjukkan kekurangan persamaan. Oleh hal yang demikian, *Pseudomonas* faj SM1 dicadangkan sebagai satu penemuan *Pseudomonas* faj yang baharu. ditemui. Justifikasi himpunan dan anotasi genom ini telah dijalankan melalui pengklonan dan ekspresi gen kapsid mengandungi perancah dalam sistem *E.coli*. Pemerhatian zarah homogen yang menyerupai morfologi seperti kapsid di bawah mikroskop elektron transmisi membuktikan kewujudan rekombinan

kapsid mengandung protein perancah. Kajian lanjut penjujukan protein peptida ke atas rekombinan protein kapsid mengandung perancah yang telah dimurnikan mengesahkan identiti protein tersebut. Kesimpulannya, analisis bioinformatik ke atas himpunan jujukan dan anotasi genom lengkap ini dapat disahkan untuk *Pseudomonas* faj SM1.



***In silico* GENOMIC STUDY OF PHAGE SPECIFIC TO *Pseudomonas*  
*aeruginosa* USM AR2**

**ABSTRACT**

Advancement in cloning and sequencing technologies are producing a massive collection of bacteriophage or phage genomes. The complete genome sequencing and annotation of phage genome constitute important information in the discovery of highly diverse phage population. To better understand the biology of phage, *in silico* genomic analysis would allow the identification of viral genes and proteins. In this study, *Pseudomonas* phage SM1 genome was completely sequenced and composed of 93 191 base pairs in length, with G+C content of 55.2 %. Genome analysis of the double stranded DNA revealed 129 open reading frames (ORFs) present in the full genome. All of the ORFs were arranged on the positive strand except for one ORF. Based on several software analysis, 60 ORFs show no significant homology to the known proteins deposited in the databases and were considered as unique to this phage. Meanwhile, another 18 ORFs encode for known functional proteins and 51 ORFs encode for hypothetical proteins. Genome alignment of *Pseudomonas* phage SM1 against *Pseudomonas* phage complete genome sequences that were deposited in NCBI (National Center For Biotechnological Information) showed lack of similarities. Thus, suggesting *Pseudomonas* phage SM1 as a potential newly discovered *Pseudomonas* phage. The verification of the assembled and annotated results was performed by cloning ORF 1, encoding capsid containing scaffold protein in the *E.coli* system. The detection of homogenous particles resembling capsid-like morphology of *Pseudomonas* phage SM1 under transmission electron microscope (TEM) gave assurance that the recombinant capsid containing scaffold proteins was successfully expressed. Further peptide sequencing of purified capsid containing

scaffold protein result confirmed the identity of the recombinant protein. Hence, the bioinformatics analyses of the genome sequence and annotation were verified for *Pseudomonas* phage SM1.

## CHAPTER 1: INTRODUCTION

Bacteriophages (phages) are viruses that specifically infect bacteria (Black, 2012). With an estimated phages population more than  $10^{31}$  or approximately ten million per cubic centimeter of any environmental niche where bacteria or archaea thrives, phages could be the most ubiquitous species on earth (Clokier *et al.*, 2011). Therefore, it is not surprising phages play significant roles in shaping the natural population and the evolution of the bacteria (Hatfull, 2015).

In general, studies on phages were fundamental to the emerging fields of molecular biology and genomics (Grath and van Sinderen, 2007; Petty *et al.*, 2007). In 1976, phage MS2 was the first RNA genome completely sequenced and determined (Fiers *et al.*, 1976). The following year, the first DNA genome completely sequenced was also a phage, phi X174 which is a single stranded DNA (ssDNA) with approximate 5400 nucleotides (Sanger *et al.*, 1977). After ten years later, the first double stranded DNA (dsDNA) complete genome sequence reported was a mycobacteriophage L5, a tailed phage which infecting a non *Escherichia coli* host (Hatfull and Sarkis, 1993). Since then, the numbers of sequenced phage genomes have exploded as DNA sequencing methodologies have advanced.

At present, over 6000 phages have been reported by International Committee on Taxonomy of Viruses (ICTV) (Ackermann, 2011). But, out of that number, only 1953 complete phages genome sequences have been deposited in the National Centre for Biotechnological Information (NCBI) phage genome database. These sequenced phages just represent a small portion (32%) of overall reported phages in ICTV. The slow increase in the number of phage complete genomes being published shows that there are many more phages are waiting to be sequenced (Hatfull, 2008). For this

reason, the study of complete genome sequence of a lytic and locally isolated phage infecting *Pseudomonas aeruginosa* USM AR2 designated as *Pseudomonas* phage SM1 could contribute to the increase of phage genomic sequence data.

According to the Nur Ashiffa (2005), *Pseudomonas aeruginosa* USM AR2 is a good hydrocarbon utilizing microbe and a good rhamnolipid producer. Previously, *Pseudomonas* phage SM1 showed to have a unique feature as an inducer for rhamnolipid in infected *Pseudomonas aeruginosa* USM AR2 (Wan Khairunisa, 2012). Furthermore, there was also no scientific study or publication on phage capable as rhamnolipid inducer ever being reported. Therefore, the complete genome could provide biological information in explaining the unique characteristic of *Pseudomonas* phage SM1.

In this work, understanding the complete genome sequence of the *Pseudomonas* phage SM1 could address the following questions: (i) Did the *Pseudomonas* phage SM1 carry any special gene(s) in the genome that contribute to this special ability as rhamnolipid inducer? (ii) If so, how novel the gene compared with reported genes in databases? Thus, this work was carried out with the following specific objectives:

- i. To sequence and annotate the complete genome sequence of *Pseudomonas* phage SM1.
- ii. To verify the genome annotation sequence of *Pseudomonas* phage SM1 by expression of selected gene.

## CHAPTER 2: LITERATURE REVIEW

### 2.1 Discovery of phage

Phage was discovered independently by two scientists, bacteriologist William Twort and microbiologist Felix d'herelle, in 1915 and 1917 (Twort, 1915; d'Hérelle, 1917; Fong *et al.*, 2017). Twort discovered that phage displayed the ability to kill bacteria via bacterial lysis which was determined by the appearance of plaques on the bacterial lawn of *Micrococcus* and led to the lytic phage concept (Twort, 1915). After two years later, Felix d'herelle then found the same phenomenon of bacterial lysis in *Shigella* cultures and consequently termed the anti microb particles as “bacteriophage” which means bacterial eater (d'Hérelle, 1917).

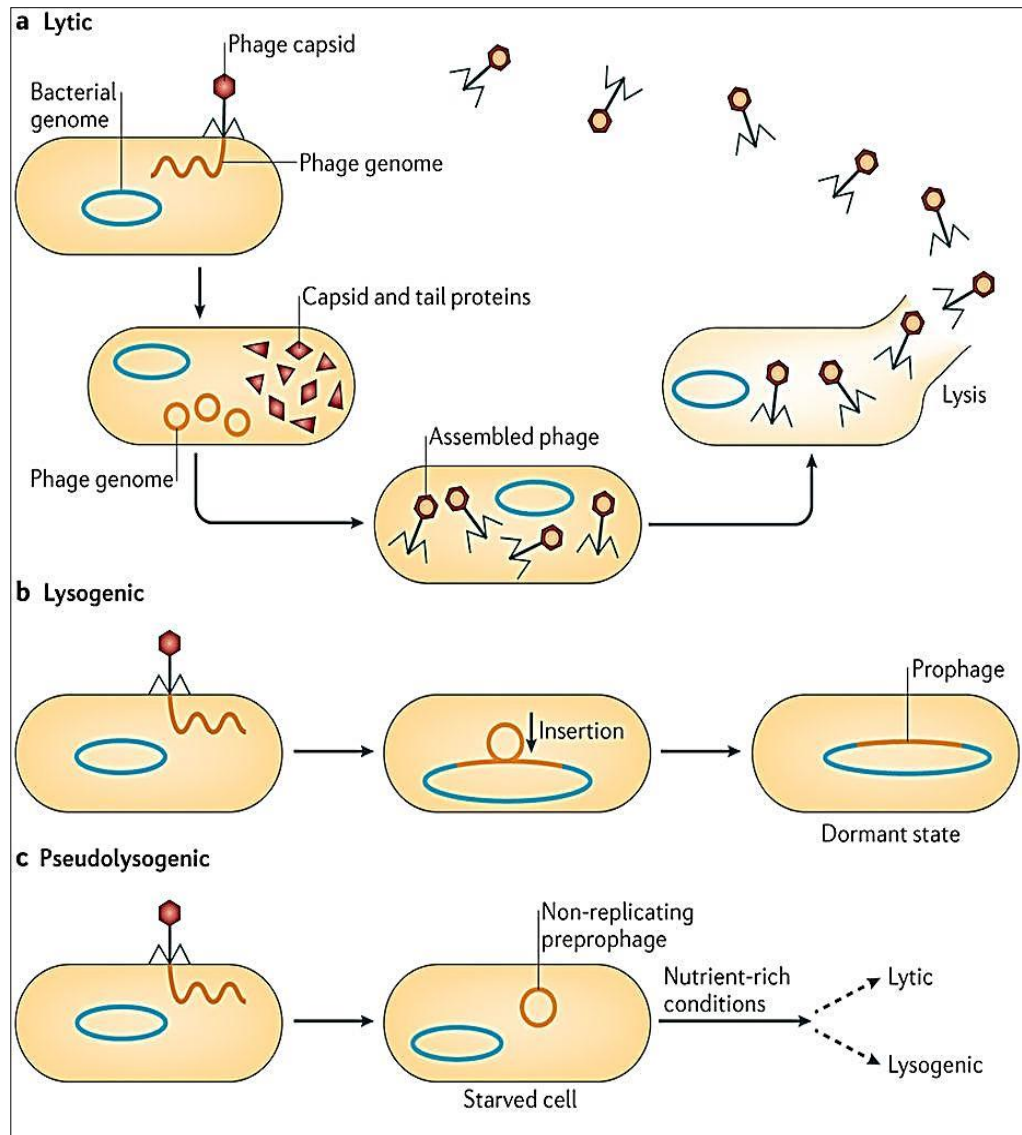
### 2.2 The nature of phage

Phage is a virus infects bacterium. Like other type of viruses, phage is an obligate intracellular parasite which multiplies by invading the host cells and manipulates the cells metabolic machinery (Black, 2012). To establish phage infections, the process requires specific protein recognition between phage receptor that bind to the specific receptor sites on the surface of a susceptible host (Brock, 2012). Therefore, the specificity of phage attacking only the targeted species or in few cases a single strain of bacterium consequently influences the bacterial host range (Sulakvelidze *et al.*, 2001; Weinbauer, 2004).

Phage can exhibit one of two types of life cycle: lytic or lysogeny (**Figure 2.1**) (Lenski, 1988). Virulent phage undergo lytic infectious cycle, with infection resulting in rapid lysis and death of the host within a short period, releasing new infectious phage progeny within minutes or hours (Young, 1992; Harper *et al.*, 2014).

Meanwhile, temperate phages basically exist as dormant DNA (prophage) in their life cycle (Lwoff, 1953). Upon infection, prophages are replicated together with the bacterial host chromosome and this lysogenic state is maintained by the repression of the phage lysis genes. The lysogenic-lytic switch can be induced by stressor such as DNA damage (Quinones *et al.*, 2005) or spontaneous conditions (absence of stressor) (Ranquet *et al.*, 2005; Nanda *et al.*, 2015). Some prophages may persist as low copy number plasmids instead of integrate into the host chromosome such as P1 and N15 phages (Edlin *et al.*, 1977; Ravin *et al.*, 2000). The crucial steps in the onset of the lytic and lysogenic cycles are 1) phage genome excision and 2) integration which are mediated by phage encoded DNA recombinases such as integrases and excisionases (Nash, 1981). This integration process occurred at a site-specific recombination between 'attachment sites' in the phage (attP) and the host (attB) (Nash, 1981). Besides that, some temperate phages also are capable to integrate randomly within their host genome and thus increase variation and possible mutations within the bacterial population such as phage Mu (Harshey, 2012).

Other than those two most commonly described phage life cycle, there is another less common phage life cycle termed as pseudolysogeny, inactive form of the phage genome which neither replicate as in lytic production nor integrates into the host chromosome as in lysogeny state (Ripp and Miller, 1997). It is hypothesized that pseudolysogeny mode occurs mostly due to the bacterial cells starved state, an unfavorable growth condition which make the host cells cannot support DNA replication (Ripp and Miller, 1997; Łoś and Węgrzyn, 2012). However, once the nutritional status is restored, the pre-prophage will eventually be triggered to enter into the lytic or lysogenic life cycle (Feiner *et al.*, 2015).



**Figure 2.1: Phage replication cycles** (Feiner *et al.*, 2015). In lytic cycle (a), phage multiplies inside the host bacterium and released by lysing the host bacterial cell wall. However, in lysogeny cycle (b), the phage genome gets integrated into the bacterial DNA without causing lysis. In pseudolysogeny (c), phage becomes either virulent or temperate (without integrate into the host genome) upon the addition of sufficient nutrient concentrations.

### 2.3 Phages are ubiquitous

Phages are ubiquitous everywhere wherever their bacterial host exist (Allen *et al.*, 2013) even in extreme thermal environments such as hot spring and halo-alkaline habitats (Gudbergsdóttir *et al.*, 2016; Van Zyl *et al.*, 2016). Few decades ago, phages were first discovered highly abundant in marine ecosystem by epifluorescent microscopy approach following nucleic acid staining and suggested that the phage to bacterium ratio of more than 10 phages per microbe which correlated with the density of the phage and host bacterial populations (Marie *et al.*, 1999; Suttle, 2005). Phages concentration in seawater was also reported varied at different seasonal and depth in response to variations in environmental parameters including temperature, salinity, dissolved oxygen and nutrient concentrations (Bergh *et al.*, 1989; Brum *et al.*, 2015; Fuhrman *et al.*, 2015). Advances on metagenomic study once again justified the claims that phages are abundant in marine ecosystem (Sepulveda *et al.*, 2016).

Besides that, phages are also revealed extremely vast in a diverse range of soil types and locations. In a survey of Delaware soils which included wetlands and agricultural soils, viral abundance ranged from 8 to  $39 \times 10^8$  particles per gram dry weight of soils, while in cryptic Antarctic soils viral abundance was 23 -  $64 \times 10^7$  particles per gram (Williamson *et al.*, 2005). With estimation of  $1.5 \times 10^8$  particle per gram of soil, this is equivalent to 4 % of the total population of bacteria, therefore making them important integral part of soil bacterial ecology (Ashelford *et al.*, 2003).

Next, majority of published human faecal metagenomes revealed that phages are highly abundant and stable in the human gut (Reyes *et al.*, 2010; Dutilh *et al.*, 2014; Waller *et al.*, 2014). The significant numbers of phages found in feces can be expected because the numbers of bacterial in the gut are also overwhelming. Gut



phages are most likely contributed to maintaining human health and control the pathogenic microbe in human colon (Lusiak-Szelachowska *et al.*, 2006; Stern *et al.*, 2012). This was followed by study based on analysis of DNA sequence dataset of active phages in healthy individuals and compare to the available metagenomic datasets of phage community from healthy individuals (Manrique *et al.*, 2016). The results show that the shared phages between healthy individuals with patients suffering gastrointestinal disease were small percentage. Therefore, existence of human gut phages suggests that the net influence of phages in the human gut is not deleterious, but rather beneficial (Manrique *et al.*, 2016).

The metagenomic analysis of the viral communities in fermented foods also illustrates the phage abundance in foods such as fermented shrimp, kimchi, and sauerkraut (Park *et al.*, 2011). In this study, it was found that *Caudovirales* phages dominated the fermented product; *Siphoviridae* family was abundant in the fermented shrimp (53.55 %) and sauerkraut (60.07 %). Meanwhile, the most abundant viral family in kimchi was the *Podoviridae* (52.82 %). Similar figure has been shown in dairy fermentations by lactic acid bacteria (Marcó *et al.*, 2012). The study estimated that phage concentrations up to  $10^9$  phages per ml of cheese whey or per gram of product.

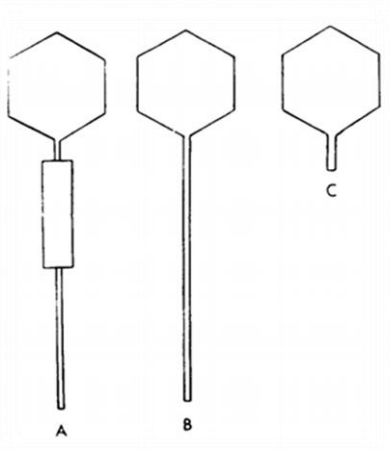
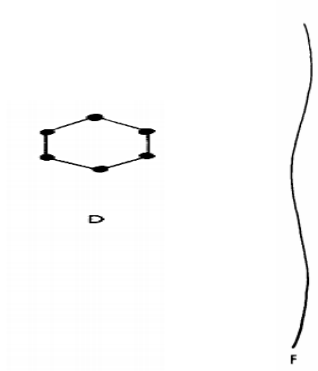

## **2.4 Classification of phage**

More than 5500 phages have been examined under electron microscope after the technique of negative staining was introduced (Ackermann, 2007). Microscope image provide interesting information about the virion morphological features but identification of phages solely based on the microscopic studies are difficult because micrograph often confuse phages with other filamentous particles such as bacterial

flagella, pilli and broken tail phages (Ackermann and Tiekotter, 2012). Among 5500 examined phages, most phages (96%) are tailed and belonged to order *Caudovirales* constitute of three families named *Myoviridae*, *Siphoviridae* and *Podoviridae*. Meanwhile only 208 (3.7%) phages reported belong to pleomorphic, filamentous and polyhedral group (Ackermann, 2007).

Traditional classification of phages outlined six basic morphotypes based on morphology and nucleic acid (**Table 2.1**) (Bradley, 1967). According to the Bradley scheme, groups A-C composed of double stranded DNA (dsDNA) genome enclosed within capsid and a tail attached to it. These tailed phages are classified depend on the nature of their tail, which is long and non-contractile long (*Siphoviridae*) and contractile (*Myoviridae*), or short (*Podoviridae*) (Bradley, 1967). On the other hand, all phages in group D and F contains single stranded DNA (ssDNA) but their morphology is different in which phages in Group D represented by grand size nucleocapsid with fibrous or spiky surface structures without tail whereas phages in group F are defined by its filament-like shape (Bradley, 1967). Other Group E includes only non-tail and nucleocapsid phages containing RNA genome (Bradley, 1967)

**Table 2.1:** Bradley's phage classification scheme (Bradley, 1967)

Nucleic acid type	Group	Morphology
Double stranded DNA		<p>A- Icosahedral head with contractile tail</p> <p>B- Icosahedral head with long and non-contractile tail</p> <p>C- Icosahedral head with short tail</p>
Single stranded DNA		<p>D- No tail, nucleocapsid with fibrous or spiky surface structures</p> <p>F- No head, flexible filament</p>
Single stranded RNA		<p>E- No tail, nucleocapsid</p>

Nowadays, the classification and nomenclature of viruses is the responsibility of the International Committee on Taxonomy of Viruses (ICTV) (Adams *et al.*, 2015; Adams *et al.*, 2017). Accessible phage genome sequencing has resulted in an increase of sequences deposited to databases and the need to classify them into new species, genera, subfamilies and families (King *et al.*, 2011). Previously, Ackermann (2011) reported that ICTV phage taxonomy includes one order *Caudovirales* constitutes of 3 tailed phages families (*Myoviridae*, *Podoviridae*, and *Siphoviridae*), 7 families of pleomorphic, filamentous and polyhedral phages and 17 genera. The order *Caudovirales* at present includes six subfamilies, 80 genera and 441 species (Krupovic *et al.*, 2016). This phage classification was expanded dramatically which reflects genome based relationship between phages because recent phage classification efforts are based on sequence similarity analyses for identification and comparative study of closely related phages using program such as provided by Basic Local Alignment Search Tool Nucleotide (BLASTN) (Altschul *et al.*, 1990; Ceysens *et al.*, 2011; Mizuno *et al.*, 2013). But, classifying phages based on their genome sequence alone is often difficult due to the highly mosaic organization of their genome (Krupovic *et al.*, 2011). Furthermore, phage also has no universal gene marker that can be used for analysis taxonomy and phylogeny compared to their host which have 16S rRNA and various other conserved genes that are useful for their phylogenetic classification (Rohwer and Edwards, 2002; Clokie *et al.*, 2011).

## 2.5 Phage genomic

### 2.5.1 Sequences diversity

Phages have been acknowledged as dynamic parasite and possibly the most diversified microorganisms on the planet (Breitbart *et al.*, 2003; Thurber *et al.*, 2017). The frequencies of novel genes that are portrayed in newly characterized phage genomes give a better picture of the phage genetic diversity (Nolan *et al.*, 2006; Juarez *et al.*, 2016). Such remarkable feature is mostly driven by their survival strategies when facing selective pressure created by anti-phage barriers and not surprisingly phages are now widely recognized as vehicles of horizontal gene transfer (Labrie *et al.*, 2010).

In general, there are essentially two types of approaches available to define phage diversity. First, complete genome analyzing of individually isolated phage (Hatfull, 2008). This approach takes advantage of the small size of phages genome to determine its sequence, annotating gene predictions and comparing the sequence to that of other known viruses and therefore reveals the complexities of the relationship among phages (Hatfull *et al.*, 2006; Jordan *et al.*, 2014). For example, recent analysis of whole genome comparison of individually isolated 627 phages infecting *Mycobacterium smegmatis* revealed that phages infect the same host can exhibit genetic diversity and enormous mosaic genome. Perhaps somewhat more surprising is that high frequency of novel genes from other sources are constantly added to the phage genome suggesting that phages underwent gene influx for phage survival in certain host or environments (Morris *et al.*, 2008; Pope *et al.*, 2015).

Next general approach is viral metagenomics in which large amount of phages genome sequences data were generated at random from environmental samples (Hatfull, 2008; Simmonds *et al.*, 2017). Since 2002, applications of culture independent sequence analysis of viral assemblages have been useful to provide insights into viral functions, community composition and structure in the environment (Breitbart *et al.*, 2002). Using this high throughput sequencing method, accessing of virosphere diversity is possible (Mizuno *et al.*, 2013). Furthermore, Dutilh *et al.*, (2014) stated that metagenomic sequencing opened new windows in understanding the dynamics of phage populations and diversity. However, the weakness of this method is that no biological resources for further experimentation are offered and therefore has raised the concern that virus classification based on metagenomic sequence alone would result in a taxonomy of sequence rather than of viruses (Van Regenmortel, 2016).

### **2.5.2 Mosaic genome architecture**

In the 1990s, Hendrix *et al.*, revealed that double stranded DNA (dsDNA) phages and prophages genomes are highly mosaic with different segments having distinct evolutionary histories. This event reflected the high degree of horizontal genetic exchange occurred within the phage communities and linked to phage evolution (Hendrix *et al.*, 1999).

This exchange event also prevalent in bacteria as bacteria also acquired gene through horizontal genetic transfer (i.e transformation, transduction and conjugation) (Retchless and Lawrence, 2007) but the extent of mosaicism in phage genomes is much greater and can be seen even clearer through the phage whole genome comparative analysis (Hatfull, 2008; Belcaid *et al.*, 2010; Hatfull and Hendrix, 2011).

Phage genome mosaicism can be observed by comparative analysis of whole genome nucleotide sequences (Hatfull, 2008). In sequence comparison analysis, different genes, groups of genes, or segment of genes which all share different ancestor can be revealed and these represent the exchangeable genetic modules present in phage mosaic genome (Hatfull and Hendrix, 2011). For example, recent comparative genomic analysis of *Pseudomonas aeruginosa* phage PaMx25 (accession number JQ067084) revealed that this phage shared homologous ORFs with those different phage-host taxonomic classes (phylum Proteobacteria) (Flores *et al.*, 2017). Commonly, genetic exchange among phages is restricted by their host range, but it is unusual for phages infecting different hosts to display extensive nucleotide sequence identity (Hatfull and Hendrix, 2011). Nevertheless, these events possibly arise from homologous and non-homologous recombination mediated by host or phage-encoded recombination machineries and tend to create mosaicism typically observed in phage genomes (Hendrix, 2002). Therefore, the picture of phage genomic architecture and evolution becoming clearer by addition of more phage genome sequences (Flores *et al.*, 2017). In addition, phamily circle can also be used as an alternative tool to construct the phylogenetic relationship and subsequently illustrate their evolution routes taken by each constituent module (Hatfull *et al.*, 2006; Belcaid *et al.*, 2010).

### **2.5.3 Overlapping gene**

Overlapping genes (OGs) are common feature in DNA and RNA viruses (Barrell *et al.*, 1976; Lamb and Horvath, 1991; Firth, 2014). Besides that, OGs also are found in the sequenced genome of several prokaryotic and eukaryotic including human genomes (Johnson and Chisholm, 2004; Nakayama *et al.*, 2007; Sabath *et al.*, 2008). In overlapping genes, the same nucleotides sequence is described as a region

coding for more than one polypeptide using different reading frames (Johnson and Chisholm, 2004).

The existence of overlapping phage genes was first identified in viruses through the first sequenced of DNA phage  $\Phi$ X174 (Barrell *et al.*, 1976). Barrell, *et al.*, suggested that the evolution of overlap genes occurred possibly due to the size of the  $\Phi$ X174 genome is limited by packaging of protein capsid or other unknown constraints which consequently compressed the genome and cause the genes overlap. Thus, allowing the virus to increase its repertoire of proteins without increasing its genome length (Scherbakov and Garber, 2000; Chung *et al.*, 2008; Chirico *et al.*, 2010). However, this point of view has weakened by the discoveries of inverse correlation between overlapping rate and genome size in viruses (Brandes and Linial, 2016). But instead, overlapping gene is proposed as an effective method for generating novel gene printed on top of an already compact genome which led to the increasing of genome coding capacity and allowing the virus to overcome the host and environmental limitation (Brandes and Linial, 2016).

Since then, several explanations have been proposed by many authors for the abundance of overlapping gene in viruses. According to Holmes (2009), gene overlap is probably due to the extremely high mutation rate in viruses (especially RNA viruses) because of the absence of proofreading mechanism in their replication enzyme (RNA polymerase) which indirectly will limit the genome length and can result the new genes from overlaps region. Although this theory links the mutation rates with genome size of a virus but it also has been reported that the genome size is negatively correlated with mutation rate (Båtshake and Sundelin, 1996; Makalowska *et al.*, 2005; Sanna *et al.*, 2008).



Second, Krakauer and Plotkin (2002) study shows that overlapping gene could be one of many safety mechanism to remove the mutant genomes from the viral population through antiredundancy mechanism; mechanism that recognizes changes to the genome and repair them by amplifying the deleterious effect of mutations, thereby benefits their fitness traits. Hence, overlapping genes that are exhibited by viruses reflects their high rates of recombination and frequent positive selection (Duffy *et al.*, 2008).

Overlap gene is also argued to play role in the regulation of gene expression by coordinating gene expression within a cluster of functionally linked genes (Krakauer 2000; Dreher and Miller, 2006; Boldogkői, 2012). For example, RNA phage MS2 genome revealed that the beginning of the lysis gene was overlapped with end coat gene, and the translation of both reading frames is coupled; synthesis of lysis gene depends on translation of coat protein termination (Berkhout *et al.*, 1987).

## **2.6 Host**

### **2.6.1 *Pseudomonas aeruginosa***

The genus *Pseudomonas* is characterized by gram negative, rod shape with polar flagella belongs to the family *Pseudomonadaceae* and obligate aerobic chemoorganotroph (Brock, 2012). They are oxidase positive, but they are not strict aerobe because *Pseudomonas* sp are also found under anoxic condition (Zhao *et al.*, 2015). *Pseudomonas* sp are capable to degrade the agrochemicals and xenobiotic compound such as pesticides and toxic chemicals (Kulkarni and Kaliwal, 2014). *Pseudomonas* sp are ubiquitous (Green *et al.*, 1974) in environment because of their

versatile metabolic capacity to use various carbon sources and electron acceptors to adapt to diverse environments (Williams and Worsey, 1976).

*Pseudomonas aeruginosa* is an opportunistic pathogenic strain commonly associated with nosocomial infection (Rosenthal *et al.*, 2016). It rarely affects healthy people but can cause disease in immunocompromised patients like whom associated with cystic fibrosis, burn infection, acquired immune deficiency syndrome (AIDS) or cancer (Bendig *et al.*, 1987; Franzetti *et al.*, 1992; De Bentzmann and Plésiat, 2011). These bacteria are also naturally resistant to a wide range of antibiotics, so the treatment of infections is often difficult (Livermore, 2002) and they are capable of biofilm production which contributes to their multidrug resistant strains (Dunne 2002; Manson *et al.*, 2017). Hence, these causes *P. aeruginosa* becomes more virulent and cannot be eradicated easily (Rahim *et al.*, 2017).

Interestingly, most of the *Pseudomonas aeruginosa* strains also have an ability to produce compounds of biotechnological importance such as biosurfactant rhamnolipid which can be used as a replacement of synthetic surfactant and has many potential applications in environmental and industrial applications (Caiazza *et al.*, 2005; Pacwa-Płociniczak *et al.*, 2011). In this study, *Pseudomonas aeruginosa* USM AR2 was used as a host for the phage SM1.

## **2.7 Rhamnolipid**

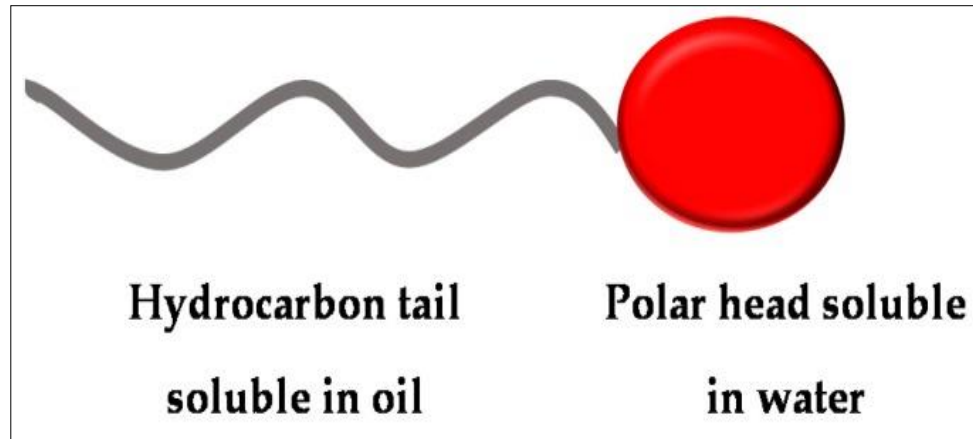
Biosurfactants are “green” chemical (bio origin), amphiphilic compounds that display surface activity (**Figure 2.2**) (Desai and Banat, 1997). These surfactants are extracellular compounds produced by various microbes such as fungi, bacteria and yeast (Healy *et al.*, 1996). According to Rosenberg and Ron (1999), biosurfactants are divided into two categories; 1) high molecular weight and 2) low molecular

weight molecules. Low molecular weight surfactants include glycolipids, lipopeptides and phospholipids, while high molecular weight surfactants include polymeric and particulate surfactants (**Table 2.2**) (Shoeb *et al.*, 2013; Silva *et al.*, 2014). As an excellent surface tension reducer chemical, biosurfactants are attracting much attention over the chemical surfactants due to their ecological acceptance that they have low toxicity levels, produced from renewable sources (fungi, bacteria and yeast) and biodegradable compound (Kosaric 1992; Mulligan and Wang, 2006). Due to the biosurfactants diverse environmental and industrial applications (**Table 2.3**) (Pacwa-Płociniczak *et al.*, 2011), many patents have been filed on biosurfactants (**Table 2.4**) (Sachdev and Cameotra 2013; Silva *et al.*, 2014).

Up to now, rhamnolipid produced by *Pseudomonas aeruginosa* are among the best studied glycolipid biosurfactants (Desai and Banat, 1997; Soberón-Chávez *et al.*, 2005). Rhamnolipid is a surface active glycolipid containing 1-rhamnose and  $\beta$ -hydroxy fatty acids moieties (Lang and Wullbrandt, 1999). Due to their amphiphilic nature, these surface-active molecules are capable to foaming, form emulsions, and solubilize oil (Desai and Banat, 1997; Van Bogaert *et al.*, 2007). Thus, these molecules can be used as emulsifiers, de-emulsifiers, wetting and foaming agents, functional food ingredients and as detergents in petroleum, petrochemicals, environmental management, agrochemicals, foods and beverages and cosmetics and pharmaceuticals industries (**Table 2.5**) (Singh *et al.*, 2007; Souza *et al.*, 2014; Chong and Li, 2017).

From an economic perspective, capability to produce rhamnolipid from low-cost substrates (Rahman *et al.*, 2002) and agro-industrial waste (George and Jayachandran, 2013) are among important criteria for economical large scale

biosurfactant production (Marchant and Banat, 2012). Besides that, another problematic issue in the biosurfactant large scale production compared to synthetic surfactant is the high cost of purification and recovery process (Mukherjee *et al.*, 2006). According to Heyd *et al.*, (2008), the downstream process often problematic due to the microbial strains used are pathogenic and they produced product mixtures instead of single product which consequently lead to higher costs and problems with scaling up the downstream processing. Hence, for commercialization purpose, the improvement and alternative technology must be implement (Sekhon Randhawa and Rahman, 2014).



**Figure 2.2: Structure of surfactant.** Surfactant molecule with apolar (hydrophobic) and polar (hydrophilic) moieties. The apolar moiety is often a hydrocarbon chain, whereas the polar moiety may be ionic (cationic or anionic), non-ionic or amphoteric (Santos *et al.*, 2016)

**Table 2.2:** Main type of biosurfactants (Silva *et al.*, 2014)

<b>Class</b>	<b>Type of Biosurfactant</b>	<b>Microorganisms</b>
<b>Glycolipids</b>	Rhamnolipids	<i>Pseudomonas aeruginosa</i>
	Sophorolipids	<i>Torulopsis bombicola</i> , <i>T. apicola</i>
	Trehalolipids	<i>Rhodococcus erythropolis</i> , <i>Mycobacterium</i> sp.
<b>Lipopeptides and lipoproteins</b>	Peptide-lipid	<i>Bacillus licheniformis</i>
	Viscosin	<i>Pseudomonas fluorescens</i>
	Serrawettin	<i>Serratia marcescens</i>
	Surfactin	<i>Bacillus subtilis</i>
	Subtilisin	<i>Bacillus subtilis</i>
	Gramicidin	<i>Bacillus brevis</i>
	Polymyxin	<i>Bacillus polymyxa</i>
<b>Fatty acids, neutral lipids and phospholipids</b>	Fatty acid	<i>Corynebacterium lepus</i>
	Neutral lipids	<i>Nocardia erythropolis</i>
	Phospholipids	<i>Thiobacillus thiooxidans</i>
<b>Polymeric surfactants</b>	Emulsan	<i>Acinetobacter calcoaceticus</i>
	Biodispersan	<i>Acinetobacter calcoaceticus</i>
	Liposan	<i>Candida lipolytica</i>
	Carbohydrate-lipid-protein	<i>Pseudomonas fluorescens</i>
	Mannan-lipid-protein	<i>Candida tropicalis</i>
<b>Particulate surfactant</b>	Vesicles	<i>Acinetobacter calcoaceticus</i>

**Table 2.3:** Application of biosurfactants in bioremediation of contaminated sites  
(Pacwa- Płociniczak *et al.*, 2011)

<b>Biosurfactants</b>		<b>Applications in Environmental Biotechnology</b>
<b>Group</b>	<b>Class</b>	
<b>Glycolipids</b>	Rhamnolipids	Enhancement of the degradation and dispersion of different classes of hydrocarbons; emulsification of hydrocarbons and vegetable oils; removal of metals from soil
	Trehalolipids	Enhancement of the bioavailability of hydrocarbons
	Sophorolipids	Recovery of hydrocarbons from dregs and muds; removal of heavy metals from sediments; enhancement of oil recovery
<b>Fatty acids, phospholipids and neutral lipids</b>	Corynomycolic acid	Enhancement of bitumen recovery
	Spiculisporic acid	Removal of metal ions from aqueous solution; dispersion action for hydrophilic pigments; preparation of new emulsion-type organogels, superfine microcapsules (vesicles or liposomes), heavy metal sequestrants
	Phosphatidylethanolamine	Increasing the tolerance of bacteria to heavy metals
<b>Lipopeptides</b>	Surfactin	Enhancement of the biodegradation of hydrocarbons and chlorinated pesticides; removal of heavy metals from a contaminated soil, sediment and water; increasing the effectiveness of phytoextraction
<b>Polymeric biosurfactants</b>	Emulsan	Stabilization of the hydrocarbon in water emulsions
	Alasan	
	Biodispersan	Dispersion of limestone in water
	Liposan	Stabilization of hydrocarbon-in-water emulsions
	Mannoprotein	

**Table 2.4:** Patents on biosurfactants produced by microorganisms (Silva *et al.*, 2014).

<b>Microorganism / Type of Biosurfactant</b>	<b>Patent Holder</b>	<b>Title of Patent</b>	<b>Publication No.</b>
Sophorolipid producer	Borzeix F	Sophorolipids as stimulating agent of dermal fibroblast metabolism	US 6057302 A
Sophorolipid producer	Borzeix F, Concaix	Use of sophorolipids comprising diacetyl lactones as agent for stimulating skin fibroblast metabolism	US 6596265 B1
New strains of hydrocarbon-degrading bacteria capable of producing biosurfactants	Robin L. Brigmon, Sandra Story, Denis Altman, Christopher J. Berry	Surfactant biocatalyst for remediation of recalcitrant organics and heavy metals	PI 0519962-0 A2
Sophorolipid producer	Gross RA, Shah V, Doncel GF	Spermiocidal and virucidal properties of various forms of sophorolipids	WO 2005089522 A2
<i>C. albicans</i> , <i>C. rugosa</i> , <i>C. tropicalis</i> , <i>C. lipolytica</i> , <i>C. torulopsis</i>	Awada S, Spendlove R, Awada M	Microbial biosurfactants as agents for controlling pests	US 20050266036 A1
<i>Pseudomonas aeruginosa</i>	Silvanito Alves Barbosa, Roberto Rodrigues De Souza	Biosurfactant production for development of biodegradable detergent	PI 1102592-1 A2
Sophorolipid producer	Cox TF, Crawford RJ, Gregory LG, Hosking SL, Kotsakis	Mild to skin, foaming detergent composition	WO2011120776 A1



**Table 2.5:** Applications of rhamnolipid (Chong and Li, 2017)

Applications	Examples
Oil recovery	Microbial enhanced oil recovery (MEOR) Increase amount of recoverable oil aided by rhamnolipid producing microorganisms
Bioremediation	Bioremediation of petroleum at contaminated sites Addition of rhamnolipids improves solubility of hydrocarbons to facilitate degradation
	Bioremediation of heavy metal at contaminated lands or in water treatment plant; rhamnolipids can be applied in foaming-surfactant technology to remove heavy metal contaminants
	Bioremediation of pesticides at agricultural fields; addition of rhamnolipids can enhance degradation of chemical pesticides
Pest control	As emulsifier, spreaders and dispersing agent in pesticide formulations
	As bio-pesticide against agricultural pests; rhamnolipids have insecticidal activity against green peach aphids and <i>Aedes aegypti</i> larvae
Crop protection	As biocontrol agent against several phytopathogenic fungi; addition of rhamnolipids or rhamnolipid-containing cell-free broth are effective in inhibiting growth of phytopathogens, e.g. <i>F. oxysporum</i> , <i>B. cinerea</i> , <i>Mucor</i> spp. and many more
	As stimulant for plant immunity; induced genes involved in plant's defense system in tobacco, wheat and <i>Arabidopsis thaliana</i> ; induced biosynthesis of plant hormones important for signaling pathways involved in plant immunity
Food processing	As food ingredients or additives functioning as emulsifier, solubilizer, foaming and wetting agent

	As antimicrobial agent preventing food spoilage and for sanitization; rhamnolipids inhibit growth of foodborne pathogenic bacteria, e.g. <i>L. monocytogenes</i> , <i>B. subtilis</i> ; rhamnolipids also prevent formation of biofilms due to their anti-adhesive nature
Medical use	As biofilm control agent to prevent medical device-related infections; inhibit biofilm formation; synergistic effect with caprylic acid to inhibit biofilms of more resistant pathogens, e.g. <i>P. aeruginosa</i> and <i>S. aureus</i>
Protein folding	Aid in folding of outer membrane protein A
Microbial fuel cells	Improve power density output of microbial fuel cells
Synthesis of nanoparticles	As structure directing agent in nanoparticles synthesis