GENOME-WIDE METABOLIC AND EVOLUTIONARY ANALYSIS OF *PAENIBACILLUS POLYMYXA* ATCC 15970

SIM KEE SHIN

UNIVERSITI SAINS MALAYSIA

2015

GENOME-WIDE METABOLIC AND EVOLUTIONARY ANALYSIS OF *PAENIBACILLUS POLYMYXA* ATCC 15970

by

SIM KEE SHIN

Thesis submitted in fulfillment of the requirements

for the degree of

Master Science

AUGUST 2015

ANALISIS METABOLIK DAN EVOLUSI MENYELURUH GENOM *PAENIBACILLUS POLYMYXA* ATCC 15970

Oleh

SIM KEE SHIN

Tesis yang diserahkan untuk

memenuhi keperluan bagi

Ijazah Sarjana Sains

OGOS 2015

ACKNOWLEDGEMENTS

In the course of my project, I faced many obstacles. Nevertheless, I was able to overcome these challenges with the support and motivation of fellow friend.

First and foremost, my deepest gratitude and thankfulness to my supervisor Professor Nazalan as well as Professor Razip for everything. Without their guidance, I would not be able to complete this thesis.

To Dr. Mohd Noor Mat Isa from Malaysia Genome Institute (MGI), thank you for your bio-informatics training and support. I would also like to thank Pn. Ayu Sapian and Pn. Enizza for providing hand on training on next generation sequencing library preparation.To Ahmad Yamin, thank you for the assistance in Linux and Bio-informatics support. To Lab 414 members, thank you for offering me guidance, helps and advice.

To Ahmad Yamin, thank you for the assistance in Linux and Bio-informatics support.

To Lab 414 members, thank you for offering me guidance, helps and advice.

Last but not least, I would like to send my special acknowledgement to my mother and friends for their support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

˚C	Degree Celsius
%	Percentage
µl	Microliter
ADP	Adenosine diphosphate
AMOS	A modular Open source assembler
ATP	Adenosine triphosphate
ATCC	American type culture collection
BGI	Beijing genome institute
BLAST	Basic local alignment search tool
Bp	Base pair
CA	Celera assembler
cas9	CRISPR associated protein 9
CDS	Coding sequence
CLR	Continuous long read
Cm	Centimeter
CoA	Coenzyme A
COG	Orthologous groups of protein
CRISPs	Clustered regularly interspaced short palindromic repeats
D	Dextrorotatory
Da	Dalton
DBG	De bruijn graph
DNA	Deoxyribonucleic acid
EC	Enzyme commission
EDTA	Ethylenediaminetetraacetic
et al.	et alia (and others)
Fe	Iron
FeMo-co	Iron molybdenum cofactor
Fld	Ferrodoxin
g	Gram
gDNA	Genomic DNA
GIs	Genomic island
GO	Gene onthology
HMM	Homology hidden markov model
kb	Kilo base pair
KEGG	Kyoto encyclopedia of genes and genomes
KO	KEGG orthology
L	Levorotatory
LB	Luria bertani
M	Molarity
Mbp	Mega base pair

mer	Repeat unit length
mins	Minutes
ml	Mili-liter
Mo	Molybdenum
Mo-Fe	Molybdenum iron protein
Mr	Relative molecular mass
NCBI	National Center for Biotechnology Information
ncRNA	Non-coding RNA
NEB	New england biolabs
NGS	Next generation sequencing
NRPs	Non-ribosomal peptide synthetase
OD	Optical density
OD_{230mm}	Optical density at 230mm
OD_{260mm}	Optical density at 260mm
OD_{280mm}	Optical density at 280mm
OEM	Original equiment manufacturer
OLC	Overlap layout/consensus
oxo	Oxo alcohols
PacBio	Pacific bioscience
PCR	Polymerase chain reaction
PEG	Polyethylene glycol
PGPB	Plant growth promoting bacteria
PHAST	Phage search Tool
PKSs	Polyketide synthetase
Pmol	Pico molar
Psi	Pound per square inch
G+C	Guanine + cytosine
Rpm	Rotation per minute
rRNA	Ribosomal ribonucleic acid
RSAT	Rapid annotation using subsystem technology
s^{70}	Sigma factor 70
SDS	Sodium dodecyl sulphate
sec	Second
sff	Standard flowgram format
SSR	Simple sequence repeat
Ta	Temperature of annealing
TAE	Tris acetic EDTA
TCA	Tricarboxylic acid cycle
TEN	Tris edta sodium chloride buffer
Tm	Temperature of melting point
tRNA	Tranfer ribonucleic acid
TSS	Transcriptional start site
UV	Ultra-violet

V	Voltage
w/v	Weight per volume
WGS	Whole genome sequence
x	Times
α	Alpha
β	Beta
δ	Gamma
Δ	Delta

ANALISIS EVOLUSI DAN METABOLIK MENYELURUH GENOM *PAENIBACILLUS POLYMYXA* ATCC 15970

ABSTRAK

Paenibacillus polymyxa ATCC 15970 merupakan bakteria Gram positif hidup bebas dan berkupayaan untuk mengikat nitrogen. Bakteria ini menggalakkan pertumbuhan pokok, penghasilan sebatian antimicrobial dan perembesan enzim hidrolisis. Teknologi penjujukan generasi terkini telah digunakan untuk penjujukan genom bakteria : Roche 454 FLX, Hiseq2000 Illumina Solexa dan Pacbio RS II CLR. Draf genome telah dihimpun dalam tige bersaiz jumlah 6,202,583 bp dengan purata kandungan GC sebanyak 45.6 %. Anotasi genom meramalkan kewujudan 5342 gen pengekodan protein. Genom ini telah mendedahkan satu operon pengikat nitrogen yang disusun sebagai *nifBHDKENXV* dan *hesA*. Analisis pembinaan semula metabolic menunjukkan *Paenibacillus polymyxa* ATCC 15970 mempunyai gen-gen untuk biosintesis lanthionine (metabolit sekunder), hormon pertumbuhan pokok, butanol dan metabolism glikogen. Penyelidikan genom bandingan menunujukkan persamaan yang tinggi dengan genom *Paenibacillus polymyxa* (strains *E681*, *SC2*, *M1*, CR1, ATCC12321), *Paenibacillus durus*, *Paenibacillus macerans* and *Paenibacillus terrae HPL-003*. Spesies ini meramalkan sebanyak 7 bacteria virus dan 6 CRISPR lokus dengan saiz antara 23 hingga 45 bp. Ia juga mempunyai 18 GIs diramalkan oleh Sigi-HMM. Analisis promoter mencadangkan pengikatan nitrogen boleh dikawal selia oleh promoter yang ditemui di hulu *nifB* dan *nifH*. Analisis hubungan filogenetik NifHDK menunjukkan *Paenibacillus* dan *Frankia* dikelompokkan didalam kelompok monofiletik dan analisis philogenom menyarankan pengikatan nitrogen telah hadir dalam *Paenibacillus durus*.

GENOME-WIDE METABLIC AND EVOLUTIONARY ANALYSIS OF *PAENIBACILLUS POLYMYXA ATCC 15970*

ABSTRACT

Paenibacillus polymyxa ATCC 15970 is a Gram positive free-living nitrogen fixing bacterium known for its ability to promote plant growth, production of antimicrobial compounds and secretion of hydrolytic enzymes. Genome sequencing of this bacterium was performed using next generation sequencing technology: Roche 454 FLX, Hiseq2000 Illumina Solexa and Pacbio RS II CLR. The draft genome assembled into 3 contig with total bases of 6,202,583 bp with a mean GC content of 45.6%. Genome annotation predicted a total of 5342 protein-coding genes. The genome revealed a single nitrogen fixation operon arranged as *nifBHDKENXV* and *hesA*. A metabolic reconstruction analysis indicated that *P. polymyxa* ATCC 15970 has the genes for the biosynthesis of *lanthionine* (a secondary metabolite), plant growth hormones, butanol and glycogen metabolism. A comparative whole genome investigation demonstrated high similarity among the genomes of *P. polymyxa* (strains *E681*, *SC2*, *M1*, CR1, ATCC12321), *P. durus*, *P. macerans* and *P. terrae HPL-003*. The analysis was result a questionable phage out of 7 predicted regions. This species possess also 6 CRISPR loci with size ranging from 23 to 45 bp. It also possessed 18 GIs predicted by SIGI-HMM. The promoter analysis suggested the nitrogen fixing process could be regulated by the promoter found at upstream of *nifB* and *nifH*. The phylogenetic relationship analysis of concatenated *nifHDK* revealed *Paenibacillus* and *Frankia* were clustered into monophyletic group. In addition, phylogenomic analysis suggested nitrogen fixing had been present in *P. durus* ATCC 35767.

CHAPTER 1

INTRODUCTION

1.1 General Introduction

Land for agricultural purposes is greatly reduced due to the blooming human population and the subsequent development of residential areas, exploitation of land for industrial activities, deforestation as well as pollutions. Nitrogenous and phosphorus sources are usually the limiting factor in agricultural activities. Hence, an enormous effort is needed to solve the continuous depletion of nitrogen sources before it becomes a significant issue (Matson *et al.*, 2002, Matson *et al.*, 1997).

Nitrogen is the most abundant element on earth and it accounts for 78% of the earth's atmosphere. For human, it is an essential element which is primarily derived from plant and animal dietary. Amino acids are the building blocks of proteins and they have nitrogen in their components. An increased proportion of nitrogen in human nutrition was come from ammonia that was fixed through industrial production. However, approximately half of the 23 million metric tons of nitrogen sources come from biological nitrogen fixation by bacteria was consumed as human food (Halbleib and Ludden, 2000). The nitrate form of nitrogen is usually found in soil and this is the kind usually acquired by plant.

Nitrogen fixing organisms are widely distributed across bacterial and archaeal phylum. These organisms were found rely solely on the biological nitrogenase enzyme system in which the formation of molecular hydrogen is accompanied by the production of ammonia. This nitrogenase enzyme complex is made up of Fe Protein and MoFe protein components. Nitrogen fixing microbes are also known as

diazotrophs and their diversity includes various species of cyanobacteria, green sulfur bacteria, azotobacteraceae, rhizobia and frankia (Dixon *et al.*, 1977).

The most of the biological nitrogen fixation study is conducted on Gram negative nitrogen fixation bacteria which are *Klebsiella pneumoniae* and *Azotobacter vinelandii*. Various studies on gene characterization and regulations related to these bacteria were performed (Schmitz *et al.*, 2002, Streicher *et al.*, 1974, MacNeil *et al.*, 1981, Hamilton *et al.*, 2011). Besides that, there are a few research on Gram positive diazotrophs such as *Clostridium pasteurianum* (Wang *et al.*, 1988) and *Frankia sp* (Harriott *et al.*, 1995).

Since 2005, there has been a remarkable emergence of parallel DNA sequencing process in the form of next generation sequencing (NGS) technologies. The cost-per-base offered by the NGS technologies has dramatically reduced. Thus, this enables more cost-effective options for many experimental approaches. It has fundamentally changed high-throughput genomic research and opened up many novel applications which allows experiments that were previously not technically feasible or not affordable to be carried out (Jacquier, 2009).

Bacteria of the genus *Paenibacillus* belong to the *Firmicutes* phylum. In this study, the genome of *Paenibacillus polymyxa* ATCC 15970 was sequenced using NGS technologies and the resultant data were used for a comparative genomic analysis to obtain an evolutionary perspective of the nitrogen fixation system within the *Paenibacillus* species. To date, there are only 4 strains of *P. polymyxa* with complete genomes available: *P. polymyxa* E681, *P. polymyxa* M1, *P. polymyxa* SC2 and *P. polymyxa* CR1. There are also several draft genomes of *Paenibacillus* species which includes the Type strain *P. polymyxa* ATCC842. However, only *P. terrae* HPL-003, *P. sp*. WYP 78, and *P. aloe* strain 11 are nitrogen fixing bacteria and are

available for comparative analysis. Therefore, this *Paenibacillus* study will be contributed towards a better understanding of the evolutionary inference of the Gram positive *Paenibacillus* species in terms of their unique nitrogen fixation system.

1.2 Problem Statement

Nitrogen fixation in Gram positive bacteria is less characterized.

1.3 Research Objective

1) To analyze the *Paenibacillus polymyxa* ATCC 15970 genome and use it as a database.
2) To construct the genome organization and composition of *Paenibacillus polymyxa* ATCC 15970.
3) To elucidate the evolution of nitrogen fixing genes of *Paenibacillus polymyxa* ATCC 15970 using comparative genomic analysis.

CHAPTER 2

LITERATURE REVIEW

2.1 Nitrogen fixation

Biological nitrogen fixation activity is a process in which atmospheric nitrogen is reduced to ammonia. The process plays an important role in agriculture as well as to the global nitrogen cycle. The reduction process is performed by the enzyme nitrogenase in a group of prokaryotes known as diazotrophs. These nitrogen-fixing prokaryotes occur widely across both the bacterial and archaeal domains. These include families of *Firmicutes, Proteobacteria, Cyanobacteria, Chlorobi* and *Actinobacteria.* The organizations of nitrogen fixation operons and the related genes are significantly different among the nitrogen-fixing organisms from different phylum or family (Falkowski, 1997, Dos Santos *et al.*, 2012). A wide range of habitat exists for free living nitrogen fixing organisms such as *Klebsiella pneumoniae* (Arnold *et al.*, 1988, Streicher *et al.*, 1974), *Azotobacter vinelandii* (Hamilton *et al.*, 2011), and *Paenibacillus spp.* while some such as *Ensifer meliloti* and *Frankia sp.* live in symbiotic association with a number of plants (Burns and Hardy, 1975).

The nitrogenase enzyme complex is the crucial enzyme system which determines the ability to fix atmospheric nitrogen (Simpson and Burris, 1984). Biological nitrogen fixation activity is catalyzed mostly by the molybdenum-based nitrogenase enzymes. However, alternative forms of the enzyme are found in various bacteria such as *Azotobacter vinelandii*. The nitrogenase enzymes that are present among these microorganisms can be characterized into five different groups: (1) The Mo-Fe based nitrogenase enzyme which is usually found in *proteobacterial* phylum and *cyanobacterial* phylum; (2) The anaerobic Mo-Fe based nitrogenase which is

found to occur in a wide range of anaerobic organisms including *clostridia*, acetogenic bacteria and several of methanogens bacteria; (3) The alternative nitrogenase which included the Mo-independent *anf* and *vnf* genes; (4) The uncharacterized *nif* homologous group which is detected only in methanogens bacteria and some of the anoxygenics photosynthetic bacteria; and (5) The bacterio-chlorophyll and chlorophyll biosynthesis group of genes found common to all phototrophs (Raymond *et al.*, 2004).

The capacity to perform nitrogen fixation activity in these organisms relies solely on the nitrogenase system, the most metabolically expensive process in biology that utilizes 16 ATPs per nitrogen fixed. There are three stages of electron transfer involved in the nitrogenase substrate reduction process (Kim and Rees, 1994): Firstly, the Fe-protein is reduced by electron carriers of flavodoxin or ferredoxin. Subsequently, the single electron is transferred from the Fe-protein to the MoFe-protein and this process is dependent on MgATP hydrolysis. Finally, the electron is transferred to the substrate. This process is summarized in Figure 2.1.

The stoichiometry of nitrogenase enzyme complex nitrogen fixation process is shown in the following equation (Rees and Howard, 2000).

$$N_2 + (6+2n)H^+ + (6+2n)e^- + p(6+2n)ATP > 2NH_3 + nH_2 + p(6+2n)ADP + p(6+2n)P_i$$

In the standard model proposed by Simpson and Burris (1984), a single molecule of dihydrogen is coupled by the reduction of one molecule of dinitrogen with an average of hydrolyzed two ATP molecules per electron transferred, so that n=1 and p=2. Under a typical experimental condition, the stoichiometric equation of the dinitrogen reduction process is found to utilize more than theoretical amount of

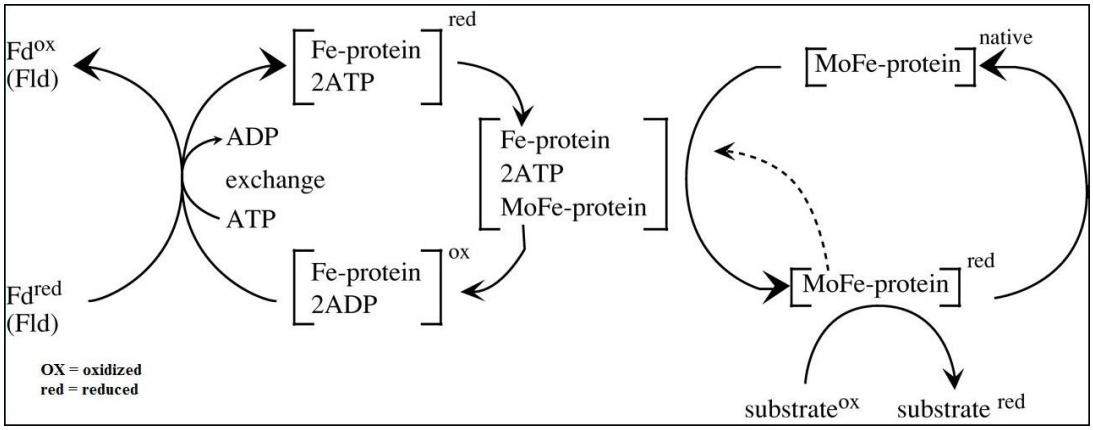

Figure 2.1 A schematic figure of the flow of electrons in a nitrogenase complex.

The figure illustrates the flow of electron of nitrogen fixation process from ferrodoxin (Fd) or flavodoxin (Fld) electron carriers to the Fe-protein. Then, the transfer of electron from the Fe-protein to the MoFe-Protein is coupled by the hydrolysis of ATP which subsequent by the reduction of substrates that is coupled with a return of the MoFe-protein to the resting redox state (Adapted from Burgess and Lowe, 1996).

ATP. The ATP hydrolysis is appeared uncoupled from the expected electron transfer system giving n>1 and p>2 (Rees *et al.*, 2005).

2.2 Property of nitrogenase enzyme

Based on their cofactors, there are three types of nitrogenase systems that has been identified and characterized in nitrogen fixing bacteria. The molybdenum-dependent nitrogenase (Mo-nitrogenase) is the most well characterized nitrogenase enzyme in which the sequence and structure are highly conserved throughout the vast diverse nitrogen-fixing bacteria. This Mo-nitrogenase consists of two metalloproteins. The molybdenum-iron protein (FeMo-protein) protein contains the site that possesses the function of substrate reduction and it is sometimes termed as dinitrogenase. Meanwhile, the iron protein (Fe-protein) functions as an obligate electron donor to the FeMo-protein and is also known as nitrogenase reductase. The nitrogen fixation regulon is encoded dinitrogenase and nitrogenase reductase and also electron transfer protein, metal cluster synthesis and regulation as accessory genes (Peters *et al.*, 1995).

Extensive studies on the molybdenum nitrogenase revealed that there are alternative forms of nitrogenase. The molybdenum base nitrogenase is not the only essential element for the functionality of the nitrogenase enzyme complex. Iron and other related proteins are also required to stabilize the enzyme complex. The alternative forms of nitrogenase which lacked molybdenum were first discovered in *Azotobacter vinelandii* (Bishop *et al.*, 1980) and the evidence for their existence had been subsequently reported in variety of other organisms (Robson *et al.*, 1986, Schneider *et al.*, 1991, Lehman and Roberts, 1991).

2.2.1 Fe-protein of Mo-nitrogenase

The Fe-protein of nitrogenase enzyme is encoded by *nifH*. The protein forms a dimer with two identical α subunits. Each subunit is folded with a single α-helix-β sheet type domain and covalently connected to [Fe_4S_4] cluster at one of the dimer end (Georgiadis *et al.*, 1992). Moreover, each of the single [Fe_4S_4] cluster is symmetrically coordinated to each subunit by the Cys97 and Cys132 (Kim and Rees, 1994). This [Fe_4S_4] cluster is the redox active center that is directly involved in the FeMo-protein electron transfer process. The Fe-protein of [Fe_4S_4] cluster is cycles between the reduced (+1) state and the oxidized (+2) states during transferring the electron to FeMo-protein. Each of the Fe-protein dimer binding two nucleotide molecules of electrons at the site that is distal from the [Fe_4S_4] active site (Burgess and Lowe, 1996). The binding of MgATP compound caused the Fe-protein conformational changed. These two subunits rotate toward each other with the extruding of the [Fe_4S_4] cluster toward the protein surface by a distance of 4 Å. This in turn conclude the interaction with the P-cluster of FeMo-protein (Schindelin *et al.*, 1997). This conformation change is thought to be the key step in the catalytic cycle of nitrogenase.

2.2.2 FeMo-protein of Mo-nitrogenase

The FeMo-protein is a $\alpha_2\beta_2$ tetramer with 230 kDa molecular mass. These α and β subunits are coded by the genes *nifD* and *nifK*, respectively. Each FeMo-protein consists of two metalloclusters pairs, which are molybdenum-iron-sulfer-homocitrate clusters (FeMo-co) and [Fe_8S_7] clusters (P-cluster). The FeMo-cofactor is composed of two partial cubanes [Fe4S3] and [MoFe3S3] as well. They are bridged by three sulfides with the homocitrate moiety that is coordinated to the

molybdenum atom (Kim and Rees, 1994). The FeMo-cofactor moiety is completely encompassed by a 3 α-subunit domain and it is assumed to be the substrate reduction site (Howard and Rees, 1996). The P-cluster ($[Fe_8S_7]$ clusters) consists of a redox dependent structure that is made up of two iron-sulphur partial cubanes (Peters *et al.*, 1997). The P-cluster paired are bridged by two cysteine thiol ligands, Cys α88 and Cys β95 and a disulfide bond from the sulfur cluster (Kim and Rees, 1994). The P-cluster is located at the interface of α and β subunit. This finding suggested it as the intermediates of electron transport pathway between Fe-protein and FeMo-cofactor (Howard and Rees, 1996).

2.2.3 Alternative nitrogenases

Several diazotrophs such as *Azotobacter vinelandi* contain alternative nitrogenase systems in which the molybdenum is replaced by either iron or vanadium (Bishop *et al.,* 1980). However, these alternative nitrogenase systems are found in a limited number of diazotrophs and they are present as secondary nitrogenase subsystems. These alternative nitrogenase enzymes are expressed only when the molybdenum concentration is found to be limited (Joerger and Bishop, 1988).

The 'Vanadium' nitrogenase protein consists of single α and β subunit that are homologous to 'MoFe' nitrogenase protein, plus an additional δ subunit with a molecular mass of 14 kDa. Together, it is forming a hexameric structure, $\alpha_2\beta_2\delta_2$. The MoFe protein is found to be more efficient and specific in reducing dinitrogen to ammonia when compared to the alternative nitrogenase enzyme system (Miller and Eady, 1988). Furthermore, the Mo-dependent nitrogenase is found to be capable of reducing acetylene to ethylene but the *anf* and *vnf* encoded nitrogenase system are

found to reduce acetylene to a mixture of ethylene and ethane, unlike the Mo-dependent nitrogenase (Dilworth *et al.*, 1988, Scott *et al.*, 1990).

The 'Iron' nitrogenase (Nitrogenase-3) consists of neither Molybdenum nor Vanadium as the heterometal cofactor atom and instead the cofactor is replaced by an iron atom (Muller *et al.*, 1992). In *Azotobacter vinelandii*, the iron nitrogenase protein is coded by a single operon *anfHDGKOR* (Joerger *et al.*, 1989, Mylona *et al.*, 1996). The α and β subunit of dinitrogenase are coded by *anfD* and *anfK*, whereas *anfH* coded for dinitrogenase reductase. The anf system (iron-iron nitrogenase) is unusual as there are no additional *nifENX* or *vnfENX* homologs genes in *A. vinelandii*. Base on genomic analysis of *A. vinelandii*, the iron nitrogenase maturation was based on *vnfEN* gene *(Wolfinger and Bishop, 1991, Hamilton et al., 2011)*. The **Table 2.1** shows a summary of the nitrogen fixation genes and their functions.

Table 2.1 Nitrogen fixation genes products and functions.

Gene	Function	Citation
NifQ	Incorporation of molybdenum into nitrogenase.	(Imperial *et al.*, 1984)
NifB	FeMo-cofactor synthesis.	(Martínez-Noël *et al.*, 2011, Dixon *et al.*, 1977, Harriott *et al.*, 1995)
NifA	Positive regulation.	(Buchanan-Wollaston *et al.*, 1981, Dixon *et al.*, 1977)
NifL	Negative regulation.	(Milenkov *et al.*, 2011, Schmitz *et al.*, 2002)
NifF	Flavodoxin (electron carrier).	(Dixon *et al.*, 1977, Hill and Kavanagh, 1980)
NifM	Nitrogenase reductase processing.	(Paul and Merrick, 1989)
NifZ	Maturation and activation of FeMo-protein.	(Harriott *et al.*, 1995, Paul and Merrick, 1989, Stricker *et al.*, 1997)
NifW	Maturation, activation and protect FeMo-protein from oxygen.	(Harriott *et al.*, 1995, Paul and Merrick, 1989)
NifV	FeMo-cofector synthesis (homocitrate synthase or homoaconitate synthase).	(Stricker *et al.*, 1997, Wang *et al.*, 2013)
NifS	S activation in metallocluster synthesis.	(Zheng *et al.*, 1993)
NifU	FeMo-protein processing.	(Fu *et al.*, 1994)
NifX	FeMo cofactor synthesis and negative regulation.	(Gosink *et al.*, 1990, Harriott *et al.*, 1995)
NifE	Synthesis and insertion of FeMo cofactor into dinitrogenase protein.	(Dean and Brigle, 1985, Fani *et al.*, 2000, Martínez-Noël *et al.*, 2011)
NifY	Processing of MoFe protein.	(Homer *et al.*, 1993)
NifT	Nitrogenase maturation.	(Stricker *et al.*, 1997)
NifK	FeMo protein β subunit.	(Dixon *et al.*, 1977, Fani *et al.*, 2000, Schneider *et al.*, 1991)
NifD	FeMo protein α subunit.	(Dean and Brigle, 1985, Fani *et al.*, 2000)
NifH	FeMo protein subunit.	(Dixon *et al.*, 1977, Schneider *et al.*, 1991)
NifJ	Electron transfer (pyruvate flavodoxin oxidoreductase).	(Hill and Kavanagh, 1980)

2.3 *Paenibacillus*

2.3.1 *Paenibacillus* species

The *Paenibacillus* species is an endospore-forming Gram-positive soil dwelling bacterium and plant growth promoting bacteria (PGPB) that is agriculturally beneficial. They have a variety of benefit on plants especially promoting plant growth and suppression of diseases caused by pathogens. The enhancement of growth could be due to its nitrogen fixation activity. Most of the *P. polymyxa* strains are free-living nitrogen fixing bacteria (Grau and Wilson, 1962, Seldin *et al.*, 1983) and they are well known for their ability to promote plant growth activity by root-association (Holl *et al.*, 1988). Their production of antimicrobial compounds (polymyxin and polyketide) and secretion of a wide range of hydrolytic enzymes make these *P. polymyxa* strains important for the pharmaceutical and food industries (Ehrlich, Priest, 1993). In terms of its ecological importance as a plant growth probiotic rhizobacterium, *P. polymyxa* is known for producing two types of antimicrobial peptide (Beatty and Jensen, 2002) as well as synthesizing plant growth hormones such as auxin (Lebuhn *et al.*, 1997) and cytokinin (Timmusk *et al.*, 1999).

Paenibacillus was actually first isolated and characterized by Prazmowski but it was named *Bacillus* (Levine, 1975) until it was further reclassified based on 16S rRNA sequences taxonomy classification (Ash *et al.*, 1993). The *Paenibacillus* genus is a member of family '*Paenibacillaceae*' and nitrogen fixers such as *Paenibacillus durus* (formerly *P. azotofixans*), *P. polymyxa* and *P. macerans* are the oldest species that were formerly labeled as *Bacillus*.

An earlier study showed that the strain *P. polymyxa* ATCC 15970 possesses a single copy of the *nifH, nifD* and *nifK* homologues which code for the nitrogenase enzyme subunits (Yam, 2007). It was suggested that the expression of the nitrogen

fixing genes in *P. polymyxa* could be possibly controlled by a unique set of regulation system based on distinguishable promoter sequences for *nifB* and *nifH* found in this *Paenibacillus* species.

To date, the publicly available database of the complete *Paenibacillus polymyxa* genomes consist of 5 strains: *P. polymyxa* CR1, *P. polymyxa* E681, *P. polymyxa* M1, *P. polymyxa* SC2 and the Type strain *P. polymyxa* ATCC 842. However, these *P. polymyxa* strains are non-nitrogen fixing except *P. polymyxa* CR1. There is also a draft genome sequence of the strain *P. polymyxa* ATCC 35971 which is also capable of fixing nitrogen. This strain was found to possess high potential in the industrial, medical and agricultural application (Raza *et al.*, 2008a). All the complete or draft genome sequences are publicly available in Genbank (ftp://ftp.ncbi.nlm.nih.gov/genomes/).

2.3.2 Applications of *Paenibacillus* species

The ability of *Paenibacillus polymyxa* to survive in the phyllosphere along with its antimicrobial properties makes it suitable to be a biofertilizer (Xu *et al.*, 2014). Excessive usage of chemical fertilizers, increased agricultural costs, increasing environmental problems as well as food safety are current concerns in modern agriculture (Godfray *et al.*, 2010, Carvalho, 2006). These issues support the need for practical applications of beneficial bacteria such as *P. polymyxa*. For example, excessive usage of chemical fertilizers results in a large amount of nitrate accumulating in the soil. The former later leaches into water bodies and causes contamination of the surface and ground water (Dinnes *et al.*, 2002, Ju *et al.*, 2004, Hooda *et al.*, 2000). Therefore, utilization of plant growth promoting bacteria (PGPB) as biofertilizers emerged as an attractive environmentally safe alternative solution to

chemical fertilizers (Singh *et al.*, 2011). This could improve the quality of sustainable agro ecosystems and also increase crop yield (Lucero *et al.*, 2014, D'Hose *et al.*, 2014).

The application of biofertilizers reduces the need of introducing more biotic and abiotic substances into the soil environment. Consequently, biofertilizers increase crop growth and yield without raising any pollution. In addition, the use of biofertilizers could lower the production cost by reducing expensive chemical inputs such as pesticides and organic and mineral fertilizers (Xu *et al.*, 2014).

Paenibacillus polymyxa also produces antibiotics and the main type of peptide antibiotic produced is *polymyxin,* a colistin peptide from the circulin family. Other strains produce different peptides including *polypeptins*, *jolipeptin*, *satavalin*, *gavaserin*, *saltavidin*, *fusaricidins* and *polyxin*. There are also strains of *P. polymyxa* that also produced antimicrobial and antifungal compounds whose function remains undefined (Liu *et al.*, 2011, Beatty and Jensen, 2002, Mageshwaran *et al.*, 2011). For example, an unknown antibiotic with a molecular weight less than 3,500 Da was found active against most of the Gram-positive bacteria. However, this antibiotic was found to lack antimicrobial activity against *Pseudomonas aeruginosa*. Then, an unknown antimicrobial peptide (2,983 Da) was shown to possess a high degree of post-translational modification active against Gram-positive spoilage and food borne bacteria (Rosado and Seldin, 1993).

Interestingly, two compounds (Mr 1,184 and 1,202 Da) with a cyclic heptapeptide moiety attached to a tripeptide side chain and to the fatty acyl residue was found to exhibit more hydrophobic activity compared to than *polymyxin* B. The 1,184 Da peptide contained a 2,3-didehydrobutyrine residue with a size of Mr 101

Da that replaced the threonine at position 2 of the polymyxin side chain. Extensive characterizations of these two peptides confirmed their broad range antimicrobial activity against Gram-positive bacteria. These chemical modifications resulted in a wider range of antagonistic activity compared to *polymyxin* B (Selim *et al.*, 2005).

The molecular and genetic aspects of antibiotic and non-ribosomal peptide synthesis by the *Paenibacillus* species are still completely unknown. Non-ribosomal peptide synthesis are independent of mRNA and ribosomes (Schwarzer *et al.*, 2003). According to Martin *et al.* (2003), a mutagenesis study of the antibiotic biosynthesis by *Paenibacillus kobensis* M revealed a set of genes involved in non-ribosomal peptide synthesis. A majority of the mutants show a decreased in production but with no complete loss of the gene functions. This is not consistent with another study which involved mutagenesis using *Tn917*. Complete loss of antibiotic production was observed when the transposon was inserted into the *iturin* and *fengycin* biosynthesis genes (Chen *et al.*, 1995, Tsuge *et al.*, 2001). The antibiotic biosynthesis genes of *P. polymyxa* showed low homology toward the sequences in the NCBI database indicative of how diverge the system is in this bacterium.

According to Seldin *et al.* (1999), four of *P. polymyxa* strain SCE2 with plasmid pTV32 (Ts) mutant lost it capability to inhibit *Micrococcus* sp. and *Staphylococcus aureus* RN450. However, it was found out that this mutant continued to inhibit the growth of *Corynebacterium fimi* NCTC7547 and *Escherichia coli* HB101. This suggested that there is more than one antimicrobial substance that is produced by this strain (Seldin *et al.*, 1999).

According to Li *et al.* (2007), the first two modules of putative fusaricidin synthetase (*fusA)* of DNA fragment was isolated from *P. polymyxa* PKB1. The mutated *fusA* sequenced resulted in the complete loss of antifungal activity against

Leptosphaeria maculans. This mutation suggested the *fusA* gene played an important role in the non-ribosomal peptide synthesis of fusaricidin (Li *et al.*, 2007).

Various *P. polymyxa* strains could produce different types of antimicrobial compounds reflecting their genetic diversity (Beatty and Jensen, 2002, Lal and Tabacchioni, 2009, Raza *et al.*, 2008b). It is arduous to genetically modify and manipulate the non-ribosomal peptide synthetase (NRPSs) and polyketide synthetases (PKSs) to possess new properties provided there is a powerful tool for development (Weber *et al.*, 2015). However, with an increase in knowledge on the regulation of their genes and an advancement in technology, improvement on their efficacy and host range can be made possible (Dixon, 2001, Payne *et al.*, 2007, Compant *et al.*, 2005). Thus, further research is required to fully understand antibiotics biosynthetic assembly and discover new secondary metabolites and (Goto *et al.*, 2010).

Various *P. polymyxa* strains potentially play important roles in industry by producing a variety of hydrolytic enzymes (Dijksterhuis *et al.*, 1999). For example, the strain *P. polymyxa* 72 produced an amylase of 48 kDA that comprised 1,161 amino acids (Uozumi et al. (1989). *P. polymyxa* could also produce proteases (30 kDa) that are thermo stable. Depending on the strain, this thermo stable protease was constitutively produced or partially inducible in nature. The production was either strongly influenced by the presence metabolisable sugars or the carbon over nitrogen ratio (Alvarez *et al.*, 2006). For example, the thermo stable protease produced by *P. polymyxa* B-17 was found to be optimum at 50 °C and it was able to retain a significant enzymatic activity at 70 °C. This enzyme was found to be active in a wide range of pH ranging from 5.5 to 10.0 as well. However, this thermo stable protease

could be inhibited by metal chelating agents such as zink and EDTA (Matta and Punj, 1998).

2.4 Genome Assembly

2.4.1 Algorithm (String graph, De-bruijn graph, Overlap by consensus)

The next generation graph-based sequencing assembler algorithms can be organized into three categories, namely, Overlap Layout/Consenses (OLC) (Sutton and Dew, 2006), de Bruijn Graph (DBG) (Pop, 2009, Zerbino and Birney, 2008), and Greedy Graphs (Zhang *et al.*, 2000). A graph is a term that is used widely in computer science or mathematic. This graph contains a set of nodes with a set of edges at each modes. The nodes and edges of graph can also be referred as vertices and arcs, respectively. The graph is known as a direct graph when the edges of the graph traversed in one direction only. The graph can also be conceptualized as balls in space with arrows connecting each of the edges. There is also a special kind of path which is known as a simple path. This simple path contains only distinct nodes in the graph. An overlap graph means the sequencing reads data is overlapping and it must be pre-computed by a series of pair-wise alignments to identify the true overlaps. Hence, the mode of the graph representing the sequencing reads and the edges is representing the overlaps of reads (Miller *et al.*, 2010).

The whole genome sequencing data actually possess problem that is raised from overlaps graphs and K-mer graphs during the assembly process. This is because sequencing error of reads contribute the dead end divergences and short spurs of the graph or path while the path solving the overlaps of each reads. This is also induced by coverage decline in the data coverage as shown in **Figure 2.2a**. Second, the presence of bubbles (**Figure 2.2b**) in the assembly graph can cause a path that

diverges and then converges. This bubble problem is contributed by sequencing error that is present in the middle of read or caused by polymorphism (Fasulo *et al.*, 2002). The third problem is the path that converges and then diverges (as shown in **Figure 2.2c**). This problem is induced by repeats region that is always present in the target genome. Lastly, a phenomenon known as 'cycles' in which several paths converges on themselves. This problem is induced by short tandem repeats or other type of repeats in the target genome sequence (Miller *et al.*, 2010, Treangen and Salzberg, 2012).

In general, branching and convergence will lead to an increasing complexity of graphs generated (Nagarajan and Pop, 2009). This complexity is difficult to be resolved by short reads length or limited reads distance information. However, most of the complexity occurring during graph construction is caused by repeats in the genome sequence itself. Sequencing error originating from either the machine or during the library preparation can also cause this complexity. The software Assembler relies on two algorithms, which are heuristic algorithms and approximation algorithms to resolve the path redundancy or repair errors in order to reduce complexity of the graph. Hence, a simple path can be enlarged while simplifying the graph (Miller *et al.*, 2010).

One approach known as the greedy algorithm applies a basic operation in which any read or contig or add is added in for consensus building. This basic 'find and extend' operation function is repeated until the read or contig cannot be extended. This basic operation utilizes the next highest-scoring overlap scheme in its process to find and joint the read or contig. This scoring scheme functions by measuring the number of matching bases in the overlaps of reads or contigs. Then, the contig is

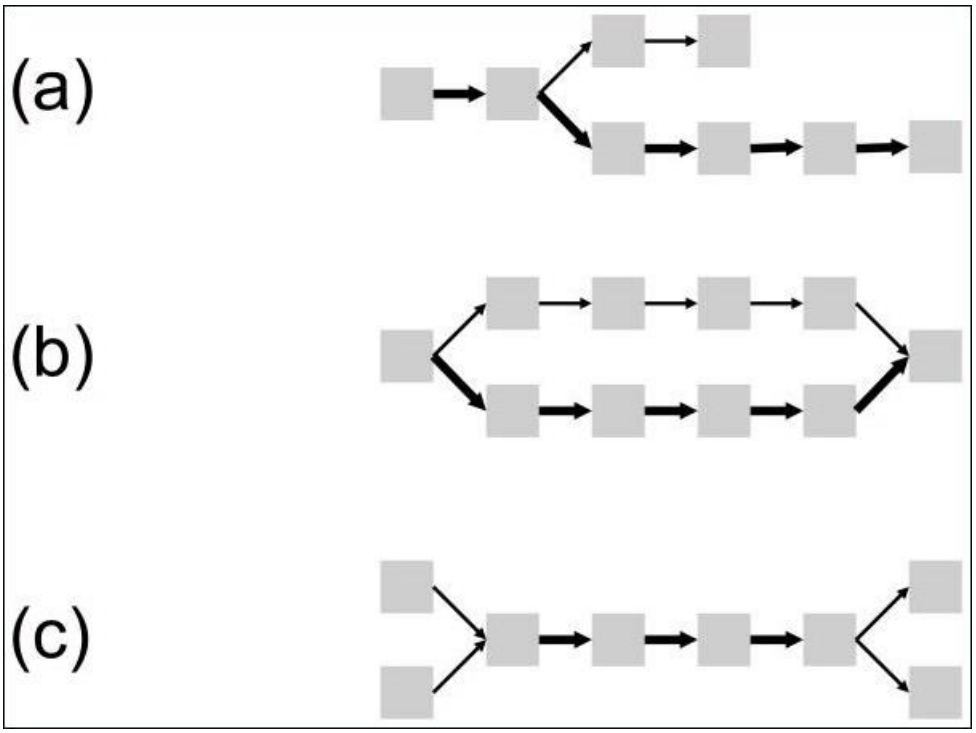

Figure 2.2 Complexity in K-mer graphs can be diagnosed with read multiplicity information (Adapted from Miller *et al.*, 2010).

extended and grown in a greedy manner by taking next highest-scoring overlap read into consideration. However, the limitation of this greedy algorithm is that it may get stuck especially at the local maxima. The pause occurs while it is processing the contig or read that is usable to extend other read or contig into a larger consensus. In other words, this greedy algorithm requires mechanisms to avoid the incorporation of false-positive overlaps of read or contig during contigs extension process. Moreover, the repetitive sequence may induce a higher score compare to the score induced by common position. Hence, the assembly built by the greedy algorithm can be erroneously based on the false positive overlaps at either side of the repeat and produce chimera (Pop and Salzberg, 2008, Miller *et al.*, 2010).

The Overlap Layout/Consensus (OLC) algorithm has the same objective to optimize large genomes assembly. For example, the software known as Celera Assembler uses this algorithm (Myers *et al.*, 2000). Basically, the OLC assembler can be separated into three phases. The first phase is known as "Overlap Discovery" phase, which involves all against all and pair wise read alignment comparison by utilizing efficiently the seed and extend heuristic algorithm in this process. This software basically pre-computes the K-mer graph across all the sequencing reads, selects the best overlap candidates that shared K-mers path, and computes pair wise reads alignments by using the K-mers as alignment seeds. This overlap discovery phase is sensitive to K-mers size which will greatly affect the minimum overlap length of read and affect the minimum present identity that is required for an overlap of read. It is also affected by base calling error and low sequencing coverage. This suggests that larger parameters of k-mers size leads to shorter but a more accurate contig. The second phase is the phase of overlapping graph construction and manipulation. This phase functions to estimate the reads layout base on the k-mer

constructed graph. This overlapping graph may fit into a practical computational memory for large genome as it does not include sequence base calls. The third phase is called multiple sequence alignment (MSA) and it functions to determine the overall overlap layout and the precision of sequence consensus. Due to lack of an efficient method to compute optimal multiple sequence alignment, the consensus buildup utilizes a progressive pair wide alignment process that is guided by approximate read layout but with the sequence base call integrated into the graph. This sequence base call integration will result in more memory consumption during the process and might not be practical for large genome data (Miller *et al.*, 2010, Sutton and Dew, 2006, Myers *et al.*, 2000).

The software known as Newbler was developed and used by 454 Life Sciences, Inc. (USA) for its sequencing platform (Margulies *et al.*, 2005). Two OLC algorithms are implemented on Newbler at two different phases. The first phase functions to generate unitigs from 454 reads. This unitigs represent a mini-assembly that is built from reads that overlap with other unitigs. This unitigs function as a high confidence and conservative consensus or contigs that later will be used as a seed throughout the Newbler assembly pipeline. In the second phase, larger contigs generated by the unitigs are joined among the unitigs by a simple pair-wise alignment process. The unitigs may be split into prefix and suffix and then being aligned to other different contigs as well. However, the reads may become chimeric at the boundaries of repeats in which unitig splitting causes the individual reads to split leading into multiple contigs (Margulies *et al.*, 2005, Myers *et al.*, 2000).

If possible, the assembler software Newbler exploits coverage in order to repair the base calling errors generated by the 454 sequencing machine. In particular, this software utilizes the instrument metrics from the flow cell to overcome the long

repeats of homo-polymer base calling inaccuracy. The Newbler software is programmed to calculate consensus of unitig and contig in a ‘flow space’ format by utilizing signal strength information of each particular nucleotide flow cell. The normalized signal is proportionally corrected to the number of repeat of that particular position of the nucleotide reads. The consensus is calculated based on ‘base space’ which is equivalent to the average signal and this sacrifice the base calling precision. The Newbler software also uses the average signal by rounding up from each of the MSA column followed by calculating the consensus from it (Miller *et al.*, 2010).

The OLC algorithm is applied to short reads from Solexa and SOLiD platforms. The algorithm of the Edena assembler software functions to discard duplicate reads during the assembly and this enable the algorithm to finds all perfect and error-free overlaps (Hernandez *et al.*, 2008). This software applied the transitive overlap reduction algorithm to remove individually redundant overlaps with pairs of other overlaps. Moreover, Edena software also prunes to spurs and bubbles assembly graph problem (Myers, 1995, Hernandez *et al.*, 2014).

The De Bruijn graph relies on the K-mer function which makes it suitable for large number of short reads data assembly. The advantage of K-mer graphs is that it all “mer” overlaps are not required for the “overlap discovery”. Individual reads or their overlaps are not required for storage and redundant sequences are compressed. The great disadvantage of K-mer graph is that it exhausts a lot of memory as the actual read sequence is loaded in the memory during the K-mer graph construction. The k-mer graph construction utilizes a constant hash table for heuristic search for the existence of each K-mer data in order to speed up the overall heuristic search process. Even though the hash length table consumes a lot of computational memory,

this K-mer graph stores the occurrence of each read once even if a single read possesses multiple and different k-mer values in the graph (Pevzner *et al.*, 2001, Miller *et al.*, 2010).

There are three main factors that complicate the application of K-mer graphs in genome assembly. First is due to the fact that genomic DNA is double stranded. Any given sequencing read may come from the forward or reverse strand DNA. Therefore, the nodes and edges of the K-mer graph need to carry the information for both forward and reverse strands to prevent their assembly twice (Zerbino and Birney, 2008, Simpson *et al.*, 2009). Secondly, the real genome possesses complex and repeated structures. Any repeat with a size larger than the k-mer size will complicate the assembly process. The constructed K-mer paths tend to converge based on the repeats length and then diverged. Thus, K-mer graph had to separate the converging paths in order to achieve a successful assembly. The graph that lacks information from the reads will be difficult to be used to resolve the repeat problem. To resolve these complex regions, more information from the sequence reads is needed to resolve these complex regions such as the utilization of the mate pair distance information (Miller *et al.*, 2010). Thirdly, a palindromic DNA sequence with its reverse complement on its own will induce a K-mer path that will give a fold back pattern of itself. To avoid this palindromic complexity, the K-mer size used is odd numbered as an odd number cannot have a match with it reverse complement (Zerbino and Birney, 2008). Lastly, sequencing data may possess sequencing errors caused either by machine or the library making process. Sequencing reads can be pre-processed to remove error. The graph can be weighted by the number of reads that support the edges to find the lightly supported paths (Miller *et al.*, 2010).

Alternatively, the path can be converted into a sequence alignment to identify the correct assembly path (Miller *et al.*, 2010).

In short, most assembler softwares share common algorithms. They possess an error detection module and also correct the error bases on the read sequence and composition. Each assembler uses reads and sequence information to construct heuristic. The softwares also implement modules or functions to simplify the non-interacting path to single node path by removing error induced paths such as spurs or bubbles. They also function to resolve polymorphism-induced complexity as well. Lastly, they all aim to convert success paths to contigs, then into scaffolds, and finally into consensus (Miller *et al.*, 2010).