WHOLE GENOME SEQUENCING AND CHARACTERIZATION OF THREE NOVEL *MANGROVIMONAS*-LIKE STRAINS ISOLATED FROM MANGROVE FORESTS SEDIMENT IN PERAK, MALAYSIA

by

DINESH A/L BALACHANDRA

Thesis submitted in fulfilment of the requirements
for the degree of
Master of Science

April 2017

ACKNOWLEDGEMENT

First and foremost, I owe my highest gratitude to Dr. Lau Nyok Sean, Dr. Go Furusawa and Professor Alexander Chong Shu-Chien. They have been tirelessly helping in my research and have been always available when needed. Their hands on approach has always been very helpful for me in conducting my research. Besides that, I would also like to thank other CCB lab members such as Muhd. Khairul Luqman and Priya Gunaselvam who have assisted me in tough times.

Furthermore, I would like to thank Jess Loh Swee Cheng, Sim Pei Fang, Gavin Moh Tsu Horng and Sam Ka Kei in assisting my research by providing me lab consumables and commercial products when needed. I would also like to thank the entire CCB administration staff including Mrs. Tengku Zalina Tengku Ahmad, Mrs. Nurul Farhana Che Hassan, Mrs Nazira Zainal Abidin, Ms. Amira Mohd. Ali, Mrs Uswatun Bahirah, Ms. Nyuk Pui and Mr. Izam Fairus Kamaruddin. The administration staff have certainly made my stay in CCB a pleasant and to all my CCB labmates, thanks for the wonderful experience.

I would like to thank my family for their endless support and motivation. They have been my pillar of strength and their never ending motivation has been driving me for the past 2 years. Lastly, I would like to thank the Malaysian Ministry of Education for funding my Master’s study by providing the MyMaster (MyBrain15) scholarship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

°C	Degree Celsius
µL	Microlitre
µm	Micrometre
AL	Lysis Buffer
ANI	Average Nucleotide Identity
ATC	Tissue Lysis Buffer
ATM	Amplicon Tagment Mix
AW1	Wash Buffer 1
AW2	Wash Buffer 2
bp	Base pairs
CMC	Carboxymethylcelluose
DDBJ	DNA Databank of Japan
ddH_2	Deionized distilled water
DNA	Deoxyribonucleic acid
ds	Double stranded
EMBL	European Molecular Biology Laboratory
g	Gram
GSC	Genomics Standard Consortium
H4	Strain TPBH4
HPLC	High Performance Liquid Chromatography
INSDC	International Nucleotide Sequence Data Collaboration
JCM	Japan Collection of Microorganisms

kb	Kilo base pairs
L	Litre
L12	Strain ST2L12
L15	Strain ST2L15
LMG	Laboratory of Microbiology (Belgian Coordinated Collection of Microorganisms)
M	Molar
M. yunxiaonensis	*Mangrovimonas yunxiaonensis*
MIGS	Minimum, Information About A Genome Sequence
MIS	Microbial Identification System
mL	Mililitre
mM	Milimolar
ng	Nanogram
nm	Nanometre
NPM	Nexter PCR Master Minx
NT	Neutralize Tagment
PCR	Polymerase chain reaction
PDBG	Paired de Bruijn graph
RNA	Ribonucleic acid
rpm	Revolutions per minute
RSB	Resuspension buffer
sp.	Species
ss	Single stranded
TAE	Tris-acetate-EDTA

TD	Tagment DNA
TEM	Transmission Electron Microscope
TLC	Thin Layer Chromatography
UV	Ultraviolet

PENJUJUKAN SELURUH GENOM DAN PENCIRIAN TIGA STRAIN NOVEL SEPERTI-MANGROVIMONAS YANG DIPENCILKAN DARIPADA MENDAPAN PAYA BAKAU DI PERAK, MALAYSIA

ABSTRAK

Strain TPBH4 (= LMG 28913, = JCM 30882), ST2L12 (= LMG 28914, = JCM 30880) dan ST2L15 (= LMG 28915, = JCM 30881) telah dipencilkan daripada sedimen muara di Perak, Malaysia. TPBH4 dan ST2L12 telah dipencilkan daripada sedimen muara bakau Matang Mangrove Forest manakala ST2L15 telah dipencilkan daripada sedimen muara bakau di Taman Paya Bakau. 16S rRNA jujukan gen persamaan antara TPBH4, ST2L12, ST2L15 dan *M. yunxiaonensis* masing-masing adalah 94.9%, 95.3% dan 95.2%. Sehingga kini, genus *Mangrovimonas* terdiri daripada hanya satu spesies yang telah dicirikan dan dikenalpasti sebagai *M. yunxiaonensis*. Strain ini telah dipencilkan daripada sedimen bakau di China dan telah mempunyai draf genom diterbitkan dalam jurnal antarabangsa. Dalam kajian ini, genom keseluruhan tiga strain (TPBH4, ST2L12, ST2L15) yang berkait rapat dengan genus *Mangrovimonas* itu telah dianalisis. Genom yang dijujuk daripada tiga strain tersebut mempunyai saiz dalam lingkungan3.56-4.15 Mb yang secara relatif lebih besar daripada spesies, *M. yunxiaonensis* (2.67 Mb) yang merupakan species terdekat dari segi taksonomi bagi ketiga-tiga strain tersebut. Gen degradasi karbohidrat xylan, xylose, L-arabinan dan L-arabinose terbongkar dalam ketiga-tiga genom strain-strain tersebut. Sebagai perbandingan, gen ini tidak terdapat di dalam genom *M. yunxiaonensis*. Selain itu, strain TPBH4 dan ST2L12 mempunyai keupayaan untuk memecahkan xylan yang dipamerkan melalui kaedah plat assay kualitatif. Selain daripada itu, ketiga-tiga strain tersebut telah dikenalpasti dan dicirikan. TPBH4, ST2L12 dan ST2L15 adalah Gram-negatif, kuning / oren / merah berpigmen dan bakteria yang meluncur. pH pertumbuhan, toleransi kemasinan dan suhu pertumbuhan tiga strain tersebut telah dianalisis. Tambahan pula, quinone pernafasan dan asid lemak di dalam strain ini telah dikenal pasti. Daripada penemuan ini, kesimpulan telah tercapai bahawa ST2L12 mewakili spesies novel genus *Mangrovimonas* dalam keluarga *Flavobacteriaceae,* dengan nama *Mangrovimonas xylaniphaga* sp. nov.

WHOLE GENOME SEQUENCING AND CHARACTERIZATION OF THREE NOVEL *MANGROVIMONAS*-LIKE STRAINS ISOLATED FROM MANGROVE FORESTS SEDIMENT IN PERAK, MALAYSIA

ABSTRACT

Strains TPBH4 (=LMG 28913,=JCM 30882), ST2L12 (=LMG 28914,=JCM 30880) and ST2L15 (=LMG 28915,=JCM 30881) were isolated from estuarine sediments in Perak, Malaysia. TPBH4 and ST2L12 were isolated from the estuarine mangrove sediment of Matang Mangrove Forest whereas ST2L15 was isolated from the estuarine mangrove sediment of Taman Paya Bakau. The 16S rRNA gene sequence similarity between TPBH4, ST2L12, ST2L15 and *M. yunxiaonensis* was 94.9%, 95.3% and 95.2%, respectively. To date, the genus *Mangrovimonas* is made up of only one type species which has been characterized and identified as *M. yunxiaonensis*. This strain was isolated from mangrove sediment in China and has had its draft genome published. In this study, the whole genome sequence of these three strains which were closely related to the genus *Mangrovimonas* were analyzed. The genomes sequenced (3.56 to 4.15 Mb) were relatively larger than that of the type species, *M. yunxiaonensis* (2.67 Mb). Carbohydrate degradation genes of xylan, xylose, L-arabinan and L-arabinose were uncovered in the genomes of the three *Mangrovimonas*-like strains. In comparison, these genes were not present in the genome of *M. yunxiaonensis*. Moreover, strains TPBH4 and ST2L12 has the ability to breakdown xylan which was exhibited through a qualitative plate assay method. Furthermore, these three strains were identified and characterized. TPBH4, ST2L12 and ST2L15 were Gram-negative, yellow/orange/red pigmented, gliding bacteria. Growth pH, salinity and temperature of these three strains were analysed. Furthermore, the respiratory quinone and fatty acids present in these strains were identified. Out of these findings, it was concluded that strain ST2L12 represents a novel species of the genus *Mangrovimonas* in the family *Flavobacteriaceae*, with the name *Mangrovimonas xylaniphaga* sp. nov.

CHAPTER 1 INTRODUCTION

1.1 Background of research

Mangrove forests ecosystems can be referred as the medial interface between land and sea in the tropical regions, as they cover about 60-70% of the coastal regions of various tropical areas (Ong et al., 1995, Taketani et al., 2009). The mangrove ecosystems are known to be responsible for reducing greenhouse gas, carbon sequestration which governs the carbon cycle, preventing coastal erosion, biomass production, mitigating the risk of tsunami and giving economic benefits reaped from mangrove resources (Ashton et al., 1999, Kathiresan and Rajendran, 2005, Kathiresan, 2012). They are also an important habitat for a wealth of species of birds, fishes, crustaceans and other animals (Nagelkerken et al., 2008). Malaysia's mangrove areas cover approximately 575,000 hectares and this ranks fifth after Indonesia, Nigeria, Australia and Mexico.

The Matang Mangrove Forest in the state of Perak on the northwest coast of Peninsular Malaysia, is referred to as the best managed sustainable mangrove forest all over the world (Jusoff, 2009, Okamura et al., 2010). This mangrove forest has gained its reputation through permanent forest reserve establishment and sustainable tree harvesting activities as cutting limits and 30 to 55 years rotation systems (Jusoff and Taha, 2008). Matang Mangrove Forest is mostly made up of *Rhizophora apiculata* trees (Ong et al., 2004). Besides *R.apiculata*, the other mangrove tree species found in the area include *R. mucronata, Bruguiera parviflora, B. cylindrical* and *B. gymnorrhiza* (Goessens et al., 2014). As for Taman Paya Bakau which is the other sampling site also located in Perak, there has yet to be a microbial study conducted in this region.

The mangrove ecosystems with consistently-changing factors such as salinity, intertidal variation and detritus build-up have allowed a huge microbial diversity to occur in the mangrove areas. The microbial community that can be found in mangrove sediments may have bacteria that can adapt themselves to various ecological, biogeochemical and anthropogenic conditions (Ghosh et al., 2010). Mangrove leaf litter and biomass that are present in the mangrove sediment are mainly decomposed by microorganisms. This environment stimulates the birth of a diverse population of organisms that are capable of utilizing the energy source presented by the detritus in the mangrove sediments (Kristensen et al., 2008). As an addition, the natural mangrove ecosystems equipped with tropical rainfall and leaf litter provide a conducive environment for microorganisms with striking polysaccharide-degrading capabilities to flourish (Rosado and Govind, 2003). With some novel species of bacteria recognised from these regions, they may prove to be invaluable in discovering new enzymes or genes to be exploited for biotechnological applications. It has been reported that mangrove *Actinomycetes* are promising sources of bioactive compounds that have potential medical applications (Hong et al., 2009, Xu et al., 2014).

Around 130 strains were isolated from Matang Mangrove Forest and Taman Paya Bakau in Perak, Malaysia in an effort to create a Marine Biodiversity Library (MBL) in Centre for Chemical Biology, Universiti Sains Malaysia from bacteria isolated from mangrove sediment. In order to identify the strains isolated, their 16S rRNA gene sequence was determined. Among these 130 strains, there were three strains, TPBH4, ST2L12 and ST2L15 which exhibited the lowest 16S rRNA gene sequence similarity to their type strain after BLAST analysis was carried out. These strains formed a unique phylogenetic lineage with the genus *Mangrovimonas* (Li et al., 2013) in the family *Flavobacteriaceae*through 16S rRNA gene sequence analysis. In recent times, only the genome of the type strain in the genus *Mangrovimonas* (*M. yunxiaonensis*) has been published (Li et al., 2014a). The family *Flavobacteriaceae* was first proposed by Jooste and currently it accommodates 114 genera (http://www.bacterio.net/-classifgenerafamilies.html) (Jooste et al., 1985). Members of *Flavobacteriaceae* are known to be active players in carbon cycling via dissolved organic matter (DOM) uptake (Cottrell and Kirchman, 2000) in the aquatic environment. In addition, members of this family are known to breakdown carbohydrates such as starch, agar, alginate, chitin and cellulose (Bowman, 2006, Tully et al., 2014).

In this study, time was spent on sequencing the genomes of recently isolated these strains TPBH4, ST2L12 and ST2L15 and focused on carbohydrate metabolism genes of three isolates in comparison with their closest taxonomic neighbour. In addition, this study sheds some light on the characteristics of the novel species represented by the strain ST2L12 (*Mangrovimonas xylaniphaga* sp. nov.) and characterization of TPBH4 and ST2L15 ascertained through the phenotypic, phylogenetic and chemotaxonomic analyses.

1.2 Objectives

The objectives of this study were:

1) To assemble, annotate and comparatively analyse the genomes of strains TPBH4, ST2L12 and ST2L15 with their closest taxonomic relative.
2) To characterize the strains TPBH4, ST2L12 and ST2L15 through phenotypic, phylogenetic and chemotaxonomic analysis.
3) To identify whether these strains, TPBH4, ST2L12 and ST2L15 are novel in terms of taxonomy and nomenclature.

CHAPTER 2 LITERATURE REVIEW

2.1 Mangrove forests: A broad overview

Mangroves, the main woody halophytes living at the conversion of land and ocean, have been vigorously utilized generally for sustenance, timber, fuel and medication, and it involves around 181,000 km^2 of tropical and subtropical coastline in between the latitudes of 25°N and 30°S. In the course of recent years, roughly 33% of the world's mangrove woods have been lost, yet most information show extremely variable rates of loss between evaluations. Mangroves are a profitable natural and monetary asset, being critical nursery grounds and rearing destinations for birds, fish, scavengers, shellfish, reptiles and warm blooded animals; a renewable wellspring of wood; and offer protection against waterfront erosion (Alongi, 2002).

The rates at which mangrove forests are being lost is alarming and concern about this has been raised in specialized scientific literature (Saenger et al., 1983, Spalding et al., 1997). The obliteration of mangrove forests is a worldwide phenomenon and the increase in sea level caused by for instance, climate change may harm mangroves (Ellison, 1993, Field, 2016). Moreover, it has been human activities have changed the face of mangroves through aquaculture, forestry, agriculture and urbanization (Saenger et al., 1983, Fortes, 1988, Twilley, 1988, Primavera, 1991, Marshall, 1994). The magnitude of the destruction of mangrove area is easily comprehended through data presented in the form of satellite images and this information is also readily accessible through the internet.

2.2 The current status, ecology and ecological management of Malaysian mangroves

In Malaysia, Sabah accounts for the largest area of mangroves with an estimated area of 577,500 ha which accounts for 59 % of the country’s total. Sarawak ranks in at second place with a mangrove area of 132,000 ha which is 23 % of the total whereas 18 % of the country’s mangroves are found in Peninsular Malaysia (Jusoff, 2013). Malaysian mangroves are generally characterized intro 3 species-specific zones namely *Avicennia-Sonneratia*, *Bruguiera-Rhizophora* and the back mangrove zones (Jusoff, 2013). In terms of biodiversity, Malaysian mangrove forests harbour over 60 different tree species. Mangrove trees usually grow up to a height of 7-25 m. Mangrove trees have unique habitat adaptation features such as hardy root systems and unique bark and leaf structures. Moreover, mangrove trees provide house building material, charcoal and material to build fish traps for local people (Jusoff, 2013).

In a span of 10 years between 1980 and 1990, 12 % of Malaysian mangrove forests was lost to mariculture, agriculture, deforestation and urbanization (Spalding et al., 1997). Although mangrove forests are decreasing globally, Malaysia’s mangroves are generally still intact under a mangrove forest management hierarchy system (Jusoff, 2009). The mangrove forests’ ecological management system has experienced tremendous changes from administering its wood produce to running a management system that sees the involvement of multiple roles, protection, and conservation. Systematic ecological management of the Malaysian mangrove forests began in1904, with the adoption of the first working plan for mangrove forests in Matang, Perak. The Matang mangrove is known to be the best described managed mangrove forest in the world and it stands prominently as one of the sustainably ecologically managed mangrove forests. The Matang mangrove is still unflawed, and it still delivers various goods and services, sustainably. This is proof of the successful ecological forest management practices that lends Matang mangroves the credible reputation as the best managed mangrove forest in the world. A special emphasis on the mangrove forest protection is duly recognized and given particular attention in the National Forestry Act 1984, and revamped in the National Forest Policy 1978 (revised 1992). Future ecological management of mangrove forests in Peninsular Malaysia will envisage the adoption of an integrated approach by further refining the current management approach and incorporating the latest findings and updating various helpful and relevant information through more active research and development (R&D), timely scientific expeditions, and ecological studies on mangrove forests. The National Forestry Policy and other policies that have to do with mangrove forests need to be revised every now and then to match the prevailing conditions and requirements, to ensure that the multiple functions can further be realised (Jusoff, 2013).

2.3 Biogeography of Matang Mangrove Forest and Taman Paya Bakau, Perak, Malaysia

Matang Mangrove Forest is located in the district of Matang-Larut-Selama and it covers approximately 40,000 ha. This forest is primarily dominated by the mangrove tree species, *Rhizophora apiculata* (Ong et al., 2004). Moreover, approximately 28 true mangrove species have been found in this forest in terms of flora (http://www.unepscs.org/Mangrove-Training/20-Matang-Management.pdf). On the other hand, Taman Paya Bakau (Mangrove Forest Park) located in Manjung, Perak has yet to be studied extensively ecologically or biologically but there has been a study done in terms of ecotourism in this region (Sahazali and ErAh, 2013).

2.4 Soil microbial diversity

To this end, the microbial diversity in soil ecosystems has exceeded that of the eukaryotic organisms (Torsvik and Øvreås, 2002). One gram of soil may harbour up to 10 billion microorganisms of perhaps, thousands of different species (Rosselló-Mora and Amann, 2001). As not more than 1% of the microorganisms observed under the microscope is cultivated and characterized, soil ecosystems are, mostly uncharted. Microbial diversity entails the complexity and variability at different levels of biological organization. It deals with the genetic variability within taxons (species) and the number (richness) and relative abundance (evenness) of taxons and functional groups (guilds) in communities. Important aspects of diversity at the ecosystem level include the many processes, the complexity of interactions, and the number of trophic levels(Torsvik and Øvreås, 2002). In recent times, researchers have been interested to assess the bacterial diversity of mangrove soil because there has been very few studies done to study the aspect of bacterial diversity in mangrove soil (Sahoo and Dhal, 2009, Ghizelini et al., 2012, Basak et al., 2016).

2.5 Bacterial whole-genome sequencing (overview)

It has been twenty years since the first bacterial genome was successfully sequenced (Fleischmann et al., 1995, Fraser et al., 1995), and interestingly, the technical improvements and subsequent increases in biological knowledge have been just as spectacular in the second decade as they were in the first one. The most influential factor behind this scientific progress was, predictably, the great reduction in the sequencing price, following the exponential technical developments. Coupled with the cost reduction, second-generation sequencing techniques reduced the average read length dramatically; in contrast, the third-generation (single molecule) sequencing gives way to longer read lengths, although at the time of writing, these methods are still a novelty (Land et al., 2015). The drastic reduction in the cost of sequencing has made bacterial genome sequencing quite bearable, financially to many labs, leading to what is termed as sequencing democratization (Shendure and Ji 2008). The explosive growth of data has led to the shift of costs from sequencing to assembly, analysis, and managing data (Land et al., 2015).

2.5.1 DNA quality and library preparation for whole genome sequencing

Whole-genome sequencing, or in specific long-insert size libraries, requires high-quality, intact, non-degraded DNA at a sufficient amount (Wong et al., 2012). For sequencing, a full genome with a set of different libraries requires ~1 mg of DNA as starting material (about 6 μg for short-insert libraries, about 40 μg for 2–10 kb libraries, about 60 μg for >20 kb libraries). Before being involved in genome sequencing, it is essential then, to get a large amount of high-quality DNA of the target species. This can greatly be an encumbrance to many species with conservation concern. If captive animals are available, such samples can work as a source of high-quality DNA, but it is noteworthy that the genomic variation identified from such sources may not represent all wild populations. Before sending a DNA sample, its integrity should be checked on a high-resolution gel (e.g. pulse-field electrophoresis; the fragments are usually >100 kb).When opting for the necessary raw read depth, it should be realised that for the time being, most technologies include several PCR steps which can lead to a non-negligible number of duplicated reads. While single reads can occur in duplicate by chance should the coverage be high enough, duplication will tend to be an artefact for identical read pairs which are very impossible to occur by chance (as they follow a length distribution). As duplicated reads are deemed trivial and duplication artefacts can damage the coverage-based quality validation, they should be removed before the assembly. Duplicates generally make up a few percentage of short-insert size libraries (<500 bp), but they can reach more than 95% for long-insert libraries (>10 kb). Another important question concerns with the insert sizes to use (Ekblom and Wolf, 2014). Generally, a good advice would be to have a good mix of sizes in the range of 0.2–40 kb with the shorter libraries being sequenced to remarkably higher depth (Gnerre et al. 2011). Insert sizes of >20 kb make a large difference to the final contiguity and scaffold size of the assembly, but we can never undermine their production at high quality and which currently constitute a limitation of many sequencing centres. Library preparations are dissimilar in quality and in how well they depict the different parts of the genome. Therefore, ideally, more than one library should be generated per size class. Some assembly programs (such as ALLPATHS-LG) would anticipate a predefined mix of sequencing libraries as input data. Another remarkable issue for downstream analyses that comes with library preparation would be the read orientation. We need to observe the technology adopted, before we know if the reads can face inwards or outward in regard of the original DNA fragment. Mis-oriented reads with unexpectedly short insert sizes can arise because of the pair sequencing from within the original DNA fragment rather than at its ends. Also, mate-pairs with aberrant insert sizes and orientation normally depict the chimeric sequences from nonadjacent genomic regions. For the majority of the assembly methods, such artefacts need to be screened out during the preassembly steps, leaving only a small fraction of usable, unique read pairs for assembly after the trimming. To accurately process the data, the bioinformatician who handles the data needs to be consistently ‘library-aware’ (Ekblom and Wolf, 2014).

2.5.2 Sequencing platform

In reference to the first decisions made when starting, a genome sequencing project makes the choice of sequencing platform, the type and amount of sequence data to generate. The latter is often constrained by project funding, and the former may have to rely on the sequencing technology which never fails to be available. Judging from recently completed whole-genome sequencing projects, there is a clear trend that shifts from traditional Sanger sequencing (approximately 1kb sequence reads) and Roche 454 sequencing (can be up to 800 bp) towards short read technologies such as Illumina HiSeq (usually 150 bp) and SOLiD (usually 50 bp). To date, there has been an undeniable progress in producing longer reads at high throughput; several technologies offering this, such as Pacific Biosciences (can be up to 5 kb), IonTorrent (aprroximately 500 bp) and IlluminaMoleculo (can be up to 10 kb), are penetrating into the market, and we anticipate looking at a broader spectrum of read lengths. Although this development blurs the early dichotomy of short reads (e.g. 35 bp Illumina reads) versus long reads (~1 kb Sanger reads), read length still has some significant bioinformatic implications, as assembly algorithms optimized for long reads are essentially different from approaches which target short reads (Ekblom and Wolf, 2014). Recent studies begin to merge the data of different read lengths and from several different sequencing platforms (Koren et al., 2012). This strategy makes intuitive sense as one can counterbalance the drawbacks of each method, although it has yet to be decided whether such hybrid assemblies always outperform single data type approaches (Bradnam et al., 2013). At this point, we adhere to the principle of current common practice and largely base our considerations on the sequencing of Illumina libraries of different lengths. Many of the following reflections, however, more generally relate to the assembly problem and they are not reliant on the specific choice of sequencing library (Ekblom and Wolf, 2014).

2.5.3 Genome assembly

Before the assembly, a proper evaluation must be done on the quality of the sequencing data, overall GC content, repeat abundance or the proportion of duplicated reads. Tools such as FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) which can give summary statistics would make a useful starting point (Ekblom and Wolf, 2014). Trimming low-quality data and reads resulting from PCR duplications can be performed with variety diversity of software and scripts (Smeds and Künstner, 2011) for example seqtk (https://github.com/lh3/seqtk). In terms of assembly, the SPAdes assembler incorporates four stages during the assembly process. The first stage would be assembly graph construction which simplifies the de Bruijn graph. The second stage would be k-bimer adjustment which estimates the distance between k-mers in the genome. The third stage would paired assembly graph which incorporates the paired de Bruijn graph (PDBG) approach. The fourth stage would be contig construction which maps the reads to their respective contigs (Bankevich et al., 2012).

2.5.4 Genome annotation

To exploit the full potential of a genome sequence, it has to be annotated with biologically relevant information that can range from gene models and functional information, such as Gene Ontology Consortium (Primmer et al., 2013) or 'Kyoto encyclopedia of genes and genomes' (KEGG) pathways(Ogata et al., 1999), to microRNA and epigenetic modifications (Consortium, 2012). As far as the genetic non-model organisms are concerned, a more general annotation is often confined to protein-coding sequence (CDS) or transcripts (Ekblom and Wolf, 2014). Although there is a considerable challenge to annotate genes in newly sequenced species where pre-existing gene models are mostly inadequate, in principle the automated gene annotation has become possible for individual research groups (Yandell and Ence, 2012). For example, the RAST genome annotation server goes through several basic steps to annotate the genome which is submitted. These steps include tRNA and rRNA gene calling, protein-encoding gene calling, establishment of phylogenetic context, FIGfam search, metabolic reconstruction and processing the remaining unassigned protein encoding genes (Aziz et al., 2008).

2.6 Minimum Information about a Genome Sequence (MIGS)

The analysis of genomic information leaves an impact on every area of the life sciences and beyond. A genome sequence is a prerequisite to comprehending the molecular basis of phenotype, how it grows through time and how we can manipulate it to offer new solutions to critical issues. Such solutions include therapies and cures for disease, industrial products and many more.

With improvements evident in sequencing technologies, the growing interest in metagenomic approaches and the proven power of comparative analysis of groups of related genomes, we can envision the day when it will not be strange anymore to sequence tens to hundreds of genomes or more as part of a single study(Field et al., 2008). Considering the importance of the growing genome collection, the capital investment in its creation and the benefits of leveraging its value through various comparative analyses, every effort should be exerted to describe it as accurately and comprehensively as possible. There is a heightened interest from the community in doing so, for these following reasons, the first is the interest in testing hypotheses about the features observed in genomes using through the adoption of the evo- and eco-genomic approaches (Hughes Martiny and Field, 2005). The second is the requirement to supplement the content of various databases with high-level descriptions of genomes that give allowance for useful grouping, sorting and searching of the underlying data. The third is the growth in the genome sequence data from environmental isolates and metagenomes which have very large data sets of DNA fragments from environmental samples (Edwards et al., 2006, Rusch et al., 2007, Handelsman et al., 2007). The data generated by such studies will dwarf the current stores of genomic information, rendering the improved descriptions of genomes even more important. Presently, both top-level descriptors and genome descriptions are incomplete for multiple reasons. Ultimately, we now know the minimum quality and quantity of information that is required to make each description reliable and precise (Field et al., 2008). Based on some empirical observations, we are broadening our view of the types of information that are important to test particular hypotheses, look into new patterns and quantify the innate sampling biases (Hughes Martiny and Field, 2005, Haft et al., 2005).

The Genomics Standards Consortium (GSC) is making a harmonized effort to help and expedite the process of gathering relevant metadata which would clearly mitigate any ongoing replication of efforts and maximize the sharing ability and integrate data within the genomics community. The clear-cut solution is to develop a consensus-based approach. The GSC seeks to define a set of core descriptors for genomes and metagenomes in the form of a MIGS specification. MIGS makes further use of the minimum information already captured by the INSDC. The MIGS checklist can be accessed from the consortium's website (http://gensc.sf.net). The information required to follow the MIGS is included routinely in primary genome publications (or is referenced therein) (Field et al., 2008). However, this information has to be in a formal mode and made available in electronic form to develop its accessibility (Field and Hughes, 2005). As it was originally proposed (Field and Hughes, 2005), the MIGS specification has been made simpler and changed by the GSC through an iterative revision process to contain (i) only curated information that cannot be computed from raw genomic sequence and (ii) core descriptors which are especially catered for the major taxonomic groups (eukaryotes, bacteria and Archaea (Pace, 2006), plasmids, viruses, organelles) and metagenomes. MIGS is built upon as an 'Investigation' composed of a 'Study' and an 'Assay', in reference to the Reporting Structures for Biological Investigations (RSBI) working group's recommendation for the modularization of checklists (Sansone et al., 2006, Taylor et al., 2008). Under 'Study' are the top-level concepts 'Environment' and 'Nucleic Acid Sequence' and under 'Assay' is a description of the sequencing technology. MIGS has the aim to support the unencumbered access to genomic reagents (Ward et al., 2001), place the complete (meta)genome collection into geospatial and temporal contexts (latitude, longitude, altitude or depth, date and time of sampling) and give the important details of the experimental method used (e.g., sequencing method)

(Field et al., 2008). MIGS also offers a framework for the capture of extra information that is deemed ‘minimum’ to specific communities.

2.7 Methods of polyphasic taxonomy

Major techniques that are adopted in terms of taxonomy are as follows:

i) Determination of the DNA base ratio (moles percent GC). Determination of the moles percent guanosine plus cytosine is one of the classical genotypic methods and is considered part of the standard description of bacterial taxa (Vandamme et al., 1996). In general, the range observed is not more than 3% within a well-defined species and not more than 10% within a well-defined genus (Stackebrandt and Liesack, 1993). It varies between 24 and 76% in the bacterial world.

ii) DNA-DNA hybridization studies. As has been mentioned, the percentage of the DNA-DNA hybridization and the decrease in the thermal stability of the hybrid are used to delineate species (Wayne et al., 1987). The percentage of DNA binding (Ley et al., 1970) or the DNA-DNA hybridization value or the relative binding ratio (Brenner et al., 1969, Grimont et al., 1980, Popoff and Coynault, 1980) is an indirect parameter of the sequence similarity between two entire genomes. There has been a verification that thermal stabilities decrease from 1 to 2.2% for each 1% of mispairing (Bautz and Bautz, 1964, Ullman and McCarthy, 1973, Stackebrandt and Goebel, 1994). It is, however, highly arguable whether data which were obtained with short oligonucleotides and experimentally induced mispairing can be extrapolated to entire genomes. As for the time being, it therefore remains impossible to convert a percent DNA-binding or DNA-DNA hybridization value into a percentage of whole-genome sequence similarity (Vandamme et al., 1996).

iii) rRNA homology studies. The rRNA is now accepted as the best target for studying phylogenetic relationships because of its presence in all bacteria, the fact that it is functionally constant, and that it is composed of highly conserved, as well as more variable, domains (Woese, 1987, Stackebrandt and Goebel, 1994). The components of the ribosome (rRNA and ribosomal proteins) have been the main focus of different phylogenetic studies for a few decades. The slow-yet-steady development of new molecular techniques enabled microbiologists to divert their attention to the comparative study of the rRNA molecules (Vandamme et al., 1996). Indirect comparison by either hybridization studies (De Ley and De Smedt, 1975, Palleroni et al., 1984) or rRNAcataloging of RNase T1-resistant oligonucleotides of 16S rRNA (Fox et al., 1977, Stackebrandt et al., 1985, Fox and Stackebrandt, 1987) have already exposed the natural relationships with and within the bacterial lineages. Later, the sequencing of the rRNA molecules gradually resulted in an rRNA sequence database of 5S rRNA(Wolters and Erdmann, 1988), which was recorded to be the first rRNA molecule to be sequenced for a lot of bacteria because of its less complex primary and secondary structures. A limited number of 16S rRNA gene sequences had been available by direct sequencing after the cloning of the genes from the bulk of the DNA (Goodfellow and Stackebrandt, 1991). The sequencing of 16S rRNA with conserved primers and reverse transcriptase (Lane et al., 1985) became a very important progress in bacterial phylogeny and it led to a spectacular increase in 16S rRNA sequences. To date, these techniques have mostly been taken over by direct sequencing of parts or nearly the entire 16S or 23S rDNA molecules by using the PCR technique and a particular range of appropriate primers. They give a phylogenetic framework which serves as the mainstay of modern microbial taxonomy. The results obtained and the dendograms constructed with data obtained from the above methods are more or less equivalent, with consideration given to the specific resolution of each method. However, obviously, the larger the conserved elements, the more information they carry and the more reliable the conclusions will be. The cataloguing method and the DNA-rRNA hybridization experiments have slowly and steadily disappeared, although the latter method boast off an important advantage that multiple strains could easily be included (Vandamme et al., 1996). International databases which comprise of all published and some unpublished partial or complete sequences have been constructed (Olsen et al., 1991, De Rijk et al., 1992).

iv) Classical phenotypic analyses. The classical phenotypic characteristics of bacteria boast off some morphological, physiological, and biochemical features. In isolation, many of these characteristics have been demonstrated as irrelevant as parameters for genetic relatedness, yet all in all, they provide some descriptive information that allows us to identify the taxa. The bacterium morphology includes both cellular (shape, endospore, flagella, inclusion bodies, Gram staining) and colonial (colour, dimensions, form) characteristics. The physiological and biochemical features also account for the data on growth at various temperatures, pH values, salt concentrations, or atmospheric conditions, growth as various substances such as antimicrobial agents are present, and data on the presence or activity of multiple enzymes, metabolization of compounds, etc (Vandamme et al., 1996). Too often, highly standardized procedures are needed to obtain reproducible results within and between laboratories (On and Holmes, 1991, On and Holmes, 1992).

v) Automated systems. Miniaturized phenotypic fingerprinting systems have been introduced and may possibly substitute for the classical phenotypic analyses in the future. These systems mostly contain a battery of dehydrated reagents, and additionally, standardized inoculums which trigger the reaction (growth, production of enzymatic activity, etc.). The results are seen to have been recommended by the manufacturer and they are readily available with a minimal input of time. The outcome of a particular test with a commercial system for example API tests (API ZYM, API 20NE, API 20E) is at times, different from that with a classical procedure, but the same is often true for two classical procedures performed in the same test. It is undeniable that the phenotypic tests must be performed under well-standardized conditions to obtain reproducible results (Vandamme et al., 1996).

vi) Cellular fatty acids. Mostly, the total cellular fatty acid fraction is extracted, but particular fractions like the polar lipids have also been studied (Embley and Wait, 1994). Fatty acids are the main constituents of lipids and lipopolysaccharides and they have been used extensively for taxonomic purposes. More than 300 different chemical structures of fatty acids have been recognised. The variability in chain length, double-bond position, and substituent groups has evidently been very useful for the bacterial taxa characterization (Suzuki et al., 1993). Cellular fatty acid methyl ester content is a stable parameter under the condition that highly standardized culture conditions are used. The method is affordable, fast and has a high degree of automation (Vandamme et al., 1996). Polar lipids serve to be the major constituents of the lipid bilayer of bacterial membranes and they have been analysed frequently for classification and identification. Other types of lipids, such as sphingophospholipids, occur in only a restricted number of taxa and were shown to be very important (Jones and Krieg, 1984). The lipopolysaccharides presented in the outer membranes of gram-negative bacteria can be analyzed by gel electrophoresis, giving typical lipopolysaccharide ladder patterns defined as variants in the O-specific side chains (De Weger et al., 1987, Siverio et al., 1993).

vii) Isoprenoidquinones. Isoprenoidquinones occur in the cytoplasmic membranes of most prokaryotes and they have an important role to play in electron transport, oxidative phosphorylation, and, possibly, active transport (Collins and Jones, 1981, Collins, 1994). Two leading structural groups, the naphthoquinones and the benzoquinones, are differentiated. The former can be categorised into two main types, the phylloquinones, which occur less commonly in bacteria, and the menaquinones. The large variability of the side chains (differences in length, saturation, and hydrogenation) can work to characterize bacteria at varying taxonomic levels (Collins and Jones, 1981).

viii) Polyamines. While the role of polyamines in the bacterial cell is ambiguous, they appear to be important in bacterial metabolism (White Tabor and Tabor, 1985). How their universal character and quantitative and qualitative variability is observed has turned them into a suitable chemotaxonomic marker that can be ascertained via gas chromatography (Yamamoto et al., 1983) or high-performance liquid chromatography (Scherer and Kneifel, 1983, Carteni-Farina et al., 1985). Based on the group of organisms studied, polyamine patterning serves to trace relatedness at and above the genus and at the species level (25, 124, 278, 370) (Busse and Auling, 1988, Hamana and Matsuzaki, 1990, Yang et al., 1993, Segers et al., 1994).