Using oligonucleotide signatures to build a system for effective detection of pathogenic bacteria in metagenomic samples

WARREN EMMETT

Submitted in partial fulfillment of the requirements for the degree

Magister Scientiae

In the Faculty of Natural and Agricultural Sciences

...

Bioinformatics and Computational Biology Unit
Department of Biochemistry
School of Biological Sciences
Faculty of Natural and Agricultural Sciences
University of Pretoria
Pretoria

..

September 2008

Declaration

I, *Warren Anthony Emmett* declare that the thesis/dissertation that I hereby submit for the degree in *Msc. Bioinformatics* at the University of Pretoria has not previously been submitted by me for degree purposes at any other university and I take note that, if the thesis/dissertation is approved, I have to submit the additional copies, as stipulated by the relevant regulations, at least six weeks before the following graduation ceremony takes place and that if I do not comply with the stipulations, the degree will not be conferred upon me.

SIGNATURE..... DATE.....

Summary

Pathogenic bacteria are responsible for millions of deaths every year with an estimated mortality of 70 million people by 2010 for *Mycobacterium tuberculosis* alone. Novel methods for identification of bacterial species in hosts, urban environments, water sources and food stuffs are required to advance diagnosis and preventative medicine. Detection of bacterial species in environmental samples is a complex task since large numbers of bacteria are present and are resistant to culturing. Therefore, the genetic content of the entire sample has to be analysed simultaneously and this constitutes a metagenomic sample.

Commonly-used methods of bacterial identification focus on detection of specific genomic regions to determine species. Currently only one percent of a metagenomic sample can be used for identification employing phylogenetic markers. This method is highly inefficient. The search for more widespread markers within each genome is essential to improve detection methods. Also, modern sequencing technologies used in these environments have short read lengths which prove difficult to assemble e.g. repeats can lead to incorrect assembly.

The use of overrepresented oligonucleotides provides a solution to both of these difficulties. Overrepresented oligonucleotides (8-14bp in length) are utilised to differentiate between species based on observed frequency of occurrence rather than presence or absence. They occur throughout the genome thereby increasing genomic coverage. Furthermore, overrepresented oligonucleotides can be easily identified in a raw metagenomic sample, bypassing the need for sequence assembly.

Raw oligonucleotide data was filtered, analysed and imported into a structured database. A program, *Oligosignatures,* allowed for creation of species and phylogenetic lineage specific oligonucleotide markers dependent on the selection of species specified by the user. For the purposes of this study, the context of bacterial identification in an unknown environment was selected. A similarity trial was then executed to determine if strains of the same species can be separated from each other using overrepresented oligonucleotides. Outcomes of this test provided a guideline for the creation of species and lineage specific oligonucleotide markers. Each species and lineage was therefore described by a marker profile which consisted of representative oligonucleotide markers. These marker profiles were then tested against artificial and experimental data to determine their effectivity. Two approaches were used for testing, namely Oligonucleotide frequency analysis and Se-

quence read analysis. Oligonucleotide frequency analysis focused on the identification of species dependent on the global frequencies of marker oligonucleotides within each marker profile. Sequence read analysis attempted to assign metagenomic reads to a specific species dependent on the number of marker oligonucleotides present within the read.

The final database contained 439 bacterial genomes from 22 different phylogenetic lineages. Interpretation of the results obtained after strain similarity testing showed that strains of the same species had highly similar markers and were not separable using this approach. All strains of a species that conformed to this premise were reduced to a single representative member. Similarly, species marker profiles demonstrated that closely related species remained difficult to separate. Twenty-one of the 22 lineages showed sufficient lineage specific markers for use in testing. This provides support for the abundance of overrepresented oligonucleotides and their potential for use as a detection method.

In general, metagenomic testing of marker profiles showed that species specific determination was prone to interference, specifically, in closely related species. However, more distantly related species could be separated using both methods. Lineage discrimination generated more reliable results proving that lineage determination was possible in both artificial and experimental datasets. Oligonucleotide frequency analysis, the most sensitive approach, showed the best results for lineage determination but poorer results for species identification. Sequence read analysis provided a more effective method of determining confidence using different thresholds for read classification.

In conclusion, the use of overrepresented oligonucleotides holds promise as a novel method for bacterial identification in a metagenomic context. Although several obstacles still prevent optimal utilization of these oligonucleotides, with further research the classification and identification of species and phylogenetic lineages from metagenomic samples can become a reality.

Acknowledgments

I express my gratitude and deep appreciation to the following people and institutions:

- Dr. Oleg Reva for his creative insight and guidance throughout my study.
- Prof. Fourie Joubert for his support and guidance during my years at the University of Pretoria.
- Colin Davenport for his guidance and advice surrounding my project.
- My parents, sister and grandparents for their encouragement and support
- Sarah for her love and interest in my work.
- Phillip Labuschagne and Werner Smidt for their valued insight and help in the fields of statistics and computer science respectively.
- My senior colleagues at the Bioinformatics and Computational Biology Unit of the University of Pretoria, Charles, Philip, Pieter and Tjaart for their support throughout my study.
- The National Bioinformatics Network (NBN) and the University of Pretoria for awarding the bursaries which allowed me to complete my masters.

Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

BAC	Bacterial artificial chromosome
CGR	Chaos Game Representation
DNA	Deoxyribonucleic Acid
ERIC	Enterobacterial Repetitive Intergenic Consensus sequences
GB	Gigabyte
kbps	kilobase pairs
OUV	Oligonucleotide Usage Variance
PAI	Pathogenicity Islands
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
pH	The power of hydrogen
PS	Pattern Skew
RAM	Random access memory
REP	Repetitive Extragenic Palindromic elements
SOM	Self Organising Map
tRNAs	Transfer Ribonucleic acid
WGS	Whole Genome Shotgun strategy

Chapter 1

Introduction and Literature Review

1.1 Introduction

Since the development of high throughput sequencing methods, genomic data has grown exponentially with the result that hundreds of complete genome sequences exist for a large number of prokaryotes and eukaryotes. This leaves the mammoth task of analyzing and processing these sequences to extract usable information. That develops insight into genomic processes and provides characteristics for identification and comparison of these organisms with their relatives.

Bacterial disease forms a subgroup of infectious disease which is one of the leading forms of mortality. It is estimated that by 2010 70 million people will die from *Mycobacterium tuberculosis* alone (WHO, 2008). As pathogenic bacteria occur within varied environments and can be present in drinking water, food stuffs and aerial environments, early detection is vital to prevent their spread. This work proposes a novel method to identify bacteria from environmental samples.

The investigation of bacteria and their environments have always been of great interest to the scientific community. Bacteria play a crucial role in all aspects of ecological interaction. Furthering the understanding of bacterial life and interactions can enhance almost every aspect of human existence. Identification of the constituent bacterial species within environmental (metagenomic) samples is a highly complex process whereby DNA is extracted and analysed directly from the sample, avoiding difficulties in culturing bacterial species. This process however, creates difficulties in determining which fragments belong to which species. Current forms of sequence and species identification have proved inadequate within this context, as a large percentage of unique genomic signatures occur only within predefined regions. In order to overcome this, short overrepresented oligonucleotides (2-4 base pairs in length) have been successfully used to identify species fragments. However, the short lengths of the oligomers decrease specificity. The current study aims to apply this technique to longer oligonucleotides that it will provide the

necessary increase in specificity to classify species reliably using shorter fragment lengths.

This review is divided into six main sections. The first section deals with bacterial pathogenicity and the genomic component of pathogenicity, defining what makes a bacterium pathogenic. Furthermore, the study and characterisation of bacteria in environmental samples is discussed. The next section deals with the different types of DNA signatures discovered within bacterial genomes. This includes coding signatures such as codon usage bias as well as non coding signatures. Experimentally determined repeat sequence patterns provide an overview of longer sequences found within different bacterial families, which in turn offer a background for the focus of the study, namely, overrepresented oligonucleotides of intermediate length.

The third section addresses different methods for identifying pathogenic bacteria in a metagenomic context using sequence data. The standard techniques for bacterial identification are reviewed (16S rRNA gene sequences and comparative genomics), followed by short oligonucleotide frequency profiles and finally oligonucleotides of intermediate length. The fourth section gives a brief overview of modern sequencing technologies and the current difficulties with the use of short sequence fragments. Previous work on oligonucleotide signatures is then briefly discussed in the fifth section, providing background on past research into overrepresented oligonucleotides. Finally, the sixth section designates the aims and problem statement of the current study in terms of identification of bacterial species using overrepresented oligonucleotides.

1.2 Identification of bacterial pathogens in environmental samples

As bacterial pathogens cause millions of human deaths each year, the ability to control and treat these diseases is focused on fast and accurate diagnosis of these pathogens. Identification of pathogens in external environments is highly beneficial as preventative medicine can be practiced. In order to identify pathogenic bacteria effectively, mechanisms of pathogenicity and the existence of bacteria within an environmental context must be understood.

This section discusses bacterial pathogens and the metagenomic context surrounding the identification of bacteria. It focuses on the identification of bacteria within these communities based on the different approaches used and the different types of environments currently under study. From this overview the need for intervention into identifying pathogenic bacteria and the associated complexity of identification will become apparent.

1.2.1 Bacterial pathogenicity

The ability to induce disease is determined by the bacteria's pathogenicity or virulence.

There are three different categories of bacteria:

- Primary pathogens are bacteria whose function is to invade, infect and proliferate within a host. They have evolved sophisticated methods to avoid host defenses.
- Opportunistic pathogens are those that only cause disease in weakened hosts with depleted defense against such organisms.
- Non-pathogens form a group of bacteria without pathogenic effects on the host. This can, however, change due to the dynamic ability of bacteria to alter their gene expression and rearrange their genomic DNA.

In order to develop novel diagnostics, understanding the function and processes of pathogenic bacteria is essential. There are two categories which the majority of these organisms share, namely, invasiveness and toxigenesis (Baron, 1996; Todar, 2008). Invasiveness is the ability of the pathogen to gain entry and proliferate within the host organism. This involves several steps namely colonization, production of invasins and evasion of host defense mechanisms.

Toxigenesis is the ability of bacteria to produce and release toxic substances into the host organism. These toxic substances are referred to as virulence factors and can be divided into two families namely exotoxins and endotoxins. Exotoxins are soluble proteins that are released into the host environment and act on targets distant from the bacteria. These are generally specific to a particular bacteria and virulence is therefore dependent on the release of these toxins. Endotoxins are lipopolysaccharides that adhere to the outer membranes of some bacteria, participating in many essential functions for the bacteria including growth and survival. The lipid component of the endotoxin is responsible for the pathogenicity and is less potent and specific than the exotins but does not lose its toxicity over time (Todar, 2008).

Although the underlying mechanisms have been briefly reviewed the genomic component of these processes is essential to the understanding of the current study. The next section will review the genetic basis of bacterial virulence.

1.2.2 Genetic and molecular basis of virulence: Pathogenicity islands

Bacterial virulence is generally determined by several different processes which can be divided into two gene classes. The first class consists of genes responsible for the survival of the bacteria within or outside a host, these are also present in the non-pathogenic strains of a bacterial species. The second class incorporates genes directly involved in the bacteria's virulence and these are unique to pathogenic strains (Groisman and Ochman, 1996).

Virulence factors can be encoded in various places within the genome such as chromosomal DNA, bacteriophage DNA, plasmids or transposons. It is also possible that virulence factors may be mobile and found throughout the genome, as is the case for genome islands e.g. pathogenicity islands (PAI's) (Hacker *et al.*, 1990; Knapp *et al.*, 1986). These mobile genetic elements play a crucial role in the virulence of pathogenic bacteria and loss of a PAI can convert a pathogen to a non pathogen (Schmidt and Hensel, 2004). A large proportion of genes associated with virulence are encoded by mobile genetic elements and many of these are located inside PAIs.

PAIs often have complex control mechanisms that respond to specific environmental stimuli allowing the bacterium to act in a versatile fashion when shifting between environments. These control mechanisms include regulators within pathogenicity islands as well as within the core genome and other PAIs. It is well documented that PAI regulators can act on the core genes within the genome. This gives an idea of the complexity of PAIs in general and how they integrate themselves into existing genomes rather than remaining as external factors (Schmidt and Hensel, 2004).

Although the focus of the current study is on pathogenic bacteria, it must be noted that these organisms often occur in communities. For accurate detection the surrounding environment must be considered. In the next section the analysis and characterization of bacterial environments is explored in more detail.

1.2.3 Metagenomics: The study of organisms in environmental samples

Metagenomics involves the study of communities of micro organisms occurring in their natural environment. It combines the disciplines of genomics, bioinformatics and systems biology to unravel the complex collection of heterogeneous DNA collected from natural environments. Past studies focused on attempting to purify and culture bacteria independently, while in metagenomics DNA is extracted directly from bacterial cells. This is due to the fact that as much as 99% of all bacteria in an ecological environment cannot be cultured in the laboratory (Mongodin *et al.*, 2005).

One of the most exciting uses for metagenomics is that it can provide community-wide assessment of metabolic and biogeochemical function. The best example of this is the acid mine drainage system which is a simple setup of two dominant bacterial species found in a low pH, high sulfur mine dump. In this example, the ecosystem was simple enough to allow complete sequencing of one genome and a detailed annotation of its metabolic functions (Tyson *et al.*, 2004).

In an opposing example, the Sargasso Sea project involved an exceptionally complex community of bacteria present in sea water. With over one billion base pairs sequenced, there is still a staggering amount of data to be analysed - over 794,061 genes in con-

served hypothetical protein groups remain to be identified (Venter *et al.*, 2004). Although environments can differ substantially, as demonstrated by these examples, the same processes are involved in their analysis and these are described below. Metagenomic analysis involves several initial steps:

- Isolating DNA from an environmental sample
- Cloning the DNA into a suitable vector
- Transforming the clones into a host bacterium and applying a specific screening approach to the bacteria (Handelsman, 2004).

The screening approaches include using hybridization or multiplex PCR to identify phylogenetic markers (Stein *et al.*, 1996), searching for expression of specific traits (Courtois *et al.*, 2003) or random sequencing (Tyson *et al.*, 2004; Venter *et al.*, 2004). These methods can be divided into two main approaches namely sequence-based metagenomics and function-based metagenomics.

Function-based metagenomics is a laboratory process whereby randomly selected DNA is taken from environments and inserted into bacteria that can be cultured in a laboratory. These laboratory bacteria are then monitored for production of alien proteins and screened for any unique properties. This method has identified novel antibiotics (Courtois *et al.*, 2003; Venter *et al.*, 2004), antibiotic resistance genes (Diaz-Torres *et al.*, 2003), sodium transporters (Song *et al.*, 2005) as well as several biocatalytic enzymes (Lorenz *et al.*, 2002). The benefit of this approach is that it does not require any sequence data and so allows for the identification of novel classes of genes with new or known functions. One of the weaknesses of this approach is the possibility that an alien gene will be incompatible or incorrectly transcribed and expressed within a host bacterium. However, sufficient success has been achieved with this approach that it still remains feasible and integral to further research.

The sequence-based metagenomics approaches consist of two different methods. The first involves laboratory screening of metagenomic fragments using phylogenetic markers then sequencing only the resultant fragments that can be linked to specific family or taxa. This was first proposed by Stein *et al.* (1996) and resulted in the discovery of a new archeaon in their study performed on seawater. One of the most high impact discoveries was that of proteorhodopsin, a retinal-binding integral membrane protein found in marine bacterioplankton (Béjà *et al.*, 2000).

The second sequence based method involves random sequencing of a sample of metagenomic DNA followed by screening for phylogenetic markers and attempting to identify features of interest within the genomic fragment. This approach, applied on a large scale, has allowed for further insights into linkage of traits, distribution and redundancy of functions in community, genomic organization, and horizontal gene transfer. The approach

has culminated in the reconstruction of uncultured bacterial genomes in an acid mine drainage community analysis and creating new linkages between phylogeny and function in complex environments (Tyson *et al.*, 2004; Handelsman, 2004).

Phylogenetic markers fulfill an integral function within sequence-based approaches by associating DNA fragments with specific species. The most widely used phylogenetic markers are the 16S rRNA gene sequences which have enabled the identification of novel bacteria and allowed for their classification by comparative genomics. However, as little as 0.5-1% of the reads within a metagenome are identifiable using current marker systems and the probability of finding a novel functional gene on these fragments is even more unlikely. Fragment identification via laboratory techniques such as genome-walking is possible if reads are large and overlap with 1-2kbp. GC content and codon usage have also been used to classify fragments but have been found to be unreliable in complex environments due to insufficient variation. Reasonable success has been obtained with initial studies using tetranucleotide frequencies to identify metagenomic fragments but, since modern sequencing technologies produce read lengths of only a few hundred base pairs, the efficiency of this technique has yet to be proved in this context (Teeling *et al.*, 2004*a*). This study endeavors to contribute to the field by investigating a novel method for identification of bacterial species using raw metagenomic fragments. The technique can then be extended to the subsequent identification of genomic fragments.

An increasing number of metagenomes associated with human health are under study, such as the human distal gut metagenome (Gill *et al.*, 2006) and the human oral genome (Dewhirst and Chen., 2008). Further related studies include the analysis of drinking water (Schmeisser *et al.*, 2003) as well as aerial metagenomes taken from various urban areas such as shopping malls and hospitals (Tringe *et al.*, 2008). Each of these environments show the potential for metagenomic analyses to impact positively on human health.

Metagenomics opens new doors to the study of bacterial populations and their complex interactions. Such studies will continue to have a profound effect on both scientific knowledge and everyday life as the secrets of these organisms are utilised for biotechnological purposes. This is still an emerging discipline and as such current approaches are incapable of keeping up with the amount of data available for analysis. This study intends to add to this body of knowledge by providing a new method for organism identification and species fragment grouping. In the following section, different methods of characterising bacteria using sequence information will be discussed.

1.3 Oligonucleotide signatures

Genomic data contains a wealth of information that describes the organism and its processes completely. The most obvious and well studied aspect in this area are genes which are often used as genomic markers due to their functional constraint. As more research

has been carried out it has been found that genes alone do not encapsulate the complexity of the organism - there remains a large proportion of unexplored data in non coding regions. Some of the most notable of these include gene control regions such as promoters and transcription binding sites as well as repeat regions created via transposition, translocation, recombination or amplification.

In general the definition of a genomic signature is quite vague but can be described as a signature sequence (or probe) which accurately distinguishes between a target genome or set of genomes and all other background genomes. Phillippy *et al.* (2007) further refined this definition by concluding that a signature sequence must not only be conserved among a set of target genomes but also dissimilar to any sequence in the surrounding environment.

In the work presented by Tembe *et al.* (2007) a more formal criteria for the selection of signatures was incorporated. This process depends on several features. Firstly, a decision needs to be made on what is going to be identified: Will it be a single pathogenic strain, a group of pathogens or multiple bacteria with or without phylogenetic relationship? Secondly, all specifications must be met for the required technologies.

The first DNA signatures created to identify pathogenic bacteria were designed based on sequences linked to genes assumed to be involved in the organism's pathogenicity. This approach, although successful in certain instances, failed in others where environmental testing with multiple other bacteria yielded false positives, most often caused by the selected genes not being unique to the organism as was assumed.

The process of finding a signature sequence that adequately describes a particular genome is an involved process requiring many different criteria to be met. One of the most important attributes in this respect is sequence length. Through simple calculation, the shorter the sequence length of a signature the more likely a sequence is to occur randomly within several genomes. However, longer sequence lengths can become incompatible with practical laboratory technologies. A delicate balance has to be struck between the sensitivity, the number of genomes that contain this oligonucleotide and the specificity, the number of genomes that do not contain this signature (Slezak *et al.*, 2003).

In the current study a signature is defined by its difference in magnitude of occurrence rather than its presence. The focus is not on identifying wholly unique oligonucleotides but rather to determine frequency discrepancies between oligonucleotides in different genomes and to exploit these characteristics to differentiate a bacterial species from its neighbours.

In the remainder of this section an overview will be given of several different genomic signatures found in both coding and non-coding sequences and their utility as signatures to differentiate organisms. Codon bias and its advantages and disadvantages will be discussed first, followed by the properties of short oligonucleotide frequencies. An overview will then be given of known repeats found within non-coding DNA and lastly a description of overrepresented oligonucleotides and their application to the current study.

1.3.1 Codon residues and codon bias

The most commonly used identifier for patterns in coding DNA is the codon. The codon is a nucleotide triplet which is translated into an amino acid or acts as a terminus for translation. There are 64 different combinations possible for the four nucleotides in a triplet sequence, yet only 20 amino acids are coded from these codons. This implies that more than one amino acid can be coded by more than one triplet. Codon usage in most species is known to be highly biased to specific codons for each amino acid. This is referred to as codon bias. As can be expected, closely related organisms seem to share similar codon usage patterns while distantly related organisms have highly divergent codon usage (Ikemura and Ozeki, 1983).

Codon bias has been attributed to several factors, namely, variations in tRNAs, translational accuracy and efficiency and codon or anticodon interaction strength. These factors suggest that codon bias is an adaptation for better translational efficiency within a species (Andersson and Kurland, 1990; Gouy and Gautier, 1982; Kanaya *et al.*, 1999). Codon usage is also affected by mutational pressure, which depends on several different factors such as: function of the gene, position within the genome and level of expression (Daubin and Perrière, 2003). From these findings the conclusion was drawn that the unique codon bias of an organism was determined by selection and mutation within the species. Therefore, searching bacterial genomes will allow for the identification of foreign genes by means of their highly divergent codon bias (Karlin *et al.*, 1998). However, implicit in this theory is the realisation that the newly transferred gene will adapt to its environment and become indistinguishable from other genes. Codon bias is further limited by the discovery that different positions of the genome experience different rates of mutation (Daubin and Perrière, 2003). It was also found that certain highly expressed genes in bacteria had highly divergent codon usage from the rest of the genes (Karlin, 1998*a*). This leads to the conclusion that although codon bias is an informative tool it is not entirely reliable or accurate under different circumstances.

Although coding sequence has proved to be informative it does not constitute the entire genome as large regions of non-coding DNA are also present. The subsection that follows therefore explores genome-wide signature sequences found largely within non-coding regions.

1.3.2 Over and under representation of short oligonucleotides

Analysing genomes by utilizing the occurrence or absence of oligonucleotides was a step forward from past methods which focused on localised areas rather than the entire genome. It also departs from common methods of similarity detection such as alignment of homologous segments. It is an entirely new way of analysing the genome and its subsequent features. The properties of di- and tetranucleotide frequencies will be discussed in detail

below.

1.3.2.1 Dinucleotide frequencies

The over- or under-representation of certain dinucleotides in a genome is one of the most used comparative measures, referred to as the genome signature. This genome signature is defined as the ratios between observed dinucleotide frequencies, calculated from a specific genome, and the expected frequencies of the mononucleotide components randomly associating with each other within the genome (Karlin and Burge, 1995).

Bacterial genomes are continuously changing through transposition, transduction and recombination. Regardless of this fact, data strongly supports the presence of these genome signatures (Karlin *et al.*, 1997). The fact that fluctuations in GC content throughout the genome does not have a drastic effect on genomic signature variability indicates that this is indeed a stable characteristic (Karlin and Burge, 1995). Interestingly, these dinucleotides (referred to as doublets) are not only highly stable but appear throughout the genome in regions of varying complexity (Karlin, 1998*b*). The relative abundance of certain dinucleotides also reflect structural features such as super coiling and it has been noted that DNA and RNA binding proteins are affected by dinucleotide arrangements (Travers, 1997).

This phenomenon is not restricted to coding regions and cannot be explained by codon usage. This leads to the conclusion that these sequence features are the product of global mechanisms such as repair and replication. Replication machinery has been known to generate context-dependent mutations rates, hence generating these dinucleotides. Repair machinery generally operates more effectively on specific sequences and therefore preferentially selects these sequences.

Dinucleotide frequencies have also been utilized as an alternative to conventional methods of phylogenetic reconstruction. The primary advantage of this approach is the use of the entire genome's sequence data. Furthermore, the low variance present in genome fragments of length 50kbp or greater effectively means that as little as 50kbp of sequence can be used to generate a genomic signature for a species.

Other nucleotide frequencies such as tri- and tetranucleotides also show high correlation and similar properties to dinucleotide frequencies. This indicates that DNA conformational arrangements are determined by base-step dinucleotide arrangements. Therefore, these dinucleotide arrangements allow for the prediction of longer oligonucleotides. Tetranucleotide frequencies and their proposed benefits will be discussed next (Karlin and Burge, 1995).

1.3.2.2 Tetranucleotide frequencies

Tetranucleotide frequencies have been utilized quite successfully as a method to determine evolutionary distance, generating phylogenetic trees similar to 16 rRNA gene studies. The use of short oligonucleotides as a measure to describe genomes and their evolutionary distances has been identified as a feasible approach (Pride *et al.*, 2003). Although tetranucleotide frequencies have been shown to carry a phylogenetic signal, this signal fades rapidly in moving from species level to higher order taxa and cannot be used to resolve distant relationships (Teeling *et al.*, 2004*a*). This alludes to the further use of longer oligonucleotides in order to unearth more species information needed for these differences. The advantage of tetranucleotide frequencies over shorter di- and tri- nucleotide frequencies rests on the increasing uniqueness of sequence. This fact has been exploited via the creation of several statistical characteristics and advanced analyses discussed below.

Pattern skew (PS) is a tetranucleotide frequency statistic which focuses on the strand symmetry of genome signatures. The collected data shows that all bacterial chromosomes have a low PS. Interestingly, higher PS values were obtained when calculating local regions within these bacterial chromosomes and the value was also high in conjugative genome islands and in various bacteriophages. In addition, it was found that an increase in the length of oligomers increased PS. Further research shows that oligonucleotides and their reverse complements occur at similar frequencies throughout a bacterial genome as they share structural features. This type of strand symmetry seems to be necessary for genome stability as incorporation of foreign DNA is balanced by global shifts to minimize PS. In conclusion, PS acts as an identifier of foreign DNA within an organism. Furthermore, highly skewed PS values indicate the presence of sequence regions highly divergent from the genomic core, such as ribosomal operons (Reva and Tümmler, 2004, 2005; Ganesan *et al.*, 2008).

Another characteristic of genome signatures is oligonucleotide usage variance (OUV). This refers to the numerical variation in oligonucleotides in a 10kb region where a low value can indicate a repeat region. OUV is strongly dependent on GC content and is sensitive enough to indicate differences between organisms, although strains show similar OUV. From these findings the conclusion is drawn that the higher the OUV, the less random the sequence (Reva and Tümmler, 2004, 2005; Bohlin *et al.*, 2008).

Di- and tetranucleotide frequencies have been identified as useful parameters for the characterisation of bacteria. These characteristics show the versatility of nucleotide frequencies as they enable identification of foreign DNA sequences within bacterial genomes. These short signatures lay the foundation in the search for longer overrepresented nucleotides that can provide more unique identifiers for each genome. In the current study an attempt is made to elucidate more specific information from genomes by using the same basic technique on longer oligonucleotides.

In the search to identify longer oligonucleotides a brief review of the literature is performed to determine known sequence features present within bacterial genomes. The following section will encompass a more detailed analysis of longer sequence features known to exist within bacteria.

1.3.3 Repeat sequence patterns

Repeated sequence patterns are oligonucleotides found reoccurring within the non-coding DNA of bacteria. This subsection outlines the different types of repeats, their similarities and differences, and the potential for use of these sequences in characterising bacteria.

Repeated sequence patterns can typically be found occurring within the non coding regions of a large number of eukaryotes. These elements have been well categorized, one of the best known is the *alu* element in humans. This element also occurs throughout many different mammalian lines and have been used as a basis for the development of PCR laboratory techniques to identify unique DNA sequences within different mammals (Amariglio and Rechavi, 1993).

In general, prokaryotes contain a much smaller amount of non-coding DNA due, not only to the smaller genome sizes, but also to selective pressures which result in efficient nucleotide usage (Nesin *et al.*, 1987). There are some extreme cases where genes' open reading frames overlap to conserve space (Suzuki *et al.*, 1986). It seems counter intuitive that patterns would be present throughout a bacterial genome, but researchers have found that various short interspersed repeating units do exist in bacteria (Dimri *et al.*, 1992; Mancuso *et al.*, 2007). Of these the most well known are repetitive extragenic palindromic (REP) elements (Dimri *et al.*, 1992) and the Enterobacterial Repetitive Intergenic Consensus (ERIC) sequences (Hulton *et al.*, 1991). These were first identified in *Escherichia coli* and *Salmonella typhimurium* respectively.

REP elements identified as palindromic sequences were thought to be linked to regulatory function. The 38bp consensus sequence was used to identify further relations between bacterial species. REP elements can be of any length between 21 and 65 bases and have been linked to various functions including mRNA stability (Khemici and Carpousis, 2004), binding sites for various transciption factors including DNA Polymerase I (Gilson *et al.*, 1990), as well as DNA targets for the transposition of mobile insertion elements (Tobes and Pareja, 2006).

ERIC sequences, roughly 126 bp in length, also contain a palindromic sequence within their highly conserved consensus which appears to be unrelated to the REP elements (Hulton *et al.*, 1991). Both ERIC and REP elements seem to be related to Gram-negative enteric bacteria and closely related species within the phyla. Their evolutionary conservation suggests that their existence precedes the formation of the Gram-negative enteric bacterial lineage (Versalovic *et al.*, 1991).

The reasons for the conservation of these sequences has been well researched. The first is that there is a strong selective pressure due to the importance of these sequences in essential protein interactions. An example of this is *E.coli* transcriptional machinery where DNA Polymerase I binds to REP elements (Gilson *et al.*, 1990). A second hypothesis asserts that these sequences' only function is rapid self-replication (referred to as 'selfish' DNA) and that gene conversion may play a role in the evolution and maintenance of REP sequences (Hulton *et al.*, 1991).

Although REPs and ERICs have been well researched other types of repeats have been identified in bacteria. Recently a group of repeats referred to as Short Regularly Spaced Repeats (SRSRs) have been identified. These sequences are between 24-40 base pairs in length and contain partial inverted repeats of roughly 11 base pairs in length arranged evenly in clusters. These units seem to be widespread throughout different phylogenetic groups, being present in the majority of Archaea and in several members of the cyanobacteria and proteobacteria lineages. These sequences appear to be highly similar within most genomes and between closely related species indicating a common origin of these sequences (Mojica *et al.*, 2000). Previously isolated occurrences of these repeats have been well documented, one example in a study done on *E.coli* by Nakata *et al.* (1989).

Furthermore, four to six base pair palindromic restriction sites were investigated for a collection of different bacteria. It was found that these sites were highly variable within bacteria but that there was a definite inverse relation between the presence of restriction enzymes and the respective palindromic sites that they cut. Reference is also made to these short palindromic sequences being part of longer binding sites, some as long as 14 base pairs (Karlin *et al.*, 1992).

These findings emphasize the wealth of information present within each genome as the diversity of repeat elements continues to expand with closer investigation. These examples strongly support the hypothesis for the presence of longer overrepresented oligonucleotides and their abilities to distinguish phylogenies as well as provide information on cellular processes such as restriction enzyme systems. Against this background a formal introduction to current research concerning overrepresented oligonucleotides of intermediate lengths follows.

1.3.4 Overrepresented oligonucleotides of intermediate length

Overrepresented oligonucleotides are defined as oligonucleotides of intermediate lengths (between 8 and 14 base pairs) occurring in high frequencies throughout bacterial genomes. They are central to the current study which investigates the use of overrepresented oligonucleotides in the identification of bacterial species by analysing frequencies of occurrence of these oligonucleotides in metagenomic datasets and comparing observed results to ex-

Table 1.1: Overrepresented oligonucleotides of 8-14bp in length were derived using sigma cutoff values designated above for each length of oligonucleotide.

Oligomer	Minimum number in 100kbp	Corresponding Sigma value
8mer	20	15
9mer	15	24
10mer	12	36
11mer	8	52
12mer	6	78
13mer	6	78
14mer	6	78

pected values.

In earlier research a command line java application OligoCounter was utilized to identify overrepresented oligonucleotides from unannotated FASTA files obtained from the NCBI FTP site in April 2007 (Benson *et al.*, 2007). These files included over 738 bacterial strains including sequenced bacterial genomes, plasmids and partially sequenced bacterial chromosomes (Davenport *et al.*, 2008).

After analysing thirteen Pseudomonas strains, overrepresented oligonucleotides were discovered for all strains at the stringent chi-squared threshold of 3000. This program was then run on a randomly generated genome with GC content of 50% and no oligomers were found at the same level of overrepresentation. This proves that there are indeed highly overrepresented oligonucleotides of longer lengths within bacterial genomes and that these oligonucleotides are usable for scientific research and exploitation (Davenport *et al.*, 2008).

For use with the current study overrepresented oligonucleotides were identified using relaxed cutoffs to ensure the maximum data was available for analysis. This was done using two different methods, sigma values and chi square statistics. Sigma statistics, which are derived from an assumed normal distribution were determined empirically according to the minimum number of oligomers expected to be found within 100kbp of sequence. Due to the increasing specificity of longer oligonucleotides and their associated reduction in frequency these values were adjusted accordingly to avoid bias. table 1.1 shows the values used for each length of oligonucleotide.

The second method for determination of overrepresented oligonucleotides is chi square. Chi square statistics involve the use of a zero order Markov model to derive expected values of occurrence of an oligonucleotide. This value is dependent on the percentage GC within the genome and the base composition of the oligomer. These expected values were then compared against the observed frequencies using the chi squared formula and several threshold values were then set based on empirical findings.

A java viewing program JCircleGraph was developed to visualise different statistics for each genome including the overrepresentation of oligonucleotides. Figure 1.1 illustrates

the genome of *Psuedomonas aeruginosa* and allows for an overview of all statistics present within the diagram (Davenport *et al.*, 2008).

Psuedomonas aeruginosa is a highly dynamic opportunistic pathogen capable of infecting plants, animals and humans as well as several different tissue types within humans (Campa and Friedman, 1993). This organism is well known for its metabolic flexibility and its antibiotic resistance that has lead to complications in medical treatment. *P. Aeruginosa* has been shown to contain several pathogenicity islands that have been classified as adding to its ability to infect various tissue types and increase its pathogenicity (Larbig *et al.*, 2002).

Within the JCircleGraph visualization the thresholds are defined as three standard deviations above and below the calculated mean values for each statistic. The statistics start from the four innermost rings which are tetranucleotide parameters derived using the program OligoWords (Ganesan *et al.*, 2008). The next two outer rings show the percentage occupancy of 5kbp regions of bases of 8-14mer oligonucleotides at different chi squared thresholds (3000 and 7000 respectively). The outermost ring shows deviation between the smaller 4mer oligos and the longer 8-14 mer oligomers. The deviation class illustrates how much the class (i.e. colour) diverges between the OUV 4mer ring and the 8-14mer ring. If both have a value of 1 (pink or orange respectively) then there is no deviation and grey will be printed (deviation of 0). The deviation ranges between plus 10 and minus 10. The above information is summarised below from the innermost to the outermost ring:

- GC content
- Distance measure, the distance of a local 10kb pattern relative to genome pattern
- Pattern skew, the ratio of occurrence of a 4mer oligonucleotide and its reverse compliment
- Oligonucleotide variance, the numerical variance of oligomers where a lower value indicates a relatively small variety of tetrameters are used in the specific region in the genome (for example in repeat regions)
- Percentage occupancy of the 5kbp regions bases by overrepresented 8-14bp oligos at chi square level 3000
- Percentage occupancy of the 5kbp regions bases by overrepresented 8-14bp oligos at chi square level 7000
- 4mer-8mer correlation class derived from rings 4 (OUV, 4mer) and 5(% occupancy, 8-14mers)

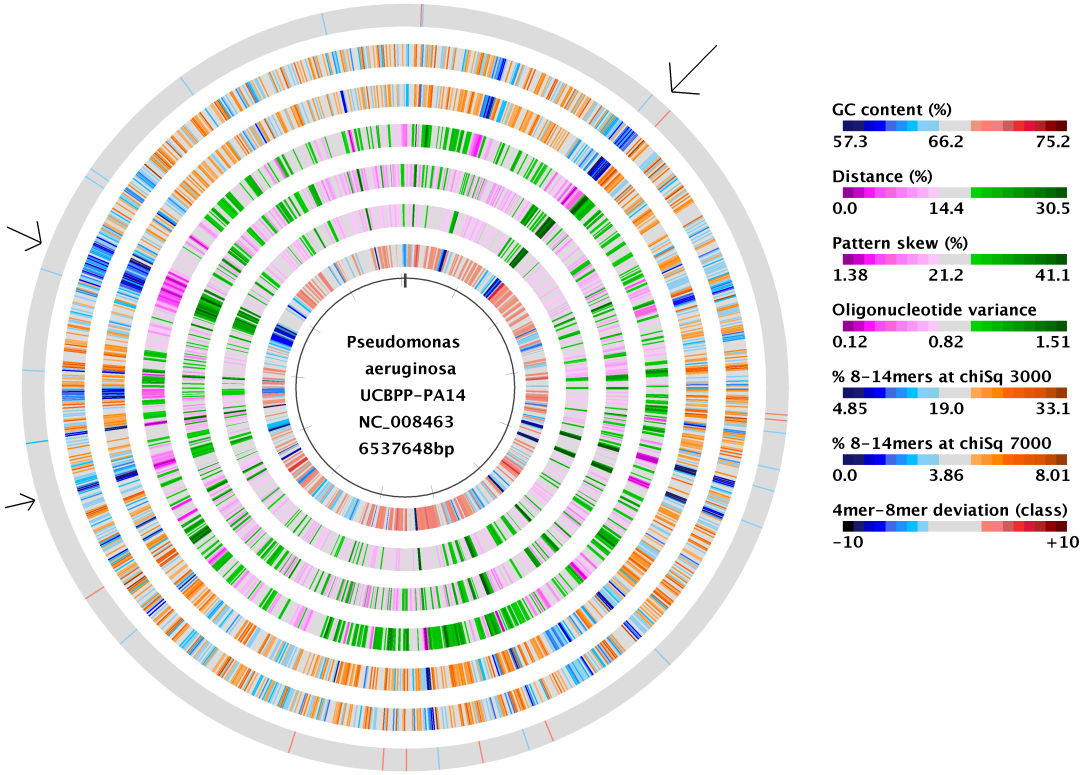


Figure 1.1: The JCircleGraph (Davenport *et al.*, 2008) graphic displaying all the discussed genome features of *Psuedomonas aeruginosa*. Three pathogenicity islands are identified in the diagram by arrows.

(Reva and Tümmler, 2004; Davenport *et al.*, 2008)

It is interesting to note that several foreign DNA pathogenicity islands (PAIs) are present within the *P. Aeruginosa* genome and are clearly visible in different locations throughout the genome. They can be identified as bands radiating outward from the innermost circle. All PAIs express several features described below:

1. Low GC content
2. High distance measure
3. Large discrepancy in pattern skew
4. Very low oligonucleotide variance
5. Low occurrence of oligomers found throughout the rest of the genome
6. A general deviation from 4mer to 8mer

From Figure 1.1 it is clear how interrelated the different statistics are and how the intermediate oligonucleotides still exhibit the same discriminating characteristics as the other statistics and will help add to the body of knowledge surrounding bacteria and their genomic features.

There have been various case studies that have found over- or underrepresented oligonucleotides with different distributions throughout bacterial genomes (Robinson *et al.*, 1995). This has produced much speculation about the occurrences of these oligonucleotides, and possible explanations for oligonucleotide distributions include that they are transcription binding sites for common transcription factors, structural sequences or transposable elements.

In the next section the methods and technologies used in identifying bacteria will be discussed in more depth, including the use of the sequence features described within this section.

1.4 Methods and technologies for identifying bacteria using sequence data

The effective identification of pathogenic organisms using DNA sequence has become an important mechanism in the diagnosis and treatment of pathogenesis. However, identification of pathogenic organisms has been complicated by the presence of several closely related species which are impossible to distinguish via clinical signs, pathogenesis or seroreactivity. These organisms often infect similar hosts and share resistances and genetic material making it very difficult to distinguish between them. In many cases control and treatment of these organisms is often similar and distinguishing between them is merely academic. In other cases where organisms are highly similar on a genomic level but display varied phenotypes, identification can be crucial to correct treatment. This is the case with the *Bacillus* genus (Draghici *et al.*, 2005).

B. anthracis, *B. cereus*, and *B. thuringiensis* are species so closely related that there has been a proposal to name them as a single species (Helgason *et al.*, 2000*b*), yet they display highly divergent phenotypic properties. *B. anthracis* is a virulent pathogen for mammals and has been used as a biological weapon (Check, 2004). *B. cereus* is a food contaminant and an opportunistic human pathogen, while *B. thuringiensis* is used as a biological pesticide (Helgason *et al.*, 2000*a*). To distinguish these organisms their genomic DNA must be closely inspected. There are a number of different methods that can be employed towards this purpose.

There are two main approaches to identifying organisms using DNA sequences. The first is laboratory assays. This includes several methods such as amplified fragment length polymorphism (AFLP) (Ticknor *et al.*, 2001), suppression subtractive hybridization (SSH) (Diatchenko *et al.*, 1996) and custom DNA microarrays (Kingsley *et al.*, 2002). SSH is a PCR based technique whereby DNA differences are determined by a process of subtraction of common sequences. This approach is limited in that created libraries apply only to the driver and target populations and cannot be generalised to identify genomic

signatures (Diatchenko *et al.*, 1996). AFLP analysis is a fluorescence technique based on identification of differences in fragment lengths that indicate polymorphisms in the DNA sequence (Vos *et al.*, 1995).

Laboratory processes have the advantage of not requiring the entire genomic sequence to identify a specific species, both SSH and AFLP have been successfully used for such purposes (Akopyants *et al.*, 1998; Ticknor *et al.*, 2001; Helgason *et al.*, 2000*b*). The disadvantage of these techniques is that they can only identify differences between two organisms and cannot be used to identify global genome signatures.

The second approach to identifying organisms using DNA sequences centers on using bioinformatics tools to analyse genomic sequence. This is done to determine unique sequence features which can be used as markers. The most commonly used technique in the identification of bacteria from sequence data is the identification of 16S rRNA genes within the sequence. Criticism leveled against this and other common methods rests on the low percentage of the genome used to create genome signatures and the sparse selection that this technique provides.

An alternative method is comparative analysis which is widely used to identify unique regions within genomes. Genomes are aligned to each other in order to isolate unique regions of low similarity. These regions are then more closely inspected to identify highly divergent sequences. In other cases, unique polymorphisms are identified from gene sequences and are used as markers to identify and distinguish between different species. This technique, however, is computationally intensive and only identifies unique islands of DNA (Draghici *et al.*, 2005).

Recently, short oligonucleotide signatures of different lengths have also been used to identify bacteria from unknown sequences. This offers a background to understanding and building on the possibilities available for longer oligonucleotides and their advantages over current techniques.

This section discusses the use of various techniques utilised in the identification of bacterial sequences in a metagenomic context. Firstly, the use of 16S rRNA genes and the current limitations of this approach will be discussed. This is followed by discussion of the use of comparative genomic techniques for the creation of oligonucleotide signatures, the use of short oligonucleotide frequency profiles and lastly the proposed use of oligonucleotides of intermediate length for the identification of metagenomic fragments and bacterial species.

1.4.1 16S rRNA identification

The gold standard in identifying organisms is the use of the small ribosomal subunit 16S gene. This sequence seems to be conserved enough that it stays constant within a species and varied enough that it makes identification between species possible. It is also unlikely

to be transferred between species as it is part of the core genome and hence avoids false positives.

Since the 1980's many studies have been done using 16S rRNA genes to classify different species via their evolutionary relationships (Fox *et al.*, 1980). From this point onward phylogenetic studies have used the 16S gene extensively to classify unknown organisms into their appropriate phylogenetic groupings. Identification of organisms via their 16S genes has become a standard in laboratory analysis, and lately extensive effort has been put into developing rapid identification systems for bacteria. *Zhang et. al* (Zhang *et al.*, 2002) reported that there are large numbers of oligonucleotide signatures that can be created from these gene sequences. The limiting factor for this method is the length of the signatures as some are too short for use with modern technologies. Another drawback of this system is that several organisms such as *Bacillus anthracis* cannot be differentiated from their nearest neighbours using 16S genes, as their sequences are highly similar. As a result the alternative laboratory method, Variable-Number Tandem Repeat (VNTR) Analysis, was developed to overcome this difficulty (Keim *et al.*, 2000).

In several cases organisms can have highly similar 16S rRNA genes and still be highly divergent in genomic sequence and in functional roles (Jaspers and Overmann, 2004). This leads to questioning of the overall validity of the use of this technique on a metagenomic sample. Further criticism is based on the fundamental principle that in order to understand divergence of related living organisms one gene does not provide enough information and does not compensate for the differences in evolutionary rates between different parts in the genome (Koonin *et al.*, 2000). This can lead not only to an organism being incorrectly identified but also to the incorrect classification of novel bacteria.

As pointed out, there are several fundamental difficulties with this approach. The most blatant is the low occurrence of the 16S gene, covering only 0.05% of the prokaryotic genome. Therefore loss of this segment would result in loss of signal (Rodriguez, 2002). Even in the partial presence of the gene, roughly half of the gene would be unusable as a sizable portion of the sequence is lost when making an informative alignment (Karlin *et al.*, 1997).

An attempt has been made to avoid these difficulties by using oligonucleotides that occur uniformly throughout the genome, allowing for identification regardless of partial genome loss.

From the above discussion both the strengths and weaknesses of this approach can be assessed. Although the approach is currently still widely used, an alternative will be necessary as the complexity of bacterial environments increases. In this regard methods will need to be found for generating new signature sequences. The next subsection therefore examines a common method for identifying bacterial signatures using comparative genomics.

1.4.2 Comparative genomics: genomic alignments

Comparative genomics is defined as the investigation of the relationship of genome structure and function across varied biological strains (Benson *et al.*, 2007). In the context of this study it refers specifically to the identification of similarity or difference in the genomic sequence of two of more organisms. Comparative methods have been used extensively to identify unique regions within genomes that can be used to identify species in a global context. The practical method used is that of sequence alignment, where genomes are aligned to each other to isolate unique regions of low similarity. These regions are then inspected more closely to identify highly divergent sequences. The alignment of genes is another common method whereby unique polymorphisms can be identified and used as markers (Draghici *et al.*, 2005; Slezak *et al.*, 2003; Phillippy *et al.*, 2007).

In order to identify unique DNA signatures there has to be comparison between the genome in question and all other genomes. This is a very computationally expensive process and unfeasible at the present time. An alternative to this approach uses the phylogenetic background of the organism by searching for differences with only closely related species. These differences are then tested against all other organisms. This approach provides an elegant solution as the largest similarities will be with nearest neighbours and fewer similarities will be present with distant relatives. The significant reduction in computation makes this a commonly used method in comparative genomics (Draghici *et al.*, 2005).

Modern comparative approaches attempt to decrease computation further by reducing search space when looking for unique signatures. In the approach proposed by *Slezak* et al. *(2003)* their first step was to find regions of high similarity between various organisms and exclude these from further searches. This was done by whole genome alignment against the target pathogen to determine which regions of the genome will be least likely to contain unique sequences. This group then created an automated system used for predicting pathogen DNA signatures using their comparative genomics approach. This system is known as KPATH. The insignia signature identification system (Phillippy *et al.*, 2007) is an open source alternative to the KPATH system and performs similarly using open source software and a web interface as a front end (Slezak *et al.*, 2003).

Comparative genomics approaches have been successful in characterisation of unique signature oligonucleotides. However, this method is expensive in both time and computational power. Furthermore, by focusing on *regions* of dissimilarity smaller sub regions are overlooked. This results in signature sequences localized to specific areas on the genome, and implies that the organism can only be identified if these regions are present, thereby drastically reducing usability in a metagenomic context (Draghici *et al.*, 2005).

Given this major disadvantage, an alternative method capable of using the entire genome would be far more effective. In the following subsection the use of short oligonu-

cleotide frequencies in the identification of bacterial sequences is reviewed.

1.4.3 Short oligonucleotide frequencies

Short oligonucleotide frequencies provide a method for complete genome characterisation and analysis. This is done by incorporating information from oligonucleotides spread throughout the genome rather than focusing on specific regions of importance. The different methods for identifying not only genomes but genomic features will be discussed in this subsection.

Research on short oligonucleotide frequencies between 2-4 nucleotides in length have found that they carry strong species specific signals (Karlin and Cardon, 1994; Karlin, 1998*b*; Karlin and Burge, 1995). These signals were detected by a host of different techniques namely neural networks (Abe *et al.*, 2002, 2003), chaos game representations (Goldman, 1993; Deschavanne *et al.*, 1999) and naive Bayesian classifiers (Sandberg *et al.*, 2001).

In their study *Abe* et al. *(2003)* used an unsupervised neural network algorithm self organising map (SOM). The SOM was used to analyse di-, tri- and tetranucleotide frequencies in a large group of prokaryotes to attempt to identify species specific sequence characteristics for these genomes. Their algorithm had success with identifying sequence fragments as short as 1kb and attributing them to a specific species. This gives an indication of how relatively short sequences, present at any place within the genome can be attributed reliably using oligonucleotide frequencies (Abe *et al.*, 2003).

The Bayesian classifier used by Sandberg *et al.* (2001) used a principal component analysis (PCA) of oligonucleotide frequency profiles to identify and visualise the differences in oligonucleotide frequencies between the species. PCA is a vector space transform used to reduce the number of dimensions in a dataset. This enables the classification of data using several of the best dimensions, which in turn reduces unnecessary noise in the dataset. PCA was able to distinguish between 25 different species (Figure 1.2). Sequences as short as 400bp could be correctly classified with an accuracy of 85%. Remarkably, with 60bp sequences the classification accuracy was still able to identify nearly half the fragments correctly with a score of 46%. It was also found that the performance of the classifier and the accuracy with which it classifies largely depends on the length of the oligonucleotide used for the signature. In their tests on oligonucleotide frequencies of different lengths it was found that the longest oligomers, eight-mer and nine-mer gave the best results (Figure 1.3). This leads to the conclusion that the use of longer oligonucleotides can allow for even greater resolution to be attained and could potentially prove a more effective solution.

Another conclusion drawn is that classification does not depend on a few species-specific motifs, but rather on the whole set of motifs. This leads to the understanding

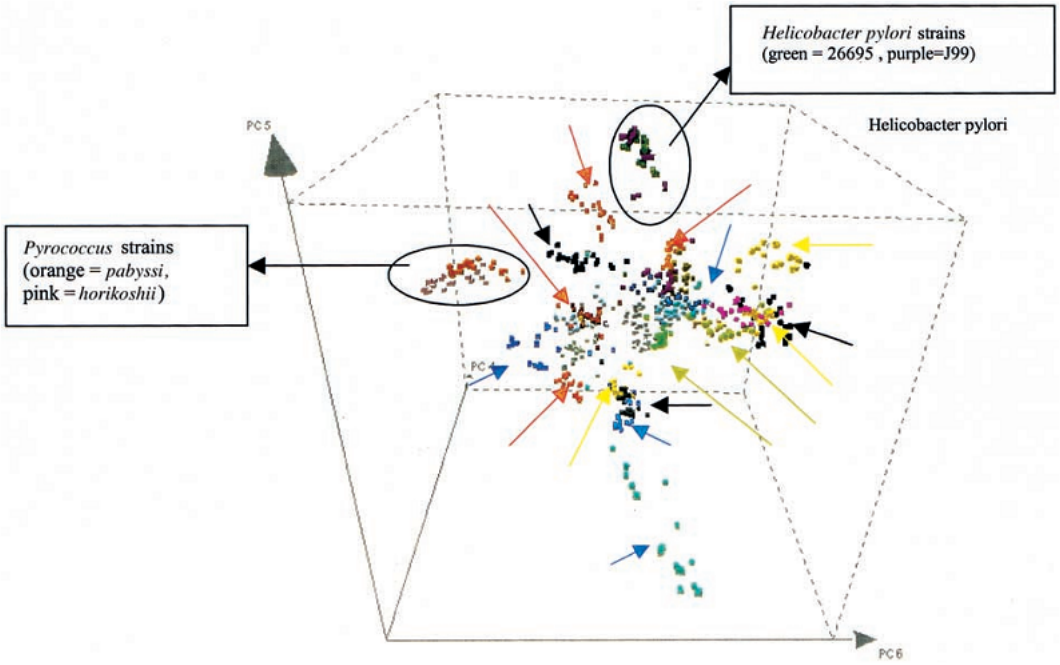


Figure 1.2: Principal components analysis (PCA) was performed on the motif frequencies of 25 genomic species. (Sandberg *et al.*, 2001)

that the oligomers need not be entirely specific for one organism to be part of a functional profile. Tests done to separate closely related strains and species sequences as short as 200 nucleotides, gave a 90% correct classification. This, however, is largely due to the fact that oligomer motifs tend to define closely related species better and that only two different classes need be discriminated (Sandberg *et al.*, 2001).

One of the most promising features for the use of oligonucleotide frequencies rests on the fact that accurate characterisation of the genomic oligonucleotide profile requires only a portion of the genome. This is based on the assumption that intragenomic profile differences are smaller than intergenomic differences (Karlin and Burge, 1995). In an experiment preformed by Sandberg *et al.* (2001) genomic regions were excluded from the training of their algorithm, and random portions of the excluded regions were then used to assess the algorithm's accuracy. The proportion of the excluded sequence was then systematically increased. Surprisingly, when as much as 90% of the genome was excluded, the classifier still produced reliable results (Figure 1.4). These results show the flexibility and power of oligonucleotide profiles in their ability to be generated and used on new, partially-sequenced bacterial species, as well as annotated and completely sequenced species (Sandberg *et al.*, 2001).

Tetranucleotide frequencies have also been employed in the task of identifying genome fragments in a metagenomic context. This technique was tested against the standard sequence methods of fragment identification such as differences in GC content, phylogenetic information and codon bias in functional genes. It was found that tetranucleotide

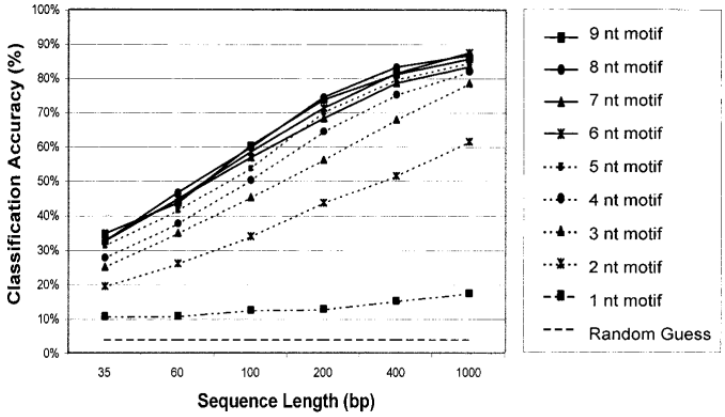


Figure 1.3: The dependence of classification accuracy on oligomer length. *(Sandberg* et al.*, 2001)*

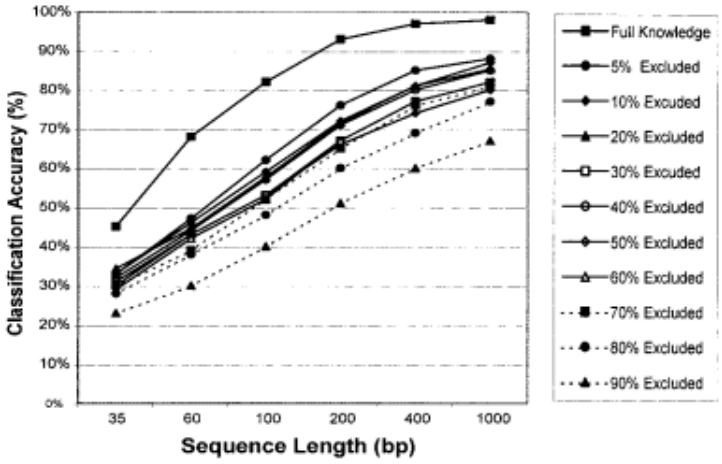


Figure 1.4: Lack-of-knowledge experiments performed by *Sandberg et al* (Sandberg *et al.*, 2001). The genomic percentage of the genome was reduced to only 10% which still provided an effective classifier.

frequencies outperformed these methods. *Teeling* et al. *(2004*b*)* stated that "the discriminatory power of correlations of tetranucleotide-derived z-scores is by far superior to that of differences in (G + C)-content " (Teeling *et al.*, 2004*a*). However, there are several restrictions to this technique. The first is that the intragenomic variation within the genome must be low. This is also one of the main criticisms against GC content. If the GC content varies greatly within a genome it makes identification impossible. The second restriction is in regard to the complexity of the dataset. If there are more than 100 organisms present at equal ratios it makes identification more difficult. However, in certain metagenomic environments with a dominant minority, the noise may be low enough to accurately identify fragments. Lastly, restrictions to this technique are widely affected by fluctuations in base composition from foreign DNA recently incorporated into the genome or highly polymorphic genomes (Teeling *et al.*, 2004*a*).

This again shows the inherent power of oligonucleotide frequencies to identify organ-

isms and the potential to solve some of the current difficulties in metagenomics. It is also clear that several limitations are placed on the use of short oligonucleotides for this purpose. In order to achieve a statistically significant result a sequence must be in excess of 1kbp for tetramers or 400bp for nonamers and octamers. This is, however, not accurate enough, as modern sequencing techniques generate raw sequence reads of only a few hundred base pairs in length. This concludes in a search for longer oligonucleotides that can be more effective in the classification of species via their metagenomic sequencing fragments.

The current study attempts to build on the foundation created by research into short oligomers by utilizing longer oligonucleotides which can be used to identify species specific characteristics in shorter sequences. This method is also computationally inexpensive as the bacterial genome is stored as a frequency table and does not require identification of a specific position within the genomic sequence as is done in sequence alignment methods. It is also possible to generate oligomer profiles using only a portion of the target genome (Sandberg *et al.*, 2001). In the next subsection therefore use of overrepresented oligonucleotides of intermediate lengths is discussed.

1.4.4 Overrepresented oligonucleotides of intermediate length

In the current study longer oligonucleotides, between 8 - 14 bp in length, are analysed to find a solution to the specificity problems of shorter oligonucleotides. The use of overrepresented oligonucleotides provides an alternative approach to identifying bacteria within metagenomic datasets. One of the major benefits of using overrepresented oligonucleotides is that it does not require the investment of time or effort usually associated with the error prone and unreliable task of sequence assembly. This method therefore allows for swifter analysis than currently available methods and can thus provide more timely identification of bacterial pathogens in situations where time-efficient detection is essential. The sequencing technology necessary to utilize overrepresented oligonucleotides is discussed in the next section.

1.5 Sequencing technologies

The sequencing of genomic DNA has always been one of the pivotal points in the furthering of genetic studies. The demand for improvements in sequencing technology are greater now than ever before. This section aims to provide a basic insight into sequencing and the modern technologies currently used.

There are three types of sequencing approaches in use today namely Microelectrophoretic sequencing, Hybridization sequencing and Cyclic-array sequencing on amplified molecules.

The best known of these methods is the Sanger method that falls within the Microelec-

trophoretic sequencing approach. This method involves the creation of a complementary strand of DNA from an existing template. The process uses normal nucleotides together with fluorescent dideoxynucleotides which terminate strand synthesis. This allows for the elucidation of sequence by electrophoresis of various lengths of the synthesized strands with their fluorescent nucleotide markers (Sanger *et al.*, 1977). Several improvements have been made to the Sanger method in the following areas: Fluorescence detection, enzymology, fluorescent dyes and capillary array electrophoresis (Metzker, 2005). Even with the advances to this technology it is too expensive, labour intensive and time consuming to meet the needs of modern researchers.

Hybridization sequencing involves the inferring of sequence by the extent to which a large sample of short oligonucleotides bind the sequence. This can be used to obtain a large amount of sequence data. The drawback to this approach is the possibility of cross-hybridization caused by sequences with high numbers of repeats or chance occurrences of similar sequence (Chan, 2005). This method, although promising, is not currently commercially viable and has not been successfully employed with stringent statistical reliability.

Lastly, the cyclic-array sequencing on amplified molecules method involves multiple cycles of enzymatic reactions on a slide spotted with oligonucleotide features. This method covers millions of features but only a few bases making it useful with multiplexing (Shendure *et al.*, 2004). One of the center pieces and the most successful technology developed so far is pyrosequencing. This approach measures the release of inorganic pyrophosphate from incorporated nucleotides as a proportional enzymatic release of light (Ronaghi *et al.*, 1996). Rather than introducing modified nucleotides that terminate DNA synthesis this approach adds limited amounts of nucleotides, thereby controlling the speed of the reaction. This is done by reintroducing new nucleotides using different enzymatic cycles. The light recorded from the release of inorganic phosphate is then visualized as a set of peaks referred to as a pyrogram. This corresponds to the order in which nucleotides were added and identifies the underlying DNA sequence (Metzker, 2005).

The company 454 Life Sciences has created a commercialized sequencer capable of whole genome sequencing by integrating pyrosequencing with the PicoTiterPlate platform. The PicoTiterPlate platform is a fiber optic faceplate containing thousands of minuscule wells each 40 micrometers wide. The reactions take place inside micro reactors consisting of separose beads containing the DNA molecule and needed enzymes for the reaction. One of the greatest difficulties with this approach centers on homopolymer repeats and the inability of this technique to measure repeats of longer than 5 nucleotides. Another flaw is the possibility of asynchronistic extensions resulting in highly error prone sequencing.

Cyclic-array sequencing on single molecules is a sub category of methods which involves removal of the need for PCR amplification or cloning steps in sequencing. These methods rely on the extension of a primed DNA template on a solid surface using special fluorescent

nucleotides to allow for signal detection. Generally, these methods have very short read lengths, roughly 20-50 base pairs in length (Illumina, 2008). The company Solexa has created such a technique using reversible termination nucleotide bases. This technique allows for continuous elongation of the DNA strand after each base has been read. A drawback to this approach is the overall cycle efficiency which is largely dependent on the chemistry of the reversible terminators being used. For an increase in sequence length significant improvements will have to be made to these reversible terminators (Shendure *et al.*, 2004; Illumina, 2008).

At present the short read lengths (25-300bp) generated by modern sequencing technologies remain a pervasive problem. The modern sequencing approaches have sacrificed read length for an increase in coverage of the base pairs being sequenced. The effect on sequence assembly (without a clone map), however, is dramatic. The greatest difficulty in assembly is the presence of repeat regions, specifically where these regions are longer than the read lengths. As read lengths decrease there is also a higher possibility of finding two highly similar reads from different regions in the genome which cannot be differentiated. This can result in further errors in assembly (Whiteford *et al.*, 2005; Chaisson *et al.*, 2004).

Repeats can result in fragmentation of the overlap-consensus sequence leading to a loss of information and larger number of contig fragments being generated. Chaisson *et al.* (2004) described these difficulties in connection with the assembly of a large, repeat rich bacterial genome using only the sequenced fragments. They stated that "with short reads, assembling a BAC became as complicated as assembling a bacterial genome with normal reads, even when reads had no sequencing errors." They conclude that "substantial (if not prohibitive) finishing efforts are required for resolving entirely all but the simplest of genomes [using short read lengths]" (Chaisson *et al.*, 2004:2068). This outlines the challenge faced by assembly of short reads in bacterial genomes.

Recently some success has been attained in assembling short read lengths, but this remains tentative (Smith *et al.*, 2008). Most modern assembly programs still require large cloned inserts for mapping of short fragments. Although this is a step in the right direction there is no simple solution to the assembly of short reads (Sundquist *et al.*, 2007).

From these findings it is clear that one of the largest drawbacks to modern sequencing techniques is the difficulty in assembly of the short read lengths. This study aims to use raw sequence reads without the necessity of arranging them into contigs. This results in a highly efficient system for the use of modern sequencing techniques and allows for extraction of information without the effort of sequence assembly.

The current study centers on the development of new data-mining approaches to give broader use to high-throughput techniques. Ultimately the goal is to scan raw sequence reads for oligonucleotide signatures to determine which organisms are present within the sample, rendering a marked decrease in computational time. A brief description of previ-

ous work done in this field is given below.

1.6 Previous work on overrepresented oligonucleotide signatures

Previous work on overrepresented oligonucleotide signatures has focused mostly on short oligonucleotide sequences between two and four nucleotides in length. The development and analysis of overrepresented oligonucleotide signatures of between 8-14 base pairs in length is a novel approach which is currently limited to researchers collaborating in the current study. However, Colin Davenport of the Hanover Medical School is the author of several programs involved in overrepresented oligonucleotide discovery which will be discussed below (Davenport *et al.*, 2008).

1.6.1 OligoCounter and Oligoviz: Massively overrepresented oligonucleotides

OligoCounter is a Java command line program which identifies overrepresented oligonucleotides between 8-14 base pairs in length from DNA sequence by counting oligonucleotide occurrences. This program and its methods of identifying overrepresented oligonucleotides have been discussed in Section 1.3.4.

OligoViz and JcircleGraph are visualization techniques created to aid in the identification and the comparison of global sequence features. OligoViz creates a graph showing the presence or absence of a large number of oligonucleotides throughout the genome. This is done using a dot representation to indicate presence or absence of an oligonucleotide within a 10kb window. Patterns can then be identified as horizontal lines or bands within Figure 1.5.

JcircleGraph creates a circular representation of different genome statistics. The four innermost rings are tetranucleotide parameters with the remaining rings identifying characteristics in overrepresented 8-14mers (Figure 1.1) (Davenport *et al.*, 2008).

These programs provide evidence for the existence of overrepresented oligonucleotides and further visual proof that they correlate well with other genomic sequence features. This alludes to the further use of overrepresented oligonucleotides as a genomic signature that will provide a reliable and sensitive diagnostic.

1.6.2 Conclusion

Pathogenic bacteria are still one of the leading forms of mortality in the world today. Millions of deaths each year are connected with *M. tuberculosis* alone. Although there are different mechanisms of pathogenicity and different degrees of virulence, the majority of

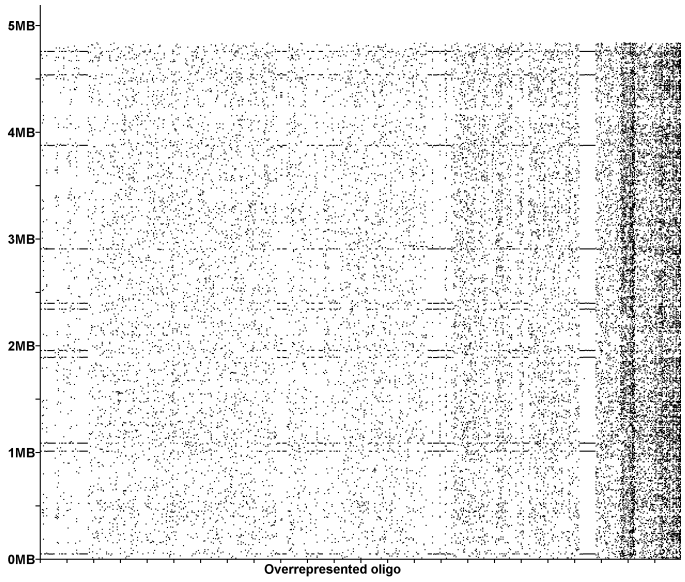


Figure 1.5: The top 600 overrepresented oligos in the *Mycobacterium avium K10*. Overrepresented oligos are represented on the X-axis and sorted by descending overrepresentation index (chi-squared), with genome position on the Y-axis. Note the absence of oligos otherwise found distributed throughout the whole *M. avium* genome at 4.2MB. This corresponds to the position of strongly divergent GC content and may indicate a genome island. Repeat regions are present as horizontal bands containing many similar oligos. (Davenport *et al.*, 2008)

pathogens have identifiable pathogenic sequence features. These are termed pathogenicity islands (PAIs) and are found within their core genomes.

Pathogenic bacteria are present in diverse environments where non pathogenic bacteria also reside. By analysing DNA taken directly from an environment, complications in culturing bacteria can be avoided. However, this comes at the cost of having to sort and filter different unknown genomic fragments. In order to improve diagnostic and preventative measures as well as advance understanding of bacterial communities, new ways of identifying bacteria using metagenomic sequence fragments must be discovered.

Bacterial genomes contain a vast diversity of sequence patterns and will take many text books to describe comprehensively. Sequence patterns can be utilised to extract a large amount of information from a genome. The simplest example is the bias each species has for specific codons to code certain amino acids. However, sequence patterns are present outside of coding regions and these can indicate genomic features such as transcription binding sites or structural complexes. Short oligonucleotide frequencies (2-4bp) are a pervasive and informative method of retrieving information about surrounding sequence features and classifying genomes and their sequenced fragments. This study will attempt to apply this process to longer oligonucleotides of 8-14bp in length in an attempt to identify bacterial genomes within a metagenomic context.

Several different methods currently exist to identify bacteria from metagenomic data.

The most well known method is the use of 16S rRNA phylogenetic gene markers. These markers can be specifically attributed to a single species and allow for classification of unknown organisms according to comparative analysis of the gene. Although this approach has proved invaluable its disadvantages center around the reliance on the short localised region that contains this gene. This is especially the case in a metagenomic context where only 1% of the reads contain these sequences (Handelsman, 2004). It is clear that further signatures must be sought in order to provide better genome coverage.

Comparative analysis is a common technique used to identify regions of similarity or dissimilarity and has been used extensively to identify unique oligonucleotide signatures. However, this approach suffers from the same disadvantage as the former. Signatures are localised to specific regions of low similarity and hence do not provide the coverage necessary for optimum identification in a metagenomic context.

Short oligonucleotide frequencies (2-4bp) are an alternative to these approaches. This technique relies on sequence features maintained throughout the genome and considers the combined occurrence of all oligonucleotides as a signature. Application of the technique results in a departure from a conventional definition of a genome signature. Instead of relying on wholly unique sequences the profile of occurrence of several less specific oligonucleotides is used. This sacrifices sensitivity for a much broader ability to identify genomic fragments regardless of genomic position. The approach has been used successfully in the identification of species fragments from 1kb to as short as 400bp in length. The limitation of this method, however, is the lack of specificity which can be seen in the use of sequences shorter than 1kb.

With the evolution of sequencing, the decrease in cost and increase in the number of bases being sequenced has greatly aided the feasibility of modern metagenomic techniques. However, this exponential increase in speed and volume has come at the cost of sequence length. At present the most modern sequencing technologies generate short sequenced reads between 25 to 300 base pairs in length. This leads to further complication, in that for these reads to be useful they have to be attributed to a specific species or assembled via sequence alignment techniques. This becomes problematic in the case of repeats and highly similar sequences which can result in several different species fragments being attributed to a single genome. This ultimately results in the incorrect assembly of genomes.

The current study centers around the use of longer (8-14bp) overrepresented oligonucleotides to recognise bacterial species in unknown sequence. The study proposes an approach to identify bacteria within a metagenomic sample by optimum utilization of modern sequencing technologies without the need for sequence assembly.

Having outlined the foundations upon which the current study has been based, this chapter concludes by setting out the problem statement and aims of the study.

1.7 Aims and problem statement

1.7.1 Problem statement

The identification of pathogenic bacteria within environmental samples is a complex and laborious task. Sequencing of environmental samples results in a mixed bag of sequences from different species with further constraints such as read length limitations hampering attempts to separate and assemble these sequences. Current marker methods for this purpose are deficient largely due to the localization of markers. There is no guarantee that these markers are completely unique or accurate in identifying species as they refer only to specific regions and cannot provide an overall assessment of the genome. It is essential that alternative approaches are found which allow for identification of bacterial species regardless of partial genome loss and will enable classification of the majority of genomic fragments pertaining to a species. Overrepresented oligonucleotides (8-14bp) provide a solution to this problem as they occur throughout the genome and can be used without the need for sequence assembly.

1.7.2 Aims

In order to develop the proposed method for testing and analysis of bacterial genomes within a metagenomic sample the following aims were set:

- To create and populate a structured, relational database, including parsing and analysis of raw overrepresented oligonucleotide data (Chapter 2)
- To create a program for querying the database and performing analyses (Chapter 2)
- To identify overrepresented oligonucleotide marker profiles for the purpose of classifying each bacterial genome in an unknown metagenomic sample (Chapter 2). This aim includes the following objectives:
  - To identify closely related strains of the same species using overrepresented oligonucleotide data, thereafter, the identification of species and lineage specific overrepresented oligonucleotide marker profiles
- To test species and lineage specific overrepresented oligonucleotides on artificial and real metagenomic datasets (Chapter 3)

These aims will allow for the filtering and analysis of raw data to define the best overrepresented oligonucleotide candidates for elucidation of species within a metagenomic context. Furthermore, testing of overrepresented oligonucleotides using metagenomic data will help

determine the usability of this approach. These results can then be used in further investigation into formal methods of identification and classification using overrepresented oligonucleotides of intermediate lengths.

Chapter 2

Database development and implementation

2.1 Introduction

In the previous chapter three main aims were presented for the development and implementation of a database and program in order to achieve the final objective of testing overrepresented oligonucleotides on metagenomic datasets. The aims required to reach this goal include: development of a structured database and subsequent creation of a program and the identification of overrepresented oligonucleotide marker profiles to enable the classification of bacteria from an unknown metagenomic sample. This chapter deals with the methods used to realize these aims and the outcomes associated with their implementation.

The need for creation of a database and corresponding program was due to the nature of the available data. Overrepresented oligonucleotide information was available for a collection of species within each phylogenetic grouping. However, this information remained in an unstructured form, containing not only species specific overrepresentation but widespread overrepresentation. This information was not usable to identify bacterial species because:

- Many overrepresented oligonucleotides were widely shared between species and therefore could not aid in the separation of species
- There was no measure in place to identify the extent to which each oligonucleotide was uniquely overrepresented in a species
- No statistical measures were available to estimate the occurrence of an oligonucleotide within a random metagenomic sample
- There was no way to define if a bacterial species is indeed present in a sample using this data

In order to address these obstacles well planned structures and methods must be put in place to integrate and compare oligonucleotide information. In order to generate useful results, the raw information was therefore analysed, filtered and stored. This not only enabled recognition of candidates for species identification but allowed for efficient searching methods as well as complex comparisons. In this chapter a program, *OligoSignatures*, for the discovery of species and phylogenetic-lineage specific oligonucleotides is presented. Over 439 completely sequenced genomes are available for analysis in this system's database. The focus of the study is placed on species identification in unknown metagenomic environments.

The first step in the process of database development was the parsing and handling of raw data. This allowed for an evaluation of available information and calculation of statistics to better describe each oligonucleotide. A database was then constructed, containing 22 phylogenetic lineage tables, in order to provide an infrastructure for the handling and integration of the parsed data. Further oligonucleotide discovery was then implemented using genomic sequences as oligonucleotide data lacked consistency over different species.

After population of the database it was possible to identify a list of overrepresented oligonucleotide markers for species and lineage detection. A question was raised as to the effect of different strains of the same species being present in the same lineage. This would impact the statistics used to identify candidate markers not only within the species but throughout the lineage group. A test was then performed to gauge the similarity in oligonucleotide occurrence to quantify this hypothesized effect.

Species specific oligonucleotide markers were selected by calculation of a score for each oligonucleotide within a species. From the score it was possible to examine how uniquely overrepresented an oligonucleotide was in a lineage. To avoid redundancy, oligonucleotides were checked for similarity to ensure that they did not form part of the same repeat region.

The next analysis done was the identification of lineage specific oligonucleotides. This involved the identification of the most overrepresented oligonucleotides within the lineage. The resultant oligonucleotides were then filtered to remove redundancy.

Although analyses performed were by no means extensive, they did allow for an insight into the use of overrepresented oligonucleotides. The ability to identify species specific overrepresented oligonucleotides provides an opportunity to explore a new method of species identification. Furthermore, by identifying lineage specific oligonucleotides an attempt can be made to identify a broader phylogenetic group within an environmental sample. With the creation of the proposed program, these techniques can be readily applied to any metagenomic situation to provide the researcher with the versatility needed to discover methods for the identification of species using overrepresented oligonucleotides.

2.2 Collection and input of raw data

OligoSignatures was constructed in Python 2.4.4. Each analysis was run under a central class which allowed for customization of all parameters and returned information which could be further manipulated or stored by the user. Every analysis could be run on a standard personal computer (using either Linux or Windows operating systems), as part of the central program or as a standalone program. The larger analyses were submitted to BLART, a Linux server system with 8GB of RAM.

In this section, the initial creation and population of the database is discussed, beginning with a description of the data followed by basic parsing and analysis methods used. The construction of the database, its division into smaller tables and further analyses is then explained.

2.2.1 Data source

The first step in creation of the database was to parse and analyse the raw information received from OligoCounter (Davenport *et al.*, 2008). These files consisted of overrepresented oligonucleotides found within unannotated FASTA files obtained from the NCBI FTP site in April 2007 (Benson *et al.*, 2007). These files contained over 538 sequenced bacterial genomes.

A new format was utilized for the representation of oligonucleotides namely denary format. This format converted the letters of an oligonucleotide into easily stored decimal numbers where the order of bits encodes the initial sequence. This was done to increase the speed of database searches as digits are far more efficiently retrieved than characters.

Each OligoCounter output file contained: a FASTA heading, the overrepresented oligonucleotide in normal text and in denary format, its frequency within the genome and its positions of occurrence throughout the genome. This data was parsed into the python script and imported into the database.

2.2.2 Data parsing

The initial step in database creation involved parsing and preliminary analysis of the data. The raw OligoCounter files were parsed into a python script where the complements of each oligonucleotide were identified and integrated into a single oligonucleotide entry, this decreased redundancy within the database. For each oligonucleotide entry two statistics were calculated, namely; expected value per 100kbp and coefficient of variation. The expected value per 100 kilo base pairs (100kbp) (Algorithm 1) of an oligonucleotide was an attempt to estimate the average occurrence of an oligonucleotide, for a particular species, within 100 kbp of randomly selected genomic sequence. The following factors were taken into account

- The frequency of the oligonucleotide within the specified genome
- The spread of the oligonucleotide throughout the genome

Algorithm 1 Expected value per 100kbp.

Expected value per 100kbp $= \frac{100000}{\sqrt{\bar{\mu}^2+\sigma^2}}$

$\bar{\mu}$ indicates the average fragment lengths
σ^2 indicates the standard deviation of the fragment lengths

This algorithm relied on the positions of occurrence of each oligonucleotide by creating genomic fragments that span from the start of one occurrence to the start of the following occurrence. These fragment lengths were then used to determine the uniformity of spread (the standard deviation of the fragment lengths) and the average distance between fragments.

The coefficient of variation was then calculated and the formula is given in Algorithm 2.

Algorithm 2 Coefficient of variation.

Coefficient of variation $= \frac{\sigma}{\mu}$

μ indicates the average fragment lengths
σ indicates the standard deviation of the fragment lengths.

This identified the amount of variation in fragment lengths and hence the spread and uniformity of the oligonucleotide throughout the genome. This statistic allowed for assessment of oligonucleotide distribution at a glance as large values indicate low spread while smaller values indicate uniform distribution. Once the data had been parsed into the python script and recomputed it was imported into the database.

2.2.3 Database creation

There are multiple applications for management and construction of a database. The most commonly used systems include PostgreSQL, Oracle and MySQL. The database constructed in the current study was created using MySQL Enterprise Server 5.1. MySQL is an open source SQL relational database management tool available at http://www.mysql.com/. MySQL operates on all major operating systems and has a large community of users providing comprehensive support. MySQL has a number of interfaces to different programming languages including Python.

In this study the database interface module SQLAlchemy 2.4.3 was used for the majority of communication with the MySQL database. This module enhances the usability

Table 2.1: The number of genomes per lineage table.

Database Lineage Number	Lineage Description	Number of genomes
1	*Acidobacteria*	2
2	*Actinobacteria*	35
3	*Alphaproteobacteria*	54
4	*Aquificae*	1
5	*Bacteroidetes/Chlorobi*	10
6	*Betaproteobacteria*	36
7	*Chlamydiae/Verrucomicrobia*	11
8	*Chloroflexi*	2
9	*Crenarchaeota*	7
10	*Cyanobacteria*	19
11	*Deinococcus-Thermus*	4
12	*Deltaproteobacteria*	14
13	*Epsilonproteobacteria*	11
14	*Euryarchaeota*	23
15	*Firmicutes*	96
16	*Fusobacteria*	1
17	*Gammaproteobacteria*	100
18	*Nanoarchaeota*	1
19	*Other Bacteria*	1
20	*Planctomycetes*	1
21	*Spirochaetes*	9
22	*Thermotogae*	1

of databases by creating a simple and powerful interface that allows for dynamic and complex queries in a simple and elegant manner. These feats are achieved through relational mapping of the database to constructs created in Python.

Three tables were created to house all the data retrieved from OligoCounter, namely, the Accessions table, Oligos table and Instances table. The Accessions table contained information regarding characteristics of each genome such as genomic sequence length, genome description and GenBank accession number. The Oligos table consisted of information regarding oligonucleotide characteristics such as oligomer melting temperature and GC content. The Instances table contained each oligonucleotide entry with its positions of occurrence for all genomes within the database.

2.2.4 Database division, analysis and completion

Although the Instances table housed records for all genomes within the database further division of data was required to allow for efficient access to information. The Instances table was divided according to 22 phylogenetic lineages as identified by GenBank (Benson *et al.*, 2007). The description and the number of genomes within each lineage is listed in table 2.1.

The number of unique oligonucleotides in each lineage was referred to as the lineage oligonucleotide template. To allow for effective comparison between species within the lineage each species had to conform to the lineage oligonucleotide template. This template was applied to each genome within the lineage so that each genome contained an entry for each oligonucleotide found in the template. If an oligonucleotide was not present, the genome file was searched for these oligonucleotides and entries were then appended to the lineage table of the corresponding genome.

After completion of this process 22 lineage tables were present along with the two additional general information tables, the Accessions and Oligos tables. The completed structure of the database is shown in Figure 2.1. Each table will now be briefly described.

The Accessions table contained information relating to the genome such as: the genbank accession number (Accession_number), the species description (Genome_description), genome sequence length (Genome_length) and lineage in which the genome occurs (Lineage_id).

Each Lineage table consisted of the following information: A number used to describe each entry (Instance_id), the oligomer recorded in denary format (Oligomer_denary), the positions of occurrence of the oligomer (instances), the accession_number of the genome it originates from (Accession_id), the frequency of occurrence of the oligomer within the genome (Frequency), the coefficient of variation (Coefficient_variation), the expected value per 100kbp (Expectation_per_100kbp) and whether the reverse compliment had been included in this entry (Reverse_compliment).

The Oligos table is comprised of the following information: the oligomer recorded in denary format (Oligomer_denary) ,the oligomer recorded in normal characters (Oligomer_str), the length of the oligomer (Oligo_length), a hash-table value for calculation of oligomer neighbours differing by several nucleotides (descriptor), the GC content of an oligomer (GC_content), the melting temperature of an oligomer according to two different methods, the Wallace rule (Melt_temp_wall) and the nearest neighbor method (Melt_temp_nn). The Wallace rule (Wallace *et al.*, 1979) was easily calculated but was only usable on 13-20mers while the nearest neighbor method (Wetmur, 1991) could be used on all oligonucleotides within the 8-14mer length range.

2.3 Approaches for database analysis

This section explains the statistical parameters and development of scoring mechanisms for species specific and lineage specific oligonucleotide analysis. These statistics were critical to the effective functioning of each analysis. A great deal of effort was invested to find the most accurate and consistent methods for use in species and lineage specific analysis.

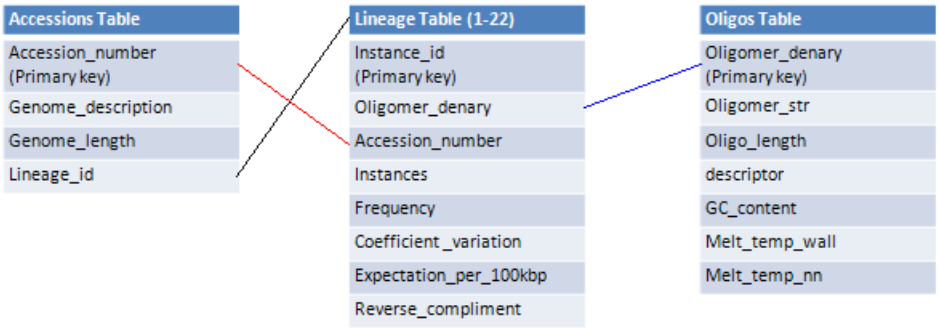


Figure 2.1: Structure of the created database. The Accessions table included all general genome information while the Oligos table included the properties of each oligonucleotide found within the database. There are 22 lineage tables containing information on over-represented oligonucleotide frequency and positions of occurrence within each genome in the lineage. The links between tables indicate the foreign keys.

2.3.1 Scoring function for calculation of species specific oligonucleotides

In order to identify whether an oligonucleotide was species specific a score was assigned based on its ability to separate one species from its lineage neighbours. This value aimed to identify the degree of difference between the expected value of the oligomer in the present genome and other expected values for that oligomer in the lineage in order to rank oligonucleotides which are overrepresented in only a few species.

In an attempt to calculate this score the expected values of several oligonucleotides were plotted from *Actinobacteria* (Lineage 2) (Figure 2.2). The resultant graph approached the shape of an exponential distribution. Nonetheless, the graph shapes were inconsistent and after hypothesis testing it was concluded that the data did not fit any standard distributions and an alternative statistic would have to be found.

Four candidate algorithms were evaluated to identify the best measure. The first two measures (Algorithm 3) both relied on the variance of the data to differentiate between candidates. This approach had several short comings in that the data was not normally distributed and did not fit any standard distribution. The first algorithm (Algorithm 3, Scoring function 1) favored datasets where several values were abnormally higher than the others and so did not select the words which best described a small sample of genomes. The second scoring function (Algorithm 3, Scoring function 2) tried to decrease the effect of the standard deviation but ultimately provided poor results.

The third and fourth scoring functions (Algorithm 4) relied on a ratio based approach. The selection criteria for these two functions worked correctly and awarded the most specific oligomers the highest score and the less specific oligomers with lower scores depending on the number of species containing these abundant oligonucleotides. The fourth algorithm (Algorithm 4, Scoring function 4), however, provided a statistical meaning which

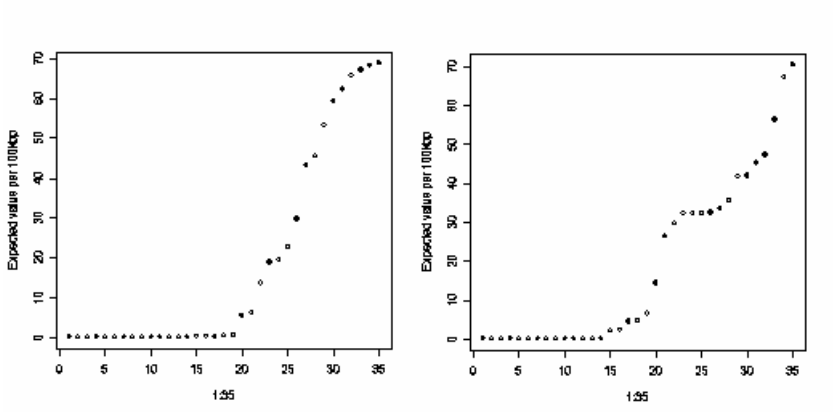


Figure 2.2: Expected value plots for two overrepresented oligonucleotides. Figure A and B show the inconsistent shape of the dataset plots for oligonucleotides "GCGCGCGCC" and "GCAGCGCGG" respectively. The X-axis shows the expected values for 35 genomes within the *Actinobacteria* lineage sorted in ascending order. The Y-axis shows the numerical value of the sorted expected values.

Algorithm 3 Species specific algorithm 1 and 2.

$$\text{Scoring function 1} = X_j \times \left(\frac{\sum_{i=1}^{n} X^2}{N} - \left(\frac{\sum_{i=1}^{n} X}{N} \right)^2 \right)$$

$$\text{Scoring function 2} = X_j \times \sqrt{\left(\frac{\sum_{i=1}^{n} X^2}{N} - \left(\frac{\sum_{i=1}^{n} X}{N} \right)^2 \right)}$$

X indicates the expected value for a genome
X_j indicates the expected value for species j
Assume N is the number of species in each lineage
i indicates each species

could be used to help quantify results. In statistical terms this value denoted the difference in numerical value from the current expected value to the average expected value calculated for the dataset.

2.3.2 Lineage specific oligonucleotide analyses

In order to identify oligonucleotides which occur abundantly within a single phylogenetic lineage two criteria needed to be satisfied. Firstly, the most generally overrepresented oligonucleotides needed to be identified. This was achieved using the inverse of Scoring function 1 for species specific oligonucleotides and is shown in Algorithm 5.

The second criterion was to determine whether the most commonly overrepresented oligonucleotides for a specific lineage were in fact highly overrepresented in other lineages. For this purpose oligonucleotides were retrieved from all the lineage tables within the database. A dataset of expected values was then created for each oligonucleotide in each lineage, if the oligonucleotide did not occur zero values were added.

Several scoring functions were created to select the oligonucleotides which best repre-

Algorithm 4 Species specific scoring function 3 and 4.

$$\text{Scoring function 3} = \frac{\left\{\sum_{j=1_{j\neq i}}^{n}\left(\frac{X_i}{X_j+1}\right)\right\}}{n-1}$$

$$\text{Scoring function 4} = \frac{X_i \times \left\{\sum_{j=1_{j\neq i}}^{n}\left(\frac{X_i}{X_j+1}\right)\right\}}{n}$$

X indicates the expected value for a genome
i indicates the species for which the score is being calculated
Assume N is the number of species in each lineage
j indicates a species with the lineage

Algorithm 5 Common oligonucleotide algorithm. This algorithm identifies oligonucleotides that are wholly overrepresented in all genomes within the lineage. This algorithm is not statistically significant and functions merely as an indicator as the data distribution is not normal.

$$\text{Common oligonucleotide algorithm} = \frac{\bar{X}}{\left(\frac{\sum_{j=1}^{N} X^2}{N} - \left(\frac{\sum_{j=1}^{N} X}{N}\right)^2\right)+1}$$

X indicates the expected value of an oligonucleotide in a specific species
$\bar{X}$ indicates the average expected value of an oligonucleotide within the lineage
j describes all genomes in the lineage
N indicates the total number of values within each dataset

sented each lineage. The first two algorithms (Algorithm 6) were an attempt to utilize the standard deviations of the expected values within one lineage and that of all the expected values from all other lineages to provide an overall ratio (Algorithm 7). The first algorithm proved inconsistent and could not provide well ranked results. This is due to the overbearing effect of the standard deviation. In the second scoring function an attempt was made to subdue this effect but results did not improve dramatically. The last two algorithms (Algorithm 8) used closely resemble the ratio algorithms used to identify species specific oligomers and use the average expected values for each lineage to determine the most descriptive oligomers for each lineage. Scoring function 4 was therefore selected due to similar reasons as in Section 2.3.1.

2.4 Database analyses

In this section the primary use of the database is discussed via different analyses each performed with a different goal in mind. The first analysis was the confirmation of strains of the same specie, via a euclidean distance approach. This endeavored to determine whether strains belonging to the same species were similar enough in terms of overrepresented oligonucleotide data that only one representative member should be included in further analyses. Another purpose of this analysis was to provide insight into the

Algorithm 6 The standard deviation calculations.

A. $\sigma_i^2 = \frac{\sum_{j=1} X_{ij}^2}{n_i} - \left(\frac{\sum_{j=1}^{ni} X_{ij}}{n_i}\right)^2$

B. $\sigma^2 = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} X_{ij}^2}{m \times \sum_{i=1}^{m} n_i} - \left(\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} X_{ij}}{m \times \sum_{i=1}^{m} n_i}\right)^2$

σ_i^2 *denotes the variance of the expected values in a specific lineage*
σ^2 *denotes the variance of all the expected values in all lineages*
Assume m *lineages*
Assume lineage i *has* n_i *genomes*
X_i *denotes the the average of lineage* i
X_{ij} *denotes the expected value for genome* i *in lineage* j

Algorithm 7 Lineage scoring function 1 and 2.

Lineage scoring function 1 $= \bar{X}_i \times \frac{\sigma^2}{\sigma_i^2}$

Lineage scoring function 2 $= \bar{X}_i \times \frac{\sqrt{\sigma^2}}{\sqrt{\sigma_i^2}}$

$\bar{X}_i$ *denotes the average of lineage* i
σ_i^2 *denotes the variance of the expected values in a specific lineage*
σ^2 *denotes the variance of all the expected values in all lineages*

limitations of overrepresented oligonucleotides to distinguish between species.

The second analysis performed was the determination of species specific oligonucleotide candidates. This step required the information regarding representative strains from the previous analysis to offer an unbiased estimate for all species within the lineage table.

The third analysis available was the determination of lineage specific oligonucleotide marker candidates. This identified which oligonucleotides best described each lineage and could therefore be used to distinguish lineage members from non-members. This approach also provided insight into the most commonly used sequences within the lineage and this in turn could hint at lineage specific sequences as a topic for further research. Each analysis method is discussed below in more detail.

2.4.1 Strain distance analysis

Strain analysis was undertaken as the database contained multiple strains and several lineages contained poorly labeled genomes which could not be identified as either separate species or strains of the same specie. This method involved the determination of how closely related the proposed strains in the database were to each other, thereby limiting the burden of multiple strains decreasing word effectivity. This process focused solely on the similarity of occurrence of overrepresented oligonucleotides between two species. It also allowed for accurate selection of the most common strain on the basis of oligonucleotide content. An initial preparation phase involved identification of all strains of the same species (including poorly labeled strains) within the database and the analysis step

Algorithm 8 Lineage scoring function 3 and 4.

$$\text{Lineage scoring function 3} = \frac{\sum_{j=1_{j\neq i}}^{m}\left\{\bar{X}_i - \bar{X}_j\right\}}{m-1}$$

$$\text{Lineage scoring function 4} = \frac{X_i \times \left\{\sum_{j=1_{j\neq i}}^{m}\left(\frac{\bar{X}_i}{\bar{X}_j+1}\right)\right\}}{m}$$

Assume m number of lineages
$\bar{X}_i$or $\bar{X}_j$ denote the the average of lineage i and j respectively
i denotes the lineage for which the score is being calculated

determined how similar these strains were using initial species specific oligonucleotide results. After a group of strains was confirmed, the most common strain was selected as the species representative.

2.4.1.1 Analysis of strain similarity

For each lineage the following processes were executed: For each genome within the lineage, the 1000 top overrepresented genome specific oligonucleotides (identified using the species analysis method, Section 2.4.2) were selected as its oligonucleotide profile. Each genome was then compared to every other genome to quantify the difference in overrepresentation of their oligonucleotide profiles. This calculation was performed using an euclidean distance measure.

The final selection list used to calculate the euclidean distance measure (Algorithm 9) was generated by selecting the top 500 oligonucleotides from the first genome's oligonucleotide profile. Following this, 500 oligonucleotides were selected from the second genome's oligomer profile, not present in the 500 oligonucleotides already selected. This process allowed for an efficient assessment of similarity between two genomes' oligonucleotide profiles by including the best oligonucleotides from each genome.

Algorithm 9 Euclidean distance measure.

$$\text{Euclidean distance measure} = \frac{\sum(X_A - X_B)^2}{1000}$$

$X_A$and X_B indicate the expected values for an oligonucleotide from organism A and B respectively.

For each lineage a table of distance measures was constructed (Table 2.2). This was then visualized by plotting of scores using a python module, Rpy 1.1, which interfaced with the statistical language R. These diagrams (Figures 2.3, 2.4,2.5 and 2.6) showed the relation of strains to the background data enabling an accurate estimation of overall distance. From the plots it is possible to visually identify how similar each genome pair is based on their calculated distances. In the majority of cases strain distances were found with a value lower than 50 distance units. For this reason the cutoff was set empirically at

Table 2.2: A reduced example of a distance table for *Chlamydiae/Verrucomicrobia*. Strain group 1 represents the *Chlamydia trachomatis* species. Strain group 2 represents the *Chlamydophila pneumoniae* species.

	Strain Group 1	Other Specie	Strain Group 2	Strain Group 2
Strain Group 1	0	82.66353345	67.01371382	67.53690649
Other Specie	82.66353345	0	59.97550241	59.73495635
Strain Group 2	67.01371382	59.97550241	0	2.266643156
Strain Group 2	67.53690649	59.73495635	2.266643156	0

50 to distinguish whether genomes were in fact similar enough to remove for the purposes of species identification.

2.4.1.2 Strain distance plots

For each lineage within the database a euclidean distance table was created. This table enabled the visualization of the distances between strains as a one-dimensional graphical plot where the Y axis was only used to distribute the data points for easy viewing. From these plots it was possible to learn a great deal not only about the strains being studied but about the relationships within the lineage as a whole.

Four plots are described in this section. Each plot contains several colored squares indicating distances between strains of the same species IE. each colored square represents a species. The interspecies distances are represented as red crosses in the data. The X-axis displays the calculated euclidean distance values.

The strain plots for *Chlamydiae/Verrucomicrobia* (Figure 2.3) and *Firmicutes* (Figure 2.4) showed a clear distinction between strains and other species within the lineage. In *Chlamydiae/Verrucomicrobia* there were 3 distinct levels of phylogeny present. Starting from the left there was a concentrated collection of strains with values between 0-20 representing mainly intraspecies distances, followed by a main body of organisms from 40-90 and a minority of distantly related species around 250 representing interspecies distances.

A similar pattern could be seen in *Firmicutes* (Figure 2.4) although the interspecies/intraspecies separation was far more vague. This could be explained by the large difference in genome count. The *Chlamydiae/Verrucomicrobia* lineage table contained only 11 genomes where *Firmicutes* contained 96. There was still, however, a clear separation between the strains from 0-40, the main body of species between 50 - 400 and the distant organisms between 500-800. An observed overlap in distances between intraspecies and interspecies could be explained as closely related organisms or misidentified bacterial species. It was also interesting to note the difference in distances between *Chlamydiae/Verrucomicrobia* and *Firmicutes*. *Firmicutes* depicted almost treble the distance demonstrating a much larger diversity of organisms.

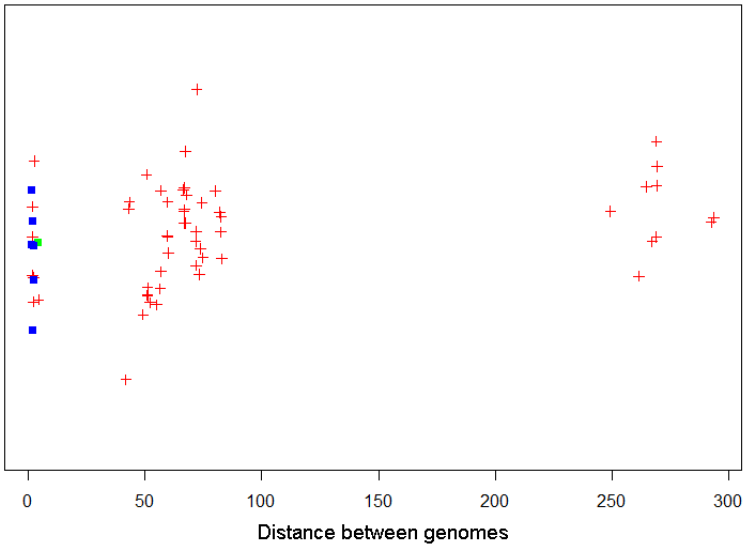


Figure 2.3: Plot of euclidean distance measures for strains in *Chlamydiae/Verrucomicrobia*. Colored squares indicate distances between strains of the same species. The interspecies distances are represented as red crosses.

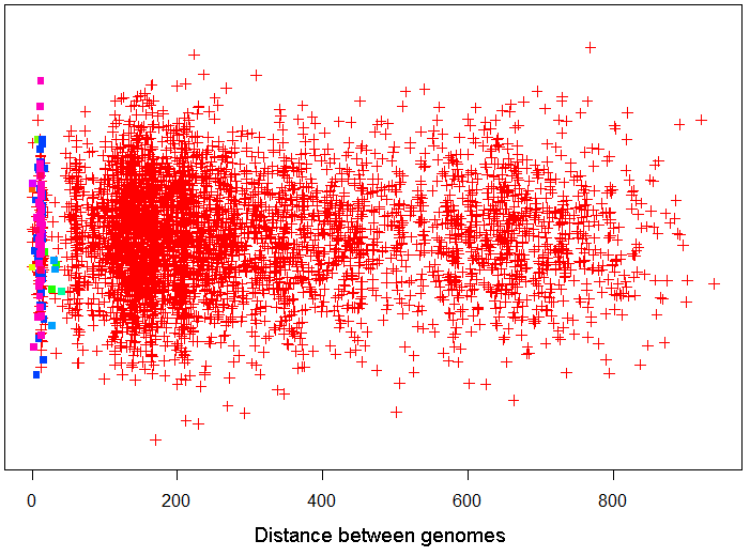


Figure 2.4: Plot of euclidean distance measures for strains in *Firmicutes*. Colored squares indicate distances between strains of the same species. The interspecies distances are represented as red crosses.

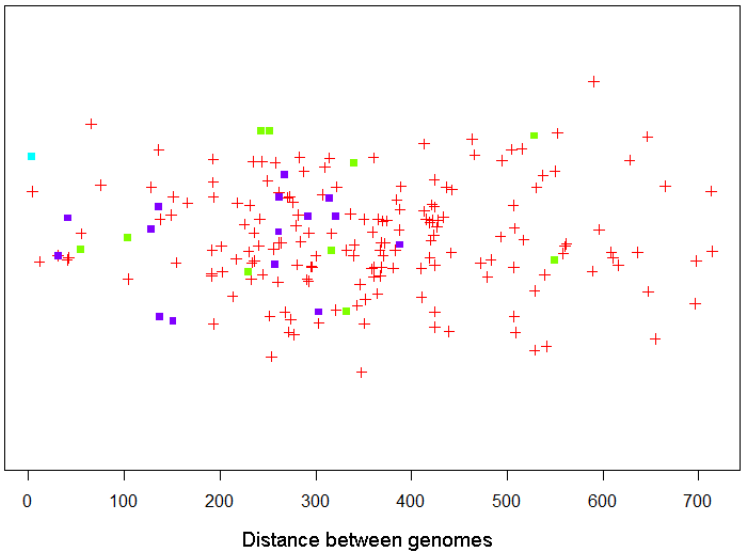


Figure 2.5: Plot of euclidean distance measures for strains in *Cyanobacteria*. Colored squares indicate distances between strains of the same species. The interspecies distances are represented as red crosses.

However, the *Cyanobacteria* (Figure 2.5) and *Gammaproteobacteria* (Figure 2.6) strain plots showed definite irregularities. *Cyanobacteria* displayed a confused representation of strain and species data. This indicated that the proposed strains were clearly not closely related enough and may not be strains of the same species. Therefore these genomes could not be excluded from analysis. A note must be made of the magnitude of distances present within this plot. *Cyanobacteria* contained only 19 species but showed almost as much variation as *Firmicutes* with the general body of organisms between 150-500 and distant relatives with values as high as 700. Reasons for this can include the vast taxonomic diversity of *Cyanobacteria*.

The plot describing *Gammaproteobacteria* showed no clear separation between strains and the main body of species. This can be partially explained by the large number of organisms present within this lineage. Furthermore, the lack of separation between intra- and interspecies distances could also indicate several highly similar species were present within this group such as *Escherichia coli, Shigella flexneri* and different *Salmonella* species. Particular note should be taken of the length of the tail in the plot. There were at least four groups of distantly related species clustering together. This could be further highlighted by the excessive distance of the most distant group with values from 2000 to 2500. As the largest table in the database with over 100 genomes, it stands to reason that this would also be the most cluttered and diverse. This diversity could be related to the diverse environments and large variety of phenotypic traits held by different families of these bacteria.

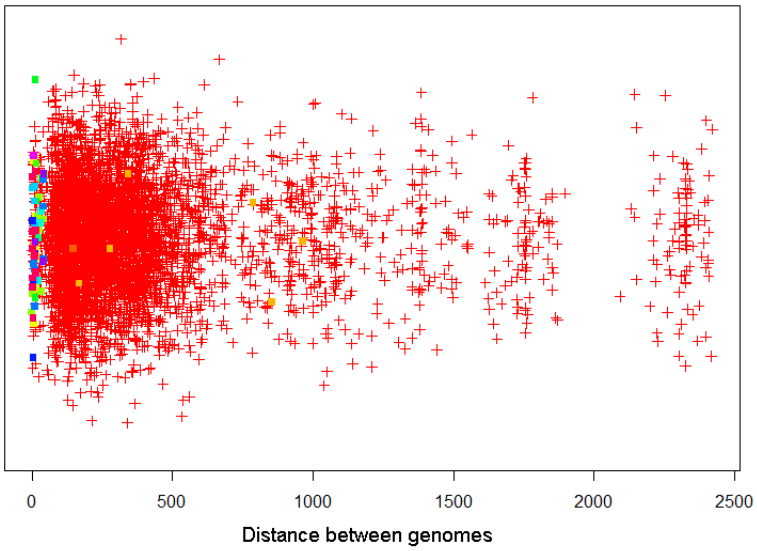


Figure 2.6: Plot of euclidean distance measures for strains in *Gammaproteobacteria*. Colored squares indicate distances between strains of the same species. The interspecies distances are represented as red crosses.

2.4.2 Determination of species specific oligonucleotide markers

In order to identify species specific overrepresented oligonucleotides comparisons must made with all species within the lineage table (Algorithm 4, scoring function 4). This could provide an accurate insight into differences between members, allowing identification of species within an unknown environmental sample.

The approach used in this analysis relied on the frequency and distribution of marker occurrence to identify bacterial species in genomic sequences. This required that the marker sequences be highly overrepresented in a given species over and above that expected in all other species in the lineage. Highly similar strains were removed from the analysis to ensure the selection of oligomers were unbiased and the maximum number of overrepresented oligomers were selected for each species.

The first step in identification of species specific markers involved scoring of each oligonucleotide. The score for each oligonucleotide aimed to identify the degree of difference between the expected value for the oligomer in the present genome and expected values for this oligomer in other genomes throughout the lineage. Several different scoring functions were tested in this respect and were reviewed in Section 2.3.1. For a score, any numerical value above zero indicated that this oligonucleotide was overrepresented within a specific genome above the average value of all other expected values for this oligonucleotide within the lineage. The magnitude of this difference was indicated by the value of the score.

A scoring dictionary was created containing a score-based ranked list of all oligonucleotides and their expected values for each specie. These values were then evaluated based on several criteria. The first criteria was that the expected value of an oligomer

must be well above the probability of finding this oligomer randomly within 100 kbp of sequence. Algorithm 10 gave greater confidence in results and did not bias results as the threshold value was calculated based on the word length and not an arbitrary value. The calculated value was then doubled to correct for any discrepancies due to the incorrect representation of nucleotide ratios as GC content in bacterial species varies greatly.

Algorithm 10 Expected value threshold. This algorithm shows the calculation of the threshold for expected values for each oligonucleotide of different length to determine their viability as overrepresented species specific candidates.

Expected value threshold $= \left(\frac{1}{4}^{l} \times 100000\right) \times 2$

l indicates oligonucleotide length

The second criterion involved filtering highly similar oligonucleotides. This limit was imposed due to the observation that several of the unfiltered oligonucleotide markers in a genome had highly similar sequence patterns with only one or two nucleotide differences. This could be caused by the oligonucleotide being part of a longer repeat region. This did not suit the purposes of the current study as these oligonucleotides described the same position within a region and were, in effect, the same oligomer. This threshold was applied to increase the diversity of markers allowing for a greater probability of identifying each genome as more regions were included in the marker set.

A progressive heuristic approach was used to check sub sequences within each oligonucleotide against all other oligonucleotides. Oligonucleotides were identified as highly similar if sub sequences matched. All highly similar oligonucleotides were removed from the dataset. If an oligonucleotide formed part of a longer oligonucleotide, only the longer oligonucleotide was kept.

Oligomers that met the above criteria were then recorded into two sets of files. The first was a set of marker files for use with *MarkerCounter*, a program developed by collaborators in Germany. These marker files were generated according to a cumulative expected value threshold, which was based on the sum of the expected scores of the best oligomers. For each oligomer appended to the candidate marker list its expected value was added to a total which had a threshold set at 500. The total was referred to as the cumulative expected value threshold. In theory this quantified how many marker oligomers were expected to be found per 100 kbp of sequence. Different sets of markers were then calculated with different oligonucleotide length cutoffs. The first set incorporated all the best 8-9mer oligomers and the second set included all the best oligonucleotides found from 10-14 base pairs in length.

A second set of marker files was generated displaying the score, expected value and empirically determined frequency for each oligomer and for all genomes within the lineage. The oligonucleotides were ranked according to the highest scoring oligonucleotides for a specific genome, this allowed for comparison both programatically as well as visually.

2.4.3 Determination of lineage specific oligonucleotide markers

The initial step in identification of lineage specific markers involves identifying and ranking the most commonly overrepresented oligonucleotides within each lineage. This was done with the use of a common oligonucleotide algorithm (Algorithm 5). The 1000 highest scoring oligonucleotides were selected for a particular lineage while being checked for similarity so as to avoid using highly similar oligonucleotides from the same repeat region. The final, filtered oligonucleotide list was then used to retrieve all oligomers found within each of the remaining 21 lineages.

From this data, the lineage specific scoring function (Algorithm 8, Scoring function 4) was used to rank the data, the highest ranked oligonucleotides indicated the greatest lineage specificity.

The final list of lineage specific oligonucleotides was then created by summing all expected values of the highest scoring oligonucleotides to the cumulative expected value threshold of 500. All oligonucleotides that fell within this range were included as lineage markers.

For each lineage within the database an output file was recorded, containing the selected markers, their associated score and average expected value. Furthermore, the average expected values of all other lineages was also included to provide an opportunity to visually inspect the results. Marker files were also generated for use with *MarkerCounter*.

2.5 Database evaluation and interface

2.5.1 Program interface

Both species and lineage analyses could be accessed and configured through the Python command line application *Oligosignatures*. This application allowed for the editing of parameters to generate custom marker lists for selected bacterial species. Figure 2.7 A showed the main menu for *Oligosignatures*. From this menu the designated task and the genomes to analyse could be selected. The two analyses available were: Signature words and Lineage Oligo Analysis. Signature words and lineage oligo analysis refer to the generation of species and lineage specific oligonucleotide markers respectively. In addition, it is possible to perform both analyses at once. This could be done by selecting either analyses and then selecting the appropriate option in a following menu (Figure 2.7 B). Furthermore, genomes could be added or removed from the analysis and the location of the database could be changed. Lastly, the database could be searched for oligonucleotides of interest. This allowed for investigation into specific oligonucleotides and their occurrences throughout the database. Signature words, lineage oligo analysis and database searching are described below in more detail.

A.

```
Signature Oligonucleotides

Settings for this run:
  A   Add genome?        : []
  R   Remove genome?
  T   Current task?      : Signature words
  S   Search database for an oligonucleotide
  D   Database to use: localhost

Please press Y to accept the settings, select an option from the menu or press Q
 to quit
?
```

B.

```
Performing different lineage analysis (This will take some time!)
What analysis would you like to do ?
    S  Species specific words
    L  Lineage words
    B  Both Analyses
?
```

Figure 2.7: The main menu for *Oligosignatures*. Figure A shows the main screen view of *Oligosignatures*. Figure B shows the menu after task selection with options for multiple analyses.

2.5.1.1 Species marker analysis

Figure 2.8 shows the steps in the creation of species specific oligonucleotide markers using the *Oligosignatures* interface. Figure 2.8 A shows the main menu for *Oligosignatures* where the Signature words analysis has been selected. Two genomes have been added for analysis. These genomes are from different lineages, this program has the capacity to analyse genomes from the same as well as different lineages. In the case of different lineages a temporary table is created. This can allow for a personalized environment to be generated for creation of unique marker lists.

After confirming the genomes to be used the parameters for the selection of species specific oligonucleotide markers is addressed in Figure 2.8 B. Each of the menu options is discussed briefly. A score threshold may be set to determine specificity of markers. Markers can then be filtered to remove highly similar oligonucleotides. For this purpose a limit can then be set on the number of oligonucleotides to be returned if time is of the essence. The cumulative expected value threshold can also be adjusted, this allows for further correction of the stringency of marker oligonucleotides by decreasing or increasing the size of final marker list.

Following completion of the analysis, a general menu for visualization and saving of the processed data is viewed (Figure. 2.8 C). From this menu the selection can then be printed to screen (Figure. 2.8 D) or saved to a file.

2.5.1.2 Lineage marker analysis

In order to create lineage specific marker oligonucleotides the task must be selected and the lineage name must be input (Figure. 2.9 A). Parameters can then be edited to filter oligonucleotides based on similarity and the number of markers to be returned can then be described (Figure. 2.9 B). There is also an option to edit the cumulative expected

A.

```
Signature Oligonucleotides

Settings for this run:
  A   Add genome?        : ['NC_008009.1', 'NC_008711.1', 'NC_008642.1']
  R   Remove genome?
  T   Current task?      : Signature words
  S   Search database for an oligonucleotide
  D   Database to use: localhost

Please press Y to accept the settings, select an option from the menu or press Q
 to quit
?
```

B.

```
Species specific words

Settings for this run:
  R Remove all redundant strains from this calculation? : N
  S   Score Threshhold : 0
  F   Filter oligos based on similarity :N
  X   Expected Value threshold: 500
  T   Limit result for filtered oligos : 100
Y to accept these settings
?
```

C.

```
Select next action:

 C       Continue
 P       Print to console
 S       Save to file
 What to show:
          E      Expectation per 100 kb? Yes
          A      Absolute word counts?   No
          F      Co-efficient of variation?      No
 M       Mode of word selection?         Signature words
 L       Minimal word length?            8
 T       Top the best words?             ALL

? ...
```

D.

Selected words for NC_008009.1

File

Number	Oligo	Score	Expected Value	Frequency	Score	Expected Value	Frequency
1	CGACTTCG	9.92717346677	11.7985883118	980	1.41175258793	6.01139587302	868
2	CTTCGTCG	9.5321344249	12.6279016683	1033	1.98990637	7.36454320911	1087
3	CTTCTTCG	9.45275382612	13.8173829199	1150	2.79352296725	9.09864819644	1293
4	CGCGACGA	9.13429235235	12.080917417	992	1.86710014371	6.98904611362	998
5	CTTCGCGA	9.00117049326	12.1785776477	963	1.98808552797	7.23880369959	1038
6	TCGCGACG	8.78997734078	10.7933485778	838	1.34225223676	5.62665948986	796
7	CGCCGACG	8.54781276368	17.0968731772	1408	7.16006280633	16.0981209181	2375
8	GCGACTTC	8.25619522801	9.78598851875	772	1.06788271024	4.7996188707	666
9	CGAAGCCG	8.22787860365	16.9250245297	1401	7.50938339136	16.4076739054	2392

Figure 2.8: Interface parameters and options for the generation of species specific oligonucleotides.

A.

```
Signature Oligonucleotides

Settings for this run:
  A   Add genome?          : []
  R   Remove genome?
  T   Current task?        : Lineage Oligo Analysis
  S   Search database for an oligonucleotide
  D   Database to use: localhost

Please press Y to accept the settings, select an option from the menu or press Q
 to quit
?y


Please input a lineage in the format LIN_3, for any lineage from 1- 22:
LIN_5
```

B.

```
Lineage Oligonucleotide analysis

Settings for this run:
  F   Filter oligos based on similarity :N
  X   Expected Value threshold: 500
  T   Limit result for oligos : 1000
Y to accept these settings
?_
```

C.

Selected words for LIN_5

File

Number	Oligo	Score	Expected Value
1	TTCTTTTT	240.928489843	16.2217893017
2	AAAAGAAA	212.862170726	15.2464233921
3	TTTTCTTT	195.811025036	14.622173948
4	TTTTTTAT	184.188392749	14.1809500172
5	TCTTTTTT	173.527639427	13.7638313851
6	TTTTTCTT	171.428350043	13.6801972535
7	TCTTTTTC	161.658604174	13.2840646091
8	TGAAAAAA	158.369393084	13.1480169475
9	AATAAAAA	151.315744434	12.8514192434
10	TATTTTTT	147.974699663	12.7085209958

Figure 2.9: Interface parameters and options for the generation of lineage specific oligonucleotides.

value threshold to increase or decrease the number of markers included.

Finally, the results can be viewed or saved and the resultant output is displayed in Figure 2.9 C.

2.5.1.3 Database searching

Database searching offers insight into the presence of specific oligonucleotides within the database and retrieves expected values for all genomes containing this oligomer. Different parameters can then be set for each search (Figure 2.10 A). The reverse compliment can be included, as can all oligonucleotides of the same length with a single permutation (known as horizontal neighbours). In addition, all longer oligonucleotides containing this sequence (upper neighbours) and all shorter oligonucleotides containing a sub sequence of this oligonucleotide (lower neighbours) can be searched. Finally the output for this analysis can be printed or saved. The printed output is described in Figure 2.10 B.

A.

```
Searching DB

Settings:
OLIGO:: GGCGCTGAA
  C   Change oligo
  R   Include reverse compliment        : Y
  H   Include Horizontal neighbours     : N
  U   Include Upper neighbours   : Y
  L   Include Lower neighbours   : N

Y to accept these settings, type the letter for one to change or Q to quit
?
```

B.

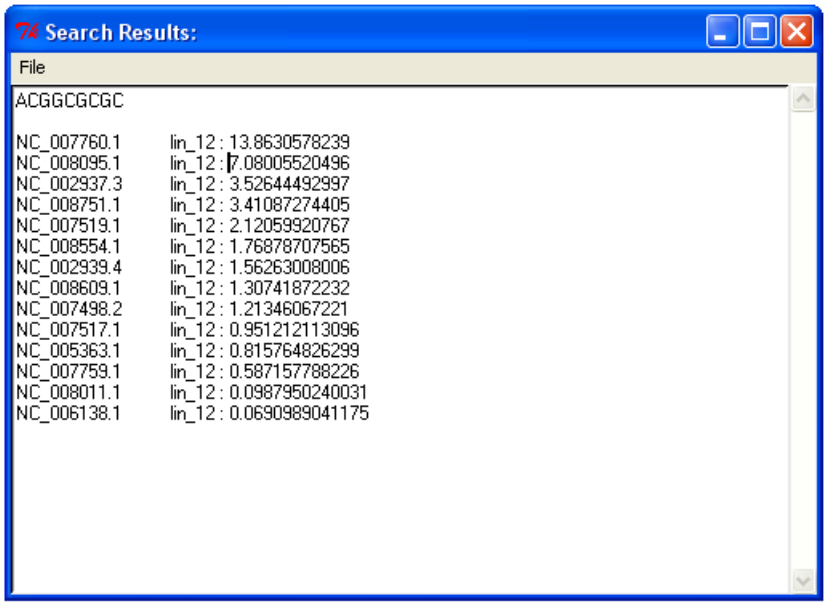


Figure 2.10: Searching the database for an oligonucleotide or other highly similar variants.

2.5.2 Lineage identification

Each of the 22 lineage tables processed using lineage specific analysis yielded varying results. Several lineages showed very few lineage specific oligonucleotides and others showed large numbers (Table 2.3). This was to be expected as several bacterial families have been found to contain far less global repeats than others. The comparison was hampered by the significant differences in genome number for each lineage. Thus, the lineages with the least number of organisms often had the lowest scores. This form of bias was ascribable to the decrease in power of the statistical test due to the insufficient length of the dataset. This was not the only reason for the low number of overrepresented oligonucleotides. In the *Chloroflexi* lineage, species within the lineage showed very few specific oligonucleotides. One of the central limitations of this approach could therefore be seen as the lack of sequenced genomes in certain lineages. Certain characteristics of lineage species such as richness in genomic repeats could impose further limitations.

From an attempt to include 8mer oligonucleotides in lineage analysis it was seen that specificity was too low and lineages could not be distinguished. This lead to the exclusion of 8mer oligonucleotides in lineage marker lists. Furthermore, results showed that many lineages did not contain sufficient oligonucleotides for lengths 10-14mer, in this instance 9mer oligonucleotides were included to enable lineages to reach the desired cumulative threshold.

The number of marker oligonucleotides to be included per lineage was calculated by using the cumulative expected value threshold. This value was selected as several lineages did not contain enough oligonucleotide entries for a combined expectation of more than 500. Furthermore, several lineages contained an extremely large number of oligonucleotides with low expectation values. This resulted in a large increase of non-specific data as slightly overrepresented oligonucleotides were included within the marker lists to attempt to reach the desired threshold.

One of the lineages, *Chloroflexi* (Lineage 8) did not contain sufficient oligonucleotides even once the threshold was reduced and therefore could not be used for analyses. Interestingly, *Gammaproteobacteria* has one of the longest marker lists (over 500 markers), this could be due to the wide diversity of species within the lineage causing a lack of specificity. Furthermore, *Nanoarchaeota* (Lineage 18) contained only 35 markers but scores for these oligonucleotide markers were very high indicating a sizable distance between this lineage and all others.

2.5.3 Species identification

Interesting features and observations from looking at many different species examples showed that several bacteria contain a relatively small number of species specific overrepresented oligonucleotides, between 40-60. This indicated that a very high occurrence

Table 2.3: Number of markers identified per lineage. * indicates lineages without sufficient overrepresented oligonucleotides to complete analysis.

Lineage Number	Lineage Description	Number of Markers
1	*Acidobacteria*	113
2	*Actinobacteria*	118
3	*Alphaproteobacteria*	169
4	*Aquificae*	60
5	*Bacteroidetes/Chlorobi*	133
6	*Betaproteobacteria*	110
7	*Chlamydiae/Verrucomicrobia*	96
8	*Chloroflexi*	55 *
9	*Crenarchaeota*	165
10	*Cyanobacteria*	205
11	*Deinococcus-Thermus*	76
12	*Deltaproteobacteria*	178
13	*Epsilonproteobacteria*	73
14	*Euryarchaeota*	203
15	*Firmicutes*	127
16	*Fusobacteria*	48
17	*Gammaproteobacteria*	525
18	*Nanoarchaeota*	35
19	*Other Bacteria*	121
20	*Planctomycetes*	115
21	*Spirochaetes*	84
22	*Thermotogae*	65

of each oligonucleotide was expected within a genome, indicating a repeat abundant genome. However, when an exceptionally large number of oligonucleotides was found, each oligonucleotide was present infrequently and at low frequency, indicating a genome with few repeats. This could result in the inability to identify these genomes due to their lack of specific markers which can result in a large number of false positives. This approach was clearly more effective on certain genomes as is evidenced by this selection of oligonucleotides.

Several examples of species specific oligonucleotides were given in the tables below. The oligonucleotide markers were selected based on the identification of species within an unknown sample. Four pathogenic bacterial strains were selected, *Mycobacterium tuberculosis* CDC1551 (NC_002755.2), *Bacillus anthracis* str. Ames (NC_003997.3), *Pseudomonas aeruginosa* UCBPP-PA14 (NC_008463.1) and *Salmonella enterica* subsp. enterica serovar Typhi Ty2 (NC_004631.1).

Each table contains a selection of species specific overrepresented oligonucleotides for a bacterial pathogen. The calculated score (Algorithm 4, scoring function 4) for each oligonucleotide is shown as is the expected value for the pathogenic species and several closely related species. From these tables it is clear that each oligonucleotide was effec-

tive at discriminating among only a few of the shown species. As all oligonucleotides have different characteristics for discrimination of species the profile created attempted to combine their discriminative power.

For *Mycobacterium tuberculosis* (Table 2.4) relatively high scores were obtained for the best oligonucleotides. However, within closely related species the expected values were highly similar. This indicated that although *M.tuberculosis* may be distinguished effectively within the lineage its highly similar neighbours will remain difficult to differentiate. *M.bovis* and *M.avium* had highly similar expected values to *M.tuberculosis* and will likely be the most difficult to distinguish.

As the lengths of the oligonucleotides increase there was a decrease in score and expected values. The expected value for *M.tuberculosis* did not differ substantially from its closest neighbours for these longer oligonucleotides. However, more distantly related genomes were easily separable as they now have much lower expected values, tending towards zero. This gave an indication of the power of including longer oligonucleotides in this approach to improve identification of a bacterial species.

From the data below it was possible to recognise that several oligonucleotides share a strong similarity. For example, "TGGCCGCGGC" and "CGGTGGCGCC" were highly similar and appear to be from the same repeat sequence. Identification of these similarities was a highly complex and computationally intensive task. Although the best efforts were made to limit this effect several such oligonucleotides remain within each dataset.

The species specific oligonucleotides generated for *Bacillus anthracis* (Table 2.5) show a similar picture. As is expected *B.thuringiensis* and *B.cereus* both show high similarity to *B.anthracis* and will be difficult to distinguish. From the oligonucleotides selected it can be seen that a large number seem to originate from homopolymer repeats of adenine and thymine. This indicates that these sequences can be integral to the species and genus as they are widespread and highly conserved within *B.anthracis* as well as *B.thuringiensis* and *B.cereus.* *B.subtilis* and *B.licheniformis* show the most difference in oligonucleotide expected values from *B.anthracis* but remain highly similar with several oligonucleotides. Scores for the oligonucleotides indicate the difficulty in separating *B.anthracis* not only from its nearest neighbours but also throughout the *Firmicutes* lineage.

The next dataset contained *Pseudomonas aeruginosa* (Table 2.6) and several of its closest relatives. Much like the *Mycobacterium* dataset, these oligonucleotide markers had higher GC content than the *Bacillus* genus. Here *P.aeruginosa* could be differentiated far more easily than the last two pathogens. The closest genomes, *P.fluorescens* and *P.entomophila* remained dissimilar on a number of oligonucleotides indicating that *Pseudomonas aeruginosa* may be more easily separable than the previous groups. It was noted that the scores for this pathogen were much higher than the previous scores. This could be attributed to the lack of similarity with closely related species and the diversity of the lineage, thereby increasing the score.

Table 2.4: A selection of identified marker oligonucleotides for *Mycobacterium tuberculosis*. The first column shows the oligonucleotide followed by the score in the second column for that oligonucleotide within the genome. The following columns contain expected values for several closely related organisms (not all organisms in the lineage are represented in the table) allowing for assessment of efficiency. * Calculated using (Algorithm 4, scoring function 4)

	M.tuberculosis	***M.tuberculosis***	***M.leprae***	***M.avium***	***M.ulcerans***	***M.vanbaalenii***	***M.bovis***
	Score*						
GGCGGCAACG	11.42	5.00	0.92	2.11	4.03	2.13	4.91
TGGCCGCGGC	7.76	4.82	1.62	5.54	3.78	3.91	4.83
CGGTGGCGCC	4.68	3.58	1.18	3.31	3.49	3.30	3.51
TTGGCCGCCG	4.64	3.00	1.25	2.43	2.61	1.73	3.05
TGCTGGCCGG	4.48	3.44	1.70	4.91	2.95	2.70	3.45
CGTCACCGCC	4.30	3.53	0.77	3.69	2.64	3.29	3.53
GCCGCCAGCG	4.22	3.41	1.10	3.68	3.27	2.86	3.51
GGCGATCACC	4.20	3.35	1.48	4.98	2.84	3.68	3.29
TCGGCCAGCA	4.18	3.37	1.61	4.97	2.68	3.61	3.37
ACCGCCGGCG	4.05	3.16	0.93	3.12	2.01	1.92	3.10
GGCGGGGCCGGCGG	2.97	1.84	0.00	0.16	0.04	1.92	1.92
GCCGTTGCCGCCG	2.73	1.84	0.11	0.31	0.29	1.91	1.91

Table 2.5: A selection of identified marker oligonucleotides for *Bacillus anthracis*. The first column shows the oligonucleotide followed by the score in the second column for that oligonucleotide within the genome. The following columns contain expected values for several closely related organisms (not all organisms in the lineage are represented in the table) allowing for assessment of efficiency. * Calculated using (Algorithm 4, scoring function 4)

	B.anthracis	***B.anthracis***	***B.subtilis***	***B.licheniformis***	***B.thuringiensis***	***B.cereus***
	Score*					
CTTTTTTTAT	3.66	3.70	3.55	2.76	3.95	4.19
TTCTTTTACA	3.43	2.60	1.49	0.87	2.34	3.23
AATGAAAGAA	3.33	2.75	1.65	1.38	2.57	3.31
ATTTCTTCTT	3.15	3.05	1.50	1.39	3.09	3.09
GAAGAAAAAG	2.88	2.93	2.36	2.49	2.66	2.76
CTTCTTTTAC	2.82	2.30	1.36	0.78	2.26	2.80
TAAAGTGAAA	2.69	2.09	0.55	0.75	2.15	2.69
GAAAGAAAAT	2.53	2.27	1.51	0.82	2.34	3.12
ATGAAAGAAA	2.52	2.39	1.55	1.58	2.48	2.64
CTTCTTCTAA	2.52	2.35	0.49	0.39	2.33	2.84
GAAATGAAAA	2.51	2.37	1.84	1.57	2.43	2.61
TATAAAAGAA	2.49	2.35	1.14	0.99	2.12	2.17
ATAGAAGAAA	2.47	2.22	1.01	0.64	1.83	1.83
AAAAGCAATT	2.44	2.58	1.11	1.00	2.48	2.29
AAGAAAAAGG	2.21	2.30	2.13	2.21	2.33	1.94
TGAAATTGAA	1.39	1.90	1.25	1.04	1.77	1.63

Table 2.6: A selection of identified marker oligonucleotides for *Pseudomonas aeruginosa*. The first column shows the oligonucleotide followed by the score in the second column for that oligonucleotide within the genome. The following columns contain expected values for several closely related organisms (not all organisms in the lineage are represented in the table) allowing for assessment of efficiency. * Calculated using (Algorithm 4, scoring function 4)

	P.aeruginosa	***P.aeruginosa***	***P.putida***	***P.fluorescens***	***P.entomophila***	***P.syringae***
	Score*					
CGCCGGCGGC	73.58	10.00	1.72	3.70	2.50	0.53
CCTGGCCGGC	43.00	7.39	3.25	4.52	4.93	0.69
CCAGGCCGGC	35.43	6.69	3.19	4.24	4.38	0.73
GCCGCCGAGG	28.98	6.19	1.49	2.24	2.44	0.66
CGCCGAGCTG	25.44	5.92	2.36	2.94	3.52	1.71
TCGCCGCCGA	23.53	5.88	1.52	2.29	2.90	1.51
CCTCGGCCAG	22.52	5.41	3.25	4.58	5.20	1.23
CCGGCGCCGC	22.12	5.39	3.21	4.22	4.84	0.36
CGCCGAGGGC	21.98	5.24	2.23	2.86	3.49	0.80
CGCCGGCAGC	21.81	5.44	1.30	2.86	2.33	0.80
TCGGCCTGCT	21.76	5.48	1.16	1.64	1.85	1.66
AGGCCGGCGA	21.12	5.12	2.09	2.94	2.68	0.58
CCTGGCGCGC	20.34	5.18	1.85	2.25	2.58	0.83
GGCGTCGGCG	17.07	4.84	1.19	2.17	2.89	4.84
GGCGTCGCCG	13.44	4.16	0.76	1.09	1.29	4.16
TCGGCGAGCA	11.46	3.91	1.09	0.89	1.95	3.91

The final table contained species specific oligonucleotides for *Salmonella enterica* (Table 2.7). This dataset showed that several oligonucleotides could distinguish this organism from *Shigella flexneri* and *Escherichia coli* although the numerical difference was small. Several oligonucleotides seemed to have highly similar values for all organisms in the dataset and was ineffective in separation of these species. The scores for these oligonucleotide markers were lower than the *Pseudomonas* dataset but higher than the *Bacillus* dataset partially due to their ability to separate closely related species more reliably.

2.6 Discussion

The current study centers on the conversion of raw information into a structured and complete database which allows for comparisons to be easily made between species. These comparisons have allowed for the identification of species and lineage specific oligonucleotide markers, although the program and interface developed have been effectively used to this end it still provides a versatile tool for further research. The creation of custom environments and the selection of smaller subgroups of species can provide more specific results depending on the needs of the researcher. This allows for a dynamic and flexible system which can be used to analyse oligonucleotides under different situations.

From inspection of the number of organisms in each lineage (Table 2.1) it was clear that the first appearance of bias was the variation in genome counts between the different lineages. This phenomenon could be partially explained by global research focus determining which organisms and phylogenetic groups get the most attention.

The analyses performed on small lineages bias the data and made results unreliable due to the small datasets and resultant lack of statistical power. Furthermore, separation of species using phylogenetic lineage may not be the best approach. Some lineages contained a vast diversity of species which could confound results. Species specific markers from smaller lineages gained an advantage in score over those in larger lineages but in actual fact were at a disadvantage due to the lack of specificity. This could be overcome by identifying an alternative method for species separation although this would require further constraints and research into the topic to ensure a consistent result.

The current study incorporates several statistical algorithms involved in the determination of species or lineage specific oligonucleotides. The expected value plays a central role as the main statistical parameter describing each oligonucleotide and highlighting its desired properties. This however, does not provide a strictly statistical definition of expected value and cannot be considered a statistically significant parameter. The focus of this parameter on the need for an evenly distributed oligonucleotide can provide skewed results. An oligonucleotide which is highly localised but occurs with large frequency may affect results by appearing to be largely overrepresented if this region is sequenced. This variation cannot be controlled for but emphasizes the experimental nature of this approach

Table 2.7: A selection of identified marker oligonucleotides for *Salmonella enterica*. The first column shows the oligonucleotide followed by the score in the second column for that oligonucleotide within the genome. The following columns contain expected values for several closely related organisms (not all organisms in the lineage are represented in the table) allowing for assessment of efficiency. * Calculated using (Algorithm 4, scoring function 4)

	Salmonella enterica	***Salmonella enterica***	***Shigella flexneri***	***Escherichia coli***
	Score*			
CGCTGGCGCA	5.94	3.09	1.73	2.03
TCGCCAGCGC	5.66	2.96	1.73	1.77
CGCTGGCGGA	5.34	2.75	1.60	1.53
CTGGCGCAGC	5.14	2.90	1.76	1.79
CCATCCGGCA	4.38	2.32	0.70	0.65
AGCGCCAGCA	3.69	2.52	2.08	1.84
GGCGGCGCGG	3.65	2.36	0.97	1.04
GCTGGAAAAA	3.60	2.50	2.67	2.29
ACGCCAGCGC	3.57	2.29	1.05	1.32
CCGCTGGCGG	3.44	2.29	1.65	1.41
GACGCTGGCG	3.28	2.17	1.53	1.38
CGCCTGCGCC	3.26	2.29	1.47	1.52
GCTGGCGAAA	3.25	2.16	1.97	1.60
CACGCTGGCG	3.16	2.12	0.92	1.00
TTCCAGCGCC	3.01	2.16	2.09	2.22
GGCCGGATAAG	2.75	1.73	1.09	1.25
CGCCATCCGGCA	2.47	1.63	0.20	0.20
GTAGGCCGGATAAG	1.68	1.33	1.00	1.05

and the need for an oligonucleotide profile rather than single marker oligonucleotides.

The central difficulty when creating scores and thresholds was that data under observation did not conform to a standard distribution. In this case statistical significance was sacrificed in an attempt to create a logical estimation. Although several different algorithms were used the ratio based approaches provided the best results. From Section 2.5.3 it was clear that the species specific scoring algorithm was effective at scoring oligonucleotides dependent on expected value. However, this algorithm was biased towards shorter sequences and preferably incorporates shorter oligonucleotides due to their higher frequency. For this reason a minimum of 10mer oligonucleotides was used in species identification. In order to provide a more permanent solution an improved algorithm should be created that could accurately determine effective longer oligonucleotides on a more statistical criterion. This would not only increase signature diversity but improve specificity.

The ability to identify strains as highly similar using overrepresented oligonucleotides was an important result. This offered an estimation of the discriminative ability of overrepresented oligonucleotides by testing the visible differences between two theoretical strains. This gave an indication that strains with such low variation were not separable using overrepresented oligonucleotides due to their minute frequency discrepancies. Furthermore, this approach opened a new opportunity in the analysis and determination of phylogeny. From the strain plots it was possible to see how organisms within a lineage could be separated into distinct groups. Closely related strains, a central body of species and several distantly related groups could be identified in each strain plot. The range of distances present within each plot could also be taken into account to estimate diversity within each lineage. With further research it may be possible to increase resolution to classify a bacterial species by analysis of its overrepresented oligonucleotides.

Although the strain analysis incorporated mostly 8mers in its identification of closely related genomes these oligonucleotides were insufficient to separate species. From further analysis it was determined that these were unusable in species identification due to the general overrepresentation in all species regardless of the relatively large differences in frequency.

From the examples given in Section 2.5.3 the diversity of overrepresented oligonucleotides can be highlighted. The *Bacillus* genus contained homopolymer AT rich oligonucleotides while other genus' had markers with higher GC contents. It was also possible to estimate the ease of identification of a species by inspection of the oligonucleotide scores and the subsequent expected values of closely related species.

One of the greatest setbacks with determination of species words was the identification of unique oligonucleotides that do not form part of the same repeat region. Determination of sequence identity generally requires sequence alignment which is computationally expensive. In decreasing computational time heuristic methods are available but these

remain error prone. When sequence identity thresholds are too high useful data can also be excluded. Therefore removal of redundant markers requires an indepth analysis into the positions of occurrence as well as the sequence identity. This was beyond the scope of this project and would increase statistical complexity and computational time. This remained one of the pitfalls of identifying species and lineage specific oligonucleotides as a percentage of oligomers occurring within the database were highly similar due to polymorphisms.

If a more robust method for similar sequence identification were uncovered it would greatly enhance the effectivity of this approach. However, it could be expected that this would be computationally intensive and require highly specialized algorithms.

In the analysis of lineage specific oligonucleotides the threshold set for species specific oligomers at 10-14mer did not apply as far fewer oligonucleotides were included in this analysis. The threshold was therefore reduced to include 9mer oligonucleotides to provide a greater selection. All lineages were found to contain large numbers of lineage specific oligonucleotides with the exception of *Chloroflexi* (lineage 8). This lineage contained only two species both of which contained small numbers of overrepresented oligonucleotides. Another interesting find was the low scores of lineage specific oligonucleotides for *Gammaproteobacteria*. This suggested that not only were the species in this lineage highly diverse but they shared common sequences with a large number of other lineages. Inversely, *Nanoarchaeota* (Lineage 18) contained only 35 lineage specific oligonucleotides showing the distance of their relation to other bacterial species. This indicated how diverse each lineage was and how much information could be gathered from a single analysis.

From lineage specific oligonucleotides a deeper insight was gained into oligonucleotides commonly found throughout a lineage. These sequences could then be investigated to determine whether they form part of specific structural or control features to determine the reasons for such widespread overrepresentation.

2.7 Conclusion

The creation of the program *OligoSignatures* and its database provides an opportunity to investigate the use of overrepresented oligonucleotides in identification of species and their lineages. The application in this context is the identification of species in an unknown sample. However, this program allows for manipulation of data to create a custom environment according to the needs of the researcher. This provides a flexible and dynamic system which can create oligonucleotide markers according to specific criterion and can therefore be used to uncover further trends and identify bacterial species within different contexts.

The initial steps in the creation of this program involved the processing and analy-

sis of raw oligonucleotide data before importing it into a database. This database was then further analysed and divided to produce 22 lineage tables containing comprehensive oligonucleotide information for its different members. The comprehensive oligonucleotide information could then be used in the confirmation of strain groups to determine whether strains of the same species where closely related enough to be removed and represented by a single genome. These results provided valuable insight into the limitations of using overrepresented oligonucleotides as well as the use of oligonucleotides in determining phylogeny.

Secondly, species specific oligonucleotides were identified using the strain conformation results. The species specific oligonucleotide results showed that (with a few exceptions) the majority of species contained overrepresented oligonucleotides. These oligonucleotides differed substantially from genus to genus, and highly similar species tended to share a large percentage of overrepresented oligonucleotides. The scores calculated for oligonucleotides provided insight into the uniqueness of the oligonucleotide within the genome and an estimation of ease of identification in an environmental context.

Lastly, lineage specific oligonucleotides were identified. It was found that the majority of lineages had sufficient oligonucleotides for analysis, with the exception of the *Chloroflexi* lineage (lineage 8). Different properties can be determined through the analysis of lineage specific oligonucleotides. The number of oligonucleotides included in the lineage profile can indicate how widespread the oligonucleotides were and hence the relation of a lineage to other lineages. From these beginnings a further investigation can be made into annotation of candidate oligonucleotides and their function within the genome can be determined.

In conclusion, the program *OligoSignatures* provides an opportunity to investigate the effectiveness of oligonucleotides in the identification of bacterial species in a metagenomic context. This context has been formalized to identifying bacteria within an unknown sample. This does not describe all contexts where overrepresented oligonucleotides can be used for species identification. *OligoSignatures* can be used in many different contexts to determine marker profiles. It provides a powerful tool that can be used under different situations to identify species and lineage specific oligonucleotide markers.

Chapter 3

Metagenomic implementation

3.1 Introduction

Identification of bacterial species and their resultant fragments within a metagenomic sample has proved an overwhelming hindrance to the furthering of metagenomic studies. Experimental research show that as little as 1% of metagenomic fragments contain identifiable phylogenetic markers. This implies that 99% of the metagenome remains unexploited. This highlights the need for more robust and flexible approaches to be unearthed to improve genomic coverage.

The identification of closely related strains and their representation by a single representative member was the first step in determining the limitations of overrepresented oligonucleotides and subsequently how to remove unnecessary complexity from further calculation. The identification of species specific oligonucleotides was then undertaken. The resulting marker profiles, based on scores, showed the expected ease of differentiating these species from their relatives within the same lineage. Furthermore, lineage specific markers were identified as commonly overrepresented oligonucleotides within the lineage that appear at low frequency throughout the other lineages. This revealed that although each lineage has specific properties the majority of lineages contained overrepresented oligonucleotides at a satisfactory level to allow for an accurate analysis to be undertaken.

With the uncovering of species and lineage specific oligonucleotides a further step must be taken to validate their functionality in an experimental context. This chapter aims to identify how effectively overrepresented oligonucleotides can be used within a metagenomic context to identify bacterial species.

For the purpose of experimentation, four different case studies were selected based on the examples used in Chapter 2 Section 2.5.3. These are focused on the pathogenic bacterial strains, *Mycobacterium tuberculosis* CDC1551 (NC_002755.2), *Bacillus anthracis* str. Ames (NC_003997.3), *Pseudomonas aeruginosa* UCBPP-PA14 (NC_008463.1) and *Salmonella enterica* subsp. enterica serovar Typhi Ty2 (NC_004631.1). Each case study

incorporated the pathogenic bacteria, its closely related relatives and distantly related species.

The initial step in this process was the creation of artificial metagenomic samples for each case study. After creation of the datasets, experimental testing of oligonucleotide marker profiles could be undertaken. This was done via two different approaches, oligonucleotide frequency analysis and sequenced read analysis.

Two modern sequencing methods were simulated here, namely, Solexa and 454 sequencing. Solexa sequencing technology generates very short read lengths which will be used primarily for the identification of bacterial species using global oligonucleotide frequencies within the metagenome. 454 pyrosequencing technology, with longer read lengths (roughly 250bp in length), will be used to identify bacterial species by the attribution of metagenomic fragments to specific species. This method is referred to as sequenced read analysis.

Oligonucleotide frequency analysis is a method for detection of bacterial species within a metagenome based on the global overrepresentation of marker oligonucleotides. Each marker oligonucleotide forms part of a marker profile. A marker profile indicates the number of oligonucleotides found within a 100kbp region to achieve a cumulative frequency of 500 oligonucleotide occurrences. In order to make a comparison of this value to different species, an error value must be calculated. The error value is an estimation of the average false positive result that can be expected for each marker profile. This value then allows for an assessment of presence or absence of this bacterial specie. A flow diagram describing the basic steps in this approach can be seen in Figure 3.1.

Differentiating species using sequenced read analysis employs a different premise. In sequenced read analysis the focus is on each short fragment (referred to as a ”read”) within the metagenomic library. Sequenced read analysis attempts to identify oligonucleotides occurring within these fragments that can attribute them to a specific marker profile. Different thresholds are set for the number of oligonucleotides required in each read to classify the read to a specific species (referred to as a ”hit”). Specificity was then increased further by including only the unique markers for each specie. Figure 3.1 shows a flow diagram depicting basic functioning of this approach. Following this introduction a brief overview of methods used in each approach is given.

3.1.1 Analysis of oligonucleotide frequency profiles

Method

This approach centered on the summation of marker oligonucleotide frequencies found within 100kbp of sequence. This resulted in a score which can be used to determine presence or absence of the oligonucleotide within a metagenomic sample.

The first step in this procedure involved retrieval of marker profiles for each species and

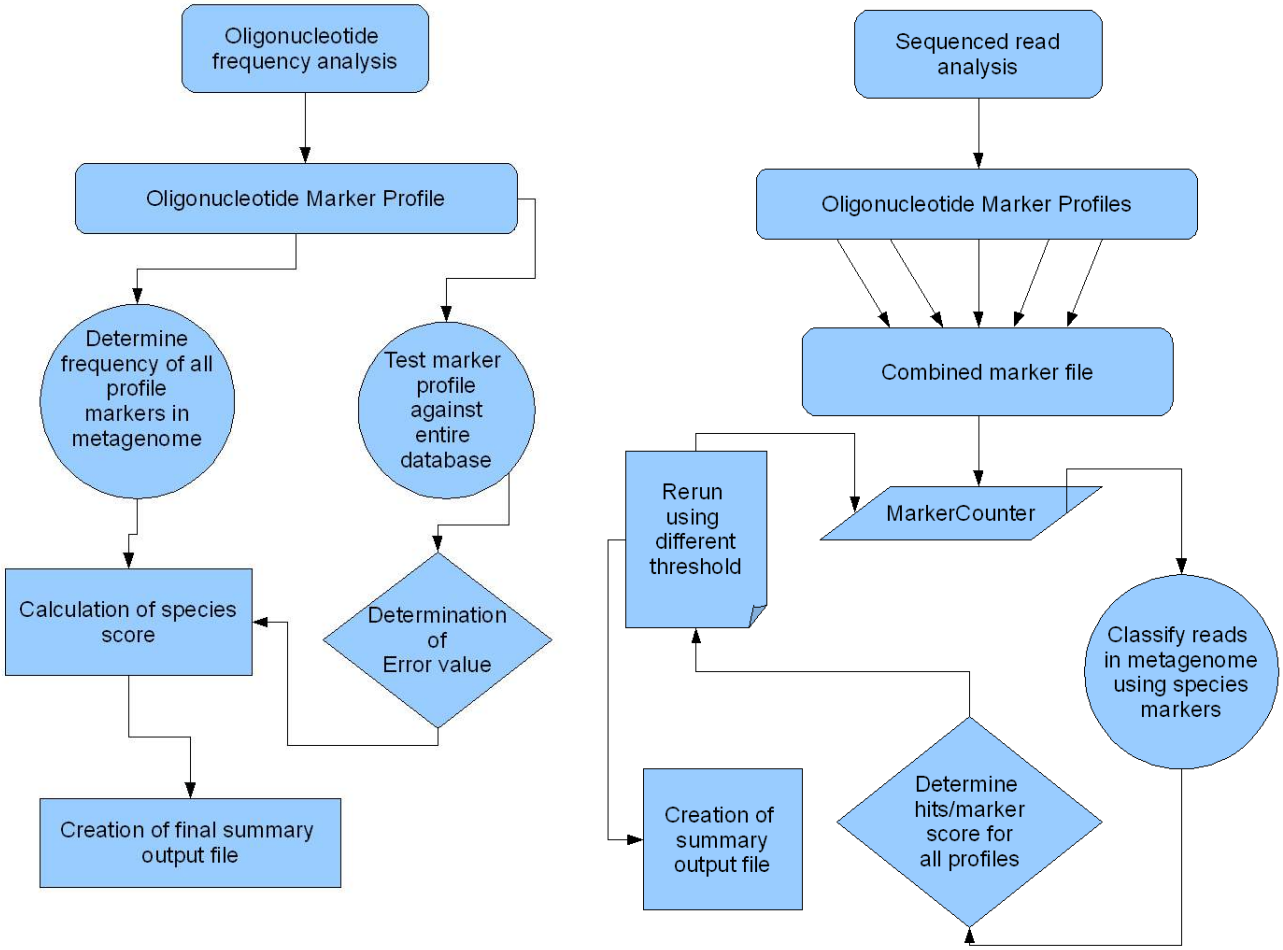


Figure 3.1: Flow diagram describing the processes for oligonucleotide frequency analysis and sequence read analysis respectively.

all lineages involved. Each marker profile was tested against every artificial metagenome created within each case study. This was done by applying a search function to identify every oligonucleotide occurrence within the metagenome for all oligonucleotides markers. A species profile value was calculated to describe the cumulative frequency of markers identified within each metagenome (Algorithm 11).

All species profile values for each metagenome were then compiled into an output file. This yielded easy readability of results and comparison amongst different species.

In order to test the effectivity of results, an error value was calculated for each species (Algorithm 12). This error value was determined by testing a species profile against a metagenomic sample of every species within the database. All species within the database had 100kbp metagenomes created and these were used for calculation.

Determination of an error value then allowed for calculation of a final species profile value. This value identified the presence of a species within a metagenome by comparing

Algorithm 11 The species profile value describes the cumulative value for a set of markers for each specie.

(Species profile value) $s = \sum_{i=1}^{n} i$

Assume n number of oligonucleotide markers for each specie
i indicates the frequency of occurrence of each oligonucleotide within the metagenomic dataset

Algorithm 12 Species profile error value. This algorithm determines the average value of a species profile tested against all other species within the database.

(Species profile error value) $\epsilon = \frac{\sum_{i=1}^{n} i}{n}$

Assume n is the number of species within the database
i indicates the species profile value for a particular metagenome

the error value to the results obtained. A value close to one indicated an insignificant result while a value greater than one indicated a species signal. However, this approach did not provide statistical significance and was considered an estimation. A further correction was made to this algorithm when dealing with an experimental metagenome as the sample contained a far larger amount of genetic material. The normalization used is indicated in Algorithm 14.

Algorithm 13 Final species profile value. This algorithm describes the presence or absence of a species within a metagenomic sample. A value less than or close to 1 indicates random noise while a value greater than one indicates a species signal.

(Final species profile value) $f = \frac{s}{\epsilon_s}$

Assume sindicates the species profile value
$\epsilon_s$indicates the error value for species profile s

3.1.2 Analysis of sequenced reads

Method

Analysis of longer sequenced reads was undertaken with the aid of the java based program *MarkerCounter*. *MarkerCounter* uses a suffix-tree based approach to search for thousands of markers from different species within a metagenome. *MarkerCounter* attempted to relegate sequence fragments within a metagenome to a specific species or lineage profile. A fragment was associated to a profile (referred to as a "hit") if a predefined number of oligonucleotide markers from a profile occurred within the fragment (referred to as the "oligos per fragment" threshold). *MarkerCounter* attempted to identify unique oligonucleotides for each species from the input marker list and used these to identify species

Algorithm 14 Final species profile value for an experimental metagenome.

(Corrected species profile value) $s_c = s \div \frac{l}{100000}$
(Final species profile value) $f = \frac{s_c}{\epsilon_s}$

l indicates the length of the experimental metagenome
Assume s_c indicates the corrected species profile value
ϵ_s indicates the error value for species profile s

fragments within a metagenomic context. Two different sets of marker profiles were used in this approach, lineage and species profile markers. Each is discussed below.

Species profile markers

The first step in analysis of sequenced reads was retrieval of species profiles from temporary marker files generated by species and lineage analyses. These profiles were then concatenated into a general marker file containing all species within the case study. The number of oligonucleotides needed to attribute a read to a profile (referred to as the "oligos per fragment" threshold) was then varied, in order to determine the different levels of significance at which reads can be attributed to a specie. For artificial case studies, threshold values from two to five were applied. In the experimental metagenome, threshold values from three to nine were used to compensate for the longer read lengths. Finally, *MarkerCounter* was executed using the general marker file on all metagenomes in each case study. A separate run was performed for each oligonucleotide threshold.

Of concern to this approach was that different numbers of markers are utilized by each species profile. The number of markers per profile is dependent on how many unique markers can be attributed to each profile. In order to correct for the discrepancy in number of markers per profile a score was calculated. This score was based on the number of reads attributed to the species compared to the number of markers used (Algorithm 15).

Algorithm 15 Hits per marker score. This score normalizes the number of reads attributed to a species by the number of markers used by the species profile.

(Hits per marker score) $hp = \frac{h_s}{m_s}$

h_s indicates the number of reads attributed to a species
$m_s$indicates the number of markers in the species profile

The resulting scores from this algorithm were then visualized using line graphs which create a concise representation of results and allow for further identification of trends.

Lineage profile markers

Lineage profiles are run using the same process as species profiles. However, the number of lineage profiles included in the general marker file was reduced after each execution.

Table 3.1: Bacterial Species Present in Case Study 1. * Indicates a more distant species within the same lineage. ** Indicates a species from a different lineage

Accession Number	Species Description
NC_008595.1	*Mycobacterium avium 104*
NC_008769.1	*Mycobacterium bovis BCG str. Pasteur 1173P2*
NC_002755.2	*Mycobacterium tuberculosis CDC1551*
NC_002677.1	*Mycobacterium leprae TN*
NC_008596.1	*Mycobacterium smegmatis str. MC2 155*
NC_008611.1	*Mycobacterium ulcerans Agy99*
NC_008726.1	*Mycobacterium vanbaalenii PYR-1*
NC_004307.2	*Bifidobacterium longum NCC2705**
NC_003450.3	*Corynebacterium glutamicum ATCC 13032**
NC_007512.1	*Pelodictyon luteolum DSM 273***
NC_000922.1	*Chlamydophila pneumoniae CWL029***

An initial analysis was done using all lineage markers for each case study but in order to avoid bias, further tests were done by removing all lineages containing less than 5 species from the general marker list. If further resolution was required, the general marker list was reduced further by removing all lineages with no score based on previous results.

3.2 Metagenomic datasets

3.2.1 Artificial metagenomic datasets

3.2.1.1 Implementation

Artificial metagenomic datasets were created for testing of species and lineage specific oligonucleotides. A Java based metagenomic simulation program, *ReadSim* was used in the creation of each metagenome (Schmid and Huson, 2006). This program randomly selects fragments from a target genome sequence while incorporating errors to simulate sequencing using the respective technologies. Each metagenome was created to contain an estimated 100kbp of sequence. In a metagenome containing more than one species each genome was given an equal portion of sequence.

In order to allow for accurate and topic-specific testing four case studies were created. Each case study contained pathogenic bacteria, their closely related relatives and randomly selected species from within the same lineage and from different lineages. The species present in each case study and their respective lineages are listed in tables 3.1, 3.2, 3.3, 3.4.

The creation of metagenomes for each case study followed the same process:

- A single species metagenome were created for each species in the case study.

Table 3.2: Bacterial Species Present in Case Study 2. * Indicates a more distant species within the same lineage. ** Indicates a species from a different lineage

Accession Number	Species Description
NC_003997.3	*Bacillus anthracis str. Ames*
NC_006274.1	*Bacillus cereus E33L*
NC_006270.2	*Bacillus licheniformis ATCC 14580*
NC_008600.1	*Bacillus thuringiensis str. Al Hakam*
NC_000964.2	*Bacillus subtilis subsp. subtilis str. 168*
NC_002162.1	*Ureaplasma parvum serovar 3 str. ATCC 700970**
NC_007907.1	*Desulfitobacterium hafniense Y51**
NC_008609.1	*Pelobacter propionicus DSM 2379***
NC_008009.1	*Acidobacteria bacterium Ellin345***

Table 3.3: Bacterial Species Present in Case Study 3. * Indicates a more distant species within the same lineage. ** Indicates a species from a different lineage

Accession Number	Species Description
NC_008463.1	*Pseudomonas aeruginosa UCBPP-PA14*
NC_004578.1	*Pseudomonas syringae pv. tomato str. DC3000*
NC_008027.1	*Pseudomonas entomophila L48*
NC_004129.6	*Pseudomonas fluorescens Pf-5*
NC_002947.3	*Pseudomonas putida KT2440*
NC_007204.1	*Psychrobacter arcticus 273-4**
NC_008570.1	*Aeromonas hydrophila subsp. hydrophila ATCC 7966**
NC_003098.1	*Streptococcus pneumoniae R6***
NC_002578.1	*Thermoplasma acidophilum DSM 1728***

Table 3.4: Bacterial Species Present in Case Study 4. * Indicates a more distant species within the same lineage. ** Indicates a species from a different lineage

Accession Number	Species Description
NC_007946.1	*Escherichia coli UTI89*
NC_004631.1	*Salmonella enterica subsp. enterica serovar Typhi Ty2*
NC_004741.1	*Shigella flexneri 2a str. 2457T*
NC_007204.1	*Psychrobacter arcticus 273-4**
NC_005126.1	*Photorhabdus luminescens subsp. laumondii TTO1**
NC_008277.1	*Borrelia afzelii PKo***
NC_007298.1	*Dechloromonas aromatica RCB***

Table 3.5: Parameters used in creation of metagenomic datasets for Solexa and 454 pyrosequencing technologies.

	Solexa sequencing	**454 pyrosequencing**
Minimum fragment length	20	200
Maximum fragment length	35	350
Mean fragment length	25	250
Read length model	Uniform	Uniform
Number of fragments	4000	400

- Each species within the case study was classified into a phylogenetic grouping. The first grouping included the pathogenic species and all close neighbours. The second included distantly related species within the same lineage. The third included randomly selected species from different lineages and the fourth included a random genome. These groups were then used to randomly select species for the combined metagenomic datasets. Each combined metagenome contained a prescribed number of each group.

A random genome consisting of randomly generated sequence reads was also included. This provided a benchmark denoting the score that can be expected under random conditions.

Two different metagenomic datasets were created for each case study to simulate Solexa sequencing and 454 pyrosequencing technologies. The parameters for these datasets is shown in Table 3.5.

3.2.2 Experimental metagenomic datasets

In order to estimate the functionality of this approach in an actual environment an experimental dataset was tested. A metagenomic sample taken from the Deep Mediterranean was selected (Martín-Cuadrado *et al.*, 2007). This metagenome, a relatively small and well annotated example provided a useful benchmark. The sequenced reads in this metagenome are long reads between 400-700bp in length. Although not optimal to testing this approach the availability of raw sequence for well annotated metagenomes are generally present in sanger sequencing or assembled contigs, making acquisition of raw sequenced reads a difficult matter.

From the work of Martín-Cuadrado *et al.* (2007) a list of the most dominant bacterial species within the deep Mediterranean metagenome was assembled. Table 3.6 shows an overrepresentation of a large number of *Alphaproteobacteria* as well as *Acidobacteria*. In order to validate the developed identification approaches a small group of bacterial species were randomly selected that do not occur within the metagenome and do not belong to lineages present in the sample (Table 3.7).

Table 3.6: Dominant bacterial species identified in the Deep Mediterranean metagenomic project.

Species number	**Species description**	**Lineage**	**Number of BLAST hits***
1	*Mesorhizobium loti MAFF303099*	*Alphaproteobacteria*	112
2	*Mesorhizobium sp. BNC1*	*Alphaproteobacteria*	84
3	*Rhodopseudomonas palustris BisA53*	*Alphaproteobacteria*	27
4	*Candidatus Pelagibacter ubique HTCC1062*	*Alphaproteobacteria*	84
5	*Solibacter usitatus ElLineage 6076*	*Acidobacteria*	56
6	*Dehalococcoides sp. CBDB1*	*Chloroflexi*	42
7	*Magnetospirillum magneticum AMB-1*	*Alphaproteobacteria*	36
8	*Pseudomonas aeruginosa UCBPP-PA14*	*Gammaproteobacteria*	36
9	*Acidobacteria bacterium ElLineage 345*	*Acidobacteria*	32
10	*Burkholderia sp. 383*	*Betaproteobacteria*	32

Table 3.7: Additional species, not found in the metagenome, added to the Deep Mediterranean dataset.

Species number	**Species description**	**Lineage**
1	*Chlamydophila pneumoniae CWL029*	*Chlamydiae/Verrucomicrobia*
2	*Pelodictyon luteolum DSM 273*	*Bacteroidetes/Chlorobi*
3	*Bifidobacterium longum NCC2705*	*Actinobacteria*
4	*Thermobifida fusca YX*	*Actinobacteria*

3.3 Analysis of oligonucleotide frequency profiles

3.3.1 Results from oligonucleotide frequency analysis

3.3.1.1 Species specific analysis

Each of the pathogenic species profiles were tested against the collection of metagenomes within each case study to produce the results in Figures 3.2, 3.3, 3.4 and 3.5.

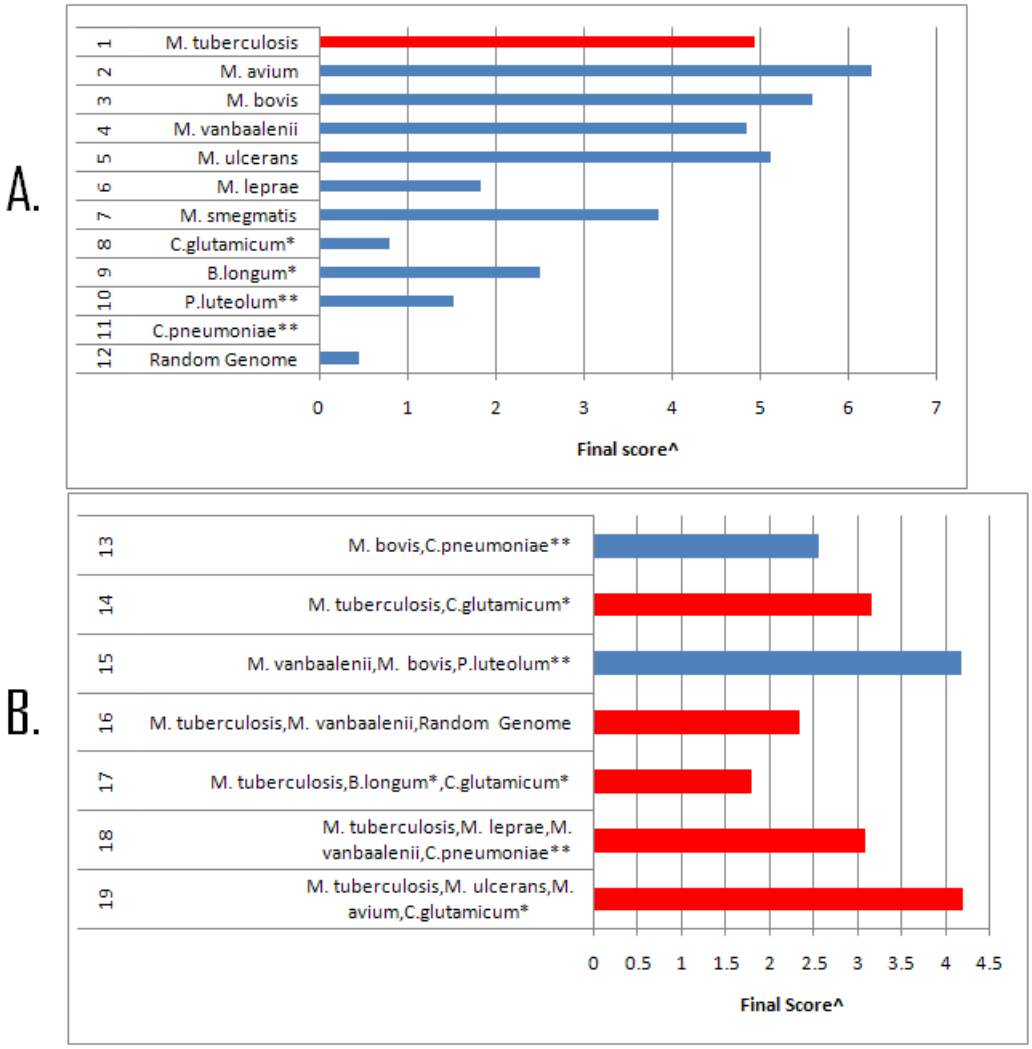
In general, the final scores for all species profiles tested against a single metagenome were above one. The species profiles clearly struggled to differentiate effectively between closely related neighbours. These results could be predicted by the difficulties in separating closely related species based on marker results from Section 2.5.3 in Chapter 2. Furthermore, the score magnitudes for each case study differed, with case study 1 (Figure 3.2) containing the highest values. This suggests the need for further corrections to be made to the score. Moreover, the presence of a large number of repeats within these genomes can help explain the difficulty in separation of these species and the inflated values in case study 1.

Pseudomonas aeruginosa (Figure 3.4) and *Salmonella enterica* (Figure 3.5) can be differentiated more reliably from their relatives, however, the relative's scores remain above one. An improved result can be expected as species within these case studies are more distantly related.

An anomaly in case study 4 was the score for *Dechloromonas aromatica* (Figure 3.5) which was much higher than expected. This supplied an example where species from different lineages may share a large number of overrepresented oligonucleotides, this will need to be taken into consideration in further research. Interestingly, *Ureaplasma parvum* (Figure 3.3), a species within the same lineage as *Bacillus anthracis*, seems to have shared a large number of repeat sequences with this specie. From this observation, a repeat rich genome such as *Ureaplasma parvum* can confuse results when identifying bacterial species using oligonucleotide frequencies.

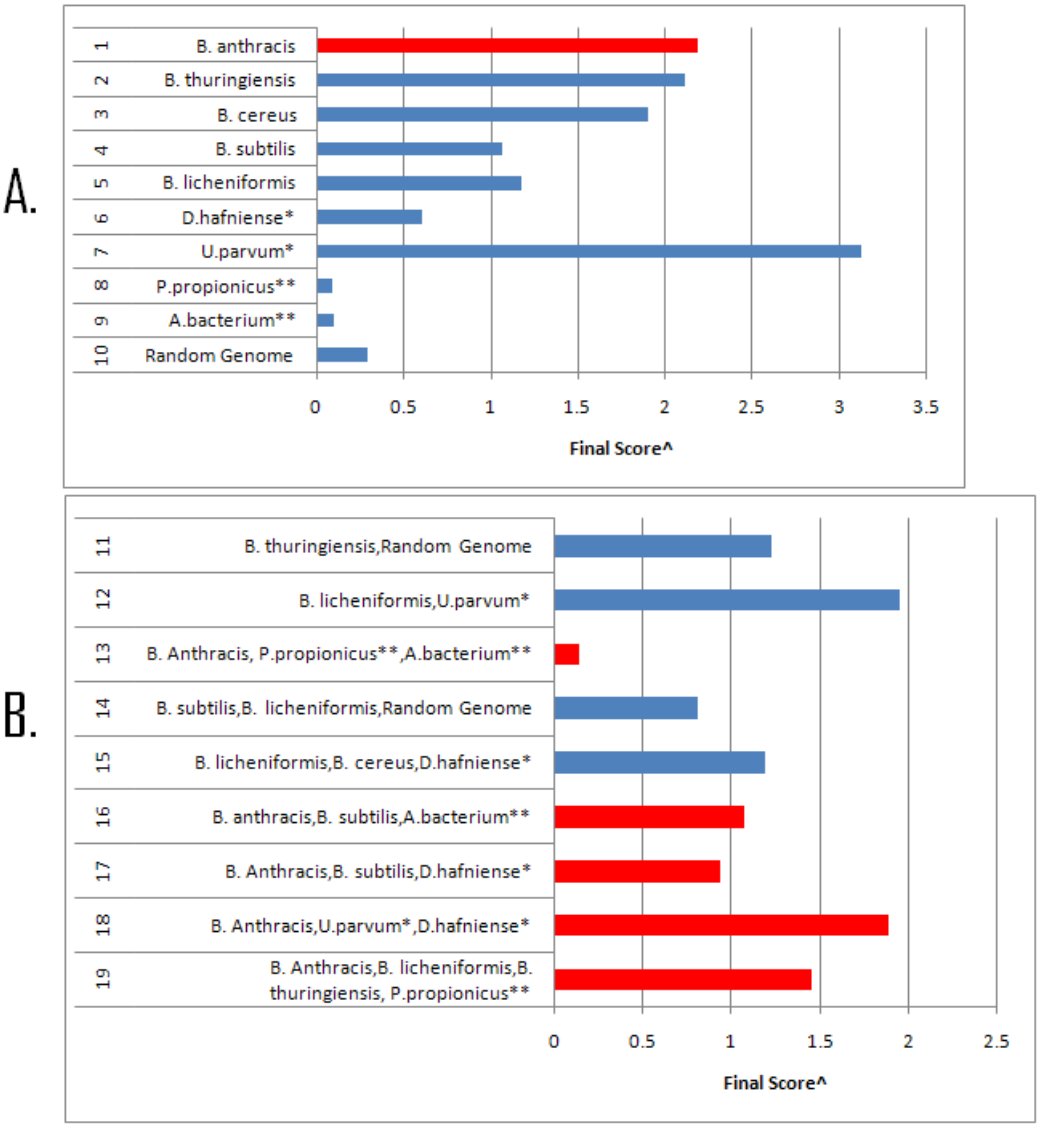
In terms of combined metagenomes lower scores for species profiles were generally obtained compared to single species metagenomes. Combined metagenomes, containing distantly related species or a random metagenome showed a significant decrease in score (Figure 3.3, metagenome 13). This indicated that signal can be easily diluted when different species are present at equal ratios. Another common discrepancy is that the presence of several closely related species found in the same metagenome can boost the signal of a closely related species profile regardless of its presence or absence. The chance of incorrectly identifying a species as one of its relatives is a definite possibility. Therefore this method cannot be reliably implemented under current conditions.

In case study 4 (Figure 3.5 B), metagenomes 17 and 18 provided an interesting insight into the testing of metagenomic sequence. Both metagenomes rely on the same three



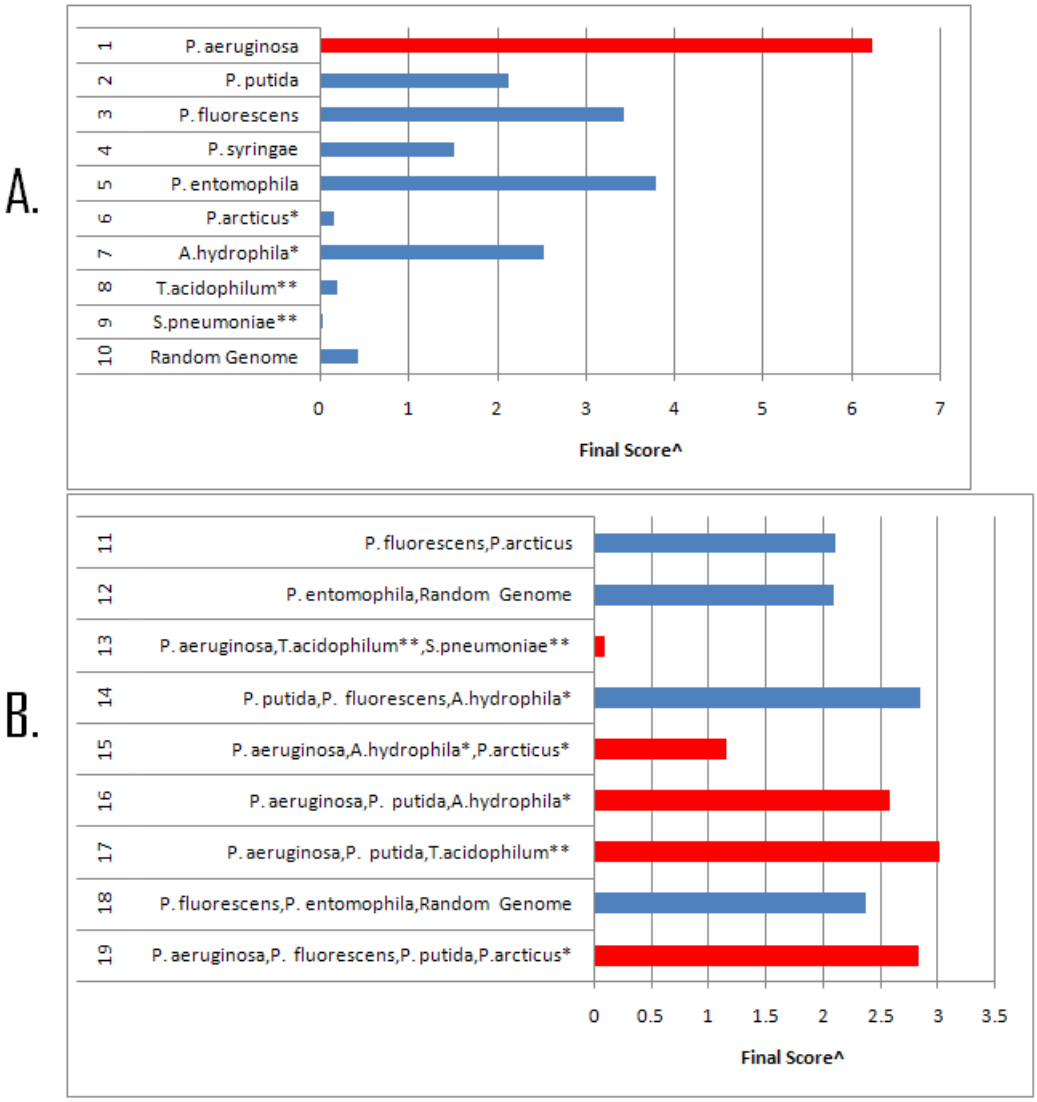
Red bars indicate metagenomes containing the organism being searched for
** Indicates a more distant species from the same lineage*
*** Indicates a species from a different lineage*
^ Score calculated using Algorithm 13
Error value used is 96.06

Figure 3.2: The use of the *Mycobacterium tuberculosis* species profile on single (Figure A) and combined metagenomes (Figure B). Score for each metagenome is plotted on the X-axis.



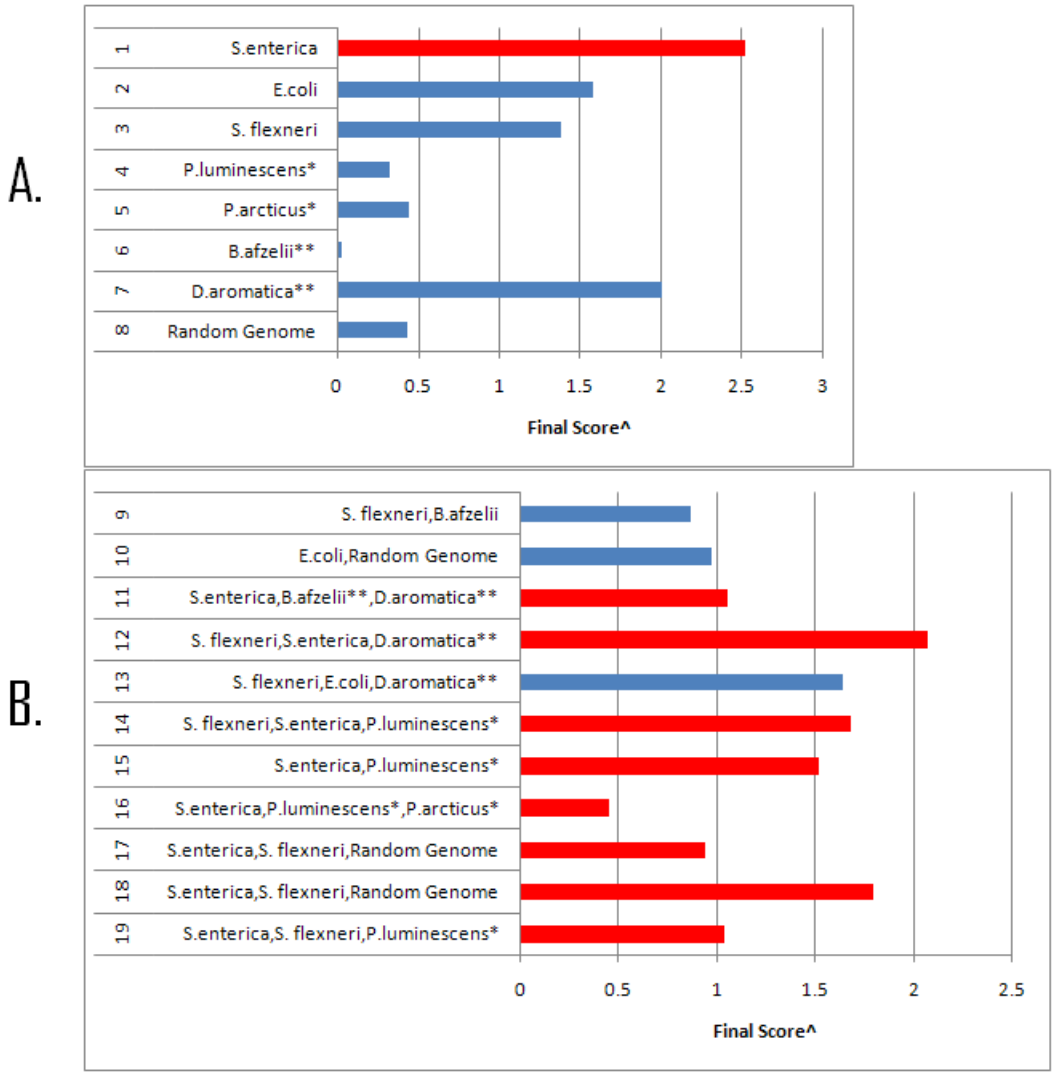
Red bars indicate metagenomes containing the organism being searched for
** Indicates a more distant species from the same lineage*
*** Indicates a species from a different lineage*
^ Score calculated using Algorithm 13
Error value used is 224.68

Figure 3.3: The use of the *Bacillus anthracis* species profile on single (Figure A) and combined metagenomes (Figure B). Score for each metagenome is plotted on the X-axis.



Red bars indicate metagenomes containing the organism being searched for
** Indicates a more distant species from the same lineage*
*** Indicates a species from a different lineage*
^ Score calculated using Algorithm 13
Error value used is 77.07

Figure 3.4: The use of the *Pseudomonas aeruginosa* species profile on single (Figure A) and combined metagenomes (Figure B). Score for each metagenome is plotted on the X-axis.



Red bars indicate metagenomes containing the organism being searched for
** Indicates a more distant species from the same lineage*
*** Indicates a species from a different lineage*
^ Score calculated using Algorithm 13
Error value used is 182.35

Figure 3.5: The use of the *Salmonella enterica* species profile on single (Figure A) and combined metagenomes (Figure B). Score for each metagenome is plotted on the X-axis.

genomic species, however, the results differ substantially. This showed the inherent variability in this approach and the difficulty in estimating presence or absence of bacterial species (Figure 3.2, metagenome 15).

The conclusion from these findings is that this method could only differentiate species that are rather distant in terms of phylogeny. Furthermore, in a combined metagenome, signal is either disproportionately increased or decreased due to low marker specificity, therefore, no reliable conclusions could be drawn.

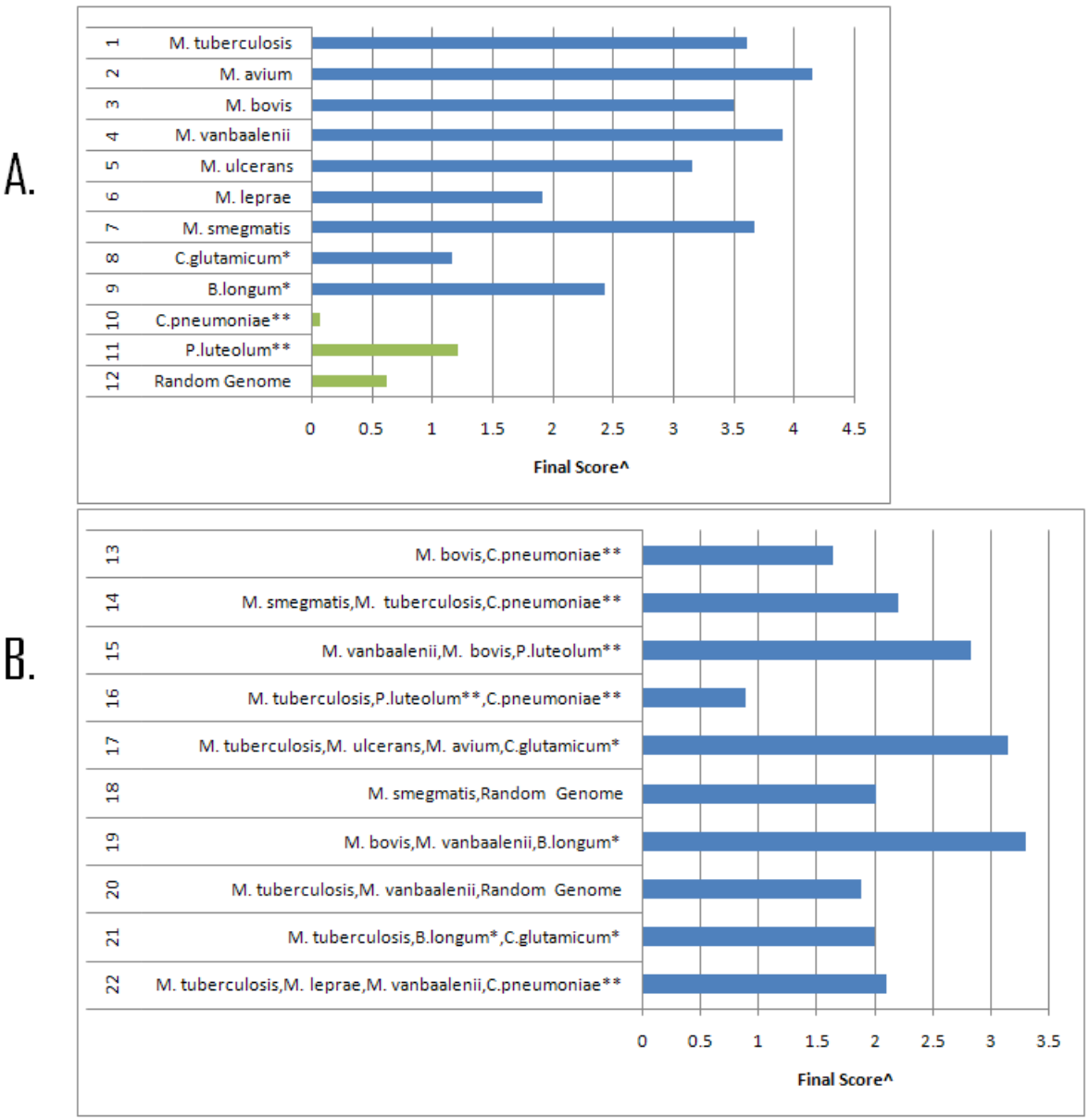
3.3.1.2 Lineage specific analysis

A lineage analysis for each case study was performed using the most prominent lineage in the case study. The figures displaying these results are 3.6, 3.7, 3.8 and 3.9.

In the majority of case studies these results showed high scores for all bacterial species within the lineage, however, there remain a few exceptions. This was due to the sizes of the lineages and the compromise in identifying a consensus list. This process resulted in the exclusion of certain oligonucleotides which may describe certain species more accurately. Distantly related species had profile scores close to one, indicating their distance from the current lineage and providing further support for this method.

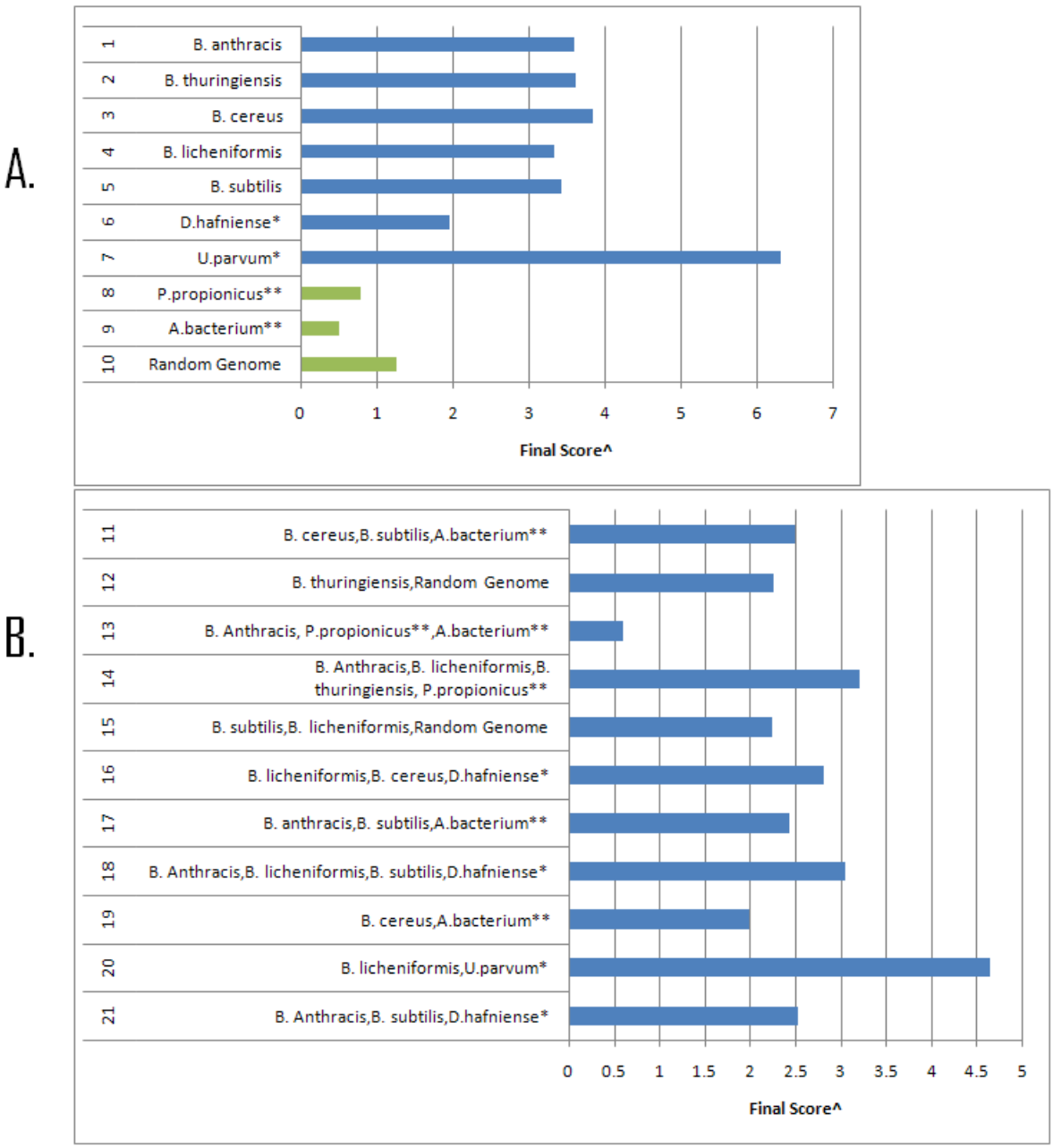
The results from Figures 3.8 and 3.9 showed the scores for the *Gammaproteobacteria* lineage profile. All bacterial species within these case studies carried little or no signal for the *Gammaproteobacteria* lineage. The pathogen and all related species displayed scores well below one with only *P. syringae* (Figure 3.8 A, metagenome 4) showing a value slightly above one. This suggested that this lineage profile does not accurately describe these families of bacteria. The presence of a false positive score for *Borrelia afzelii* (Figure 3.9, metagenome 6) equal to that of other species present in case study 4, confirmed the poor discriminating power of this lineage profile. This reiterated the findings in Section 2.4.1.2, Chapter 2. From Figure 2.6 a large variation in species can be seen within the lineage with distances reaching up to 2500 units. Creation of a consensus list for such a diverse lineage is often not satisfactory for a significant portion of members, as is evidenced by these results.

In terms of the combined metagenomes the lineage profile results looked positive. The number of species within the data strengthened or weakened the score dependent on their relation to the lineage. In case study 1 (Figure 3.6 B) metagenome 16 provided an example where two bacterial species from different lineages prevented the identification of a single species from the *Actinobacteria* lineage. This yielded an indication of the sensitivity of this method. Therefore if the majority of species are from different lineages the signal weakened considerably. In Figure 3.6 B metagenomes 13 and 18 included a single species from an alien lineage where a high score was still given. This indicated that in an environment where a lineage is represented by half of the metagenome (50kbp) detection was possible. The combined metagenomes for case study 3 (Figure 3.8 B) and 4



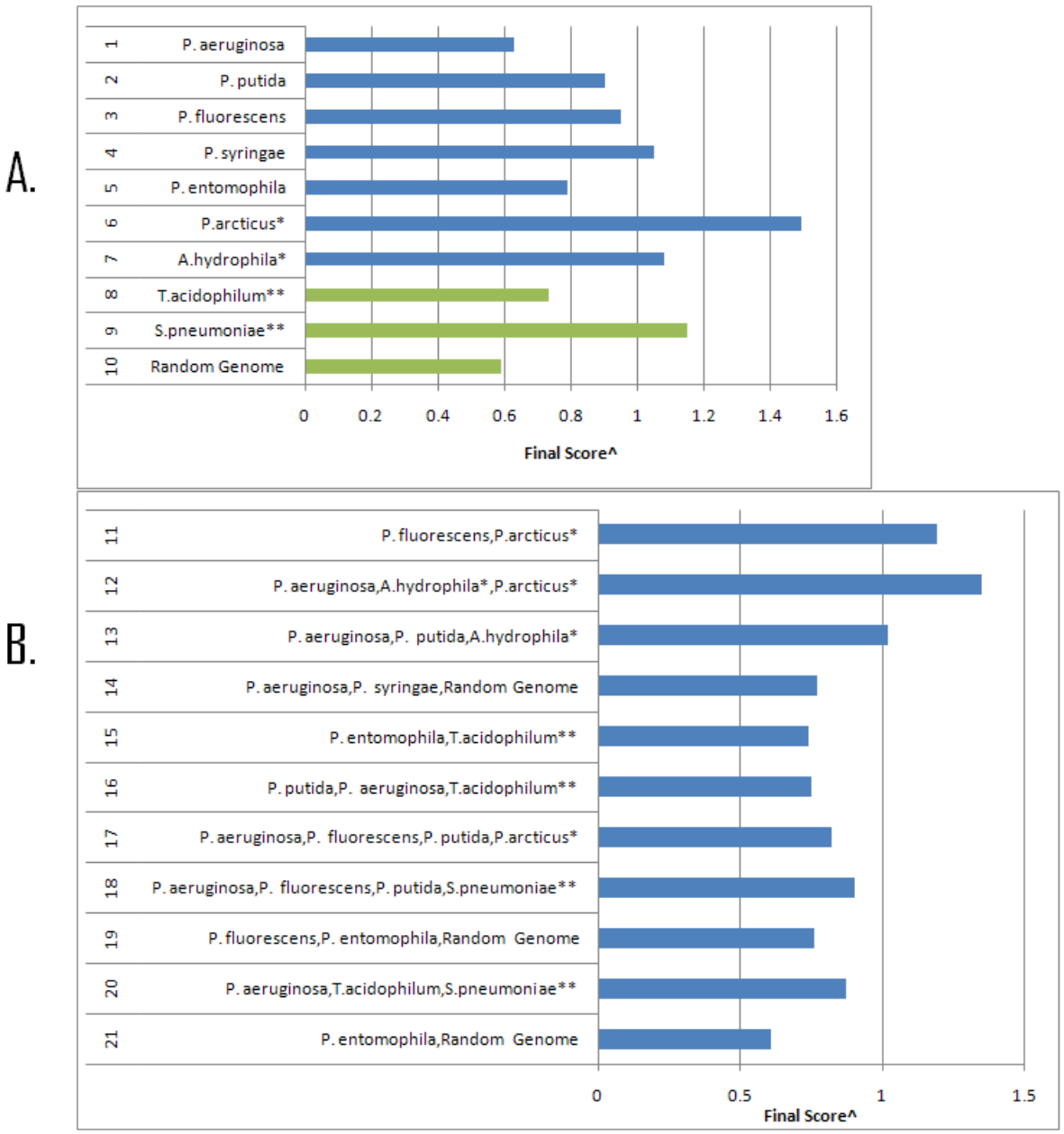
Green bars indicate single metagenomes with species not present in the lineage
** Indicates a more distant species from the same lineage*
*** Indicates a species from a different lineage*
^ Score calculated using Algorithm 13
Error value used is 174.76

Figure 3.6: Lineage profile for case study 1. This case study was evaluated using the *Actinobacteria* lineage profile. Figures A and B show single and combined metagenomes respectively. Score for each metagenome is plotted on the X-axis.



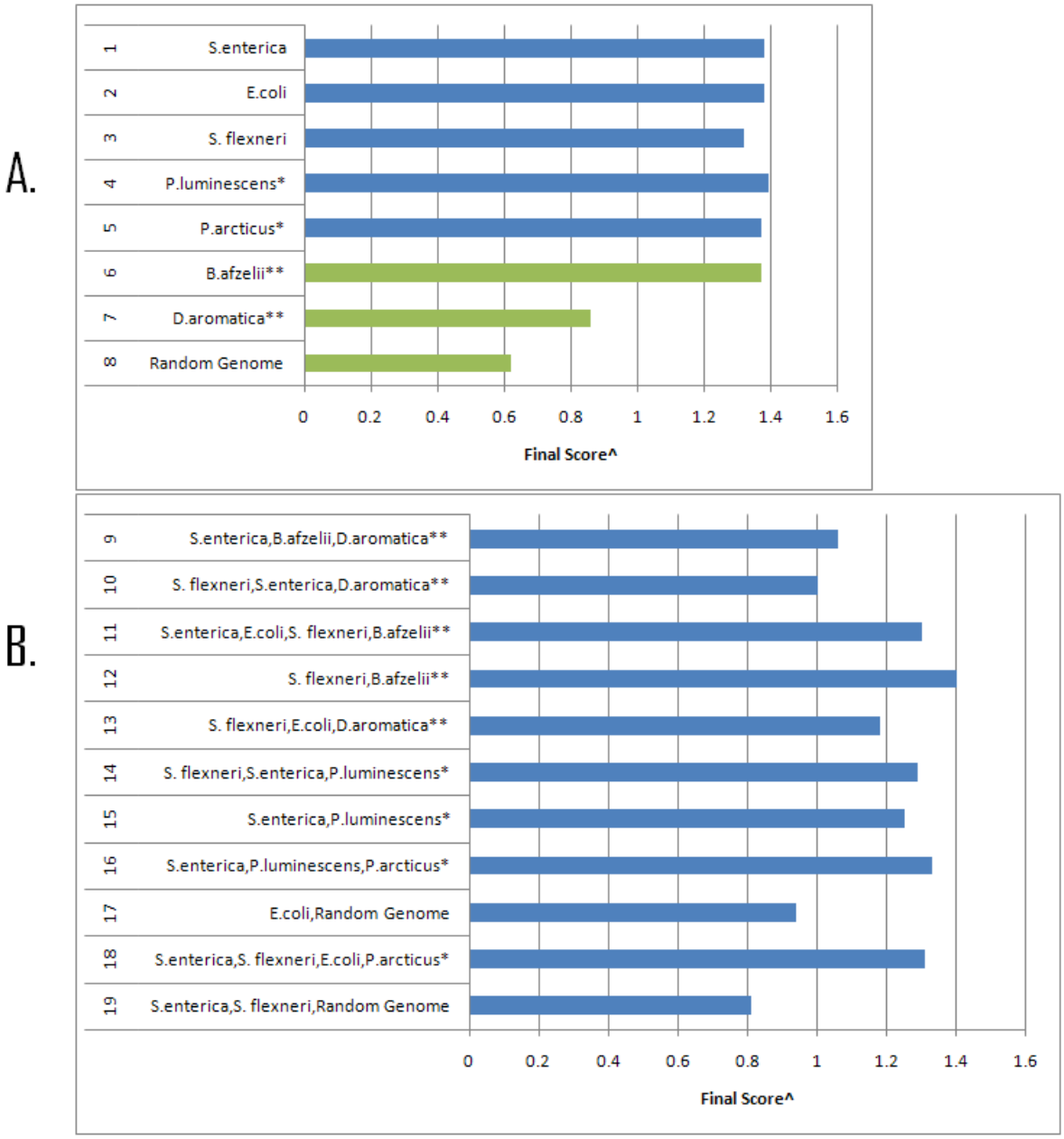
Green bars indicate single metagenomes with species not present in the lineage
** Indicates a more distant species from the same lineage*
*** Indicates a species from a different lineage*
^ Score calculated using Algorithm 13
Error value used is 127.76

Figure 3.7: Lineage profile for case study 2. This case study was evaluated using the *Firmicutes* lineage profile. Figures A and B show single and combined metagenomes respectively. Score for each metagenome is plotted on the X-axis.



Green bars indicate single metagenomes with species not present in the lineage
** Indicates a more distant species from the same lineage*
*** Indicates a species from a different lineage*
^ Score calculated using Algorithm 13
Error value used is 410.52

Figure 3.8: Lineage profile for case study 3. This case study was evaluated using the *Gammaproteobacteria* lineage profile. Figures A and B show single and combined metagenomes respectively. Score for each metagenome is plotted on the X-axis.



Green bars indicate single metagenomes with species not present in the lineage
** Indicates a more distant species from the same lineage*
*** Indicates a species from a different lineage*
^ Score calculated using Algorithm 13
Error value used is 410.52

Figure 3.9: Lineage profile for case study 4. This case study was evaluated using the *Gammaproteobacteria* lineage profile. Figures A and B show single and combined metagenomes respectively. Score for each metagenome is plotted on the X-axis.

Table 3.8: The number of markers usable by *MarkerCounter* for species in Case Study 1.

Species	Number of Markers
Mycobacterium avium	48 markers
Mycobacterium bovis	10 markers
Mycobacterium tuberculosis	14 markers
Mycobacterium leprae	640 markers
Mycobacterium smegmatis	82 markers
Mycobacterium ulcerans	242 markers
Mycobacterium vanbaalenii	34 markers
*Bifidobacterium longum**	237 markers
*Corynebacterium glutamicum**	664 markers
*Chlamydophila pneumoniae***	644 markers
*Pelodictyon luteolum***	546 markers

Table 3.9: Markers usable by *MarkerCounter* for species in Case Study 2.

Species	Number of Markers
Bacillus anthracis	22 markers
Bacillus cereus	10 markers
Bacillus licheniformis	144 markers
Bacillus thuringiensis	22 markers
Bacillus subtilis	118 markers
*Ureaplasma parvum**	103 markers
*Desulfitobacterium hafniense**	103 markers
*Pelobacter propionicus***	514 markers
*Acidobacteria bacterium***	623 markers

(Figure 3.9 B) showed a lack of confidence, even in the cases where all organisms originated from the *Gammaproteobacteria* lineage. It is clear that the *Gammaproteobacteria* lineage profile is not sufficient for the detection of lineage species using oligonucleotide frequencies.

3.4 Analysis of sequenced reads

3.4.1 Results for the analysis of sequenced reads

3.4.1.1 Species profile results

Marker tables for case studies

On inspection of Tables 3.8 and on the current page the first signs of bias can be detected. A large fluctuation in the number of markers used by each species in the case studies is present. As a result the pathogens, *Mycobacterium tuberculosis* and *Bacillus anthracis*, and their closely related species all contain a very limited number of markers. *MarkerCounter* removed all oligonucleotide markers that occurred in more than one species profile. Therefore, the most closely related organisms were most severely affected as their

Table 3.10: Markers usable by *MarkerCounter* for species in Case Study 3.

Species	Number of Markers
Pseudomonas aeruginosa	126 markers
Pseudomonas syringae	382 markers
Pseudomonas entomophila	65 markers
Pseudomonas fluorescens	91 markers
Pseudomonas putida	174 markers
*Psychrobacter arcticus**	125 markers
*Aeromonas hydrophila**	152 markers
*Streptococcus pneumoniae***	437 markers
*Thermoplasma acidophilum***	646 markers

Table 3.11: Markers usable by *MarkerCounter* for species in Case Study 4.

Species	Number of Markers
Escherichia coli	445 markers
Salmonella enterica	274 markers
Shigella flexneri	472 markers
*Psychrobacter arcticus**	125 markers
*Photorhabdus luminescens**	215 markers
*Borrelia afzelii***	163 markers
*Dechloromonas aromatica***	461 markers

marker counts were significantly reduced. Nonetheless, the removal of all shared markers dramatically decreased the false positives in the results and thereby improved signal quality.

The fluctuation in markers present within these tables also showed several species with exceptionally high numbers of markers. These were generally species that were distantly related to the pathogenic species as they do not share markers with any other organisms within the dataset. Secondly, a large number of markers could indicate a repeat poor species where hundreds of oligonucleotides were required to achieve the cumulative expected value threshold (please see Section 2.5.3 for further information). This provided a different form of bias as there was an increased probability of encountering these markers by chance. In an attempt to correct for this the hits per marker score was calculated (Algorithm 15).

On inspection of Tables 3.10 and 3.11 more balanced marker profiles were viewed. Each species within the case studies featured a much larger selection of markers, this gave an indication that marker bias will be minimized.

Single species metagenomes

From Figures 3.10 and 3.12 the species profiles tested displayed that both *Mycobacterium tuberculosis* and *Bacillus anthracis* could not be detected effectively. This is attributed

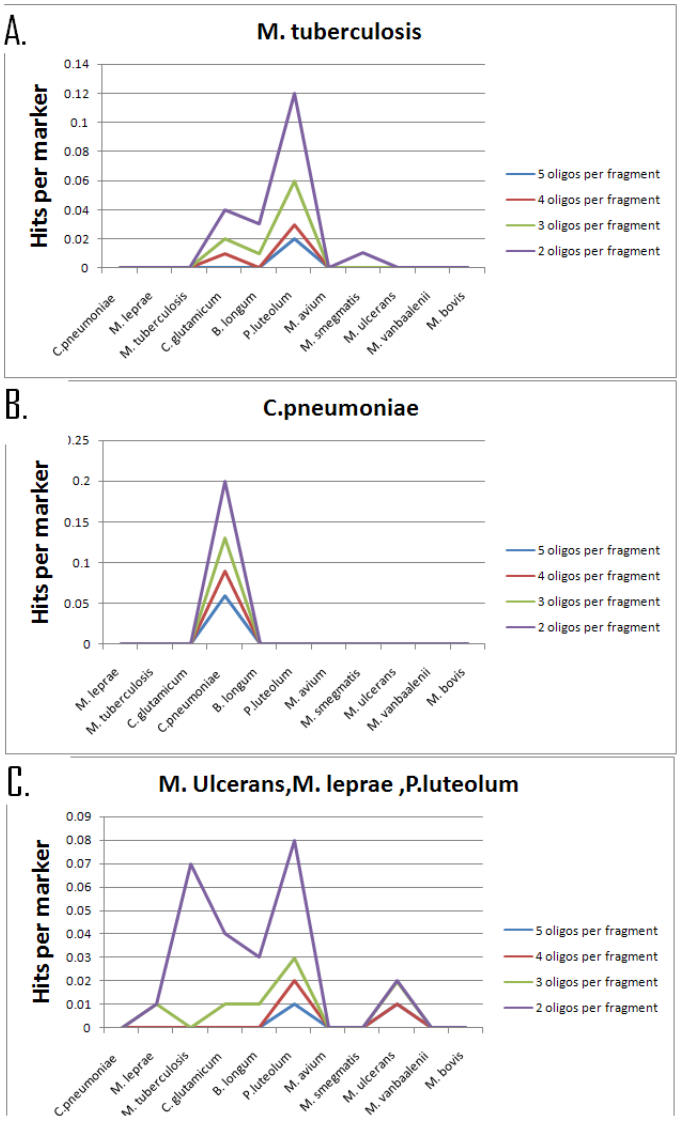


Figure 3.10: Case Study 1 species profiles. Figures A and B show marker results against metagenomes containing *Mycobacterium tuberculosis* and *Chlamydophila pneumoniae* respectively. Figure C shows marker results against a combined metagenome of species *Mycobacterium ulcerans,Mycobacterium leprae* and *Pelodictyon luteolum*.

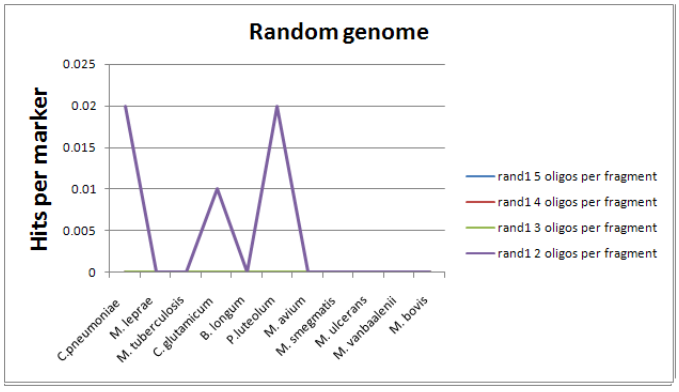


Figure 3.11: Species profile for a random metagenome taken from Case Study 1.

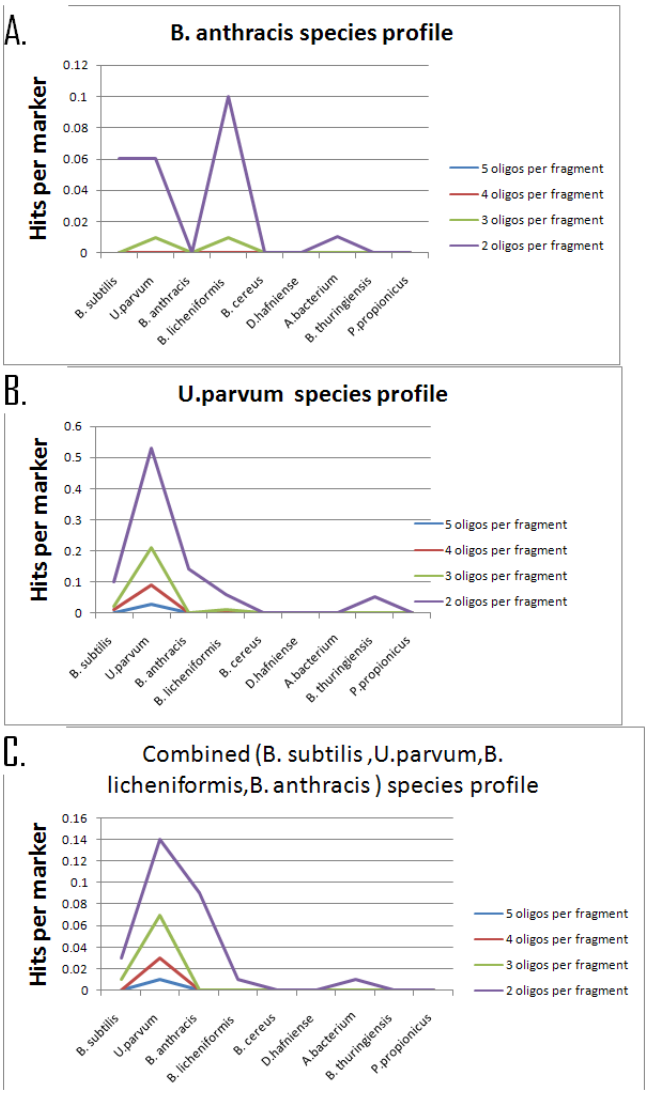


Figure 3.12: Case Study 2 species profiles. Figures A and B show marker results against metagenomes containing *Bacillus anthracis* and *Ureaplasma parvum* respectively. Figure C shows marker results against a combined metagenome of species *Bacillus anthracis, Bacillus licheniformis, Ureaplasma parvum* and *Bacillus subtilis.*

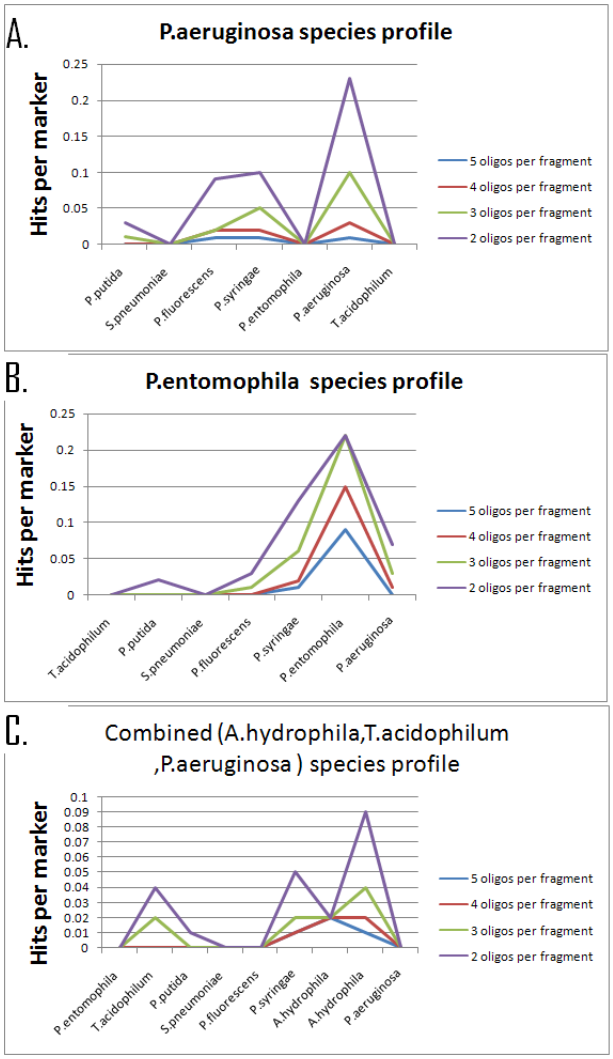


Figure 3.13: Case Study 3 species profiles. Figures A and B show marker results against metagenomes containing *Pseudomonas aeruginosa* and *Pseudomonas entomophila*. respectively. Figure C shows marker results against a combined metagenome of species *Aeromonas hydrophila,Pseudomonas aeruginosa* and *Thermoplasma acidophilum*.

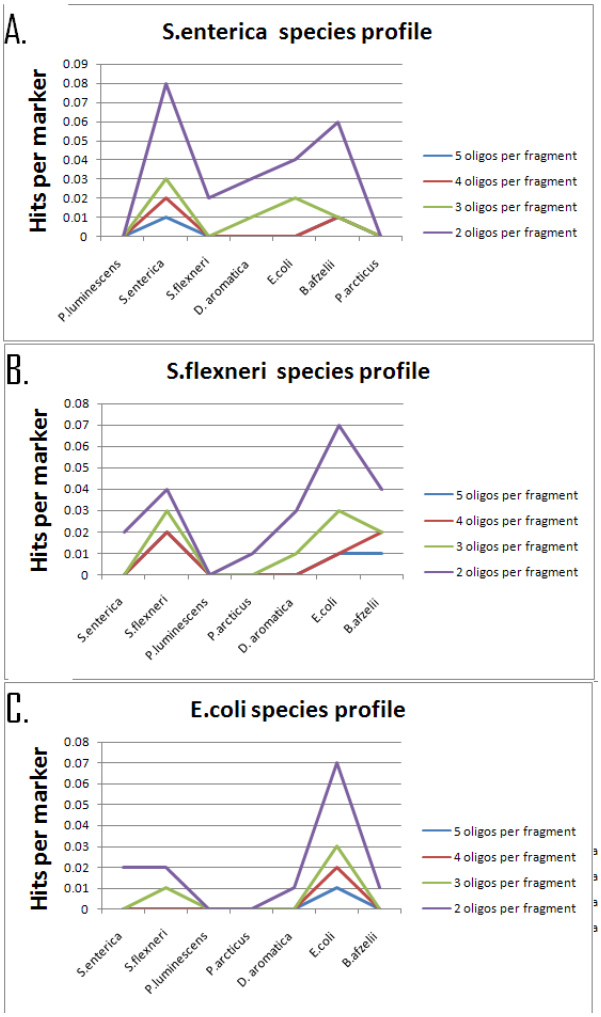


Figure 3.14: Case Study 4 species profiles. Figures A,B and C show marker results against single species metagenomes containing *Salmonella enterica*, *Shigella flexneri* and *Escherichia coli* respectively.

to the lack of viable markers. The false positive detection of all species distant from the study indicate that there was a definite bias towards these species. The calculation to normalize the number of markers per hit clearly was not effective in preventing this bias.

From Figure 3.10 B (case study 1) a very specific signal was present as *Chlamydophila pneumoniae* was represented with complete certainty. It must be taken into account that no other species from this lineage was included in this case study. This also gives a clear indication of how a large number of markers could amplify a score, this phenomenon was present in every case study. Interestingly, in Figure 3.12 B (case study 2), *Ureaplasma parvum* also indicated a strong hit. This specie, falling within the same lineage as *Bacillus anthracis,* was clearly exceptionally rich in repeat regions and confirmed results obtained in oligonucleotide frequency analysis.

Figures 3.13 and 3.14 described the species profiles for *Pseudomonas aeruginosa* and *Salmonella enterica*. There was a clear signal showing the presence of these pathogens with only mild interference from closely related species. This provided support for the differentiation of species from their neighbours on the condition that a sizable set of markers could be identified. In Figure 3.14 B *Shigella flexneri* was clearly not easily identifiable as *Escherichia coli* produced a more distinctive signal. In contrast to Figure 3.14 C where *Escherichia coli* was easily identified over and above *Shigella flexneri*. This shows that the differentiation of *Shigella flexneri* from *Escherichia coli* was clearly far more problematic and could imply that *Shigella flexneri* could not be identified effectively using the current method.

Combined species metagenomes

Generally the combined metagenomes for each case study show that the majority of species could be detected, however, there was significant interference. All species distant from the pathogen and its relatives were detected regardless of presence in the metagenome. There was also a pronounced decrease in the score of the results in combined metagenomes compared to the single species metagenomes. This could be expected as less signal from each species is present.

As identified in Figure 3.10 C (case study 1) the false positive appearance of *Mycobacterium tuberculosis* in the dataset was due to the presence of highly similar sequence from both *Mycobacterium leprae* and *Mycobacterium ulcerans*, this result was clearly not highly significant, as it was only obtained at the ”2 oligos per fragment” threshold.

Figure 3.18 A (case study 4) offered an example of the effectivity of this method. Under conditions where species markers were relatively well balanced and species within the metagenome were not highly similar a clear result could be obtained. The multiple species metagenome in Figure 3.18 A showed an optimal situation where all species present in the metagenome were identified without any background noise. However, the astronomical value for *Borrelia afzelii* (*Spirochaetes* lineage) indicated that the markers for this genome

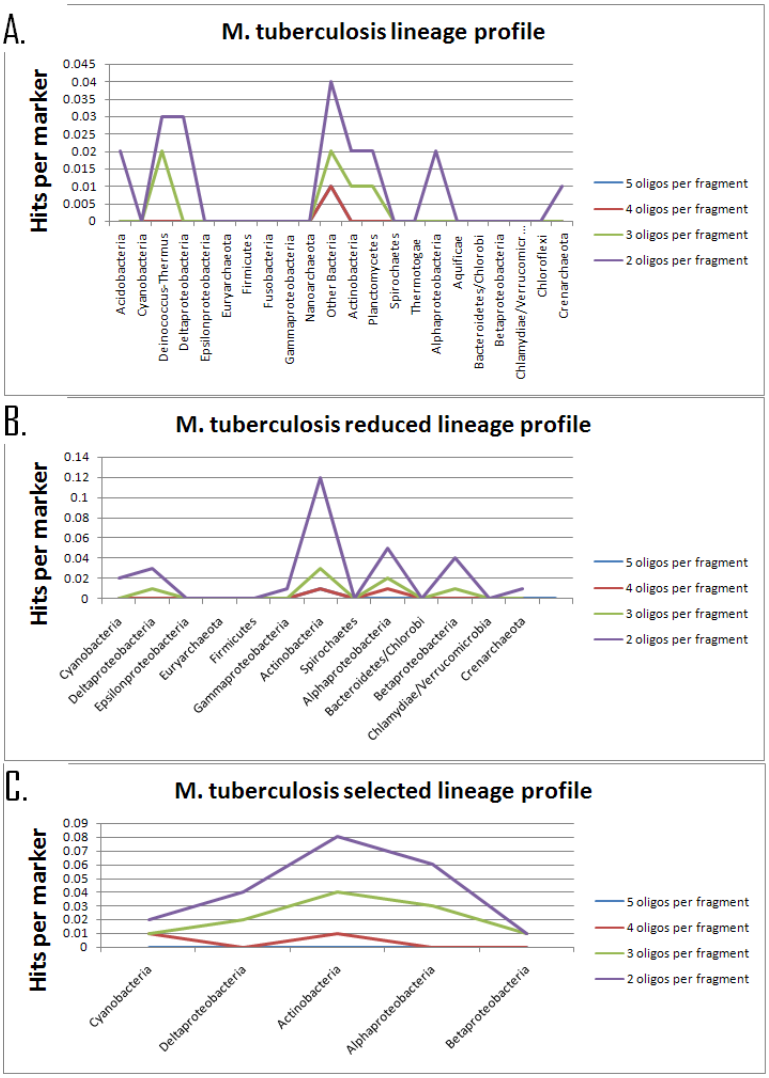


Figure 3.15: Case Study 1 lineage profiles. All Figures show the lineage profile results of *Mycobacterium tuberculosis (Actinobacteria lineage).* Figure A shows the full lineage profile for *Mycobacterium tuberculosis.* Figure B shows the reduced lineage profile. Figure C shows the further selection of several lineages with the highest scores.

were not specific. This emphasized that creation of a species profile from a small lineage could dramatically increase the number of poorly selected oligonucleotides.

3.4.1.2 Lineage profile results

Single species metagenomes

Figure 3.15 showed the results of different lineage profiles tested against a *Mycobacterium tuberculosis (Actinobacteria lineage)* metagenome. Figure 3.15 A showed the complete lineage profile containing all lineages. This figure displayed a great deal of background interference and no clear signal identifying the *Actinobacteria lineage.* Figure 3.15 B displayed a reduced lineage profile, only including lineages containing 5 or more species. This greatly improved score and produced a pronounced peak. This implied that small

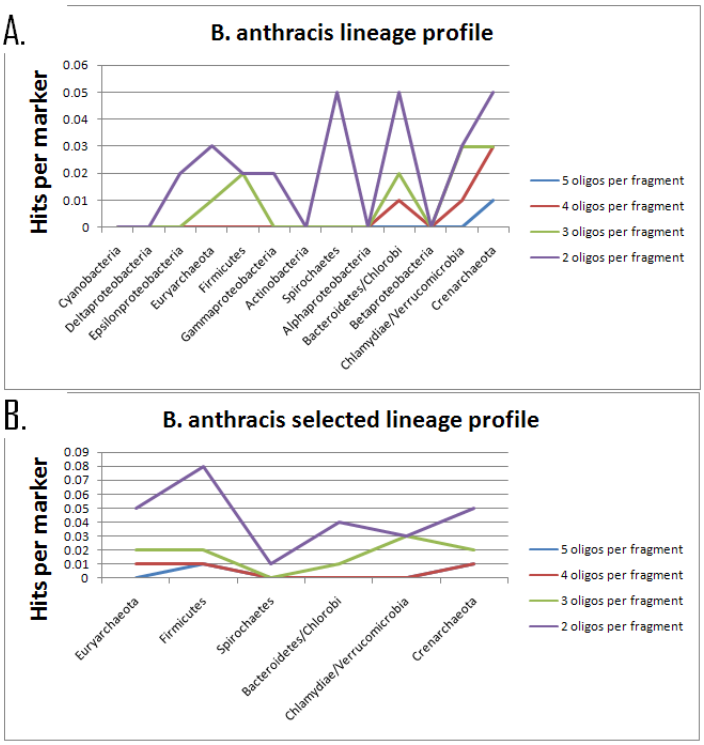


Figure 3.16: Case Study 2 lineage profiles. All Figures show the lineage profile results of *Bacillus anthracis (Firmicutes lineage).* Figure A shows the full lineage profile for *Bacillus anthracis.* Figure B shows the reduced lineage profile.

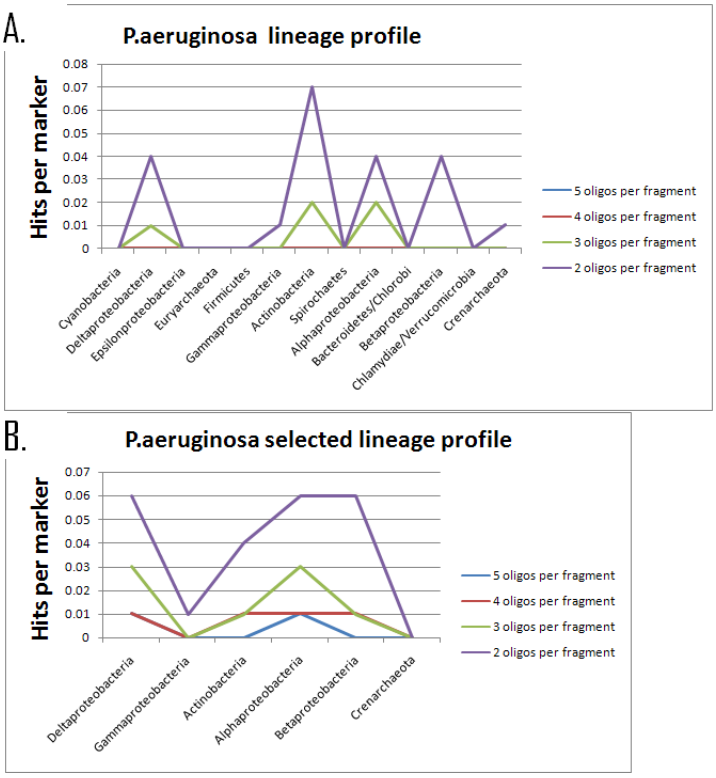


Figure 3.17: Case Study 3 lineage profiles. All Figures show the lineage profile results of *Pseudomonas aeruginosa (Gammaproteobacteria lineage).* Figure A shows the full lineage profile for *Pseudomonas aeruginosa* while Figure B shows the reduced lineage profile.

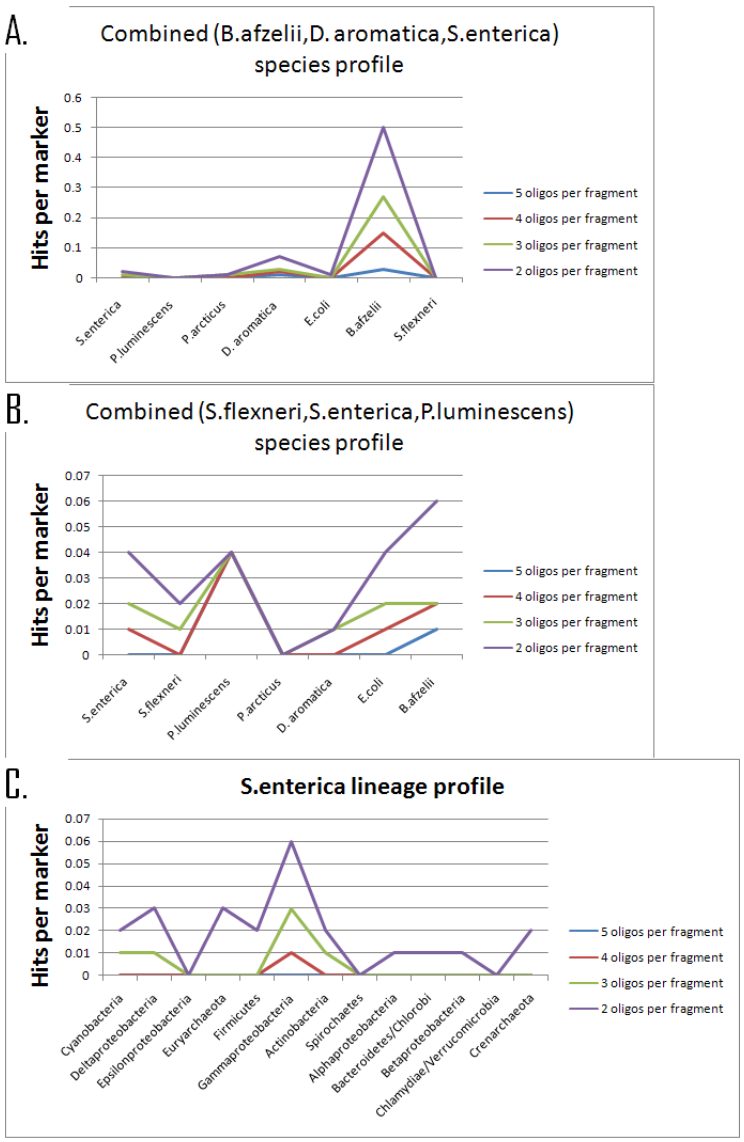


Figure 3.18: Figure A and B show marker results against a combined metagenome of species. Metagenome A contains *Dechloromonas aromatica, Borrelia afzelii* and *Salmonella enterica*. Metagenome B contains *Shigella flexneri, Photorhabdus luminescens* and *Salmonella enterica*. Figure C shows the full lineage profile for *Salmonella enterica (Gammaproteobacteria lineage)*.

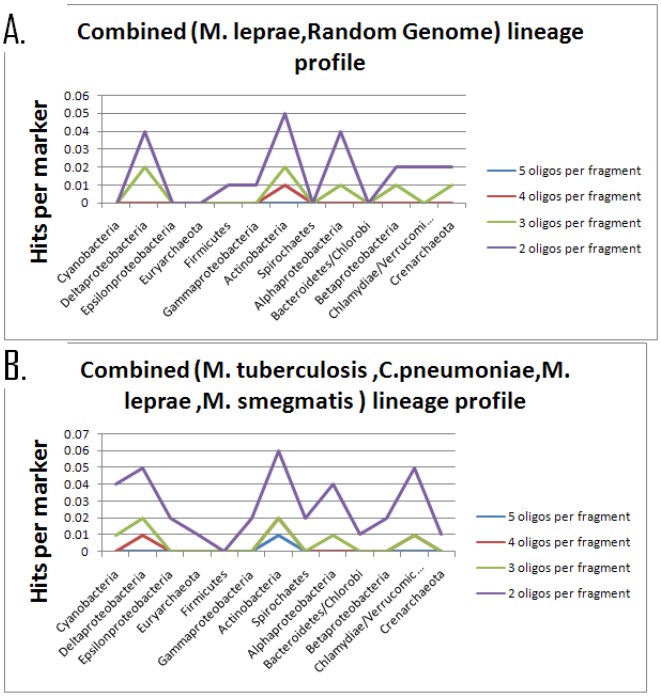


Figure 3.19: Combined metagenomes from Case Study 1, analysis is done using lineage profiles. Figure A shows a combined metagenome containing *Mycobacterium leprae* and a random metagenome. Figure B shows a combined metagenome containing *Mycobacterium tuberculosis,Mycobacterium smegmatis,Mycobacterium leprae* and *Chlamydophila pneumoniae.*

lineages caused bias which could skew results and were subsequently removed from further analysis. The same general trend was experienced in the majority of case studies. A further step was taken to improve resolution by removing all lineages with no score from Figure 3.15 B. From Figure 3.15 C it was clear that the *Actinobacteria* lineage remains the most prominent.

From Figures 3.17 A identified that the *Gammaproteobacteria* lineage profile did not effectively identify *Pseudomonas aeruginosa*. After reducing the number of lineages no improvement was visible (Figure 3.17 B). Despite this, the *Salmonella enterica* metagenomic sample achieved a surprisingly good result from the lineage profile (Figure 3.14 C). This indicated that the *Gammaproteobacteria* lineage profile could adequately describe this species using sequence read analysis.

Combined species metagenomes

The identification of the lineages present within a multiple species metagenome is now explored. From Figure 3.19 A it was clear that although the correct lineage peak was identified background noise was substantial. Furthermore, the scores for the lineage had decreased compared to the single species metagenomes. In Figure 3.19 B a more complicated metagenome, containing three species from the *Actinobacteria* lineage and one

species from the *Chlamydiae/Verrucomicrobia* lineage is shown. The increase in numbers of species has resulted in a dramatic increase in interference. Although the *Chlamydiae/Verrucomicrobia* lineage was present it does not peak above the background noise and hence could not be reliably identified. This reiterated the findings in Section 3.3.1.2 showing that lineage profiles could not be reliably identified unless species within the lineage are abundant within the metagenome.

3.4.2 Experimental Data Results

3.4.2.1 Results for the analysis of sequenced reads

From inspection of Table 3.12 it could be noted that all species in the full dataset have sufficient markers with the exception of *Dehalococcoides sp. CBDB1*. Furthermore, the differences in marker sizes between species in the full and reduced dataset showed that the species added to the core dataset had a minor effect on the number of markers for each core specie. The balanced marker profiles reduced the chance of marker bias affecting these results.

Figure 3.20 showed the scores for species profiles tested against the deep Mediterranean metagenome. The first identifiable feature in Figure 3.20 A was the astronomical value of *Chlamydophila pneumoniae,* a genome not present within this dataset. This provided yet another example of bias due not only to high marker count but unspecific marker selection.

Chlamydophila pneumoniae was then removed from the data. Figure 3.20 B displayed a scenario where a large number of false positives were still present. The highest peaks present in the data belonged to species not known to be present within the metagenome (*Pelodictyon luteolum* and *Bifidobacterium longum*). Several of the species found within the metagenome do have relatively high scoring values, however, these proved inconclusive.

An attempt was made to determine how accurately the core species within the metagenome can be identified (Figure 3.21). Figure 3.21 A displayed the overrepresentation of *Candidatus Pelagibacter ubique* within the metagenome. This species was shown to be present at high levels within the metagenome and this described a true positive. However, the excessive values associated with the species implied that the species profile may be amplified by other sequences within the metagenome. *Candidatus Pelagibacter ubique* was subsequently removed to provide a more detailed view of the remaining species.

Figure 3.21 B demonstrated the occurrence of several bacterial species found within the metagenome. Nonetheless, only several of these species (*Solibacter usitatus, Acidobacteria bacterium, Magnetospirillum magneticum* and *Mesorhizobium sp. BNC1*) appeared to provide significant values at higher thresholds.

Figure 3.22 displayed the results of testing several lineage profiles against the deep Mediterranean metagenome. The study by Martín-Cuadrado *et al.* (2007) showed *Al-*

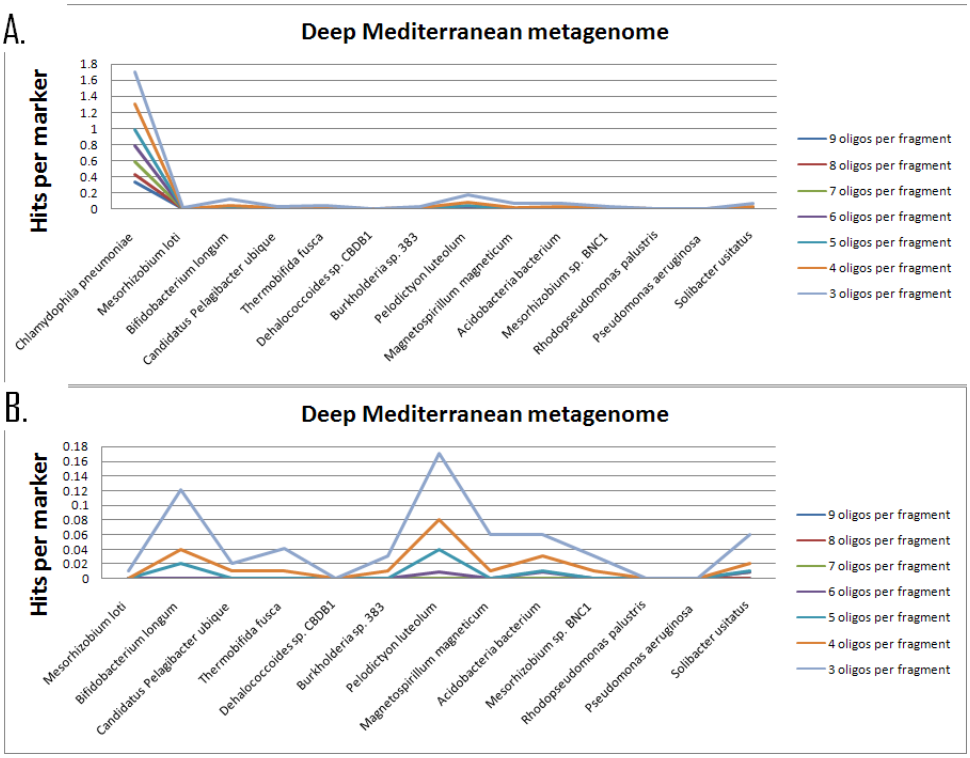


Figure 3.20: Marker results for the full deep Mediterranean metagenomic dataset including additional species (Figure A). Figure B shows the full dataset with the removal of *Chlamydophila pneumoniae*.

phaproteobacteria,Acidobacteria,Deltaproteobacteria and Betaproteobacteria as the most dominant lineages within the metagenome. Figure 3.22 correctly identified the majority of these lineages and other lineages present in the data. This signal seemed disproportionate to the expected result for *Alphaproteobacteria* which would surely have been even greater as the majority lineage. A reason for the lack of signal could be the absence of several dominant *Alphaproteobacteria* species within the constructed database. Hence the lineage profile would not incorporate these species into its consensus and would not describe them effectively.

3.4.2.2 Results from oligonucleotide frequency analysis

From Table 3.13 the same confused image can be seen as with Figures 3.20 and 3.21. In this table there was no clear pattern of identification indicating a separation between true positives and false positives. Final scores assigned are all close to one, this showed a lack of conviction in the predictions made. This lack of conviction could be expected due to the large amount of data and dilution of signal present.

From the inspection of Table 3.14 the results appeared to indicate that identification of lineages within a metagenomic sample may be a possibility. No false positives were found in this dataset, however, the values of the true positive results were generally close to one implying a lack of significance. Further statistical corrections will have to be put

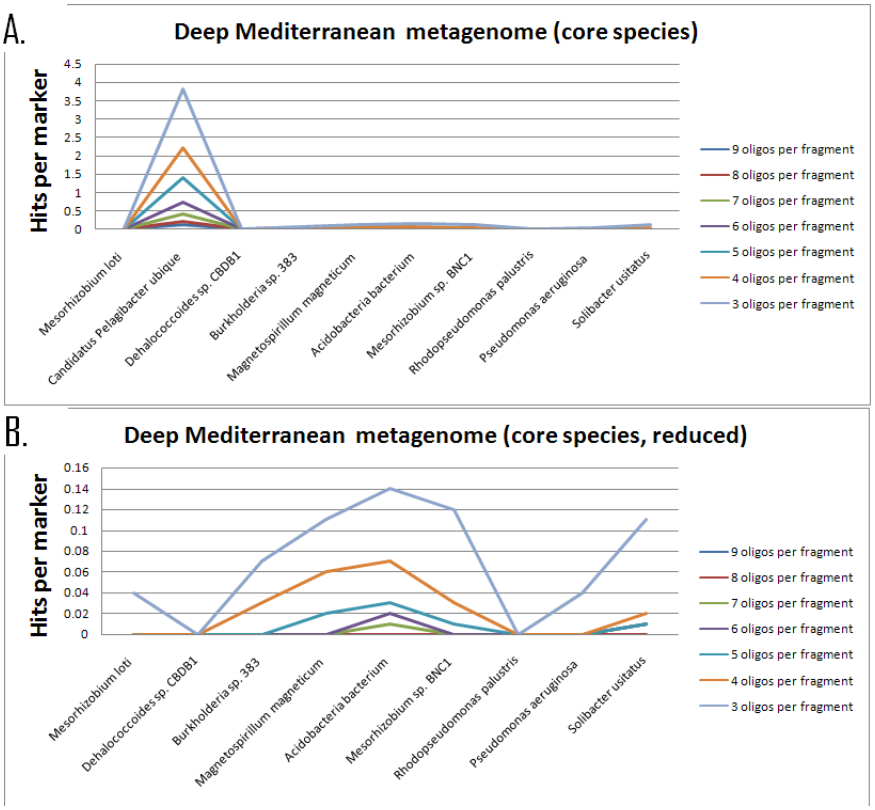


Figure 3.21: Marker results for the core species of the deep sea Mediterranean metagenomic dataset (Figure A). Figure B shows the core species with the removal of *Candidatus Pelagibacter ubique*.

in place in order to accurately estimate presence or absence of lineages in this sample. On further inspection only two false negatives were found within the dataset and these belong to minority lineages.

3.5 Discussion

3.5.1 Oligonucleotide frequency analysis

3.5.1.1 Species specific analysis

Oligonucleotide frequency analysis cannot separate closely related species effectively based on score or numerical difference. Furthermore, the scoring function did not provide consistent results for each case study. Scores appeared to be relative to specific families and their inherent genomic characteristics.

Moderate success was gained in the differentiation of more distantly related species which could be identified based largely on numerical difference in scores. There were several exceptions to this rule. Repeat rich genomes such as *Ureaplasma parvum* in case study 2 closely mimicked the profile for *Bacillus anthracis*. This needs to be taken into account in future experimentation. One solution may be to identify repeat rich genomes

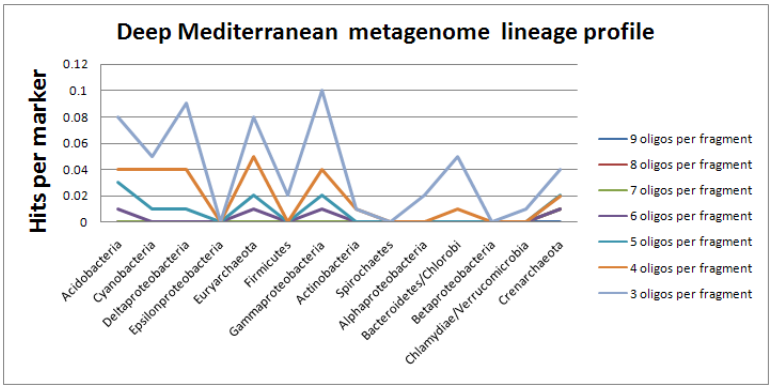


Figure 3.22: This Figure shows the full lineage profile for the deep Mediterranean metagenome. Prominent lineages found within the metagenome include; *Alphaproteobacteria, Acidobacteria,Deltaproteobacteria, Gammaproteobacteria,Betaproteobacteria and Cyanobacteria.*

first in order to correct for their affects on other species. A further exception was that distantly related species can still share common oligonucleotides. An example of this was *Dechloromonas aromatica* (Figure 3.5) with high similarity to *Salmonella enterica*, a species in a different lineage. The error value for *Dechloromonas aromatica* (121.1) did not provide any further hints as to this phenomenon. This cautions the assumption that each species or genus has highly unique signatures based on lineage comparison.

The analysis of multiple species in combined metagenomes highlighted further difficulties. Combined metagenomes were created containing several species with the total 100kbp sequence shared equally amongst them. This demonstrated a highly conservative test as a very limited amount of information was present for each species. Results for combined metagenomes showed that identification of species within this context was unreliable. A common occurrence was that two closely related species occurring in the metagenome will boost the score of a third related species absent from the metagenome. Furthermore, the dilution of signal by including foreign species greatly reduced signal and increased non-specific interference.

Combined metagenomes were highly variable. As different regions within the genomes were selected for sequencing, the different characteristics of these regions could bias results. This provided a conservative estimate of actual data where there was no guarantee for the presence of the entire genomic sequence.

In order to provide an average false positive rate for a species profile the error value was created. Nonetheless, this score is limited by the number of genomes within the database and therefore could provide skewed estimates under certain conditions.

In the final analysis of the deep Mediterranean metagenome no clear trends were visible. Large numbers of false positives and negatives were present in the results. The scores for the true positives remained close to one, indicating a lack of confidence.

3.5.1.2 Lineage specific analysis

The lineage analysis of metagenomes using oligonucleotide frequencies demonstrated a far more stable picture than that for species identification. Species within two of the four case studies were identified as belonging to their respective lineages. There were several difficulties when using this approach. The most notable was creation of a consensus that accurately described all species. This was a highly complex and often impossible task when a large group of diverse species was involved. For each case study there was a lineage profile which inadequately described a particular specie. This culminated in the *Gammaproteobacteria* lineage profile being unable to identify *Pseudomonas aeruginosa*. In this case, *Gammaproteobacteria,* a highly diverse and densely populated lineage, proved very difficult to describe using a lineage profile.

Positive results were obtained in the identification of lineages present in combined metagenomes. Several insights into the limitations of this approach and its ability to effectively detect lineages was made. Firstly, metagenomic signal seemed to be detectable when the sequence ratio is above 50%. Below this value, signal was too diluted for a positive result. In general, the score decreased and increased proportional to the number of species present within that lineage.

Testing of this method on the deep Mediterranean metagenomic dataset produced a far better result than species identification. Only one false positive was found with several false negatives. The majority of lineages were correctly identified although values remained close to 1 indicating low confidence in results. Nonetheless, a higher score was expected for the dominant *Alphaproteobacteria* lineage. This discrepancy could be due to a number of factors. The failure of the database to contain a number of species present in this environment could decrease the capacity of the lineage profile to detect these species.

Lineage specific analysis clearly provided far more promise than the species specific approach using oligonucleotides frequencies. However, further investigation will be required to produce optimal results.

3.5.2 Sequenced read analysis

3.5.2.1 Species analysis

The most limiting factor in sequenced read analysis was the number of markers designated for each specie. Case study 1 and 2 both showed the pathogen and closely related species deficient in markers. This deficiency lead to poor detection within metagenomic sequence and rendered these species profiles ineffective. Furthermore, species distantly related to the pathogen were subject to bias as their marker sets were much larger and chance occurrence of markers becomes more likely. In an attempt to control this bias the hits/marker statistic was calculated. This attempted to normalize the number of hits

against the number of markers in the species profile but did not control adequately for the bias and results were skewed towards the distant species.

Case study 4 provided an example where markers for all species profiles were well balanced and within an acceptable range. The majority of species could be easily identified in this case study. However, *Escherichia coli* produced a more distinctive signal against the *Shigella flexneri* metagenome. The species profile of one closely related species affecting another was a common occurrence. This indicated, beyond the number of markers per specie, that closely related species remained difficult to separate. This is an inherent flaw and directly hinders the effective use of overrepresented oligonucleotides in distinguishing between closely related species.

The miss-identification of *Escherichia coli* in this situation could also be due to the high number of markers. This could cause sufficient interference among closely related species that a stronger signal may be presented for the incorrect specie. This lead to the conclusion that a fine balance must be struck between these two extremes in order to produce the optimal number of markers. Careful moderation on distant species was also essential to minimize bias. A further concern was that the removal of shared markers by *MarkerCounter* was not controlled and no intelligent selection was involved. If this process was done more thoroughly, markers could be selected depending on specific characteristics as well as sequence identity.

The analysis of reads within multiple species metagenomes highlights all the points discussed above as each effect is amplified. Through this technique added confidence could be gained using each oligonucleotide threshold. Although interference from a species profile is highly likely this interference generally appeared only within the ”2 oligos per fragment” threshold. Nonetheless, for distantly related species the bias was often too great and these appeared at higher thresholds. This confounded the accurate detection of species. Multiple species metagenomes displayed a much higher rate of background interference which made detection of species unreliable.

From tests performed on the deep Mediterranean metagenome no well defined results were obtained. Several species profiles not present in the metagenome showed significant signal. This indicated that in a highly complex environment the rate of false positives is very high due to the interference between marker profiles. From these results it was clear that marker profiles lack specificity for accurate detection in such a complex environment. Further steps will have to be taken to moderate marker number as well as marker identity.

3.5.2.2 Lineage analysis

The analysis of sequenced reads using lineage profiles suggested a far more effective approach than the species specific analyses. Initial tests performed using all lineages resulted in exceptionally high interference, dwarfing the true positive signal. This lead to the realisation that lineages were not consistent, due to the variation in species number. A small

species number implied that species specific oligonucleotides were selected rather than a general lineage consensus. This could detract from another lineage which successfully described an oligonucleotide as generally overrepresented. Secondly, a very large number of species per lineage could result in such high variation that inadequate consensus can be reached. This produced an inaccurate selection of oligonucleotides which may describe only a minority of species within the lineage effectively. These difficulties all center on the issue of species separation. Although this was an essential process to enable easy computability and accurate classification it was a far more complex task than previously expected. As identified earlier, an alternative method for separating species into families or smaller phylogenetic units may improve results greatly.

In further analysis all small lineages, containing less than 5 species, were removed from the marker list. The results of these tests showed dramatic improvement, enabling reliable identification of lineages. If interference was still too prominent the marker list was reduced further.

These results lead to similar conclusions found in oligonucleotide frequency analysis. Both *Actinobacteria* and the *Firmicutes* lineages showed sufficient consensus to be capable of identifying their respective species adequately. The only significant improvement on this method was the effective use of the *Gammaproteobacteria* lineage profile in case study 4. Results displayed that this profile was indeed able to accurately identify *Salmonella enterica* and related species.

When testing lineage profiles against combined metagenomes noise increased substantially. Identification was clearly only possible when the lineage species were abundant in the metagenome. As mentioned before these experiments test the lower bounds of information as lineages could be accurately identified in as little as 50kb of shared metagenomic sequence. On inspection of the results obtained from the deep Mediterranean metagenome, several of the lineages featured were well described using this method. However, the absence of the majority lineage, *Alphaproteobacteria* created some concern about the accuracy of this approach. An explanation provided in Section 3.5.1.2 identified that species present in the metagenome may be missing in the database which could decrease the effectiveness of the lineage profile in identifying these members.

3.6 Conclusion

Two approaches are tested in this chapter. The first focused on the use of global frequencies in a metagenome to identify bacterial species and lineages. The second approach focused on classification of raw sequenced reads to marker profiles. Both of these techniques experienced similar results throughout testing and these will be summarised below.

In the identification of species within artificial metagenomes, differentiation of closely related species was unreliable. While differentiation of more distant species showed

promise, there were exceptions. Repeat rich genomes caused interference with related species making separation problematic. Unspecific species profiles created in small lineages confused results further and created a large number of false positives.

Species analysis in combined metagenomes amplified the above mentioned irregularities and made separation on a species level highly unlikely. From the analysis of the experimental deep Mediterranean metagenome a similar result was obtained. No consistency was present in the results. Furthermore, large numbers of false positives and false negatives occurred making detection of species unreliable using current methods. Sequence read analysis had marginally more success in the differentiation of species in combined metagenomes. In case studies 3 and 4 a more defined distinction can be made between species and characterisation of these metagenomes proves more accurate.

In the detection of lineages within artificial metagenomic samples lineage profiles appeared to accurately distinguish lineage members from non-members with some exceptions. Relatively low scores were obtained for certain lineage members, this was due to consensus sequences tending to describe certain species better than others. A definite irregularity was the poor results obtained from the *Gammaproteobacteria* lineage profile. This profile clearly failed to describe species adequately and results for both case studies 3 and 4 were relatively poor. This lineage is by far the most diverse and densely populated in the database and difficulties can therefore be anticipated with the accurate description of species in this instance. However, analysis of sequenced reads using lineage profiles proved more reliable in this context. After removal of lineages containing a small number of species the accurate identification of case study 4 using the *Gammaproteobacteria* lineage profile was possible.

Overall, the evaluation of combined metagenomes using lineage profiles indicated positive results. Signal for a particular lineage increases and decreases dependent on the ratio of species from said lineage present in the sample. However, lineage profiles were unable to detect signal in less than 50% of the metagenome. This can be expected as too little information is present for identification. In the analysis of the deep Mediterranean metagenome an accurate description of lineages presented within the sample was obtained. However, on closer inspection oligonucleotide frequency analysis produced a superior result by identifying more lineages correctly than sequenced read analysis.

In general, oligonucleotide frequency analysis was a far more sensitive approach than sequenced read analysis. However, this implied that it was also more susceptible to interference as was witnessed in species identification. This could be attributed to the inclusion of a larger number of oligonucleotides which ensured that the best possible score for each species is obtained, sometimes to the detriment of the analysis.

Analysis of sequenced reads provided an alternative approach which identified local trends within sequenced fragments rather than the global profile of a metagenome. This approach differed from oligonucleotide frequencies by its exclusion of shared markers

between species. This was a necessary step in reducing the excessive number of false positives. Nonetheless, this approach could also significantly reduce the ability to detect a species when too many markers were removed and hampered accurate identification in some respects. Results using this method showed that highly similar species reduced each others marker profiles dramatically, and it was clear that a more intelligent marker reduction system must be created.

One of the major advantages to sequenced read analysis was the added confidence given by the varied threshold of oligonucleotides per fragment required for a hit. At a glance this allowed for description of marker profiles not only by score but by appearance of hits at higher thresholds. This provided a significant advantage over oligonucleotide frequencies.

A further benefit was the ability to decrease marker lists hence increasing resolution for lineages present in the dataset. This could provide a powerful approach in future to aid in lessening interference within the results and showed that analyses were far more effective on a smaller target population.

In summary, the number and specificity of marker profiles was of crucial importance to the accuracy of sequenced read analysis. This approach carries great potential in its ability to provide added confidence for results, and with further investigation and research an accurate method of identification could be unearthed.

In conclusion, the identification of species within a metagenomic context is a formidable task. Both oligonucleotide frequency and sequenced read analysis are currently not capable of accurately distinguishing between species in an unknown metagenomic sample. However, lineage analysis appears to be a definite possibility as both analytical methods provided positive results in artificial as well as experimental metagenomes. Formalized use of either of these approaches will require further research. Against the background of the results obtained in this study, the potential for use of overrepresented oligonucleotides in identification within a metagenomic environment appears to be an as yet elusive but attainable goal.

Table 3.12: Markers usable by *MarkerCounter* for species in the deep Mediterranean dataset. The "Marker number (full dataset)" column describes the number of markers achieved by each genome in the dataset containing both species present and absent from the experimental metagenome. The "Marker number (core species dataset)" column contains the number of markers achieved only by species present within the experimental dataset.

Species	**Marker number (full dataset)**	**Marker number (core species dataset)**
Bifidobacterium longum	293	n/a
Burkholderia sp. 383	201	215
Dehalococcoides sp. CBDB1	1	1
Acidobacteria bacterium	356	398
Candidatus Pelagibacter ubique	116	171
Chlamydophila pneumoniae	589	n/a
Magnetospirillum magneticum	145	177
Mesorhizobium loti	100	114
Mesorhizobium sp. BNC1	312	326
Pelodictyon luteolum	483	n/a
Pseudomonas aeruginosa	95	123
Rhodopseudomonas palustris	84	96
Solibacter usitatus	179	183
Thermobifida fusca	341	n/a

Table 3.13: Identification of species in the deep Mediterranean metagenome using oligonucleotide frequencies. Total frequency count indicates the total sum of oligonucleotides found for each species profile. * indicates lineages identified within the metagenome by Martín-Cuadrado *et al.* (2007). ^ calculated using Algorithm 14

Species	**Total frequency count**	**Error Value**	**Final score^**
Burkholderia sp. 383	4447	58.91	1.03*
Thermobifida fusca YX	6446	72.85	1.21
Mesorhizobium sp. BNC1	11176	151.89	1.01*
Candidatus Pelagibacter ubique	5609	116.49	0.66*
Chlamydophila pneumoniae	21801	483.10	0.62
Rhodopseudomonas palustris	5705	75.85	1.03*
Magnetospirillum magneticum	5355	62.38	1.18*
Dehalococcoides sp. CBDB1	0	0.00	0.00*
Pseudomonas aeruginosa	5634	77.07	1.00*
Solibacter usitatus	9400	144.40	0.89*
Pelodictyon luteolum	14716	172.41	1.17
Acidobacteria bacterium	14470	215.21	0.92*
Bifidobacterium longum	10073	116.95	1.18
Mesorhizobium loti	8209	97.85	1.15*

Table 3.14: Identification of phylogenetic lineages in the deep Mediterranean metagenome using oligonucleotide frequencies. Total frequency count indicates the total sum of oligonucleotides found for each lineage profile. * indicates lineages identified within the metagenome by Martín-Cuadrado *et al.* (2007). ^ calculated using Algorithm 14

Lineage	**Total frequency**	**Error Value**	**Final Score^**
Cyanobacteria	30233	394.33	1.05*
Epsilonproteobacteria	11443	203.06	0.77
Deltaproteobacteria	23872	254.41	1.29*
Firmicutes	19314	216.22	1.22*
Euryarchaeota	30084	424.33	0.97*
Gammaproteobacteria	34342	410.52	1.15*
Spirochaetes	14256	244.60	0.80
Crenarchaeota	20805	386.44	0.74*
Acidobacteria	15463	206.90	1.02*
Alphaproteobacteria	20854	246.48	1.16*
Actinobacteria	15379	174.76	1.21
Bacteroidetes/Chlorobi	20871	354.45	0.81
Chlamydiae/Verrucomicrobia	16555	290.15	0.78
Betaproteobacteria	12162	126.23	1.32*

Chapter 4

Concluding discussion

4.1 Conclusion

Bacterial pathogens claim millions of lives each year and new mechanisms of identifying these bacteria in their natural environments as well as in their hosts have to be discovered. The current study contributes to this field by investigating a novel method for the identification of bacterial species from raw metagenomic fragments using overrepresented oligonucleotides as signature sequences. Identification of bacterial species within a metagenomic sample is a crucial step in medical diagnosis as well as in the prevention of infection.

Several complexities exist in the identification of bacteria within a metagenomic context. Currently identification of bacteria is limited to phylogenetic markers which utilize only 1% of metagenomic sequence. Furthermore, modern sequencing technologies produce reads of short length. Assembly of short reads is both computationally intensive and error prone. This study proposes a solution to these difficulties through the use of overrepresented oligonucleotide markers. Overrepresented oligonucleotides (8-14bp in length) are present throughout the genome and can be effectively employed in the identification of bacterial species in metagenomic sequence without the need for sequence assembly. Furthermore, the increased sequence length of the oligonucleotides over its predecessors (short oligonucleotide frequencies, 2-4bp in length) provides increased specificity for use on short sequence reads.

In order to identify overrepresented oligonucleotide markers for each specie, raw data had to be analysed, extended and imported into a structured database. The structured database provided a foundation for the creation of *Oligosignatures*, a program to interface with the user and query the database. *Oligosignatures* allowed for manipulation of analyses for use on any environmental context.

In the current study, the *Oligosignatures* program was applied to the identification of bacterial species within unknown metagenomic samples. The first step involved the

identification of closely related strains followed by their removal from further computation. Species and lineage specific oligonucleotide profiles were then described. Further testing was executed to determine the discriminating power of marker profiles. These tests were performed on both artificial and experimental data. Testing focused on the use of two approaches. Oligonucleotide frequency analysis employing global oligonucleotide marker frequencies were used to identify the presence of bacterial species. Concurrently, sequenced read analysis attributes metagenomic fragments to specific species using overrepresented oligonucleotide markers.

Results show that species identification is not possible under current conditions. Interference and bias prevent the detection of closely related species within metagenomic samples using both methods. However, identification of more distantly related organisms proved more reliable, improvements in this area could make species identification a distinct possibility. The determination of lineages within a metagenomic sample was far more promising. Lineages could be detected in artificial and experimental datasets using both approaches. These results, although not perfect, hint at the potential of species and lineage detection using overrepresented oligonucleotides.

4.2 Critical evaluation

The creation of a structured database from raw data offers the opportunity to search and compare data based on selected criteria. This database forms the foundation of the program *Oligosignatures*. The division of the database into phylogenetic lineage tables caused inaccuracies. Both exceptionally large and small lineages produce unreliable lineage marker profiles. The division into lineages assumes that phylogenetic relationship closely mirrors genomic similarity. There are exceptions to this rule that can confound results.

Development of *Oligosignatures* allowed for the creation of marker oligonucleotide profiles dependent on user defined contexts. Through *Oligosignatures* the opportunity to thoroughly explore the properties and uses of overrepresented oligonucleotides is now possible. The potential to identify various different uses for these oligonucleotides or to combine their use with existing identification pipelines can lead to improved identification methods within metagenomics. The identification of sequence fragments within metagenomic samples can also greatly aid research in metagenomics and is a potential future outcome for this approach.

As a test environment the detection of bacterial species in unknown metagenomic samples was selected. The identification of species in both artificial and experimental data is error prone. Closely related species cannot be reliably separated in this context. Distantly related species, however, are more easily distinguishable. Nonetheless, the identification of lineages in metagenomic samples showed promise. Results showed that lineages could

be discriminated in the majority of metagenomic samples dependent on an effective lineage profile. An effective lineage profile relies on an accurate consensus of the lineage (determined by lineage constitution) and the amount of sequence present for the lineage in the metagenome.

Two approaches were used to test the effectivity of overrepresented oligonucleotide marker profiles, namely, oligonucleotide frequency analysis and sequenced read analysis. The oligonucleotide frequency approach is clearly more sensitive than the sequenced read approach. Although the score calculated requires further mathematical correction, the magnitude differences in value hints that separation may indeed be possible with this approach. Sequenced read analysis provides a far more effective measure of confidence than oligonucleotide frequency by using multiple "oligo per fragment" thresholds. This approach holds great potential if the number and specificity of markers provided from each profile can be improved.

4.3 Recommendations for further research

The application of *Oligosignatures* to different environmental contexts will prove invaluable in the exploration of the uses of overrepresented oligonucleotides. Applying this program, the techniques employed for identification could then be improved to focus on classification of genomic fragments, realizing the potential of this technique to greatly increase the number of fragments identified within a metagenome. Overrepresented oligonucleotides may also be researched in their ability to determine phylogeny. This may provide a more robust method of classification than current approaches.

Due to the difficulties with the division of the database into lineages an alternative method should be sought. A suggestion would be to cluster genomes based on their overrepresentation of oligonucleotides. This approach would also allow for moderation of group size. Creation of small, consistent groupings may prove to be an intermediate between species and lineage identification enabling accurate classification.

Species specific methods tend to be too susceptible to interference. In future a filtering system may be put in place to decrease search space. By first identifying lineages present in the sample or the use of other exploratory methods the number of possible species with the metagenome can be significantly decreased. These results can then be used to guide species specific identification, reducing false positives.

The similarities discovered in overrepresented oligonucleotides amongst lineages or species can be used to investigate the mechanisms causing the overrepresentation of these oligonucleotides. These insights will be fundamental to further use of this approach and understanding of mechanisms involved in generating these oligonucleotides.

In terms of methods for the identification of species or lineages in metagenomic datasets the following improvements can be made. Sequence read analysis requires intelligent

marker reduction. Removal of shared markers can be based on the importance of that marker within the respective marker list. This can be further guided by the number of markers within a specific marker profile to ensure a balanced distribution of markers. Careful statistical adjustments will need to be made to the scoring system for oligonucleotide frequency analysis. The effect of genomic characteristics on the number of false positives needs to be taken into account when controlling for this phenomenon.

Bibliography

Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T. and Ikemura, T. (2002) A novel bioinformatic strategy for unveiling hidden genome signatures of eukaryotes: self-organizing map of oligonucleotide frequency. *Genome Inform* **13**, 12–20.

Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T. and Ikemura, T. (2003) Informatics for unveiling hidden genome signatures. *Genome Res* **13**, 4, 693–702.

Akopyants, N. S., Fradkov, A., Diatchenko, L., Hill, J. E., Siebert, P. D., Lukyanov, S. A., Sverdlov, E. D. and Berg, D. E. (1998) PCR-based subtractive hybridization and differences in gene content among strains of Helicobacter pylori. *Proc Natl Acad Sci U S A* **95**, 22, 13108–13113.

Amariglio, N. and Rechavi, G. (1993) Insertional mutagenesis by transposable elements in the mammalian genome. *Environ Mol Mutagen* **21**, 3, 212–218.

Andersson, S. G. and Kurland, C. G. (1990) Codon preferences in free-living microorganisms. *Microbiol Rev* **54**, 2, 198–210.

Baron, S. (1996) *Medical microbiolology* Addison-Wesley Publishing Company, Inc.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Wheeler, D. L. (2007) GenBank. *Nucleic Acids Res* **35**, Database issue, D21–D25.

Béjà, O., Aravind, L., Koonin, E. V., Suzuki, M. T., Hadd, A., Nguyen, L. P., Jovanovich, S. B., Gates, C. M., Feldman, R. A., Spudich, J. L., Spudich, E. N. and DeLong, E. F. (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289**, 5486, 1902–1906.

Bohlin, J., Skjerve, E. and Ussery, D. W. (2008) Investigations of oligonucleotide usage variance within and between prokaryotes. *PLoS Comput Biol* **4**, 4, e1000057.

Campa, M. B., M. and Friedman, H. (1993) Pseudomonas as an opportunistic pathogen *Plenum Press*.

Chaisson, M., Pevzner, P. and Tang, H. (2004) Fragment assembly with short reads. *Bioinformatics* **20**, 13, 2067–2074.

Chan, E. Y. (2005) Advances in sequencing technology. *Mutat Res* **573**, 1-2, 13–40.

Check, E. (2004) BioShield defence programme set to fund anthrax vaccine. *Nature* **429**, 6987, 4.

Courtois, S., Cappellano, C. M., Ball, M., Francou, F.-X., Normand, P., Helynck, G., Martinez, A., Kolvek, S. J., Hopke, J., Osburne, M. S., August, P. R., Nalin, R., Guérineau, M., Jeannin, P., Simonet, P. and Pernodet, J.-L. (2003) Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. *Appl Environ Microbiol* **69**, 1, 49–55.

Daubin, V. and Perrière, G. (2003) G+C3 structuring along the genome: a common feature in prokaryotes. *Mol Biol Evol* **20**, 4, 471–483.

Davenport, C., Wiehlmann, L., Reva, O. and Tümmler, B. (2008) Overrepresented oligonucleotides reveal genomic organisation and frequent coding motifs in the genus Pseudomonas *Submitted to BMC Bioinformatics*.

Deschavanne, P. J., Giron, A., Vilain, J., Fagot, G. and Fertil, B. (1999) Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol* **16**, 10, 1391–1399.

Dewhirst, J. I. B. P. A. T. W. W. W.-H. Y., F.E. and Chen., T. (2008) The Human Oral Microbiome Database.

Diatchenko, L., Lau, Y. F., Campbell, A. P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N., Sverdlov, E. D. and Siebert, P. D. (1996) Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc Natl Acad Sci U S A* **93**, 12, 6025–6030.

Diaz-Torres, M. L., McNab, R., Spratt, D. A., Villedieu, A., Hunt, N., Wilson, M. and Mullany, P. (2003) Novel tetracycline resistance determinant from the oral metagenome. *Antimicrob Agents Chemother* **47**, 4, 1430–1432.

Dimri, G. P., Rudd, K. E., Morgan, M. K., Bayat, H. and Ames, G. F. (1992) Physical mapping of repetitive extragenic palindromic sequences in Escherichia coli and phylogenetic distribution among Escherichia coli strains and other enteric bacteria. *J Bacteriol* **174**, 14, 4583–4593.

Draghici, S., Khatri, P., Liu, Y., Chase, K. J., Bode, E. A., Kulesh, D. A., Wasieloski, L. P., Norwood, D. A. and Reifman, J. (2005) Identification of genomic signatures for the design of assays for the detection and monitoring of anthrax threats. *Pac Symp Biocomput* 248–259.

Fox, G. E., Stackebrandt, E., Hespell, R. B., Gibson, J., Maniloff, J., Dyer, T. A., Wolfe, R. S., Balch, W. E., Tanner, R. S., Magrum, L. J., Zablen, L. B., Blakemore, R., Gupta, R., Bonen, L., Lewis, B. J., Stahl, D. A., Luehrsen, K. R., Chen, K. N. and Woese, C. R. (1980) The phylogeny of prokaryotes. *Science* **209**, 4455, 457–463.

Ganesan, H., Rakitianskaia, A., Davenport, C., Tummler, B. and Reva, O. (2008) The SeqWord Genome Browser: an online tool for the identification and visualization of atypical regions of bacterial genomes through oligonucleotide usage. *BMC Bioinformatics* **9**, 1, 333.

Gill, S. R., Pop, M., Deboy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M. and Nelson, K. E. (2006) Metagenomic analysis of the human distal gut microbiome. *Science* **312**, 5778, 1355–1359.

Gilson, E., Perrin, D. and Hofnung, M. (1990) DNA polymerase I and a protein complex bind specifically to E. coli palindromic unit highly repetitive DNA: implications for bacterial chromosome organization. *Nucleic Acids Res* **18**, 13, 3941–3952.

Goldman, N. (1993) Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. *Nucleic Acids Res* **21**, 10, 2487–2491.

Gouy, M. and Gautier, C. (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res* **10**, 22, 7055–7074.

Groisman, E. A. and Ochman, H. (1996) Pathogenicity islands: bacterial evolution in quantum leaps. *Cell* **87**, 5, 791–794.

Hacker, J., Bender, L., Ott, M., Wingender, J., Lund, B., Marre, R. and Goebel, W. (1990) Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal Escherichia coli isolates. *Microb Pathog* **8**, 3, 213–225.

Handelsman, J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* **68**, 4, 669–685.

Helgason, E., Caugant, D. A., Olsen, I. and Kolstø, A. B. (2000*a*) Genetic structure of population of Bacillus cereus and B. thuringiensis isolates associated with periodontitis and other human infections. *J Clin Microbiol* **38**, 4, 1615–1622.

Helgason, E., Okstad, O. A., Caugant, D. A., Johansen, H. A., Fouet, A., Mock, M., Hegna, I. and Kolstø, A. B. (2000*b*) Bacillus anthracis, Bacillus cereus, and Bacillus thuringiensis–one species on the basis of genetic evidence. *Appl Environ Microbiol* **66**, 6, 2627–2630.

Hulton, C. S., Higgins, C. F. and Sharp, P. M. (1991) ERIC sequences: a novel family of repetitive elements in the genomes of Escherichia coli, Salmonella typhimurium and other enterobacteria. *Mol Microbiol* **5**, 4, 825–834.

Ikemura, T. and Ozeki, H. (1983) Codon usage and transfer RNA contents: organism-specific codon-choice patterns in reference to the isoacceptor contents. *Cold Spring Harb Symp Quant Biol* **47 Pt 2**, 1087–1097.

Illumina, I. (2008) DNA Sequencing with Solexa® Technology *Technology Spotlight, Retrieved 28 July 2008, http://wwwilluminacom/sequencing/Technologyilmn.*

Jaspers, E. and Overmann, J. (2004) Ecological significance of microdiversity: identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. *Appl Environ Microbiol* **70**, 8, 4831–4839.

Kanaya, S., Yamada, Y., Kudo, Y. and Ikemura, T. (1999) Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene* **238**, 1, 143–155.

Karlin, M., Campbell (1998*a*) Comparative DNA analysis across diverse genomes *Annunal Review Genetics* **32**, 43.

Karlin, S. (1998*b*) Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr Opin Microbiol* **1**, 5, 598–610.

Karlin, S. and Burge, C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* **11**, 7, 283–290.

Karlin, S., Burge, C. and Campbell, A. M. (1992) Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucleic Acids Res* **20**, 6, 1363–1370.

Karlin, S., Campbell, A. M. and Mrázek, J. (1998) Comparative DNA analysis across diverse genomes. *Annu Rev Genet* **32**, 185–225.

Karlin, S. and Cardon, L. R. (1994) Computational DNA sequence analysis. *Annu Rev Microbiol* **48**, 619–654.

Karlin, S., Mrázek, J. and Campbell, A. M. (1997) Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol* **179**, 12, 3899–3913.

Keim, P., Price, L. B., Klevytska, A. M., Smith, K. L., Schupp, J. M., Okinaka, R., Jackson, P. J. and Hugh-Jones, M. E. (2000) Multiple-locus variable-number tandem

repeat analysis reveals genetic relationships within Bacillus anthracis. *J Bacteriol* **182**, 10, 2928–2936.

Khemici, V. and Carpousis, A. J. (2004) The RNA degradosome and poly(A) polymerase of Escherichia coli are required in vivo for the degradation of small mRNA decay intermediates containing REP-stabilizers. *Mol Microbiol* **51**, 3, 777–790.

Kingsley, M. T., Straub, T. M., Call, D. R., Daly, D. S., Wunschel, S. C. and Chandler, D. P. (2002) Fingerprinting closely related xanthomonas pathovars with random nonamer oligonucleotide microarrays. *Appl Environ Microbiol* **68**, 12, 6361–6370.

Knapp, S., Hacker, J., Jarchau, T. and Goebel, W. (1986) Large, unstable inserts in the chromosome affect virulence properties of uropathogenic Escherichia coli O6 strain 536. *J Bacteriol* **168**, 1, 22–30.

Koonin, E. V., Aravind, L. and Kondrashov, A. S. (2000) The impact of comparative genomics on our understanding of evolution. *Cell* **101**, 6, 573–576.

Larbig, K., Kiewitz, C. and Tümmler, B. (2002) Pathogenicity islands and PAI-like structures in Pseudomonas species. *Curr Top Microbiol Immunol* **264**, 1, 201–211.

Lorenz, P., Liebeton, K., Niehaus, F. and Eck, J. (2002) Screening for novel enzymes for biocatalytic processes: accessing the metagenome as a resource of novel functional sequence space. *Curr Opin Biotechnol* **13**, 6, 572–577.

Mancuso, M., Avendaño-Herrera, R., Zaccone, R., Toranzo, A. E. and Magariños, B. (2007) Evaluation of different DNA-based fingerprinting methods for typing Photobacterium damselae ssp. piscicida. *Biol Res* **40**, 1, 85–92.

Martín-Cuadrado, A.-B., López-García, P., Alba, J.-C., Moreira, D., Monticelli, L., Strittmatter, A., Gottschalk, G. and Rodríguez-Valera, F. (2007) Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. *PLoS ONE* **2**, 9, e914.

Metzker, M. L. (2005) Emerging technologies in DNA sequencing. *Genome Res* **15**, 12, 1767–1776.

Mojica, F. J., Díez-Villaseñor, C., Soria, E. and Juez, G. (2000) Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol Microbiol* **36**, 1, 244–246.

Mongodin, E. F., Emerson, J. B. and Nelson, K. E. (2005) Microbial metagenomics. *Genome Biol* **6**, 10, 347.

Nakata, A., Amemura, M. and Makino, K. (1989) Unusual nucleotide arrangement with repeated sequences in the Escherichia coli K-12 chromosome. *J Bacteriol* **171**, 6, 3553–3556.

Nesin, M., Lupski, J. R., Svec, P. and Godson, G. N. (1987) Possible new genes as revealed by molecular analysis of a 5-kb Escherichia coli chromosomal region 5' to the rpsU-dnaG-rpoD macromolecular-synthesis operon. *Gene* **51**, 2-3, 149–161.

Phillippy, A. M., Mason, J. A., Ayanbule, K., Sommer, D. D., Taviani, E., Huq, A., Colwell, R. R., Knight, I. T. and Salzberg, S. L. (2007) Comprehensive DNA signature discovery and validation. *PLoS Comput Biol* **3**, 5, e98.

Pride, D. T., Meinersmann, R. J., Wassenaar, T. M. and Blaser, M. J. (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* **13**, 2, 145–158.

Reva and Tümmler (2004) Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns. *BMC Bioinformatics* **5**, 90.

Reva and Tümmler (2005) Differentiation of regions with atypical oligonucleotide composition in bacterial genomes. *BMC Bioinformatics* **6**, 251.

Robinson, N. J., Robinson, P. J., Gupta, A., Bleasby, A. J., Whitton, B. A. and Morby, A. P. (1995) Singular over-representation of an octameric palindrome, HIP1, in DNA from many cyanobacteria. *Nucleic Acids Res* **23**, 5, 729–735.

Rodriguez (2002) Approaches to prokaryotic biodiversity: a population genetics perspective *Environmental Microbiology* **4**, 8.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. and Nyrén, P. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem* **242**, 1, 84–89.

Sandberg, R., Winberg, G., Bränden, C. I., Kaske, A., Ernberg, I. and Cöster, J. (2001) Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier. *Genome Res* **11**, 8, 1404–1409.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 12, 5463–5467.

Schmeisser, C., Stöckigt, C., Raasch, C., Wingender, J., Timmis, K. N., Wenderoth, D. F., Flemming, H.-C., Liesegang, H., Schmitz, R. A., Jaeger, K.-E. and Streit, W. R. (2003) Metagenome survey of biofilms in drinking-water networks. *Appl Environ Microbiol* **69**, 12, 7298–7309.

Schmid, S., Schuster and Huson (2006) ReadSim- A simulator for Sanger and 454 sequencing *Unpublished*.

Schmidt, H. and Hensel, M. (2004) Pathogenicity islands in bacterial pathogenesis. *Clin Microbiol Rev* **17**, 1, 14–56.

Shendure, J., Mitra, R. D., Varma, C. and Church, G. M. (2004) Advanced sequencing technologies: methods and goals. *Nat Rev Genet* **5**, 5, 335–344.

Slezak, T., Kuczmarski, T., Ott, L., Torres, C., Medeiros, D., Smith, J., Truitt, B., Mulakken, N., Lam, M., Vitalis, E., Zemla, A., Zhou, C. E. and Gardner, S. (2003) Comparative genomics tools applied to bioterrorism defence. *Brief Bioinform* **4**, 2, 133–149.

Smith, A. D., Xuan, Z. and Zhang, M. Q. (2008) Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics* **9**, 128.

Song, J. S., Jeon, J. H., Lee, J. H., Jeong, S. H., Jeong, B. C., Kim, S.-J., Lee, J.-H. and Lee, S. H. (2005) Molecular characterization of TEM-type beta-lactamases identified in cold-seep sediments of Edison Seamount (south of Lihir Island, Papua New Guinea). *J Microbiol* **43**, 2, 172–178.

Stein, J. L., Marsh, T. L., Wu, K. Y., Shizuya, H. and DeLong, E. F. (1996) Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J Bacteriol* **178**, 3, 591–599.

Sundquist, A., Ronaghi, M., Tang, H., Pevzner, P. and Batzoglou, S. (2007) Whole-genome sequencing and assembly with high-throughput, short-read technologies. *PLoS ONE* **2**, 5, e484.

Suzuki, H., Kunisawa, T. and Otsuka, J. (1986) Theoretical evaluation of transcriptional pausing effect on the attenuation in trp leader sequence. *Biophys J* **49**, 2, 425–435.

Teeling, H., Meyerdierks, A., Bauer, M., Amann, R. and Glöckner, F. O. (2004*a*) Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol* **6**, 9, 938–947.

Teeling, H., Waldmann, J., Lombardot, T., Bauer, M. and Glöckner, F. O. (2004*b*) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* **5**, 163.

Tembe, W., Zavaljevski, N., Bode, E., Chase, C., Geyer, J., Wasieloski, L., Benson, G. and Reifman, J. (2007) Oligonucleotide fingerprint identification for microarray-based pathogen diagnostic assays. *Bioinformatics* **23**, 1, 5–13.

Ticknor, L. O., Kolstø, A. B., Hill, K. K., Keim, P., Laker, M. T., Tonks, M. and Jackson, P. J. (2001) Fluorescent Amplified Fragment Length Polymorphism Analysis of Norwegian Bacillus cereus and Bacillus thuringiensis Soil Isolates. *Appl Environ Microbiol* **67**, 10, 4863–4873.

Tobes, R. and Pareja, E. (2006) Bacterial repetitive extragenic palindromic sequences are DNA targets for Insertion Sequence elements. *BMC Genomics* **7**, 62.

Todar, K. (2008) *Todar's Online Textbook of Bacteriology* Kenneth Todar, University of Wisconsin-Madison.

Travers, A. (1997) DNA-protein interactions: IHF–the master bender. *Curr Biol* **7**, 4, R252–R254.

Tringe, S. G., Zhang, T., Liu, X., Yu, Y., Lee, W. H., Yap, J., Yao, F., Suan, S. T., Ing, S. K., Haynes, M., Rohwer, F., Wei, C. L., Tan, P., Bristow, J., Rubin, E. M. and Ruan, Y. (2008) The airborne metagenome in an indoor urban environment. *PLoS ONE* **3**, 4, e1862.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S. and Banfield, J. F. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 6978, 37–43.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.-H. and Smith, H. O. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 5667, 66–74.

Versalovic, J., Koeuth, T. and Lupski, J. R. (1991) Distribution of repetitive DNA sequences in eubacteria and application to fingerprinting of bacterial genomes. *Nucleic Acids Res* **19**, 24, 6823–6831.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J. and Kuiper, M. (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res* **23**, 21, 4407–4414.

Wallace, R. B., Shaffer, J., Murphy, R. F., Bonner, J., Hirose, T. and Itakura, K. (1979) Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. *Nucleic Acids Res* **6**, 11, 3543–3557.

Wetmur, J. G. (1991) DNA probes: applications of the principles of nucleic acid hybridization. *Crit Rev Biochem Mol Biol* **26**, 3-4, 227–259.

Whiteford, N., Haslam, N., Weber, G., Prügel-Bennett, A., Essex, J. W., Roach, P. L., Bradley, M. and Neylon, C. (2005) An analysis of the feasibility of short read sequencing. *Nucleic Acids Res* **33**, 19, e171.

WHO (2008) World Health Organisation: World Health Statistics 2008 *Retrieved 24 July 2008 from WHO Statistical Information System (WHOSIS), http://wwwwhoint/whosis/en/indexhtml*.

Zhang, Z., Willson, R. C. and Fox, G. E. (2002) Identification of characteristic oligonucleotides in the bacterial 16S ribosomal RNA sequence dataset. *Bioinformatics* **18**, 2, 244–250.